Институт русского языка имени В.В. Виноградова

Труды Института русского языка им. В.В. Виноградова

VI

Национальный корпус русского языка: 10 лет проекту

МОСКВА
2015

Труды Института русского языка им. В.В. Виноградова. Вып. 6
— М., 2015. — 658 с.

ISSN 2311-150X

Издание основано в 2013 г.

ГЛАВНЫЙ РЕДАКТОР

А.М. Молдован, академик РАН (Москва, Россия)

РЕДАКЦИОННЫЙ СОВЕТ

Ю.Д. Апресян, академик РАН, профессор (Москва, Россия);
Бьёрн Вимер, доктор филологии, профессор (Майнц, Германия);
А.А. Гиппиус, член-корреспондент РАН, профессор (Москва, Россия);
М.Л. Каленчук, доктор филол. наук, профессор (Москва, Россия);
Туре Нессет, доктор филологии, профессор (Тромсё, Норвегия);
В.А. Плунгян, член-корреспондент РАН, профессор (Москва, Россия);
Вацлав Чермак, доктор филологии (Прага, Чехия).

ОТВЕТСТВЕННЫЙ РЕДАКТОР ВЫПУСКА

В.А. Плунгян, член-корр. РАН, профессор (Москва, Россия)

РЕДАКЦИОННАЯ КОЛЛЕГИЯ ВЫПУСКА

Е.В. Рахилина, д. филол. наук (Москва, Россия);
С.О. Савчук, к. филол наук (Москва, Россия).

Выходит 3 раза в год

Адрес редакции:
119019, Москва, ул. Волхонка, д. 18/2

Издание зарегистрировано Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций.
Свидетельство о регистрации ПИ № ФС 77-57258

ISSN 2311-150X

© Институт русского языка им. В.В. Виноградова, 2015
© Авторы, 2015

Vinogradov Russian Language Institute

Proceedings of the V.V. Vinogradov Russian Language Institute

VI

The Russian National Corpus: 10 Years of the Project

MOSCOW
2015

Proceedings of the V.V. Vinogradov Russian Language Institute. 6
— M., 2015. — 658 p.

ISSN 2311-150X

The Journal is founded in 2013

EDITOR-IN-CHIEF

Alexander M. Moldovan, D.Sc., Full Member of the RAS (Moscow, Russia)

EDITORIAL BOARD

Yury D. Apresyan, D.Sc., Professor, Full Member of the RAS (Moscow, Russia);
Václav Čermák, Ph.D. (Prague, Czech Republic);
Alexey A. Gippius, D.Sc., Professor, Corresponding Member of the RAS (Moscow, Russia);
Maria L. Kalenchuk, D.Sc., Professor (Moscow, Russia);
Tore Nesset, D.Sc., Professor (Tromsø, Norway);
Vladimir A. Plungian, D.Sc., Professor, Corresponding Member of the RAS (Moscow, Russia);
Bjoern Wiemer, D.Sc., Professor (Mainz, Germany).

CHIEF EDITOR OF THE ISSUE

Vladimir A. Plungian, D.Sc., Professor, Corresponding Member of the RAS (Moscow, Russia)

EDITORIAL BOARD OF THE ISSUE

Ekaterina V. Rakhilina, D.Sc. (Moscow, Russia);
Svetlana O. Savchuk, Ph.D. (Moscow, Russia).

Published three times a year

Address:
18/2, Volkhonka street, Moscow, 119019
E-mail: ruslang@ruslang.ru

The journal is registered by the The Federal service for supervision of communications, information technology, and mass-media.
Registration certificate ПИ № ФС 77-57258.

© by Vinogradov Russian Language Institute, 2015
© by Authors, 2015

С О Д Е Р Ж А Н И Е

Предисловие (В. А. Плунгян) 11

I. Обновление и развитие Национального корпуса русского языка

В. П. Захаров (Санкт-Петербург).
Корпуса русского языка..... 20
Е. А. Гришина (Москва).
Мультимодальный модуль в составе
Национального корпуса русского языка..... 65
А. М. Молдован (Москва).
Памятники древнерусской письменности
в Национальном корпусе русского языка 88
Е. А. Мишина, А. А. Пичхадзе (Москва).
Древнерусский подкорпус
Национального корпуса русского языка..... 99
Е. Р. Добрушина, А. Г. Кравецкий, А. Е. Поляков (Москва).
Корпус и частотный грамматический корпусный словарь
церковнославянского языка в составе
Национального корпуса русского языка..... 116
И. Б. Качинская, Д. В. Сичинава (Москва).
Диалектный подкорпус сегодня..... 142
С. О. Савчук (Москва).
Корпус региональных газет России и зарубежья..... 163
Д. В. Сичинава (Москва).
Параллельные тексты в составе Национального корпуса
русского языка: новые направления развития и результаты..... 194
К. М. Корчагин (Москва).
Поэзия XX века в поэтическом подкорпусе
Национального корпуса русского языка:
проблема репрезентативности..... 235
Е. А. Гришина, Ю. Г. Зеленков, Б. В. Орехов (Москва).
Наивная поэзия в акцентологическом корпусе..... 257

П. В. Дяченко, Л. Л. Иомдин, А. В. Лазурский, Л. Г. Митюшин, О. Ю. Подлесская, В. Г. Сизов, Т. И. Фролова, Л. Л. Цинман (Москва).
Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус) 272
А. И. Зобнин, Г. В. Носырев (Москва).
Морфологический анализатор Mystem 3.0 300

II. Национальный корпус русского языка в научных исследованиях и образовании

Е. В. Рахилина (Москва).
Степени сравнения в свете русской грамматики ошибок 310
С. О. Савчук (Москва).
Корпус как инструмент для исследования особенностей функционирования русского языка в региональной прессе 333
Г. И. Кустова (Москва).
Лексико-семантическая и социокультурная специфика русской речи Гродненщины (на материале текстов СМИ) 365
Я. Э. Ахапкина (Москва).
Рефлексивные глаголы «убираться» и «играться»: кодификация и узус 392
Д. О. Добровольский (Москва).
Корпус параллельных текстов и сопоставительная лексикология 413
Б. В. Орехов (Москва).
Еще раз об исследовательском потенциале поэтического корпуса: метр, лексика, формула 449
О. Н. Ляшевская, Е. В. Кашкин (Москва).
Типы информации о лексических конструкциях в системе ФреймБанк 464
Е. А. Гришина (Москва).
О русском жестикуляционном отрицании 556

А. А. Бонч-Осмоловская (Москва).
Культуромика в Национальном корпусе русского языка,
к постановке задачи:
три века русских дорог 605
А. И. Левинзон (Москва).
Корпусное преподавание в российской школе 641

CONTENTS

Preface (*V. A. Plungian*) 11

I. Updating and Development of the Russian National Corpus

V. P. Zakharov (St. Petersburg).
Corpora of the Russian Language 20
E. A. Grishina (Moscow).
The Multimodal Module as Part
of the Russian National Corpus 65
A. M. Moldovan (Moscow).
Old Russian Manuscripts in the Russian National Corpus 88
E. A. Mishina, A. A. Pichkhadze (Moscow).
Old Russian Subcorpus of the Russian National Corpus 99
E. R. Dobrushina, A. G. Kravetskii, A. E. Polyakov (Moscow).
Corpus and Frequency Grammar Corpus Vocabulary
of the Church Slavonic Language as Part
of the Russian National Corpus 116
I. B. Kachinskaya, D. V. Sitchinava (Moscow).
Today's State of the Dialectal Subcorpus 142
S. O. Savchuk (Moscow).
The Corpus of Regional Russian Language
Newspapers in Russia and Abroad 163
Dmitri V. Sitchinava (Moscow).
Parallel Texts within the Russian National Corpus:
New Directions and Results 194
K. M. Korchagin (Moscow).
20th-century Poetry in the Russian National Corpus:
How to Create a Representative Poetical Corpus 235
E. A. Grishina, Yu. G. Zelenkov, B. V. Orekhov (Moscow).
Naïve Poetry in Accentologic Corpus 257

P. V. Dyachenko, L. L. Iomdin, A. V. Lazursky, L. G. Mityushin, O. Yu. Podlesskaya, V. G. Sizov, T. I. Frolova, L. L. Tsinman (Moscow).
A Deeply Annotated Corpus of Russian Texts (SynTagRus): Contemporary State of Affairs 272
A. I. Zobnin, G. V. Nosyrev (Moscow).
Morphological Analyzer Mystem 3.0 300

II. The Russian National Corpus in Scientific Research and Education

Ekatetina V. Rakhilina (Moscow).
Russian Comparative Constructions in the Aspect of Error Grammar 310
S. O. Savchuk (Moscow).
Corpus as a Tool for Studying the Russian Language Usage in Regional Press 333
Galina I. Kustova (Moscow).
Lexical-Semantic and Socio-Cultural Specificity of Russian Speech of Grodno (on the Material of Media Texts) 365
Yana E. Akhapkina (Moscow).
Reflexive Verbs «убираться» (ubirat'sya) and «играться» (igrat'sya): Codification and Discourse Practice 392
D. O. Dobrovol'skij (Moscow).
The Corpus of Parallel Texts and Contrastive Lexicology 413
B. V. Orekhov (Moscow).
Once again About the Research Potential of the Poetic Corpus: Meter, Vocabulary, Formula 449
O. N. Lyashevskaya, E. V. Kashkin (Moscow).
Types of Information About Lexical Constructions in Russian FrameBank 464

A. A. Bonch-Osmolovskaya (Moscow).
Culturomics in the Russian National Corpus:
Three Centuries of Russian Roads 605
A. I. Levinson (Moscow).
Using Corpus Data in Russian Language Classroom 641

ПРЕДИСЛОВИЕ

Данный тематический номер «Трудов Института русского языка им. В. В. Виноградова» является — в несколько ином формате — продолжением серии публикаций, посвященных Национальному корпусу русского языка (www.ruscorpora.ru). Ранее в этой серии выходили следующие сборники: (1) В. А. Плунгян (ред.), *Национальный корпус русского языка: 2003–2005. Результаты и перспективы* (М.: Индрик, 2005) и (2) В. А. Плунгян, Е. В. Рахилина, Т. И. Резникова (ред.), *Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы* (СПб: Нестор-История, 2009)[1]. Если сборник (1) был самым первым изданием, посвященным Корпусу, и подводил итог трём начальным годам его существования, то сборник (2) уже содержал не только характеристику Корпуса на этапе его дальнейшего развития, но и достаточно представительную часть, состоящую из исследований русского языка, выполненных на корпусном материале. Настоящий сборник продолжает именно эту линию: в нём читатель найдет как описание архитектуры и функционирования Корпуса на современном этапе (в том числе и в сравнении с другими существующими корпусами), так и представительный раздел научных исследований, выполненных в парадигме современной корпусной лингвистики и различным образом использующих богатые возможности Национального корпуса русского языка.

Сегодня, пожалуй, исследователям языка уже не надо ни объяснять, что такое корпус, ни рассказывать об устройстве Национального корпуса русского языка «с азов». За более чем 10 лет, которые прошли с момента открытия Корпуса для свободного доступа (апрель 2004), обращение к Корпусу стало почти самоочевидным компонентом практически любого исследования русского языка; Корпус достаточно хорошо известен и в образовательной среде, его

[1] Следует упомянуть также сборник: К. Л. Киселёва, В. А. Плунгян, Е. В. Рахилина, С. Г. Татевосов (ред.), *Корпусные исследования по русской грамматике* (М.: Пробел-2000, 2009). Этот сборник, в отличие от упомянутых выше, содержит только статьи о различных явлениях русской грамматики, а не описание текущего состояния корпуса, однако с корпусной проблематикой он тесно связан; кроме того, все представленные в нём исследования выполнены именно на материале Национального корпуса русского языка.

используют редакторы, на него ссылаются журналисты в статьях и популярных передачах и т. п.

Статьи сборника отражают реалии существования Корпуса в последние годы, когда он бурно развивался — как экстенсивно, так и интенсивно. Особую роль в этом развитии сыграла программа президиума РАН «Корпусная лингвистика» (2009–2014), давшая новый импульс многим идеям и позволившая реализовать целый ряд важных проектов, которые аккумулировали значительное число исполнителей и ресурсов.

Таким образом, сегодня Национальный корпус русского языка — это уже хорошо известная и востребованная в разных областях информационно-справочная система. Об актуальных аспектах его функционирования и применения и говорится в настоящем сборнике.

Сборник состоит из двух частей: в первой части собраны статьи, посвященные обновлению и развитию Национального корпуса русского языка, во второй части — статьи, посвященные применению Корпуса в научных исследованиях и образовании.

Кратко коснемся содержания каждой части.

Первая часть открывается статьей В. П. Захарова «Корпуса русского языка». Она имеет обобщающий характер и задумана как наиболее полный на сегодняшний день обзор существующих корпусов русского языка. Обзоры такого рода публиковались в наших сборниках и раньше, но ситуация в этой области меняется так быстро, что необходимость в актуальном обзоре оказывается вновь велика. В центре статьи В. П. Захарова находится, по понятным причинам, Национальный корпус русского языка, но автор излагает и предысторию российской корпусной лингвистики, а также описывает другие корпуса русского языка, созданные в мире.

В последующих статьях данного раздела дается подробная характеристика тем особенностям Корпуса, которые появились за последние годы. В первую очередь это новые (или существенно обновленные) подкорпуса, которые теперь функционируют в качестве полноценных модулей; все они являются оригинальными типами корпусов, практически не имеющими аналогов. Так, в статье Е. А. Гришиной «Мультимодальный модуль в составе Национального корпуса русского языка» дается подробное описание открытого в 2010 г. Мультимедийного русского корпуса, содержащего размеченные аудио- и виде-

офайлы с образцами русской речи; это один из наиболее интересных, оригинальных и трудоемких проектов в составе Корпуса. Следующая группа статей описывают исторический модуль в составе Корпуса, также появившийся как единое целое в последний период: это статьи А. М. Молдована «Памятники древнерусской письменности в Национальном корпусе русского языка», Е. А. Мишиной и А. А. Пичхадзе «Древнерусский подкорпус Национального корпуса русского языка» и тесно примыкающая к ним статья Е. Р. Добрушиной, А. Г. Кравецкого и А. Е. Полякова «Корпус и частотный грамматический корпусный словарь церковнославянского языка в составе Национального корпуса русского языка».

Включение в состав Корпуса полноценного исторического модуля призвано не только способствовать более эффективному решению традиционных задач исторического исследования языка, но и позволяет поставить ряд принципиально новых исследовательских задач. Основной вектор таких исследований связан с возможностью проследить непрерывную эволюцию языковых элементов (слов, форм, грамматических категорий и конструкций) на протяжении всей засвидетельствованной письменной истории русского языка. Тем самым именно Корпус способствует превращению объекта исследований русистов в гораздо более сложный и неоднородный, но в то же время цельный объект, для описания которого привычное некогда «строгое» разделение на синхронию и диахронию (с противопоставлением методов, подходов и даже исследовательских социумов) начинает выглядеть слишком жёстким и искусственным.

Ещё одному очень важному для нас новому проекту посвящена статья С. О. Савчук «Корпус региональных газет России и зарубежья», где дается подробная характеристика массиву региональной прессы, прежде в Корпусе отсутствовавшему. Развитие подкорпуса региональной прессы (которое мы планируем продолжать) дает возможность поставить на более надежную почву целый ряд важных для современной русистики исследований: это и более полный анализ языка современных СМИ, и региональная вариативность русского языка, и социолингвистические исследования современной нормы.

Ряд статей первой части вводят читателя в курс работы над уже известными ранее подкорпусами в составе Национального корпуса русского языка, отражая изменения и дополнения в их архитектуре и составе (иногда — весьма значительные), произошедшие

в последние годы. К этой группе относятся статьи И. Б. Качинской и Д. В. Сичинавы «Диалектный подкорпус сегодня», Д. В. Сичинавы «Параллельные тексты в составе Национального корпуса русского языка: новые направления развития и результаты», а также К. М. Корчагина «Поэзия XX века в поэтическом подкорпусе Национального корпуса русского языка: проблема репрезентативности». Статья о диалектном подкорпусе (написанная совместно специалистом-диалектологом и активным участником корпусного проекта с его первых шагов) характеризует в основном новые конструктивные решения, существенно меняющие архитектуру этого модуля (хотя заметного пополнения диалектного корпуса новыми текстами на данном этапе не производилось). Напротив, параллельный корпус за последние годы крайне интенсивно расширялся, и в настоящее время представляет собой фактически целое семейство различных параллельных корпусов, включающее в общей сложности 11 самостоятельных языковых разделов (не считая отдельного мультиязычного корпуса, где текстов меньше, но зато языков гораздо больше); усложнились также состав и возможности поиска по параллельному корпусу. Все эти изменения в статье Д. В. Сичинавы подробно документированы.

Поэтический корпус, также продолжающий интенсивно развиваться и пополняться, по мере приближения к поэтическими текстам середины и конца XX века ставит перед составителями особенно нетривиальные проблемы (уводящие скорее в область литературной критики и истории литературного процесса). Основные принципы отбора поэтических текстов, сформулированные составителями в процессе работы над корпусом, обсуждаются в статье К. М. Корчагина (на данном этапе эта проблема, пожалуй, оказывается одной из наиболее нетривиальных и действительно требует пространного обсуждения). К статье К. М. Корчагина содержательно отчасти примыкает статья Е. А. Гришиной, Ю. Г. Зеленкова и Б. В. Орехова «Наивная поэзия в акцентологическом корпусе», также связанная с современными поэтическими текстами. Однако, в отличие от тех текстов, которые включены в поэтический подкорпус (и в силу этого представляют если не художественный, то хотя бы историко-литературный интерес), в этой статье рассматривается массив текстов, свободно размещаемых в интернете непрофессиональными сочинителями. Большинство этих текстов не претендует ни на ка-

кую художественную ценность и по всем критериям могут считаться скорее курьезами современной массовой культуры, однако для корпусной лингвистики они обладают одним чрезвычайно ценным качеством — будучи, в основном, написанными по правилам силлабо-тонической метрики, они сохраняют крайне важные сведения о русском ударении конца XX — начала XXI века. «Профессиональная» поэзия издавна использовалась как ценнейший материал для изучения истории русского ударения (начиная с XVIII века), однако наивная поэзия в качестве такого источника выступает впервые, и корпусные технологии автоматической разметки позволяют сравнительно легко привлечь для этой цели статистически очень существенный объем текстов. Все они входят непосредственно в акцентологический подкорпус, тем самым обогащая и диверсифицируя его состав.

О перспективах развития экспериментального подкорпуса с синтаксической разметкой (пока что единственного корпуса такого типа, доступного на сайте НКРЯ) говорится в статье П. В. Дяченко, Л. Л. Иомдина и др. «Современное состояние глубоко аннотированного корпуса текстов русского языка (СинТагРус)».

Завершает первую часть статья А. И. Зобнина и Г. В. Носырева «Морфологический анализатор Mystem 3.0». Написанная программистами компании «Яндекс», бывшими в последние годы в центре всех проектов, связанных с программным обеспечением Национального корпуса русского языка, эта статья знакомит с ходом работы над новым, усовершенствованным морфологическими анализатором, используемым при автоматической морфологической разметке текстов Корпуса.

Помимо создания новых подкорпусов и пополнения существующих, работа с Национальным корпусом русского языка предполагает и многообразные научные исследования, опирающиеся на обширный корпусный материал и использующие соответствующие методы (прежде всего, статистические). Образцы такого рода исследований (наряду со статьями, посвященными применению Корпуса в образовании) представлены в статьях второй части. Следует обратить внимание на известную симметрию между первой и второй частями: если в первой части те или иные подкорпуса вводятся в рассмотрение и детально описываются, то во второй части демонстрируется, каким образом эти подкорпуса могут быть использова-

ны для решения конкретных научных задач; зачастую эти «парные» статьи написаны одними и теми же авторами (что вполне естественно и отражает давний интерес создателей Национального корпуса русского языка к возможностям использования корпусных методов для получения новых результатов в теоретической русистике).

Вторая часть открывается статьей Е. В. Рахилиной «Степени сравнения в свете русской грамматики ошибок», в которой привлекается внимание к использованию ещё одного типа корпуса — так называемого «обучающего»; обучающие корпуса (число которых во всем мире сейчас растет) содержат тексты, написанные не-носителями языка, и предназначены как для решения дидактических задач, так и для научных исследований интерференции, механизмов порождения текста на неродном языке и т. п. Непосредственным материалом исследования Е. В. Рахилиной являются сравнительные конструкции: автор показывает, что вариативность и отклонения от нормы в этом фрагменте русской грамматики имеют неслучайный характер и свидетельствуют в конечном счете о его диахронической нестабильности.

В двух следующих статьях использован материал нового подкорпуса региональной прессы, о принципах формирования которого подробно говорилось в статье С. О. Савчук в первой части. Во вторую же часть входит ещё одна статья С. О. Савчук «Корпус как инструмент для исследования особенностей функционирования русского языка в региональной прессе» и статья Г. И. Кустовой «Лексико-семантическая и социокультурная специфика русской речи Гродненщины (на материале текстов СМИ)». Как показывают названия этих статей, в них основное внимание уделяется анализу лексических особенностей региональных СМИ (имеются в виду как анализ неологизмов и нестандартной семантики лексических единиц, так и особенности сочетаемости); материалом для обеих статей послужили газеты Гродненского региона, так как среди зарубежных СМИ они в настоящее время представлены в Корпусе с наибольшей полнотой. В статье Г. И. Кустовой также прослеживается взаимодействие между «архаизирующей» тенденций (сохранение советизмов) и многочисленными лексическими инновациями, в том числе проникновением в газетный дискурс коллоквиализмов и элементов сленга.

Языковые процессы, происходящие в устной и неформаль-

ной письменной речи, являются предметом статьи Я. Э. Ахапкиной «Рефлексивные глаголы *убираться* и *играться*: кодификация и узус». Две лексемы, фигурирующие в заголовке этой статьи, не раз подвергались осуждению пуристов; между тем, автор на обширном корпусном материале показывает, что суждения о ненормативном характере данных глаголов должны быть как минимум более нюансированными и осторожными, так как в большом числе контекстов их употребление является нейтральным и свободно допускается в том числе и образованными носителями литературной нормы.

Статья Д. О. Добровольского «Корпус параллельных текстов и сопоставительная лексикология» возвращает читателя к проблемам создания и использования параллельного корпуса, современное состояние которого было подробно описано Д. В. Сичинавой в первой части сборника. В данной же статье Д. О. Добровольский иллюстрирует впечатляющие возможности параллельного корпуса (на материале английского и немецкого разделов параллельного корпуса НКРЯ) для решения наиболее трудных задач контрастивной лексикологии, а также двуязычной лексикографии.

К проблемам изучения поэтических текстов обращается Б. В. Орехов в статье «Ещё раз об исследовательском потенциале поэтического корпуса: метр, лексика, формула». Как показывает название этой статьи, она написана на пересечении стиховедческой и лексикологической тематики и вносит вклад в известную проблему связи лексического выбора поэта с метрическими характеристиками стихотворения (поднимавшуюся в свое время в работах М. Л. Гаспарова).

Статья О. Н. Ляшевской и Е. В. Кашкина «Типы информации о лексических конструкциях в системе ФреймБанк» вводит в рассмотрение интересные возможности изучения семантико-синтаксических характеристик лексем в рамках проекта по исследованию сочетаемости «ФреймБанк». ФреймБанк, созданный на основе Национального корпуса русского языка, называется его создателями русскоязычным аналогом известной системы FrameNet, идея которой принадлежит Чарлзу Филлмору. ФреймБанк аккумулирует различные типы ситуаций («фреймы») и информацию о ролевой характеристике их участников, а также о способах их выражения. В статье показано, какого рода задачи можно решать с помощью данного словарно-грамматического ресурса.

Статья Е. А. Гришиной «О русском жестикуляционном отрицании» закономерно вырастает из работы ее автора по составлению мультимедийного корпуса русского языка. Одним из наиболее бесспорных преимуществ мультимедийных корпусов является, как известно, возможность исследования жестикуляции, сопровождающей устную речь, что прекрасно проиллюстрировано и в данной статье.

В очень своеобразной и интересной статье А. А. Бонч-Осмоловской «Культуромика с помощью Национального корпуса русского языка» рассматривается возможность социокультурных исследований через призму изучения лексической сочетаемости на обширном статистически достоверном материале Корпуса. В этом новом подходе соединяются приемы традиционного семантического анализа с учетом сочетаемости и современные методы анализа так наз. big & medium data. Непосредственным материалом статьи являются классы прилагательных, встречающихся в атрибутивной позиции при русском существительном *дорога* (не раз бывшего предметом семантических изысканий, только в рамках других исследовательских парадигм); сочетаемость прослеживается на всем массиве Национального корпуса русского языка с XVIII века по наши дни.

Завершает вторую часть статья А. И. Левинзон «Корпусное преподавание в российской школе», в которой автор обращается к очень важной для корпусной лингвистики проблематике применения корпусов в преподавании языка. Статья обращена непосредственно к тем, кто связан со школьным преподаванием русского языка, и содержит немало ценных наблюдений, практически полезной информации и рекомендаций.

Таким образом, настоящий сборник отражает, как мы попытались показать, современное состояние Национального корпуса русского языка во всей полноте, не просто характеризуя достигнутые на разных направлениях результаты, но и иллюстрируя многообразные и постоянно растущие возможности применения Корпуса в научных исследованиях и преподавательской практике. Можно надеяться, что представленные статьи будут интересны как опытным специалистам, хорошо знакомым с достижениями современной корпусной лингвистики, так и начинающим пользователям — и всем, кто в своей деятельности так или иначе соприкасается с русским языком.

В. А. Плунгян

I. Обновление и развитие Национального корпуса русского языка

I. Updating and Development of the Russian National Corpus

В. П. Захаров
Санкт-Петербургский университет
Институт лингвистических исследований РАН
(Россия, Санкт-Петербург)
vz1311@yandex.ru

КОРПУСА РУССКОГО ЯЗЫКА[1]

Статья описывает предысторию российской корпусной лингвистики и корпуса русского языка, созданные как в России, так и за рубежом. Особое внимание уделяется Национальному корпусу русского языка, специализированным корпусам и диахроническому корпусу русских текстов Google books Ngram Viewer.

Ключевые слова: корпусная лингвистика, корпуса русского языка.

Введение

В последние годы создание корпусов и корпусно-ориентированные исследования стали неотъемлемой частью деятельности лингвистов. В мире корпусная лингвистика как особое направление сложилась к началу 1990-х годов. За прошедшие годы корпусная методология становится частью лингвистической науки, и все лингвисты, работающие в самых разных направлениях, как правило, проводят свои исследования на базе корпусов. Россия встала на этот «корпусный» путь с некоторым опозданием, но движется по нему очень быстро. Появляется большое число публикаций, посвященных созданию и использованию корпусов. Начинают выходить учебные пособия по корпусной лингвистике на русском языке (Захаров 2005, Гвишиани 2008, Захаров, Богданова 2011, 2013, Грудева 2012).

[1] Настоящая статья является расширенной и обновленной версией публикации на английском языке [Zakharov 2013].

Безусловно, ключевым моментом развития корпусной лингвистики в нашей стране стало создание Национального корпуса русского языка (2004) (http://ruscorpora.ru). О востребованности корпуса свидетельствуют многочисленные исследования и публикации, подготовленные на его основе. Часть из них описывается на сайте корпуса в разделе Studiorum (http://studiorum.ruscorpora.ru/). Однако корпуса русского языка сегодня создаются не только «в недрах» НКРЯ и не только в России. И если о НКРЯ сегодня написано много, то о других корпусах информации не хватает. С корпусной проблематикой во всем ее многообразии можно ознакомиться по материалам конференций «Корпусная лингвистика» (см. на сайте http://corpora.phil.spbu.ru), проводимой кафедрой математической лингвистики СПбГУ (Труды международной конференции «Корпусная лингвистика») и «Диалог» (http://dialog-21.ru) (Компьютерная лингвистика и интеллектуальные технологии: Материалы ежегодной Международной конференции «Диалог»). Однако был бы полезен единый очерк, дающий общее представление о корпусах русского языка. Есть, правда, обзоры корпусов славянских языков, включая русский [Резникова, Копотев 2005; Резникова 2009], то они не полны — за прошедшие годы корпусная лингвистика и «корпусостроение», естественно, ушли вперед.

Задача настоящего очерка — дать краткий общий обзор различных корпусов русского языка в России и за рубежом в их многообразии. За пределами данного обзора остаются детальное описание функциональных возможностей корпусных служб, вопросы использования корпусов для исследований по русскому языку, в целях обучения, настройки лингвопроцессоров и решения разнообразных лингвистических задач на базе корпусов.

1. Первые корпуса русского языка

Первый русскоязычный корпус был создан в 1980-е гг. в Университете Упсалы (Швеция). Однако еще до первых русскоязычных корпусов в полном смысле этого слова в 1960–70-е гг. был создан ***Частотный словарь русского языка*** под руководством Л.Н. Засориной [Засорина 1977], построенный на основе примитивных текстовых файлов, фактически, электронных словарных карточек, объемом в 1 млн словоупотреблений, включавших в себя лексику четырех жанров в примерно равной пропорции: общественно-политические

тексты, художественную литературу, научные и научно-популярные тексты из разных тематических областей и драматургию.

В процессе создания этих файлов — сегодня мы бы сказали, корпуса — решались все проблемы современной корпусной лингвистики, которые обсуждались и обсуждаются при создании полноценных корпусов:

- репрезентативность,
- сбалансированность,
- графематический анализ,
- нормализация,
- лемматизация.

Так что фактически это был первый корпус русского языка, не дошедший до наших дней.

В 1985 г. в СССР по инициативе академика А.П. Ершова (доклад 1978 г., см. [Ершов 1982]) были начаты работы по созданию ***Машинного фонда русского языка*** [Андрющенко 1989; Машинный фонд русского языка 1986]. Это был грандиозный проект. В создании фонда принимали участие более 40 организаций-соисполнителей, среди них Институт русского языка, Московский, Ленинградский (Санкт-Петербургский), Харьковский, Гродненский, Сыктывкарский и Саратовский университеты и др. В задачи фонда входило накопление на машинных носителях и в базах данных текстовых, лексикографических и грамматических источников, необходимых для научного изучения русского языка и для осуществления прикладных разработок. Одновременно создавались программные средства для проведения лингвистических исследований. В 1985–1992 гг. были осуществлены разработка концепции и архитектуры Машинного фонда русского языка, разработка концепции терминологического банка данных, введены в компьютер текстовые источники русской литературы XIX–XX вв., главнейшие словари русского языка, краткая академическая грамматика, созданы текстовые корпуса поэзии, художественной прозы, общественно-политических и технических текстов. Однако с началом перестройки и в новых экономических условиях после 1991 г. работы по созданию фонда постепенно стали сокращаться и, наконец, совсем прекратились.

Уппсальский корпус русского языка (Uppsal'skij korpus russkix tekstov), созданный, как уже было сказано, в Университете Уппсалы, состоит из 600 текстов, его объем составляет 1 млн словоупотребле-

ний, поровну распределенных между образцами специальной и художественной литературы. По замыслу создателей, корпус должен был отражать современное состояние русского языка того периода. Цель формирования корпуса заключалась в том, чтобы представить, в первую очередь, письменный литературный язык.

В корпус отбирались тексты с 1985 по 1989 г. и художественные тексты с 1960 по 1988 г. Корпус составляли не фрагменты текстов, как в Брауновском корпусе, а целые тексты. В аннотации к корпусу отмечается, что среди специальных текстов особое внимание уделено более важным, с точки зрения создателей корпуса, темам, а среди художественных текстов предпочтение отдавалось более известным авторам. Тексты в корпусе записывались латиницей. Фрагмент корпуса выглядит следующим образом:

```
&Perestrojka vse glubhe zatragivaet hiznennye
interesy millionov, obqestva v celom. Estestvenno,
l~di xot„t lu†we u„snit' sut' i nazna†enie
processov obnovleni„, blihnie i dal'nie celi
preobrazovanij, opredelit' svoe otnowenie k nim
```

Уппсальский корпус сейчас входит в так называемые ***«Тюбингенские корпуса русских текстов»***, созданные в рамках работы специального научно-исследовательского сектора SFB 441 Тюбингенского университета в 1990–2000-е гг. В целом тюбингенские корпуса представляют собой несбалансированное, морфологически размеченное (частично) собрание разнородных текстов, сегодня уже не соответствующее понятию лингвистического корпуса. Разметка была осуществлена при помощи статистического теггера (TnT). Поиск может производиться как по словоформам, так и — для размеченных текстов — по морфологическим тегам. Возможен вывод текста вместе с разметкой. Для ввода поискового выражения и вывода найденного текста можно выбрать одну из следующих кодировок: кириллицу (KOI8 или Windows-1251) или транслитерацию латинскими буквами. Поиск осуществляется при помощи программы CQP, представляющей собой систему для управления большими корпусами, разработанную Институтом машинной обработки языка Штутгартского университета.

Компьютерный корпус текстов русских газет конца XX в. (http://www.philol.msu.ru/~lex/corpus/) был создан на Филологическом факультете МГУ в 2000–2002 гг. в Лаборатории общей и

компьютерной лексикологии и лексикографии под руководством А.П. Поликарпова. Подбор обширного газетного материала для корпуса (тексты общим объемом более 11 млн словоупотреблений) был осуществлен на основе принципов включения в него полных номеров 13 российских газет на русском языке за отдельные даты 1994–1997 гг. (23110 текстов), представленности в нем газет ежедневных и неежедневных («МН», «Новая газета»), «левых» («Завтра», Правда», «Правда-5») и «правых», центральных и местных, общих и профессионально ориентированных (например, «Литературная газета»). Эти принципы позволяют получить относительно объективную и надежную картину соотношения в газетном материале текстов различного типа (например, различных жанров и жанровых типов), их единиц и отношений между ними.

Корпус создан, анализируется и управляется на основе системы Диктум-1, разработанной в Лаборатории общей и компьютерной лексикологии и лексикографии МГУ. С помощью этой системы тексты и единицы корпуса автоматически и полуавтоматически маркируются различного рода маркерами: тексты – маркерами газеты-источника, объема текста, его жанра, даты публикации и т. п.; словоупотребления – маркерами грамматических, лексических, морфемных и иных категорий. С подробным описанием корпуса и результатами частотно-статистического анализа газетных текстов можно ознакомиться на сайте (http://www.philol.msu.ru/~lex/corpus/corp_descr.html).

Однако в открытом доступе исследователям доступна лишь незначительная часть корпуса. При подготовке демонстрационного варианта корпуса для Интернета был выделен фрагмент корпуса общим объемом более 200 тыс. словоупотреблений, проведена автоматическая лемматизация и морфологическая квалификация словоупотреблений корпуса, а также морфемная сегментация словоформ и лексем. В настоящее время корпус представляет лишь ограниченный интерес.

2. Современные корпуса русского языка

В начале этого века была осознана необходимость создания современного представительного корпуса русского языка. Как пилотные версии его можно рассматривать проекты корпусов Русский стандарт и БОКР (Большой Корпус русского языка) (http://

bokrcorpora.narod.ru/) (http://corpus.leeds.ac.uk/serge/bokrcorpora/index.html), который позиционировался как русский аналог Британского национального корпуса.

При этом следует упомянуть также большой корпус, созданный С. А. Шаровым в конце 1990 — начале 2000-х гг., недоступный в то время как корпус, но использованный как источник нового частотного словаря русского языка (http://www.artint.ru/projects/frqlist.php). Корпус включал в себя подборку прозы, политических мемуаров, газет и научно-популярной литературы того времени (около 40 миллионов слов, проза составляет примерно чуть больше половины объема). Все тексты корпуса были написаны на русском языке в промежутке между 1970 и 2002 гг; большинство относится к периоду 1980 1995 гг., газетный корпус охватывал период 1997–1999 гг..

И, наконец, с апреля 2004 г. в открытом доступе появился ***Национальный корпус русского языка*** (НКРЯ) (http://ruscorpora.ru/), отвечающий критерию репрезентативности и всем другим требованиям, предъявляемым к современным корпусам. Объем корпуса составляет более 500 млн словоупотреблений (по данным на сентябрь 2014 г.). Жанровое разнообразие составляющих его текстов, которые относятся ко всем основным сферам использования русского языка (научной, официально-деловой, публицистической, церковно-богословской, художественной, разговорно-бытовой, включая устную и электронную коммуникацию) обеспечивает его сбалансированность. Разметка корпуса и подкорпусов, программное и документационное обеспечение позволяют решать сложные лингвистические задачи.

В состав корпуса входят основной корпус, в том числе подкорпус со снятой омонимией, газетный корпус, диалектный, синтаксический, обучающий, поэтический, устный, мультимедийный, исторический и параллельные корпуса. Структурно-статистическое наполнение корпуса см. http://ruscorpora.ru/corpora-stat.html.

НКРЯ хорошо известен и подробно описан в многочисленных публикациях (прежде всего [Национальный корпус русского языка 2005; Национальный корпус русского языка 2009], см. также http://ruscorpora.ru/corpora-biblio.html). Здесь в числе целого ряда его особенностей, отличающих его от других корпусов и в особенности национальных корпусов, отметим в первую очередь три: *мультиме-*

дийный подкорпус (http://ruscorpora.ru/search-murco.html), *семантическая разметка* (http://ruscorpora.ru/corpora-sem.html) и *статистический инструмент проведения диахронических исследований «Графики»* (http://ruscorpora.ru/ngram.html), подобный сервису Google Books Ngram Viewer.

Мультимедийный русский корпус (МУРКО) [Гришина 2005] образован фрагментами кинофильмов 1930–2000-х гг. и другими материалами, представленными в виде параллельных видеоряда, аудиоряда и текстовой расшифровки звучащей речи, а также специально размеченных наблюдаемых в кадре жестов. Возможен поиск не только по произносимому тексту, но и по жестам (кивание головой, похлопывание по плечу и т. п.) и типу речевого действия (согласие, ирония и т. п.). В поисковой выдаче видеофрагменты доступны для просмотра и прослушивания.

Семантическая разметка [Кустова и др. 2006] приписывает единицам текста один или несколько семантических и словообразовательных признаков, например, 'лицо', 'вещество', 'пространство', 'скорость', 'движение', 'обладание', 'свойство человека', 'диминутив', 'отглагольное имя' и т. п. Используется фасетная классификация, при которой одно слово может попадать в несколько классов. В основу семантической разметки положена система классификации русской лексики, принятая в базе данных «Лексикограф», которая разрабатывалась с 1992 г. в Отделе лингвистических исследований ВИНИТИ РАН под рук. Е. В. Падучевой и Е. В. Рахилиной. И по указанным признакам, разбитым по нескольким категориям: разряд, таксономия, мереология, оценка, словообразование и т. п. — в корпусе осуществляется поиск. Можно искать контексты по основному значению слова или по вспомогательному. Можно найти все приставочные глаголы движения (отметить признак «движение» в категории «Таксономия», грамматический признак «глагол» и «тип морфемы: префикс» в категории «Словообразование») и т. д.

На *сервисе «Графики»* мы остановимся отдельно в разделе, посвященном диахроническим корпусам (п. 6.2).

Корпус русского литературного языка (1 млн словоупотреблений) (http://www.narusco.ru/) задумывался как интересный проект: тексты только второй половины XX в., сбалансированный жанровый состав, восстановленная буква «ё», проставленные словесные ударения. Однако практически корпус «не живет». Вначале разра-

ботчики не справились с задачей морфологической разметки, а затем просто перестали его развивать и поддерживать. И, как следствие, корпус доступен в сети, но не используется.

Открытый корпус (OpenCorpora) (http://opencorpora.org/) — это проект по созданию размеченного корпуса текстов силами лингвистического Интернет-сообщества. Корпус доступен бесплатно и в полном объёме (под лицензией CC-BY-SA). Это система, предназначенная для ввода текстов с лингвистической разметкой, предоставляющая интерфейс разметки, редактирования и исправления ошибок, инструмент для контроля качества и формат разметки для русского языка. Объём корпуса 1,3 млн. токенов, подкорпус со снятой омонимией — 18 тыс. токенов (сентябрь 2014 г.).

Работы начались в 2009 году, и сегодня можно констатировать, что первоначальные задачи: создать морфологически, синтаксически и семантически размеченный корпус текстов на русском языке, в полном объёме доступный для исследователей и редактируемый пользователями, — не решены. Тем не менее, корпус сыграл и играет положительную роль в развитии инструментов корпусной лингвистики. Кроме свободных ресурсов, первым среди которых следует назвать постоянно пополняемый морфологический словарь, нужно упомянуть, что корпус используется как учебная площадка студентами, обучающимися в области автоматизированной обработки текста.

Хельсинский аннотированный корпус (ХАНКО) (см. [Резникова, Копотев 2005]) создан в Хельсинском университете в начале 2000-х гг. как часть проекта «Функциональный синтаксис русского языка» (рук. проф. А. Мустайоки) и постоянно развивается. Объем корпуса — 100 тыс. словоупотреблений. Доступен в Интернете по адресу: http://www.ling.helsinki.fi/projects/hanco/. Корпус создан на основе статей первых четырех номеров журнала «Итоги» за 2001 г. В корпусе реализованы морфологическая и синтаксическая разметки и, соответственно, морфологический и синтаксический поиск. Особенность корпуса — тщательно проработанный формат лингвистического описания данных и полная визуальная (ручная) проверка результатов автоматической разметки, имеющая следствием полное снятие грамматической омонимии, там, где она может быть снята человеком. Синтаксическая разметка корпуса должна совмещать разметку в терминах членов предложения (уже реализована

см. набор синтаксических параметров на рис. 1) и в терминах деревьев зависимостей [Мустайоки и др. 2005].

Параметры предложений	**Параметры клауз**
Простое □ Простое предложение Сложное □ Предложение с сочинительной связью □ Предложение с подчинительной связью □ Предложение с бессоюзной связью	Роль □ Самостоятельная □ Главная □ Подчиненная Структура □ Двусоставная □ Односоставная □ Фразеологизированная Элиптическая □ Эллиптическая
Выбор синтаксических параметров	**Параметры клаузы для слова 1**
Части именного сказуемого □ Связочная часть □ Присвязочная часть Главные члены предложения □ Главный член односоставного предложения □ Подлежащее (= подлежащее двусоставного предложения) □ Сказуемое (= сказуемое двусоставного предложения) Второстепенные члены предложения □ Дополнение □ Определение □ Обстоятельство Не является членом предложения Обращение	Роль □ Самостоятельная □ Главная □ Подчиненная Структура □ Двусоставная □ Односоставная □ Фразеологизированная Элиптическая □ Эллиптическая

Рис. 1. Набор синтаксических параметров в корпусе ХАНКО.

Кроме того, размечены многословные устойчивые обороты (примерно 2000 единиц). Планируется семантическая разметка в терминах семантических категорий. Цели и особенности корпуса описаны на веб-странице проекта http://www.helsinki.fi/venaja/russian/e-material/hanco/index.htm. Корпус преследует в первую очередь учебные цели.

Корпуса университета г. Лидс (Великобритания). В течение 2000-х гг. в университете г. Лидс в Центре переводческих исследований С.А. Шаровым создано большое количество корпусов для разных языков (английский, арабский, китайский, французский, немецкий, итальянский, японский, испанский, польский и др.) (http://corpus.leeds.ac.uk/). Среди них имеются и корпуса русского языка (http://corpus.leeds.ac.uk/ruscorpora.html). Это, в первую очередь, две версии Национального корпуса русского языка объемом в 116 млн словоупотреблений (с разной морфологической разметкой). На основе этого подмножества НКРЯ был создан Частотный словарь русского языка [Ляшевская, Шаров 2009]. По объему эта версия уступает НКРЯ в его сегодняшнем состоянии, но интересна функциональными возможностями корпусной системы, в среде которой функционирует этот русскоязычный массив. Кроме того, на сайте Университета Лидса представлены и другие русскоязычные корпуса: корпус русских газет (2001–2004 гг., 77 млн словоупотреблений), корпус русских текстов из Интернета (198 млн словоупотреблений), бизнес-корпус (15 млн словоупотреблений), корпус на основе русской Википедии (178 млн токенов), устный корпус (русскоязычные веб-форумы, 1 млн), корпус научных текстов (5,1 млн), корпус из текстов «Живого журнала» (700 млн) и большой 2-миллиардный русскоязычный корпус, созданный на базе Интернета по технологии wacky (эта проблематика — новая технология создания корпусов — находится за пределами данной статьи; см. материалы семинаров https://sigwac.org.uk/).

Поисковый интерфейс Leeds CQP базируется на корпусном менеджере IMS Corpus Workbench и предоставляет интересные возможности (рис. 2).

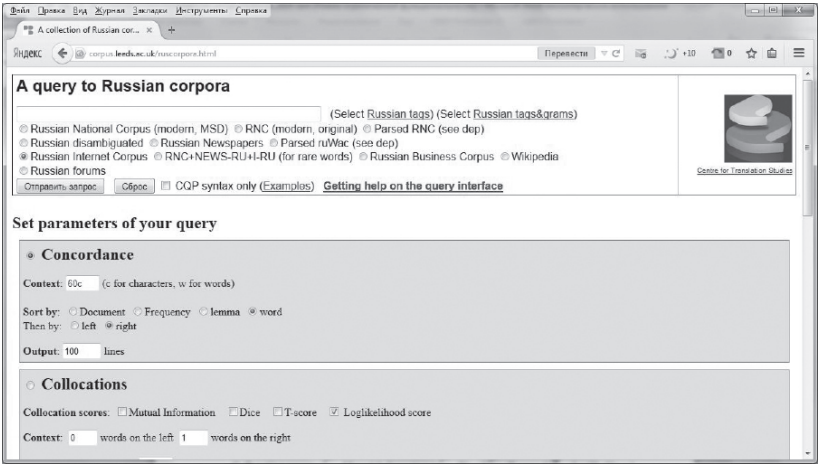


Рис. 2. Интерфейс доступа к русским корпусам в Университете Лидса.

Он позволяет вести точный лексико-грамматический поиск на основе специального языка запросов, в том числе с использованием языка регулярных выражений. Имеются способы управления выходным интерфейсом, формой представления результатов поиска. Можно также получить списки коллокаций, вычисленных и упорядоченных на основе ассоциативных мер MI, Dice, T-score и Log-likelihood. Там же имеется коллекция различных программных средств для обработки корпусных текстовых данных (http://corpus.leeds.ac.uk/tools/).

Новая система IntelliText (http://corpus.leeds.ac.uk/it) имеет более развитый и дружественный интерфейс (рис. 3, 4) и расширяет набор функций, в частности, предоставляется возможность создавать свои корпуса.

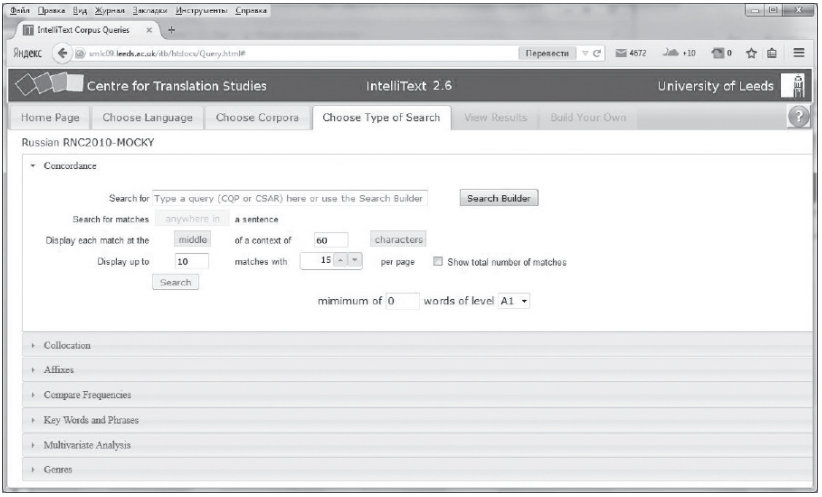


Рис. 3. Интерфейс поиска системы IntelliText: конкорданс

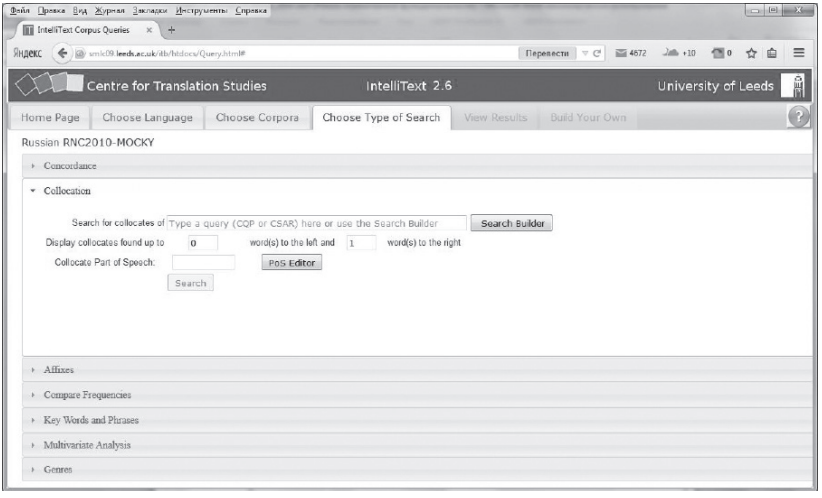


Рис. 4. Интерфейс поиска системы IntelliText: коллокации.

Корпус Библиотеки Мошкова. На сайте группы АОТ (http://aot.ru/search1.html) имеется большой корпус русских текстов объемом 680 млн словоупотреблений (844 млн токенов), созданный А. В. Сокирко по текстам из библиотеки Мошкова. Можно искать по лексическим единицам с учетом частей речи и морфологических характеристик, используя мощный язык запросов корпусного менеджера DDC (http://aot.ru/concor.html). Там же на сайте АОТ имеется сервис поиска биграмм (54 млн), вычисленных по мере MI, созданный А.В. Сокирко и А.Н. Авериным по текстам корпуса Библиотеки Мошкова (http://aot.ru/demo/bigrams.html).

Корпуса в системе Sketch Engine. Английская лингвистическая служба Lexical Computing Ltd. (A. Kilgarriff) предоставляет на коммерческой основе доступ к 329 корпусам (сентябрь 2015 г.) различных языков мира. Среди них имеется ряд корпусов русского языка и в их числе большой корпус, созданный из текстов Интернета по технологии wacky объемом 18 млрд токенов (14,5 млрд текстоформ) (ruTenTen 2011) (рис. 5, 6). Английские исследователи совместно с чешскими разработчиками из Университета им. Масарика (Брно) создали многофункциональную корпусную систему Sketch Engine (http://sketchengine.co.uk/) [Kilgarriff 2004], обладающую многими уникальными возможностями. Помимо стандартного поиска с выдачей конкорданса она выдает списки коллокаций, сформированные на основе 7 мер ассоциации, списки коллокаций по отдельным синтаксическим моделям (word sketches), формирует частотные списки слов, словоформ и тегов, группирует лексические единицы в лексико-семантические поля с внутренней кластеризацией и указанием силы связи между лексемами.

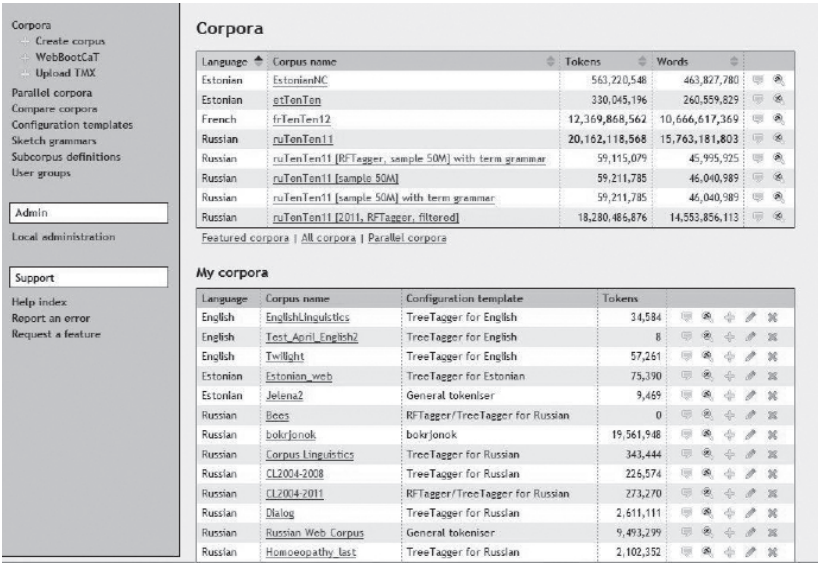


Рис. 5. Интерфейс доступа к корпусам Sketch Engine.

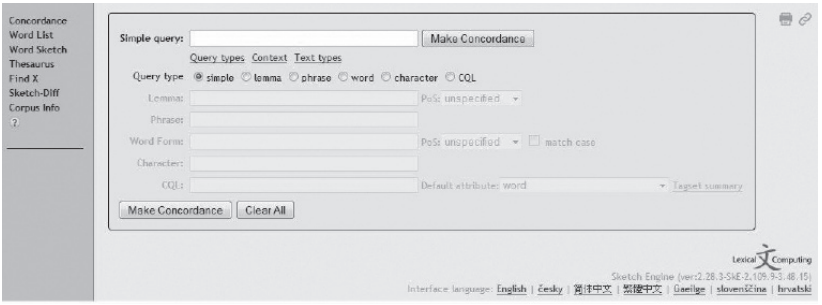


Рис. 6. Интерфейс системы Sketch Engine: конкорданс.

Корпуса в системе Aranea (A Family of Comparable Gigaword Web Corpora). Это система псевдопараллельных корпусов для 14 европейских языков, включая русский (15-м языком выступает китайский), созданная в Университете им. А. Коменского в Братиславе (http://ucts.uniba.sk/). Все языки представлены корпусами двух ти-

пов, созданными по технологии wacky: Maius (1200 млн токенов; как правило, из этого числа около 1000 млн токенов представляют собой слова, или в другой терминологии, текстоформы) и Minus (120 млн токенов, чуть более 90 млн текстоформ). Для некоторых языков имеются региональные варианты, так, англоязычных корпусов насчитывается 6: Araneum Anglicum, Araneum Anglicum Asiaticum, Araneum Anglicum Africanum, каждый соответственно, Maius и Minus. То же самое относится и к русскоязычным корпусам, которых тоже 6: Araneum Russicum Maius & Minus (русский универсальный), Araneum Russicum Russicum Maius & Minus (русский на основе текстов с сайтов с доменом .ru), Araneum Russicum Externum Maius & Minus (русский на основе текстов с доменами, отличными от домена .ru). Подробнее см. [Benko 2013]. Корпуса поддерживаются системой NoSketch Engine, отличающейся от Sketch Engine уменьшенной функциональностью, но зато бесплатно распространяемой (http://nlp.fi.muni.cz/trac/noske).

3. Лингвистические ресурсы русской устной речи

Выделяется 2 основных типа корпусов: 1) корпуса текстов (расшифровок) устной речи, которые аннотируются по правилам, схожим с разметкой текстовых корпусов, 2) мультимедийные (звуковые, речевые) корпуса или корпуса звучащей речи, единицей описания которых являются собственно звукозаписи устной речи, которые расшифровываются (транскрибируются) и далее аннотируются в соответствии с поставленными перед исследователями задачами.

Наиболее представительным ресурсом первого типа является ***Корпус устных текстов НКРЯ*** объемом более 10 млн словоупотреблений, хорошо известный по публикациям (см., например [Гришина 2005; Гришина, Савчук 2009]), поэтому здесь мы его не рассматриваем.

Корпуса звучащей русской речи, или звуковые корпуса активно создаются последние 50 лет в научно-исследовательских центрах и лабораториях для поддержки задач речевых технологий (синтеза и распознавания речи) (например, разработки НИИ «Дальняя связь», Института физиологии им. Павлова в советские годы, в наше время — Центра речевых технологий и др.)

Известны записи звучащей речи, ориентированные на проведение фундаментальных научных исследований, например, Фонети-

ческий Фонд русского языка, разрабатываемый на кафедре фонетики СПбГУ (ЛГУ) в конце прошлого века [Бондарко 1988] и многочисленные лингвистические ресурсы, содержащие аудиоматериалы по языкам малых народов Российской Федерации. Большинство таких ресурсов ориентированы на запись ограниченного набора фраз (текстов, слов), осуществляемых в лабораторных условиях или дистантно (например, по телефону).

Из разработок последних лет выделяются два ресурса — корпус звучащей речи «Рассказы о сновидениях» и корпус повседневной русской речи «Один речевой день».

«Рассказы о сновидениях» представляет собой небольшой (около 2 часов звучания) и жанрово ограниченный (рассказы детей о своих сновидениях), но хорошо проаннотированный и проанализированный ресурс, ориентированный на исследование русской устной речи и дискурса [Кибрик, Подлесская 2009]. В последние годы ресурс расширяется за счет привлечения других звуковых материалов (http://spokencorpora.ru).

Наиболее представительным корпусом повседневной устной русской речи на настоящий момент является речевой корпус ***Один речевой день*** (ОРД), разрабатываемый на Филологическом факультете СПбГУ [Asinovsky et al. 2009]. Корпус создается с целью изучения реальной речи носителей языка в естественных условиях коммуникации, и в этом его отличие от абсолютного большинства речевых корпусов, записанных в лабораторных и других специальных условиях.

Первая серия звукозаписей осуществлена осенью 2007 г. Для этого была отобрана группа информантов из 30 человек, представляющих разные социальные и возрастные слои населения Санкт-Петербурга и давших согласие прожить один день с «диктофоном на шее». Информанты получили подробный инструктаж о методике проведения звукозаписи своих речевых контактов в течение суток, заполнили социологические анкеты и прошли психологическое тестирование. Помимо речи информантов, в корпусе представлены записи их коммуникантов (родственников, друзей, коллег, знакомых и незнакомых), среди которых были люди самого разного возраста и разных специальностей. Общая длительность записанного материала — более 500 часов. Расшифровке и многоуровневой разметке подвергнуто 50 часов (по данным на сентябрь 2014 г.). Сплошное

аннотирование проводится, в частности, по следующим базовым уровням: 1) реплики (фразы), 2) говорящий, 3) невербальные речевые события, 4) фонетический комментарий, 5) фразовый комментарий, 6) мини-эпизод речевой коммуникации и др. [Шерстинова и др. 2009].

Корпус постоянно расширяется, проводятся новые звукозаписи, осуществляется их расшифровка и увеличивается количество уровней аннотирования. На конец 2016 года планируется увеличение объема текстовых расшифровок до 1 млн словоупотреблений. Поиск по анонимизированным расшифровкам и материалам корпуса ОРД будет доступен на сайте http://www.ord-corpus.spbu.ru, начиная с 2016 г., публикация самих звукозаписей не планируется ввиду большого объема личной и частной информации, представленной в данном ресурсе.

Данный звуковой корпус позволяет изучать русскую повседневную речь на всех лингвистических уровнях. Например, можно исследовать лингвистическую динамику записанного материала: изучать временные ряды количественных переменных с помощью стандартных статистических методов и анализировать частотные ряды (лексики, грамматических и, в частности, синтаксических структур, семантики или разговорных тем, тех или иных акустических явлений или просодических контуров) в зависимости от времени суток и условий коммуникации в самом широком понимании этого термина. Можно решать множество других задач, таких как анализ влияния профессии на бытовую жизнь человека, получение информации о среднем артикуляционном темпе спонтанной речи носителей русского языка.

Результаты исследований, проведенных на материале корпуса ОРД, опубликованы в многочисленных публикациях [см., напр., Асиновский и др. 2010].

Мультимодальные корпуса включают видеозапись участников коммуникации, поэтому с их помощью можно исследовать эмоции.

Мультимедийный русский корпус в составе НКРЯ мы уже упоминали, и с ним можно познакомиться по публикациям [Гришина 2005; 2009], современное состояние и перспективы развития описаны в статье Е.А. Гришиной в настоящем сборнике.

Русскоязычный эмоциональный корпус (REC), размеченный с учетом данных о мимике, движениях рук, бровей и т. д., позволя-

ет изучить стратегии эмоционального взаимодействия и конфликта, непрерывное коммуникативное поведение, хезитации и речевые сбои и др. (http://www.harpia.ru/rec) [Котов, Гопкало 2011]. Он может также использоваться как материал для обучения работников клиентских служб или как база данных эмоциональных реакций для мультипликаторов и режиссеров.

В Иркутском государственном лингвистическом университете идет работа по созданию ***Учебного Мультимодального Корпуса*** (УМКО), включающего видеозаписи неподготовленных учебных диалогов носителей и «неносителей» русского и китайского языков по определенным темам, размеченных в программе ELAN и представленных также в виде параллельных корпусов, выровненных по смысловым блокам внутри диалогов [Богданова 2013]. Например, диалог носителей русского языка на русском языке сопоставляется с диалогом на ту же тему на русском языке («Знакомство», «Регистрация в аэропорту» и др.) китайцев, изучающих русский язык. Данный корпус предназначен, в первую очередь, для лингводидактических целей, так как позволяет выявить типичные ошибки и найти пути их устранения в ходе учебных занятий и самостоятельной работы студентов.

Назовем в этом разделе еще ***Корпус устных рассказов*** на русском языке [Николаева 2009], стимулом для которых послужил 6-минутный видеосюжет, так называемый «Фильм о грушах» (“Pear film”). Об этом фильме было записано 8 рассказов, сделанных студентами МГУ. Общая длительность звучания — около 2 часов; объем корпуса — около 14 тысяч словоупотреблений. Всего в корпусе было 595 элементарных дискурсивных единиц, которые обычно совпадают с простым предложением, и 327 иллюстративных жестов, которые, в соответствии с подходом Г.Е. Крейдлина, понимаются как носители информации, выступая в качестве знаковых кинетических единиц выражения и передачи информации. На примере из корпуса устных рассказов исследователям удалось показать, как отдельные признаки жестов и положения рук могут добавлять дополнительную информацию касательно организации дискурса, состояния говорящего и процесса коммуникации. Так, изменение положения покоя рук между жестами достаточно последовательно указывает на границу между сегментами нарратива. Данный пример демонстрирует предоставляемые мультимедийным корпусом возможности изучения связи структуры устного нарратива и иллюстративных жестов.

4. Специальные корпуса текстов

Специальные корпуса текстов — это сбалансированные корпуса, как правило, небольшие по размеру, подчиненные определенной исследовательской задаче и предназначенные для использования преимущественно в целях, соответствующих замыслу составителя. Второе понимание специальных корпусов — это тексты по определенной тематике, относящиеся к какой-то предметной области, или тексты определенного типа (например, патенты).

Примером специального корпуса текстов может быть ***Санкт-Петербургский учебный корпус текстов школьников, изучающих английский язык*** (SPbEFLLC), созданный на кафедре прикладной лингвистики РГПУ им. А. И. Герцена [Камшилова 2012]. Основной целью его создания было исследование особенностей английских текстов, порождаемых русскими школьниками. Аутентичный текстовый материал был собран в школах Санкт-Петербурга в период с ноября по декабрь 2007 г. Авторами текстов являются 78 учеников 9–11 классов, предварительно прошедших тестирование. Уровень владения английским языком был определен как средний/intermediate (26%) и выше среднего/upper-intermediate (74%). Размер данного корпуса составляет около 50 тыс. словоупотреблений.

Исследование на базе корпуса показало, что систематическое предпочтение максимально простых структур развернутым и более естественным моделям стандартного английского языка приводит к так называемой «структурной бедности» речевых произведений неносителей языка. В репертуаре грамматических структур, обнаруженных в SPbEFLLC, есть такие, которые представляют собой случаи «переходной грамматики» (интеръязыка), выражающиеся, например, в нарушении правил наполнения компонентов базовых структур. Так формируется ядро грамматики EFL (English as a Foreign Language), которое не совпадает с базовыми грамматическими структурами литературного английского языка. На основании корпусных данных авторы высказывают предположение о том, что складывающиеся нормы «глобального английского» во многом опираются на «окаменевшие» модели интеръязыка.

Сложным объектом с точки зрения создания и стандартизации являются исторические корпуса, такие как, например, ***Санкт-Петербургский корпус агиографических текстов*** (СКАТ) [Герд и др. 2004], доступный на сайте http://project.phil.pu.ru/skat. СКАТ —

это электронный корпус текстов по памятникам древнерусской агиографической литературы XV–XVII вв., созданный на кафедре математической лингвистики филологического факультета СПбГУ. Язык агиографических произведений во многом обусловил судьбу и характер русского литературного языка XV–XVII вв. Отображение этого языка является первостепенной задачей создаваемого корпуса текстов русских житий того времени, что достигается, в частности, за счет широкого географического охвата территорий, где в разное время создавались памятники русской агиографии. К настоящему времени корпус охватывает 60 житий, их общий объем — более 500 тыс. словоупотреблений.

Корпус является базой широкомасштабного проекта по изданию уникальной серии текстов «Памятники русской агиографической литературы». Для представления рукописей в корпусе была разработана система отображения древнерусской графики, которая позволяет воспроизводить текст с высокой степенью приближения к оригиналу. Отображаются графические начертания всех древнерусских букв и их семантически значимых вариантов. Разработана специальная программа, позволяющая получать к введенным текстам (к каждому в отдельности или к нескольким вместе) указатели словоформ, то есть списки словоформ с их адресами (номерами листов и строк) в рукописях, фактически, конкорданс.

В настоящее время осуществляется грамматическая разметка представленных в корпусе житий [Алексеев и др. 2011]. В формальном плане разметка корпуса основывается на международных нормах оформления электронных изданий текста, в частности, на рекомендациях проекта Text Encoding Initiative (TEI), однако содержит также и дополнительные элементы, которые необходимы для адекватного отображения особенностей русского рукописного текста.

Разработан формат представления грамматической информации в структуре XML. Для каждой словоформы текста указывается ее частеречная принадлежность и приводятся значения всех релевантных грамматических категорий. При этом учитываются некоторые особенности морфологии текстов XV–XVI вв. Дело в том, что тексты житий написаны на церковнославянском языке и, с одной стороны, сохраняют архаичные формы, уже вышедшие или выходящие из употребления в древнерусском языке того периода, а с другой

стороны, отражают живые языковые процессы, такие как смешение типов склонений существительных, формирование категории одушевленности, утрата двойственного числа, перестройка системы прошедших времен глагола, образование деепричастия и т. д. Поэтому в формате грамматической разметки предусмотрена возможность отражения переходных явлений: через косую черту приводятся ожидаемое значение соответствующей категории (тип склонения, падеж и т. п.) и реально встретившееся в тексте. Представленную таким образом информацию можно подвергать разным видам автоматического анализа, например, можно оценить степень архаичности или новизны того или иного текста.

Формат синтаксической разметки текстов житий ориентируется на систему аннотирования, принятую в НКРЯ в синтаксически размеченном корпусе СинТагРус [Алексеева 2014]. Очевидно, что синтаксис церковнославянского языка, на котором писались жития в Древней Руси, отличается от современного синтаксиса. Поэтому был составлен свой набор синтаксических отношений, в который вошли 39 отношений, используемых в СинТагРус, и 25 отношений, выделенных на основании анализа конкретных текстов житий и научной литературы. Работа над форматом продолжается.

В настоящее время на сайте корпуса в открытом доступе представлена лишь часть корпуса, а именно, в формате PDF представлено 13 житий, 10 из них представлены также и в формате XML.

Специальные корпуса текстов могут быть востребованы не менее, чем национальные. Любой *отраслевой специальный корпус текстов* может пригодиться и в данной конкретной отрасли (например, кораблестроение, металлы, экология, навигация и т. д.), и в смежных областях, поскольку он дает специалисту самое главное — термины в их профессиональном конкретном окружении, позволяет отследить изменения в терминологии, включая появление новых терминов. В СПбГУ ведется работа по созданию семейства представительных специальных корпусов общим объемом более 20 млн. словоупотреблений, охватывающих широкую тематику (компьютерная лингвистика, гомеопатия, радиотехника, шахматы, футбол, пчеловодство, растения, научно-популярные (журнал «Наука и жизнь»), путешествия) [Митрофанова и др. 2014].

5. Параллельные корпуса

По критерию «параллельность» корпуса делятся на одноязычные, двуязычные и многоязычные. В одноязычных корпусах противопоставляются диалекты, варианты языка. Двуязычные и многоязычные корпуса можно разделить на два основных типа:

1) корпуса, представляющие множество текстов-оригиналов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков;

2) корпуса, объединяющие тексты из одной и той же тематической области, независимо написанные на двух или нескольких языках; такие корпуса помогают в работе с терминологией и часто используются переводчиками.

Корпуса обоих типов используются в целях разработки эффективных методов перевода, в том числе, машинного, а также для сравнительных исследований языков (в области лексикологии, грамматики, стилистики, переводоведения и т.д.). Естественно, среди параллельных корпусов есть корпуса, где в качестве одного из языков представлен русский. Не касаясь проблем создания таких корпусов (подбор материала, выравнивание и т.п.), просто перечислим известные параллельные корпуса, где есть тексты на русском языке.

Прежде всего, это ***параллельные корпуса НКРЯ***, где на сентябрь 2014 г. для русского языка имеются следующие «параллельные» языки: английский, немецкий, французский, испанский, итальянский, польский, украинский, белорусский, армянский, болгарский и латышский языки. Подробнее см. [Добровольский и др. 2005]. При поиске используется стандартный интерфейс поиска в основном корпусе.

Корпус PARASOL (PARAllel corpus of Slavic and Other Languages) — параллельный корпус современных текстов, созданный в Университете Регенсбурга [Waldenfels 2006, 2011]. По состоянию на март 2014 г. корпус насчитывал более 27 млн токенов на 31 языке. Корпус размечен морфологически для большинства языков и доступен онлайн (http://slavist.de/Ursynow/).

Корпус InterCorp — часть Чешского национального корпуса (www.korpus.cz/intercorp) [Čermák, Rosen 2012]. В корпусе содержатся тексты на 38 языках, включая русский, образующие «параллель» с текстами на чешском языке. Объем русской части — 13,4 млн токенов. Все тексты выровнены и морфологически размечены. Для

поиска используется единый поисковый интерфейс Чешского национального корпуса. Корпус доступен онлайн для зарегистрированных пользователей.

Корпус PARUS (PAralelní RUsko-Slovenský korpus) — параллельный корпус современных текстов, созданный в Институте языкознания им. Л. Штура Словацкой академии наук [Гарабик, Захаров 2006]. Корпус состоит из переводов с русского на словацкий и со словацкого на русский. Тексты выровнены по предложениям. Все тексты корпуса морфологически размечены. Версия корпуса 2.0 (январь 2014 г.) насчитывает 8,45 млн токенов, 4,2 млн токенов в словацкой части и 4,25 млн токенов в русской. Корпус доступен онлайн по адресу http://korpus.sk:8099/manatee.ks/index.

Имеются также большой ***польско-русский*** и ***русско-польский параллельный корпус*** объемом 30 млн токенов, разработанный в Варшавском университете (http://pol-ros.polon.uw.edu.pl/), небольшой ***болгарско-русский параллельный корпус*** (211 текстов, 3,3 млн токенов), разработанный в рамках работ по созданию Болгарского национального корпуса в Институте болгарского языка Болгарской академии наук (http://www.ibl.bas.bg/BGNC_parallel_bg.htm; http://search.dcl.bas.bg/bg/).

Корпус ASPAC — многоязычный корпус художественных текстов (около 70 произведений) и их переводов (иногда до 6 переводов одного и того же текста) создан в Амстердамском университете (ASPAC — Amsterdam Slavic Parallel Aligned Corpus) (http://www.uva.nl/over-de-uva/organisatie/medewerkers/content/b/a/a.a.barentsen/a.a.barentsen.html). Несмотря на название, корпус включает тексты и на 11 неславянских языках. Тексты выровнены по абзацам. Корпус доступен в режиме выгрузки по соглашению.

Параллельных корпусов достаточно много, но чаще всего они создаются для решения различных конкретных задач (переводоведение (см., например, [Добровольский 2003а, 2003б; Михайлов 2003]), литературоведение (см. Параллельный корпус переводов «Слова о полку Игореве», http://www.nevmenandr.net/slovo/pro.php), двуязычная и многоязычная терминология и др.) и не всегда доступны. Особенно много таких «целевых» корпусов создается сейчас в Европейском Союзе в рамках работ по терминологии (их обзор см. в [Steinberger R. et al. 2014]).

6. Диахронические корпуса

Выявление, описание и интерпретация изменений языка во времени — задача диахронического исследования. Еще недавно проведение такого исследования требовало больших усилий и затрат времени. Сегодня компьютерные технологии и корпусная лингвистика дают для него принципиально новые инструменты. Для этого создаются диахронические корпуса. Диахронические корпуса – понятие неопределенное. Можно сказать, что это корпуса, которые позволяют изучать развитие языка на протяжении какого-то (достаточно длинного) промежутка времени или изучать язык в его прежних состояниях. Таким образом, это понятие подразумевает представительный исторический корпус и (желательно) инструмент для диахронических исследований. В числе диахронических корпусов можно назвать уже упомянутый СКАТ, Регенсбургский диахронический корпус русского языка (древнерусские тексты) (http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new), Коллекцию древнейших и средневековых славянских и русских текстов «Манускрипт» (http://mns.udsu.ru/), Рукописные памятники Древней Руси: берестяные грамоты, летописи, рукописная книга (http://gramoty.ru/) и др.

В данной статье мы не рассматриваем особенности и наполнение разных диахронических корпусов, а остановимся лишь на программных инструментах, поддерживающих диахронические исследования, базирующиеся на корпусах. С точки зрения функциональности большой набор лингвистических модулей предлагает система «Манускрипт». Она позволяет знакомиться с текстами, указателями и осуществлять выборку данных, предлагает пользователям специализированный редактор для ввода, редактирования и фрагментирования текстов, модуль выборок и запросов, позволяющий подготовить данные для лингвистических, палеографических и текстологических исследований, морфологический анализатор для автоматического анализа и синтеза словоформ древнерусского языка и др. Здесь мы опишем два инструмента, находящихся непосредственно в рамках корпусной лингвистики и выполняющих статистическую обработку корпусных данных.

6.1. Сервис Google Books Ngram Viewer

Первой из таких специальных диахронических систем был сервис Google Books Ngram Viewer (http://books.google.com/ngrams) [Michell

2011; Захаров, Масевич 2014]. Эта система является в настоящее время наиболее мощным инструментом диахронических исследований. Доступ к ней открыт, начиная с 2009 года. Google books Ngram Viewer представляет собой информационную систему, которая содержит несколько корпусов размеченных текстов книг на 9 языках. Система также содержит отдельно корпус британского и американского английского языка, корпус всех вариантов английского языка. Самые поздние тексты корпуса, доступные для пользователей в настоящее время, относятся к 2008 году. На конец 2013 г. база данных насчитывала более 8 млн книг (текстов), что составляет около 6% всех когда-либо опубликованных печатных книг. Корпус книг на русском языке содержит 591 310 книг (текстов) или 67 137 666 353 словоупотреблений.

Интерфейс системы. Для каждой заданной лексической единицы для заданного временного интервала строится график. Каждая кривая графика маркируется цветом, в конце кривой указывается, какой N-грамме (слову или словосочетанию) она соответствует (рис. 7). Под термином N-грамма понимается последовательность от одного до пяти слов, причем N-грамма должна встречаться в корпусе не менее 40 раз. По вертикальной оси графика откладывается относительная частота встречаемости заданной N-граммы в данном году, выраженная в процентах. На горизонтальной оси показаны годы, входящие в заданный временной интервал.

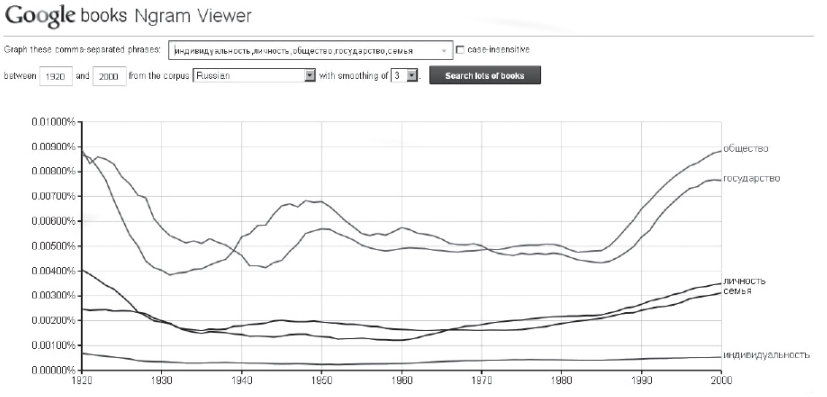


Рис. 7. Общий вид экрана с результатами выполнения запроса.

При построении графиков, показывающих динамику изменения частоты употребления, используется так называемое «сглаживание» (smoothing). При нулевом сглаживании в графике учитывается относительная частота встречаемости N-граммы за каждый год. Тенденция в динамике встречаемости слов прослеживается более отчетливо при скользящем усреднении данных. В окне “Smoothing” интерфейса, по умолчанию указано сглаживание 3, что означает, что для данного года для данной лексической единицы высчитывается среднее значение из семи чисел — число употреблений слова в данном году и число употреблений для трех предыдущих и трех последующих лет.

Возможно определение координат любой точки графика. Для этого достаточно установить курсор на любую точку над нужным годом. Система в этом случае выдаст сообщение о вертикальной и горизонтальной координатах этой точки для всех кривых. Если же установить курсор непосредственно на кривую, то исследуемая кривая будет выделена (рис. 8).

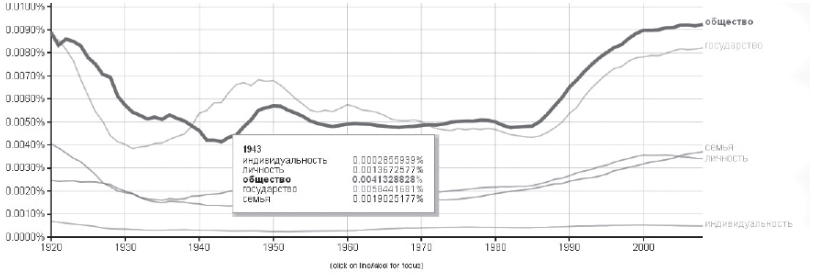


Рис. 8. Сообщение о координатах кривых для выбранного года (label for focus).

Кроме построения графиков, система представляет ссылки к текстам, найденным по запросам. Как правило, это библиографические описания книг и фрагменты текстов с выделением в них заданных N-грамм. В некоторых случаях доступен полный текст книги в графическом формате.

Поиск и лингвистические особенности системы

Для задания запроса на построение графика заполняется специальная форма (рис. 9).

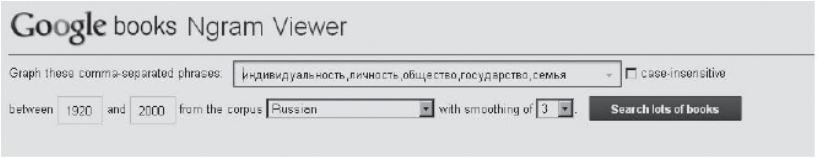


Рис. 9. Область запроса на построение графика.

Graph these case-sensitive comma-separated phrases — построить графики для этих словоформ с учетом регистра, разделять N-граммы запятыми (окно запроса на построение графика);

case-insensitive — при установке флажка в окне система не различает заглавные и строчные буквы;

between and — между … и… (окно указания временного периода, вводится год начала и конца исследуемого периода);

from the corpus — из корпуса (выбрать из выпадающего меню);

with smoothing — со сглаживанием (выбрать из выпадающего меню);

search lots of books — искать в массивах книг (кнопка команды на поиск и построение графика).

Имеется возможность при формулировке условий поиска задавать распознавание *заглавных* и *строчных букв* (case sensitive), или игнорировать различие между ними.

В системе нет *грамматической нормализации* лексических единиц, иначе говоря, поиск лексической единицы (слова или словосочетания) и построение графиков частоты ее встречаемости осуществляется для заданной словоформы.

Система предусматривает использование пользовательских тегов для модификации условий построения графиков.

Теги частей речи

Теги этой группы могут применяться изолированно, в этом случае показывается частота употребления данной части речи, а также

могут присоединяться к какому-либо знаменательному слову (больной _NOUN, больной_Adj — см. рис. 10).

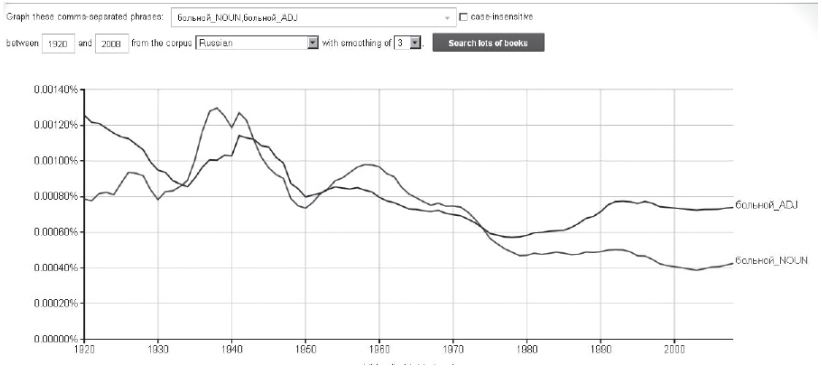


Рис. 10. График встречаемости слова «больной» как существительного и как прилагательного.

Среди прочих тегов имеется тэг _INF (*Inflections*), по которому строятся кривые для всех форм словоизменительной парадигмы данного слова (рис. 11). Следует, однако, отметить, что данная функция работает не всегда корректно, по крайней мере, для русского языка.

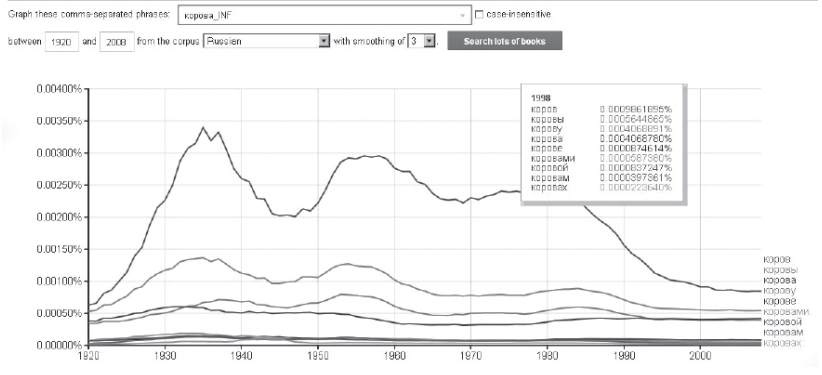


Рис. 11. Частота встречаемости форм словоизменительной парадигмы существительного «корова».

Существуют также теги позиционирования слов (начало и конец предложения, а также поиск глагола, выполняющего роль основного предиката в предложении).

Представляют интерес теги выбора корпусов. В случае их задания система строит графики по разным корпусам одновременно (рис 12).

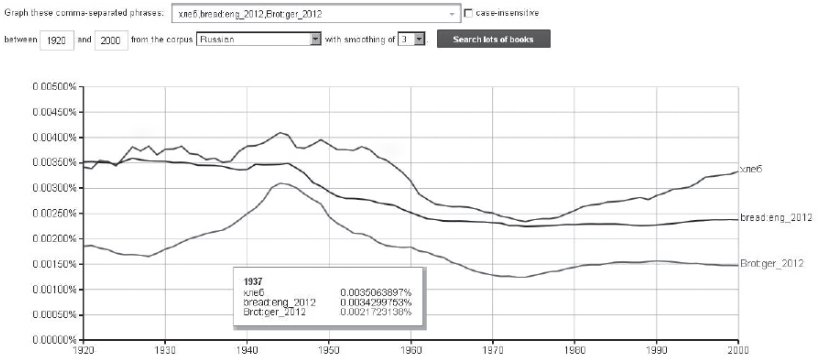


Рис. 12. График встречаемости слова «хлеб» и его английского и немецкого эквивалентов в русском, английском и немецком корпусах.

С октября 2013 года введен тег контекста — «подстановочный знак» * (wildcard). Ввод его через пробел после N-граммы (рис.13) или до неё позволяет строить график встречаемости десяти наиболее частотных сочетаний N-граммы и слова следующего за нею или ей предшествующего.

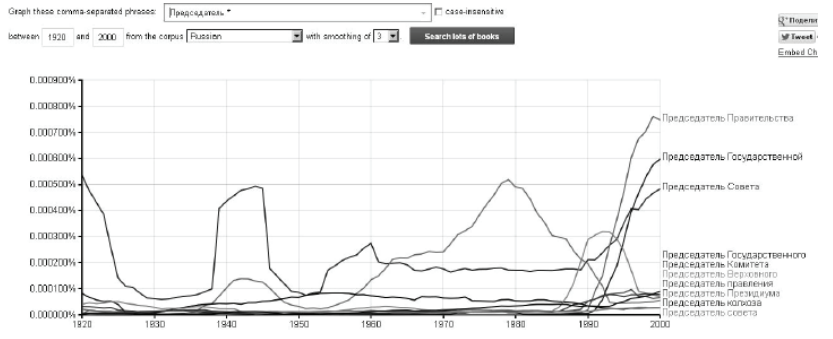


Рис. 13. Кривые встречаемости десяти биграмм с первым словом «Председатель» (использование подстановочного знака после N-граммы).

При задании запроса можно задавать *операции над кривыми графиков.*

Суммирование (сложение) кривых. Операция позволяет суммировать значения каждой точки по оси ординат двух или более кривых. Для осуществления операции поисковые слова вводятся в окно через знак +, например: *лошадь + лошади +лошадей*.

Вычитание кривых. Операция позволяет вычитать из значения каждой точки кривой по оси ординат, значение точки другой кривой для той же позиции по оси абсцисс. С помощью этой операции можно представить, насколько частота встречаемости одной N-граммы больше (меньше) другой, и как это различие менялось во времени. Для осуществления операции поисковые слова вводятся в окно через знак - (дефис), а выражение необходимо брать в круглые скобки, например: (вежливость-корректность). При этой операции вся кривая или ее часть может находиться в области отрицательных значений (см. рис. 14).

Рис. 14. Вычитание значений кривой «корректность» из значений кривой «вежливость».

Есть также умножение и деление графиков. Операция умножения позволяет умножать на число n значения всех точек графика (например, *лемматизация**100). Данная операция позволяет сделать сопоставимым вид кривых, значения которых отличаются на несколько порядков.

Старая орфография русского языка в Google Books NGram Viewer

Проблема, имеющая особую важность в диахронических исследованиях русского языка – это представление русских текстов в графике и орфографии, действовавшей до 1918 г. включительно. Наличие таких текстов очень ценно для исследования русской лексики. Однако существующие корпуса русского языка, как правило, все тексты дают в современной орфографии. Отсутствие корпусов со старой оригинальной орфографией не позволяет изучить эту сторону языка. К чести НКРЯ, следует сказать, что сейчас там появился исторический подкорпус объемом более 7 млн ссловоупотреблений.

Тексты Google Books получены посредством оцифровки и распознавания оригинальных печатных изданий. Поэтому в базе данных Google Books тексты книг, изданных до 1919 года (в определенных случаях более поздних изданиях), представлены в старой системе письма, что дает возможность разнообразных исследований. Так, слова с написанием в старой орфографии твердого знака в конце слова можно обрабатывать в двух формах, с твердым знаком и без него (*Бог+Богъ+бог+богъ*), получая на графике суммарную кривую.

Однако другие знаки старой системы русского письма, такие как і (код Unicode – 0456), ѣ (код Unicode – 0463), ѳ (код Unicode – 0473) и соответствующие заглавные буквы, обрабатываемые поисковой системой Google Books, в Google Books NGram Viewer не распознаются. Так, поиск триграммы Ѳедоръ Михайловичъ Достоевскій результатов не дает. Но поскольку собственно поисковая система Google осуществляет полноценный поиск по старой русской орфографии, хочется верить, что и в Ngram Viewer скоро эта проблема будет решена.

6.2. Сервис НКРЯ «Графики»

С 2012 г. сервис, аналогичный Google Books NGram Viewer, появился и в России. Текстовый материал НКРЯ по хронологическому материалу также может быть использован для диахронических исследований. Основной массив текстов, собранных в НКРЯ, охватывает период в 200 лет, поэтому он наиболее приспособлен для изучения коротких (несколько десятилетий) и средних (1–2 столетия) языковых изменений. Объем корпуса позволяет изучать вариативность и изменчивость достаточно частотных языковых явлений, а также получать надежные результаты по следующим направлениям:

1) изучение морфологических вариантов имен, глаголов и т. д. и их эволюции;

2) исследование словообразовательных вариантов и связанной с ними проблемы паронимов, продуктивности словообразовательных моделей и словообразовательных средств;

3) исследование изменения вариантов управления, согласования и примыкания;

4) исследование лексической вариативности.

Однако в ряде случаев, например, при изучении фразеологизмов и других устойчивых сочетаний, сервис НКРЯ дает картину мало репрезентативную по причине недостаточного объема данных.

Сервис работает на текстах основного корпуса НКРЯ и называется «Графики». Функционально сервис «Графики» подобен сервису Google Books Ngram Viewer. Он показывает хронологическое распределение заданных и найденных лексических единиц (словоформ, словосочетаний) в основном корпусе НКРЯ. Вход в этот сервис возможен как со страницы с результатами поиска по произвольному запросу к основному корпусу (ссылка *Распределение по годам*), так и из главного меню (рис. 15).

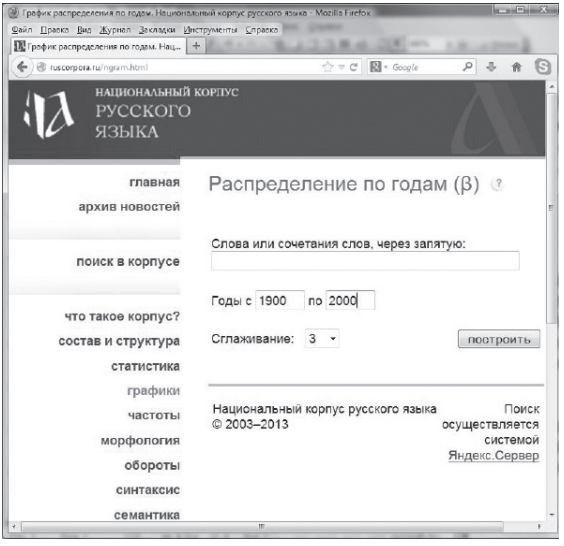


Рис. 15. Интерфейс сервиса «Графики»

Можно задать временные границы, например, с 1930 по 1960 г. Нажав на кнопку «Построить», мы получим график (рис. 16), где каждому элементу запроса (в нашем примере Черчилль, Рузвельт, Франко) соответствует линия своего цвета (см. легенду в правом верхнем углу).

По вертикальной оси графика откладывается относительная частота употреблений данной лексической единицы (в ipm). Подведя мышку к любой точке на линии, мы увидим относительную частоту употребления за определенный год (ipm). Сглаживание графиков позволяет увидеть общую тенденцию за случайными колебаниями частот. Например, сглаживание в 10 лет усредняет частоту слова с учетом предшествующих и последующих 5 лет, т.е. для данного года берется средняя частота употребления за 11 лет.

Имеется возможность показать таблицы с абсолютными и относительными частотами употреблений за каждый год. Из таблиц гиперссылки позволяют перейти к просмотру примеров из корпуса по годам.

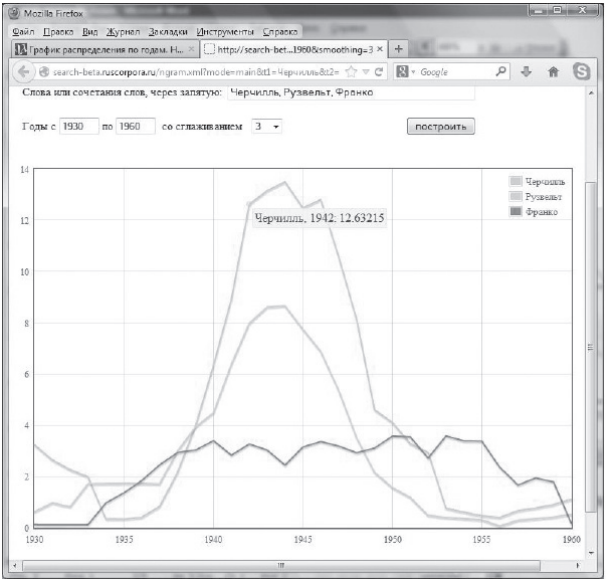


Рис. 16. График встречаемости имен Черчилль, Рузвельт, Франко в текстах, опубликованных с 1930 по 1960 гг.

Из сервиса «Графики» можно перейти на сервис Google Ngram Viewer, работающий на русскоязычной коллекции текстов Google Books. Однако при сходной идеологии, формулы подсчета относительной частоты в сервисах Национального корпуса и Google Ngram Viewer несколько отличаются.

Было бы интересно провести сравнительные исследования результатов по одним и тем же запросам, получаемым по текстам корпуса Google Books Ngram Viewer и НКРЯ.

Литература

Алексеев В. А., Алексеева Е. Л., Касьяненко С. Е. Грамматическая разметка в корпусе СКАТ // Труды международной конференции «Корпусная лингвистика — 2011». СПб, 2011. С. 69–73.

Алексеева Е. Л., Лаврентьев А. М., Азарова И. В., Захарова Л. А. Разметка корпуса древнерусских текстов // Труды международной конференции «Корпусная лингвистика 2004». 11–14 октября 2004 г. СПб., 2004. С. 16–24.

Алексеева Е. Л. Синтаксическая разметка корпуса древнерусских агиографических текстов СКАТ // Структурная и прикладная лингвистика. Выпуск 10. СПб, 2014. С. 345–351.

Андрющенко В. М. Концепция и архитектура Машинного фонда русского языка. Москва: Наука, 1989.

Асиновский А. С., Богданова Н. В., Степанова С. Б., Шерстинова Т. Ю., Маркасова Е. В., Супрунова А. В. Звуковой корпус русского языка «Один речевой день»: пути пополнения и первые результаты исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог», № 9 (16). М.: РГГУ, 2010. С. 41–47.

Богданова С. Ю. О разработке учебного мультимодального корпуса текстов // Материалы XLII международной филологической конференции. Секция прикладной и математической лингвистики. (С. Петербург, 14–19 марта 2013 г.). СПб.: СПбГУ, 2013. С. 27–32.

Бондарко Л. В. Фонетический фонд современного русского языка // Бюллетень фонетического фонда русского языка. 1988. № 1.

Гарабик Р., Захаров В. П. Параллельный Русско-Словацкий Кор-

пус // Труды международной конференции «Корпусная лингвистика–2006». СПб.: Изд-во С. Петерб. ун-та, 2006. С. 81–88.

Гвишиани Н. Б. Практикум по корпусной лингвистике: Учеб. пособие по английскому языку. М.: Высшая школа, 2008.

Герд А. С., Алексеева Е. Л., Азарова И. В., Захарова Л. А. Электронный корпус текстов по памятникам древнерусской агиографической литературы // НТИ. Сер. 2. Вып. 9. 2004. С. 16–20.

Гришина Е. А. Два новых проекта для Национального корпуса: мультимедийный подкорпус и подкорпус названий // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005.С. 233–250.

Гришина Е. А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 175–214.

Гришина Е. А. Устная речь в Национальном корпусе русского языка // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 94–110.

Гришина Е. А., Савчук С. О. Корпус устных текстов в НКРЯ: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 129–149.

Грудева Е. В. Корпусная лингвистика: Учебное пособие. 2-е изд., стереотип. М.: Флинта, 2012.

Добровольский Д. О. Корпус параллельных текстов и литературный перевод // НТИ сер.2, № 10, 2003. С. 13–18.

Добровольский Д. О. Корпус параллельных текстов как инструмент анализа литературного перевода // Компьютерная лингвистика и интеллектуальные технологии. М.: Наука, 2003. С. 126–131.

Добровольский Д. О., Кретов А. А., Шаров С. А. Корпус параллельных текстов: архитектура и возможности использования //Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 263–296.

Ершов А. П. К методологии построения диалоговых систем: феномен деловой прозы // Вопросы кибернетики. Общение с ЭВМ на естественном языке. М., 1982.

Засорина Л. Н. (ред.). Частотный словарь русского языка. М.: 1977.

Захаров В. П. Корпусная лингвистика: Учебно-метод. пособие. СПб., 2005.

Захаров В. П., Богданова С. Ю. Корпусная лингвистика: Учебник для студентов направления «Лингвистика» (1-е изд., Иркутск: ИГЛУ, 2011; 2-е изд., перераб. и дополн., СПб.: СПбГУ, Филологический факультет, 2013).

Захаров В. П., Масевич А. Ц. Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer. // Структурная и прикладная лингвистика. Выпуск 10. СПб., 2014. С. 303–330.

Камшилова О. Н. Учебный корпус текстов: потенциал, состав, структура. СПб.: ООО «Книжный Дом», 2012.

Кибрик А. А., Подлесская В. И. Рассказы о сновидениях. Корпусное исследование устного дискурса. М., 2009.

Компьютерная лингвистика и интеллектуальные технологии: Материалы ежегодной Международной конференции «Диалог» (2000–2014).

Котов А. А., Гопкало О. С. Русскоязычный эмоциональный корпус // Труды международной конференции «Корпусная лингвистика-2011». Санкт-Петербург. СПб.: СПбГУ. Филологический факультет. 2011 С. 211–216.

Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М., 2005. С. 155–174.

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

Машинный фонд русского языка: идеи и суждения. Москва: Наука, 1986.

Митрофанова О. А., Емельянова А. В., Ильина А. И., Кулажко С. А., Михайлова В. Д., Якуба Н. М. Лексико-грамматические особенности текстов специальных корпусов разной тематики // Сборник материалов по итогам XLIII Международной филологической конференции. Секция прикладной и математической лингвистики. СПб., 2014.

Михайлов М. Н. Параллельные корпуса художественных текстов: принципы составления и возможности применения в лингвистических и переводоведческих исследованиях. Тампере: Тамперский университет, 2003.

Мустайоки А., Копотев М. В., Гурин Г. Б., Саломатина М. С. Принципы синтаксической разметки Хельсинского аннотированного корпуса русских текстов ХАНКО // Труды международной конференции «MegaLing`2005. Прикладная лингвистика в поиске новых путей». СПб., 2005. С. 90–95.

Национальный корпус русского языка: 2003–2005. Сборник статей. М.: Индрик, 2005.

Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009.

Николаева Ю. В. Сегментация устного нарратива и изобразительные жесты: кинетические признаки границ и связей между сегментами дискурса // Компьютерная лингвистика и интеллектуальные технологии: Материалы ежегодной Международной конференции «Диалог». М.: РГГУ, 2009. С. 340–345.

Резникова Т. И. Славянская корпусная лингвистика: современное состояние ресурсов // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 402–461.

Резникова Т. И., Копотев М. В. Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 31–61.

Труды международной конференции «Корпусная лингвистика» (2004, 2006, 2008, 2011, 2013). Санкт-Петербург. СПб.: СПбГУ. Филологический факультет.

Шерстинова Т. Ю., Степанова С. Б., Рыко А. И. Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Мат-лы XXXVIII международной филологической конференции. Секция: «Формальные методы анализа русской речи». СПб.: СПбГУ, 2009. С. 66–75.

Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryko A., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation // Lecture Notes in Computer Science, 2009. Vol. Text, Speech and Dialogue, № 5729/2009. P. 250–257.

Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora, In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno,

Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264, ISBN: 978–3-319–108155.

Čermák F, Rosen A. The case of InterCorp, a multilingual parallel corpus. International In: Journal of Corpus Linguistics. 2012. Vol. 13, no. 3, pp. 411–427.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine // Proceedings of the XIth Euralex International Congress. Lorient: Universite de Bretagne-Sud, 2004, pp. 105–116.

Michel J. B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books science. Science 331, 176 (2011); DOI 1126/Science. 1199644 URL: http://www.sciencemag.org/content/331/6014/176.full.html

Steinberger Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski & Signe Gilbro. An overview of the European Union's highly multilingual parallel corpora. Language Resources and Evaluation Journal (LRE) 2014. DOI: 10.1007/s10579–014–9277–0.

Waldenfels Ruprecht, von. Compiling a parallel corpus of Slavic languages. In: Bernhard Brehmer, Vladislava Ždanova, Rafał Zimny (Hrsg.) Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV) 9. München, 2006. S.123–138.

Waldenfels Ruprecht, von. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In: Daniela M., and Garabík, R. (eds.), Natural Language Processing, Multilinguality. Proceedings of Slovko 2011 (Modra, Slovakia, 20–21 October 2011). Bratislava, 2011, pp. 156–162.

Zakharov V. Corpora of the Russian Language // Text, Speech and Dialogue: Proceedings of the 16th International Conference, TSD 2013, Plzen, Czech Republic, September 1–5, 2013. Ivan Habernal, Václav Matoušek (Eds.). Springer-Verlag, Berlin Heidelberg, 2013, pp. 113. (Lecture Notes in Artificial Intelligence, 8082).

V. P. Zakharov
[1]Saint-Petersburg State University
[2]Institute for Linguistic Studies
(Russia, Saint-Petersburg)
vz1311@yandex.ru

CORPORA OF THE RUSSIAN LANGUAGE

The paper describes corpora of the Russian language and the state of the art of Russian corpus linguistics. The main attention is paid to the Russian National Corpus (RNC) as the most popular one among linguists for both being the most well known and the opportunities which it presents. The author regards the subcorpora within the RNC, semantic annotation and Charts service of the RNC in comparison with Google Books Ngram Viewer. The article also presents a large number of other text corpora of Russian, among them Helsinki Annotated Corpus (HANCO), Leeds University corpora, Sketch Engine corpora and so on, speech corpora of Russian, parallel corpora, diachronic corpora and specialized corpora. Corpora of the Russian language provide corpus-based studies of both oral speech and written language using synchronic and diachronic approaches and allow us to study linguistic phenomena in typological, sociological, culturological aspects.

Key words: Russian corpus linguistics, corpora, the Russian language, specialized corpora.

References

Alekseev V. A., Alekseeva E. L., Kas'yanenko S. E. [Grammatical Markup of the "SKAT" Corpus]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika — 2011"* [Proceedings of the International Conference "Corpus Linguistics — 2011"]. St. Petersburg, 2011, pp. 69–73. (In Russ.)

Alekseeva E. L., Lavrent'ev A. M., Azarova I. V., Zaharova L. A. [Markup of the Old Russian Language Corpus]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2004"* [Proceedings of the

International Conference “Corpus Linguistics – 2004”]. 11-14 October, 2004. St. Petersburg, 2004, pp. 16–24. (In Russ.)

Alekseeva E. L. [Syntactic Markup of the Corpus of Old Russian Hagiographical Texts “SKAT”]. *Strukturnaya i prikladnaya lingvistika* [Structural and applied linguistics]. Issue 10. St. Petersburg, 2014, pp. 345–351. (In Russ.)

Andryushchenko V. M. *Kontseptsiya i arkhitektura Mashinnogo fonda russkogo yazyka* [Conception and Architecture of the Foundation for the Russian Language Machine Analysis]. Moskow, Nauka Publ., 1989. (In Russ.)

Asinovsky A. S., Bogdanova N. V., Stepanova S. B., Sherstinova T. Ju., Markasova E. V., Suprunova A.V. [“One Speaker's Day” Sound Corpus of the Russian Language: Ways of Stocking and First Research Results]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii “Dialog – 2010”* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference “Dialog–2010”]. Moscow, RSUH Publ., 2010, pp. 41–47. (In Russ.)

Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryko A., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication “One Speaker's Day”: Creation Principles and Annotation. *Lecture Notes in Computer Science, 2009*. Vol. Text, Speech and Dialogue, № 5729/2009, pp. 250–257.

Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora. Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): *Text, Speech and Dialogue. 17th International Conference,* TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264.

Bogdanova S. Ju. [On development of the educational multi-language corpus]. *Materialy XLII mezhdunarodnoi filologicheskoi konferentsii. Sektsiya prikladnoi i matematicheskoi lingvistiki* [Proceedings of the XLII International Philological Conference. Group for Applied and Mathematical Linguistic]. (S. Peterburg, 14-19 March, 2013). St. Petersburg, St. Petersburg St. Univ. Publ., 2013, pp. 27–32. (In Russ.)

Bondarko L. V. [Phonetic Foundation of the Modern Russian Language]. *Byulleten' foneticheskogo fonda russkogo yazyka,* 1988, no. 1. (In Russ.)

Čermák F., Rosen A. The case of InterCorp, a multilingual parallel

corpus. International In: *Journal of Corpus Linguistics,* 2012, Vol. 13, no. 3, pp. 411–427.

Dobrovol'skij D. O. [Corpus of Parallel Texts and Literary Translation]. *Nauchno-tekhnicheskaya informatsiya,* Ser. 2, 2003, no. 10, pp. 13–18. (In Russ.)

Dobrovol'skij D. O. [Corpus of Parallel Texts as an Instrument for Literary Translation Analysis]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Computational Linguistics and Intellectual Technologies]. Moscow, Nauka Publ., 2003, pp. 126–131. (In Russ.)

Dobrovol'skij D. O., Kretov A. A., Sharov S. A. [Corpus of Parallel Texts: Architecture and Usage Possibilities]. *Natsional'nyi korpus russkogo yazyka: 2003–2005* [The National Corpus of the Russian Language: 2003–2005]. Moscow, Indrik Publ., 2005, pp. 263–296. (In Russ.)

Ershov A. P. [On the Methodology of Dialog Systems Building: Phenomenon of Business Texts]. *Voprosy kibernetiki. Obshchenie s EVM na estestvennom yazyke* [Cybernetic Questions. Communication with Computer Using Natural Language]. Moscow, 1982. (In Russ.)

Garabik R., Zaharov V. P. [Parallel Russian — Slovak Corpus]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2006"* [Proceedings of the International Conference "Corpus Linguistics — 2006"]. St. Petersburg, St. Petersburg St. Univ. Publ, 2006, pp. 81–88. (In Russ.)

Gerd A. S., Alekseeva E. L., Azarova I. V., Zaharova L. A. [Electronic Corpus of the Old Russian Hagiographic Texts]. *Nauchno-tekhnicheskaya informatsiya,* Ser. 2. 2004, Issue 9, pp. 16–20. (In Russ.)

Grishina E. A. [Multi-Media Corpus of the Russian Language (MURKO): Problems with Annotation]. *Natsional'nyi korpus russkogo yazyka: 2006—2008. Novye rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 175–214. (In Russ.)

Grishina E. A. [Two New Projects for the National Corpus of the Russian Language: Multi-Media Subcorpus and Subcorpus of Titles]. *Natsional'nyi korpus russkogo yazyka: 2003—2005. Rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 233–250. (In Russ.)

Grishina E. A., Savchuk S. O. [Corpus of the Spoken Language in the National Corpus of the Russian Language: content and structure]. *Natsional'nyi korpus russkogo yazyka: 2006—2008. Novye rezul'taty i*

perspektivy [National Corpus of the Russian Language: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 129–149. (In Russ.)

Grishina E. A. [Spoken Language in the National Corpus of the Russian Language]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [National Corpus of the Russian Language: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 94–110. (In Russ.)

Grudeva E. V. *Korpusnaya lingvistika: Uchebnoe posobie* [Coprus Linguistics: Educational Manual]. 2nd ed. Moscow, Flinta Publ., 2012.

Gvishiani N. B. *Praktikum po korpusnoi lingvistike: Ucheb. posobie po angliiskomu yazyku* [Workshop on Corpus Lingusitics: Educational Manual on the English Language]. Moscow, Vysshaya shkola Publ., 2008.

Kamshilova O. N. *Uchebnyi korpus tekstov: potentsial, sostav, struktura* [Educational Corpus: Potential, Content, and Structure]. St. Petersburg, "Knizhnyi Dom" Publ., 2012.

Kibrik A. A., Podlesskaya V. I. *Rasskazy o snovideniyakh. Korpusnoe issledovanie ustnogo diskursa* [Description of Dreams: Corpus Research of Spoken Discourse]. Moscow, 2009.

Kilgarriff A., Rychly P., Smrz P., Tugwell D. The Sketch Engine. *Proceedings of the XIth Euralex International Congress.* Lorient, Universite de Bretagne-Sud, 2004, pp. 105–116.

Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Materialy ezhegodnoi Mezhdunarodnoi konferentsii "Dialog" [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog"] (2000–2014).

Kotov A. A., Gopkalo O. S. *Russkoyazychnyi emotsional'nyi korpus* [Emotional Corpus for the Russian Language]. *Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2011"* [Proceedings of the International Conference "Corpus Linguistics – 2011"]. St. Petersburg, St. Petersburg St. Univ. Publ., 2011, pp. 211–216. (In Russ.)

Kustova G. I., Lyashevskaya O. N., Paducheva E. V., Rahilina E. V. [Semantic Markup in the National Corpus of the Russian Language: Principles, Problems, Prospective]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [National Corpus of the Russian Language: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 155–174. (In Russ.)

Lyashevskaja O. N., Sharov S. A. *Chastotnyi slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)* [Frequency Dictionary of the Modern Russian Language (Based on the Data of the Narional Corpus of the Russian Language)]. Moscow, Azbukovnik Publ., 2009.

Mashinnyi fond russkogo yazyka: idei i suzhdeniya [Computerized Foundation for the Russian Language: Ideas and Opinions]. Moscow, Nauka Publ., 1986.

Michel J. B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books science. Science 331, 176 (2011); DOI 1126/ Science. 1199644. Available at: URL: http://www.sciencemag.org/content/331/6014/176.full.html (accessed 13.06.2015)

Mihajlov M. N. *Parallel'nye korpusa khudozhestvennykh tekstov: printsipy sostavleniya i vozmozhnosti primeneniya v lingvisticheskikh i perevodovedcheskikh issledovaniyakh* [Parallel Corpora of Fiction: Principles of Building and Possibilities of Usage through Researches on Linguistics and Translation]. Tampere, 2003. (In Russ.)

Mitrofanova O. A., Emel'yanova A. V., Il'ina A. I., Kulazhko S. A., Mikhailova V. D., Yakuba N. M. [Lexical and Grammatical Features of the Corpus of the Technical Texts of Different Aspects]. *Materialy XLII mezhdunarodnoi filologicheskoi konferentsii. Sektsiya prikladnoi i matematicheskoi lingvistiki* [Proceedings of the XLII International Philological Conference. Group for Applied and Mathematical Linguistics]. (St. Petersburg, 11–16 March, 2014). St. Petersburg, St. Petersburg St. Univ. Publ., 2014. (In Russ.)

Mustajoki A., Kopotev M. V., Gurin G. B., Salomatina M. S. [Principles of Syntactic Markup of the HANCO (Helsinki Annotated Corpus of Russian Texts)]. *Trudy mezhdunarodnoi konferentsii "MegaLing`2005. Prikladnaya lingvistika v poiske novykh putei"* [Proceedings of the Intrnational Conference "MegaLing'2005. Applied Linguistics: Looking for New Ways"]. St. Petersburg, 2005, pp. 90–95. (In Russ.)

Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy [Russian National Corpus: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005.

Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy. [Russian National Corpus: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009.

Nikolaeva Yu. V. [Segmentation of Spoken Narrative and Pictorial

Gestures: Kinetic Indication of Borders and Bounds between Segments of Discourse]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Materialy ezhegodnoi Mezhdunarodnoi konferentsii "Dialog–2009"* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2009"]. Moscow, RSUH Publ., 2009, pp. 340–345. (In Russ.)

Reznikova T. I. [Slavic Corpus Linguistics: Modern State of Recourses]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006–2008. New Results and Prospective]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 402–461. (In Russ.)

Reznikova T. I., Kopotev M. V. [Linguistically Annotated Corpora of the Russian Language (Review of the Public Resources)]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [Russian National Corpus: 2003–2005. Results and Prospective]. Moscow, Indrik Publ., 2005, pp. 31–61. (In Russ.)

Sherstinova T. Yu., Stepanova S. B., Ryko A. I. [Markup System of the ["One day of speech" Sound Corpus of the Russian Language]. *Materialy XXXVIII mezhdunarodnoi filologicheskoi konferentsii. Sektsiya "Formal'nye metody analiza russkoi rechi"* [Proceeding of the XXXVIII International Philological Conference. "Formal Methods of Russian Speech Analysis" Group]. St. Petersburg, St. Petersburg St. Univ. Publ., 2009, pp. 66–75. (In Russ.)

Steinberger Ralf, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski & Signe Gilbro. An overview of the European Union's highly multilingual parallel corpora. *Language Resources and Evaluation Journal (LRE)* 2014. DOI: 10.1007/s10579-014-9277-0.

Trudy mezhdunarodnoi konferentsii "Korpusnaya lingvistika" (2004, 2006, 2008, 2011, 2013) [Proceedings of the International Conference "Corpus Linguistics" (2004, 2006, 2008, 2011, 2013)]. St. Petersburg, St. Petersburg St. Univ., Philological faculty.

Waldenfels Ruprecht, von. Compiling a parallel corpus of Slavic languages. Bernhard Brehmer, Vladislava Ždanova, Rafał Zimny (Hrsg.) *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 9. München, 2006, pp.123–138.

Waldenfels Ruprecht, von. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. Daniela M.,

and Garabík, R. (eds.), *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011* (Modra, Slovakia, 20–21 October 2011). Bratislava, 2011, pp. 156–162.

Zakharov V. Corpora of the Russian Language. *Text, Speech and Dialogue: Proceedings of the 16th International Conference, TSD 2013,* Plzen, Czech Republic, September 1–5, 2013. Ivan Habernal, Václav Matoušek (Eds.). Springer-Verlag, Berlin Heidelberg, 2013, pp. 1 13. (Lecture Notes in Artificial Intelligence, 8082).

Zakharov V. P. *Korpusnaya lingvistika: Uchebno-metod. posobie* [Corpus Lingustics: Educational and Methodological Manual]. St. Petersburg, 2005.

Zakharov V. P., Bogdanova S. Ju. *Korpusnaja lingvistika: Uchebnik dlja studentov napravlenija "Lingvistika"* [Corpus Lingustics: Manual for Students of Linguistic Departments]. (1st ed., Irkutsk: IGLU, 2011; 2nd ed., St. Petersburg, St. Petersburg St. Univ., Philological faculty, 2013).

Zakharov V. P., Masevich A. Ts. [Diachronical Research Based on Google Books Ngram Viewer Corpus of Russian texts]. *Strukturnaya i prikladnaya lingvistika* [Structural and Applied Linguistics]. Issue 10, St. Petersburg, 2014, pp. 303–330. (In Russ.)

Zasorina L. N. (Ed.). *Chastotnyi slovar' russkogo yazyka* [Frequency Dictionary of the Russian Language]. Moscow, 1977.

Е. А. Гришина
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
rudi2007@yandex.ru

МУЛЬТИМОДАЛЬНЫЙ МОДУЛЬ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА[1]

В статье описывается состав и структура Мультимедийного корпуса русского языка (МУРКО), сложившаяся за годы его существования (2010–2015), основные способы получения информации из корпуса, а также в общих чертах обрисовываются перспективы его развития — количественное пополнение состава, создание программного обеспечения для обработки мультимедийного контента, разработка новых модулей в составе МУРКО — мультимедийных параллельных корпусов.

Ключевые слова: Мультимедийный русский корпус, Национальный корпус русского языка, мультимедийный параллельный корпус, аннотация, инструменты поиска.

1. Введение

В 2010 году в составе Национального корпуса русского языка был открыт пилотный вариант Мультимедийного русского корпуса (МУРКО). Общая идеология будущего МУРКО была изложена в работах [Гришина, Савчук 2008], [Гришина 2008], [Гришина 2009 a], [Гришина, Кудинов 2009], [Grishina 2009a, b]. После открытия пилотного МУРКО корпус рос и развивался; разные этапы его развития были зафиксированы в статьях [Grishina, Savchuk, Sichinava 2010], [Grishina 2010], [Гришина 2011]. В настоящей статье плани-

[1] Работа выполнена при поддержке РФФИ (гранты № 14-06-00245 и № 15-06-04334) и РГНФ (грант № 15-04-12018).

руется описать состав и структуру МУРКО — так, как она сложилась на данный момент (2014 год), — основные способы получения информации из корпуса, а также в общих чертах обрисовать перспективы его развития.

В течение всех лет над пополнением и развитием МУРКО работали С. О. Савчук, С. Б. Иванютин, А. А. Курсакова, Л. Д. Алексеевский, А. А. Махова, М. С. Кудинов, автор статьи и другие коллеги. Огромную благодарность команда МУРКО выражает Алексею Зобнину, Семену Давыдову и Игорю Шалыминову, программистам Яндекса, которые очень творчески, с большим интересом и ответственностью подошли к непростой задаче создания и усовершенствования МУРКО, являющегося, насколько нам известно, на сегодня одним из самых больших открытых мультимедийных корпусов в мире (обзор мультимедийных корпусов см. в статье В. П. Захарова в настоящем сборнике).

2. Состав корпуса

Единицей выдачи в МУРКО является пара «клип+текст» (кликст) или, в случае звуковой дорожки, — пара «фрагмент звукового файла + текст». В ходе подготовки материала для размещения в корпусе мы разрезаем звуковой и видеофайл на небольшие фрагменты, затем разрезаем на такие же фрагменты файлы расшифровок, — и выравниваем между собой мультимедийные и текстовые файлы посредством присвоения им одних и тех же имен. Решение разрезать мультимедийные исходники на фрагменты, а не выкладывать их целиком было принято в связи с тем, что большие мультимедийные файлы, во-первых, медленно загружаются на компьютеры пользователей, т. е. качество использования корпуса начинает в значительной степени зависеть от трафика, а во-вторых, — в больших мультимедийных файлах очень трудно ориентироваться, и, таким образом, пользователь, которого обычно интересует не просмотр исходного мультимедийного файла целиком, а лишь один конкретный, иногда очень короткий, практически точечный фрагмент, вынужден тратить слишком много времени на «перемотку» исходного файла в поисках нужного эпизода. Разумеется, принятое решение имеет ряд недостатков, прежде всего — «обрубленные», иногда весьма неудачно, начало и конец клипа, а во-вторых, — вырванные из более широкого контекста

отдельные высказывания, что для исследования устной речи, понимание которой весьма плотно связано с широкой конситуацией, может оказаться весьма неприятным обстоятельством. Для преодоления этих двух недостатков в корпусе предусмотрен переход к предыдущему и последующему эпизоду, о чем см. в § 4.

По состоянию на апрель 2015 года объем МУРКО составлял 4,3 млн словоупотреблений, или более 187 тыс. кликстов.

ОСНОВНЫЕ ТИПЫ ТЕКСТОВ В СОСТАВЕ КОРПУСА.

1. Речь кино. На сегодняшний день это самая обширная часть корпуса — около 3,4 млн словоупотреблений, 664 кинофильмов/мультфильмов. МУРКО начинал формироваться именно с кинематографической речи — ввиду ее доступности, а также в связи с тем фактом, что полные и точные расшифровки кинофильмов уже были получены в ходе работы над устным и акцентологическим корпусами в составе Национального корпуса. Сейчас перед создателями корпуса стоит задача довести остальные зоны МУРКО до сопоставимых объемов.

2. Устная публичная речь. Объем — чуть более 812 тыс. словоупотреблений. Подкорпус включает в себя телепередачи, документальные фильмы, доклады, лекции и другие образцы устной неподготовленной или слабоподготовленной публичной речи. Подкорпус не включает в себя чтение вслух написанного текста (т. е. тексты типа written-to-be-spoken).

3. Устная непубличная речь. Объем — чуть более 12 тыс. словоупотреблений, включает в себя частные беседы между людьми в разных ситуациях — рассказы, краткие беседы и диалоги. Этот тип текстов — наиболее «дефицитный», поскольку связан с ограничениями экстралингвистического порядка (в частности, нежеланием людей делать свою речь широкодоступной). Создатели корпуса планируют и в дальнейшем пополнять эту зону подкорпуса, однако рассчитывать на ее принципиальное увеличение, по-видимому, не приходится.

4. Авторское чтение. Объем — чуть более 22 тыс. словоупотреблений. Включает в себя авторское чтение художественных и иных произведений (свои рассказы и воспоминания читают Шаламов, Пришвин, Довлатов, Чуковский и др.).

5. Художественное чтение. Объем — примерно 15,4 тыс. сло-

воупотреблений. Включает в себя актерское чтение художественных произведений, в т. ч. и по ролям.

6. Театральная речь. Запись радио- и телеспектаклей, объем около 55,7 тыс. словоупотреблений.

Последние три типа текстов представляют собой в высшей степени интересный источник для орфоэпических, фонетических, акцентологических и интонационных штудий.

В общей сложности материалы МУРКО покрывают временной интервал с 1930 по 2013 год. Естественно, при относительно небольшом объеме корпуса разные периоды представлены достаточно неравномерно. Однако и нынешний объем корпуса позволяет по ряду параметров ставить и решать диахронические задачи.

3. Типы поисковых запросов

МУРКО сохраняет все типы запросов, которые приняты в основном корпусе (лексический, морфологический, семантический поиск и их комбинации), а также в устном и акцентологическом модуле. Здесь мы не будем повторяться — типы соответствующих запросов описаны как в данном, так и в предыдущих сборниках (см. [Гришина, Савчук 2009], [Гришина 2009 б]), — сосредоточимся на тех запросах, которые характерны именно для МУРКО.

3.1. ОРФОЭПИЧЕСКАЯ СТРУКТУРА СЛОВА

Наличие звуковой дорожки в выдаче позволяет пользователю прослушать, как говорящие произносят те или иные сочетания звуков. Для этого в форме поиска существует форма запроса орфоэпической структуры слова (см. рис. 1)

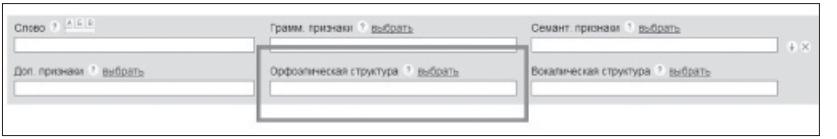


Рис. 1

Основная идея орфоэпического поиска базируется на том, что русская орфография имеет фонематический характер, что позволяет практически в один шаг переходить при корпусном поиске от орфографии слова к

его орфоэпии. Предположим, нам хотелось бы получить данные о том, как произносятся согласные [сз] перед мягкими согласными [тднрл] (очевидно, что поиск такого материала вручную чрезвычайно трудоемок). Нажав ссылку *выбрать* (см. рис. 1), мы получаем диалоговое окно, где формулируем наш запрос (см. рис. 2).

Рис. 2

После нажатия ОК и *Искать* на странице поиска мы получаем на выдаче большое количество клипов (порядка 90 тыс. вхождений), где встречаются искомые звукосочетания (например, *смыСЛе*, *еСЛи*, *поСЛедняя*, *чаСТи*, *иССЛедователей*, *миСТификатор*, *иЗДеваться*, *жиЗНи*, *ЗДесь*, *двеСТи*, *СДелать*, *чаСТичная*, *раЗНица* и многие другие). Этот материал с помощью сортировок можно упорядочить хронологически (и тогда мы получаем данные начиная с 1930 и заканчивая 2013 годом), а также выбрать подкорпус мужской и женской речи, или речи одного специального говорящего, если известно его имя. Кроме того, с помощью указания жанра можно сузить область поиска и искать материал, например, только в некинематографической речи.

Еще один пример. Предположим, нас интересует произнесение геминаты [жж] перед гласными. Орфографически гемината [жж] может быть реализована сочетанием букв *зж* и *жж*, поэтому запрос выглядит следующим образом (см. рис. 3).

Рис. 3

На выдаче мы получаем около полутысячи вхождений (*позже*, *жжение*, *жженка*, *жжет*, *жужжать*, *заезжий*, *изжелта*, *изжить*, *изъезженный*, *можжевельник*, *изжеванный*, *необъезженный*, *обезжирить*, *обезживотеть*, *попозже*, *приезжий* и др.— естественно, не только в исходных, но и в косвенных формах; материал — с 1934 по 2013 год).

Возможен поиск на границе слова. Например, если нас интересует произнесение начального *и-* после слов, заканчивающихся шумной согласной, то запрос будет выглядеть следующим образом (см. рис. 4, где _ обозначает границу слова).

Рис. 4

На выдаче получим порядка 20 тыс. вхождений в интервале с 1930 по 2013 год.

На границе слов можно запросить и сочетание групп, например, запрос на сочетание заднеязычная согласная — граница слова — заднеязычная согласная будет выглядеть следующим образом (см. рис. 5).

Рис. 5

На выдаче получаем порядка 12 тыс. примеров в интервале с 1930 по 2013 год.

3.2. ВОКАЛИЧЕСКАЯ СТРУКТУРА СЛОВА

Специфической особенностью МУРКО является возможность поиска по вокалической структуре слова. Чтобы исследовать редукцию русских гласных в разных позициях относительно ударного слога, фонетистам нужен материал, который представляет собой набор словоформ одной и той же вокалической структуры — вне зависимости от их морфологических и семантических особенностей. Понятно, что без соответствующих поисковых средств отбор такого материала весьма трудозатратен. Именно для осуществления тако-

го отбора предусмотрена опция ***Вокалическая структура слова*** (см. рис. 6).

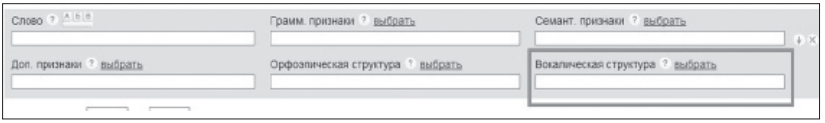


Рис. 6

Предположим, мы хотели бы подобрать фонетический материал следующей структуры: общее количество слогов в словоформе не задано, ударный слог по счету — третий от начала слова, ударная гласная *о*, первый предударный — фонема *о*, первый заударный — фонема *о*. Тогда запрос в диалоговом окне будет выглядеть следующим образом (см. рис. 7).

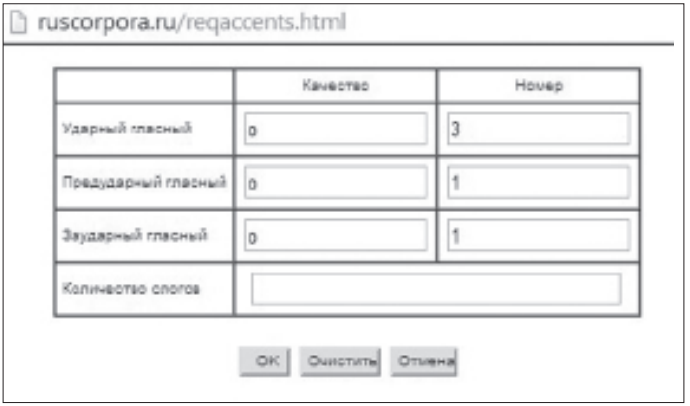


Рис. 7

На выдаче мы получим словоформы *разговорного*, *невозможно*, *паровозом*, *небольшого*, *мирового*, *руководство*, *Циолковского*, *оборотом*, *огромное*, *неспособность* и практически неограниченное, т. е. заведомо превосходящее любые исследовательские потребности число иных словоформ такой же вокалической структуры.

В том случае, если нам требуются данные не по одному, а по нескольким предударным / заударным слогам, запрос формулируется в два приема. Предположим, мы хотели бы получить мате-

риал по запросу: общее количество слогов не определено, ударный слог — пятый от начала, качество ударного слога не определено, все предударные слоги содержат фонему *о*. Сначала заполняем диалоговое окно только для первого предударного слога (см. рис. 8).

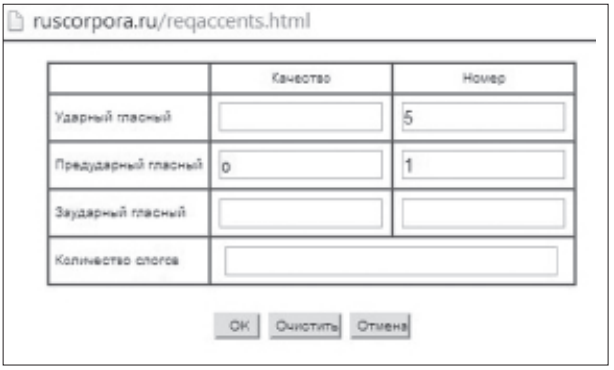


Рис. 8

После нажатия кнопки ОК в поисковой строке видим следующее (см. рис. 9).

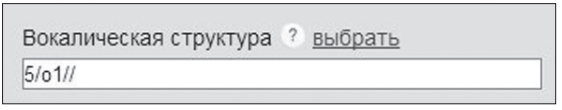


Рис. 9

После этого дописываем руками недостающие слоги прямо в строку поиска (см. рис. 10).

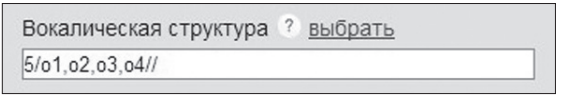


Рис. 10

На выдаче получаем 31 вхождение: *проголосовал*, *проголосовали*, *проголосовать*, *опростоволоситься*, *основоположницей*, *глоттохронологии*, *оплодотворению*, *оплодотворение*, *оплодотворяя*, *оплодотворения*, *оплодотворяются*, *многоговорящие*, *подмногообразии*, *подмногообразия*, *многокомпонентной*, *коротковолновым*,

основоположники, *основоположника*, *поосвобожусь*, *опростоволосилась*, *опростоволосились* (некоторые — по нескольку раз).

3.3. РЕЧЕВЫЕ ДЕЙСТВИЯ И ЖЕСТИКУЛЯЦИЯ

Как мы уже писали ранее, очень небольшая часть МУРКО (глубоко аннотированный МУРКО) имеет расширенную аннотацию, а именно — в нем размечены типы речевых действий и жестикуляция. Эта разметка позволяет отбирать материал и по этим параметрам.

Например, можно выбрать клипы, в которых говорящий осуществляет тот или иной речевой акт. Предположим, мы хотели бы выбрать все клипы, где говорящий благодарит кого-либо или задает общий вопрос. Для этого в разделе Речевые действия на странице поиска мы выбираем *Этикетные высказывания* и далее *Благодарность / Вопросы* и далее *Общие*. Если нам нужно узнать, какие вообще речевые действия размечены и в какую группу они помещены, мы нажимаем на вопросительный знак рядом с закладкой ***Речевые действия*** (см. рис. 11) и получаем полный список речевых актов с разбивкой по группам.

Рис. 11

С помощью разметки речевых действий можно целенаправленно отбирать выказывания определенной семантики, типа иллокуции, перлокутивных функций – для изучения интонационных контуров и иных фонетических особенностей устного высказывания (темп речи, уровень громкости, типы речевого подчеркивания – парцелляция, скандирование – и условия их функционирования, типы междометий и вокальных жестов, типы повторов).

Аналогичным образом устроен жестикуляционный модуль глубоко аннотированного МУРКО. Здесь можно отобрать жесты по их субъективным характеристикам (типу и значению), по объективным характеристикам (активному и пассивному органу, ориентации в пространстве, направлению движения и проч.). Выбрав соответствующие характеристики, пользователь получает

клипы, в которых встречаются жесты заданного типа. Таким образом, МУРКО предоставляет три основных способа получения из корпуса жестикуляционного материала:

1) обращение к глубоко аннотированному МУРКО и запрос жестов с конкретными характеристиками, вне зависимости от того, какой лингвистический ряд их сопровождает, и есть ли такой лингвистический ряд (последний случай — это жесты в режиме пантомимы или жесты слушающего);

2) обращение к тем или иным лингвистическим конструкциям, к отдельным лексемам или классам лексем — с тем, чтобы проанализировать, какие жесты их сопровождают (в частности, в нашей статье в данном сборнике, «О русском жестикуляционном отрицании» был использован именно этот тип запроса,— из МУРКО целенаправленно отбирались фрагменты, включающие в себя разные типы отрицательных конструкций).

Третий способ получения жестикуляционной информации будет нами описан в следующем разделе.

4. Страница выдачи

После формулировки запроса говорящий получает из МУРКО страницу выдачи. Каждый элемент выдачи представляет собой пару «текст расшифровки + соответствующий клип» (в относительно редких случаях, когда клип содержит только жестикуляцию, расшифровка клипа отсутствует; в случае, если элемент выдачи входит в глубоко аннотированный МУРКО, присутствует также таблица жестов, зафиксированных в данном клипе). На рис. 12 показана такая пара:

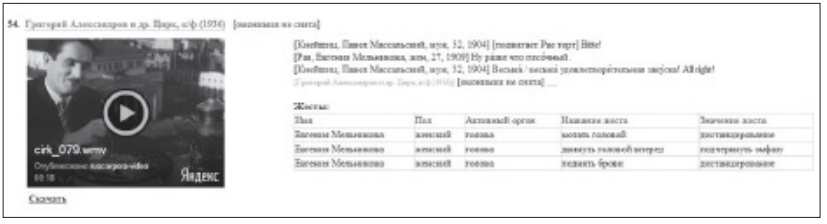


Рис. 12

Клип имеет несколько обозначений, существенных для поль-

зователя. Большая стрелка в кружочке, расположенная в центре, обозначает клавишу проигрыша клипа (при нажатии на нее клип начинает проигрываться). В левом нижнем углу обозначена длительность клипа (на рис. 12–18 сек.) и уникальное имя клипа в базе данных МУРКО (на рис. 12 — cirk_079). В случае, если пользователь захочет сохранить клип на своем компьютере для дальнейшего использования, а не просто просмотреть / прослушать его в режиме он-лайн, ему следует воспользоваться опцией *Скачать* слева под клипом — клип будет сохранен на пользовательском компьютере под своим уникальным именем.

Чрезвычайно важной является функция Расширение контекста (троеточие, которое на рис. 13 взято в кружок).

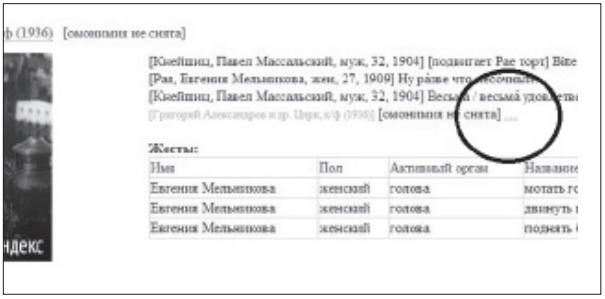


Рис. 13

При нажатии на эту гиперссылку открывается страница, на которой присутствует не набор клипов, а именно этот, выбранный пользователем единичный клип (см. рис. 14).

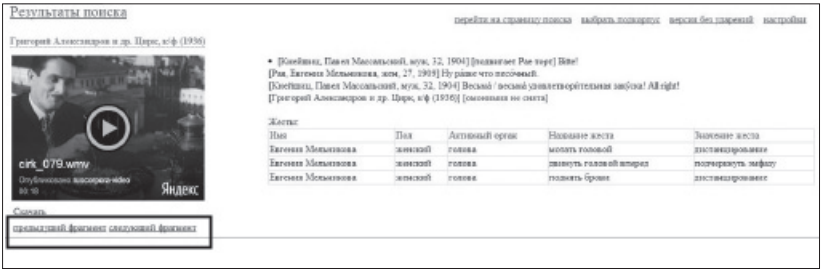


Рис. 14

Такая индивидуальная подача клипа, во-первых, предоставляет возможность получить уникальный адрес данной пары «клип+текст», а во-вторых, просмотреть предшествующий и последующий клипы — в ситуации, когда целый текст при подготовке корпуса расчленялся на небольшие фрагменты, эта проблема является весьма актуальной, поскольку без обращения к пред- и послетексту содержание данного клипа может оказаться абсолютно неясным. Для этого следует обратиться к ссылкам *предыдущий фрагмент* и *следующий фрагмент*, расположенным сразу под клипом (на рис. 14 — в прямоугольнике). Эта же опция предоставляет пользователю возможность просмотреть материал целиком — от клипа к клипу — и тем самым, среди прочего, осуществить третью возможность получения жестикуляционной информации из МУРКО, а именно, — с помощью сплошного просмотра видеоматериала.

5. Развитие МУРКО

5.1. КОЛИЧЕСТВЕННОЕ РАЗВИТИЕ

Мультимедийный модуль должен, прежде всего, развиваться количественно. Ситуация, когда бóльшая часть текстов относится к кинематографической речи, является для ряда исследований существенным ограничителем. Поэтому в перспективе некинематографическая часть МУРКО должна по крайней мере не уступать по объему речи кино. За последние два года в составе корпуса существенно увеличился объем научной речи. Это направление пополнения корпуса следует в обязательном порядке продолжать, но не менее насущным является и добавление других типов дискурса. Прежде всего это касается политической и бытовой речи. При этом при нынешнем уровне развития интернета добавление политического дискурса в МУРКО не представляет практически никаких сложностей и ограничено лишь недостатком ресурсов (людских и материальных) для обработки уже имеющегося в наличии материала. Что же касается бытовой речи, то ее помещение в корпус ограничено, так сказать, нематериально — создатели корпуса не имеют ни юридического, ни морального права помещать «личное» видео в открытый доступ без согласия участников и владельцев этого видео, даже если исходные файлы расположены на открытых площадках, например, на YouTube. Как следствие, принципиальные

изменения в пополнении подкорпуса непубличной речи в МУРКО возможны, по-видимому, только если удастся задействовать краудфандинг, т. е. добровольную передачу в МУРКО видео из личных архивов их владельцами.

Минимальным рабочим объемом МУРКО, который позволит решать бóльшую часть исследовательских задач, на наш взгляд, можно считать объем в 10 млн словоупотреблений, при том что как минимум половина из этого объема должна представлять собой неподготовленную или слабоподготовленную речь максимально широкой тематики.

5.2. КАЧЕСТВЕННОЕ РАЗВИТИЕ

5.2.1. Программное обеспечение

Такой корпус, как МУРКО, — открытый, т. е. бесплатный, и достаточно большой, — на наш взгляд, должен иметь свою программу первичной обработки и описания видео- и аудиоматериала. Разумеется, для таких целей уже созданы и широко используются мощнейшие и при этом бесплатные программы, прежде всего, ELAN, ANVIL, PRAAT, Speech Analyzer. При безусловном высочайшем качестве этих программ они имеют один существенный недостаток (это не касается Speech Analyzer): из-за огромного числа выполняемых функций программы эти слишком сложны и требуют специального обучения. Исследователю же очень часто требуется элементарная проверка собранного из корпуса материала на предмет того, — есть ли какие-то интересные для него явления в отобранном материале или нет, и если нет — то как следует поменять стратегию поиска и структуру запросов.

Предварительное техническое задание для программы мультимедийной разметки

1) Предполагается, что программа будет расположена на сайте Национального корпуса русского языка (точнее — в подкорпусе МУРКО) и будет лицензирована для свободного использования.

2) Задача — создать максимально простое и дружественное средство для разметки аудио- и видеоряда. **На входе** — видеоклип + расшифровка текста в формате xml или txt (если клип не имеет речевого фрагмента — только клип). **На выходе** — таблица Excel.

3) Программа включает в себя окно видео. Предполагается,

что все видео- и аудиоформаты, работающие на компьютере пользователя и на странице МУРКО, опознаются программой, т.е. переконвертации мультимедиа не требуется.

4) Должна быть предусмотрена опция *Кадрировать*, т.е. из видео- аудиофайла большого размера пользователь должен иметь возможность вырезать необходимый ему фрагмент прямо внутри программы. Если пользователю необходимо последовательно работать с разными фрагментами одного и того же клипа, то он должен иметь возможность, не выходя из программы, перезагружать исходный клип, чтобы его снова кадрировать.

5) При проигрывании клипа должно быть предусмотрено удобное замедление прогона (градуальное, с помощью движка).

6) Пользователь должен иметь возможность размечать клип в четырех ипостасях:

— видео

— waveform

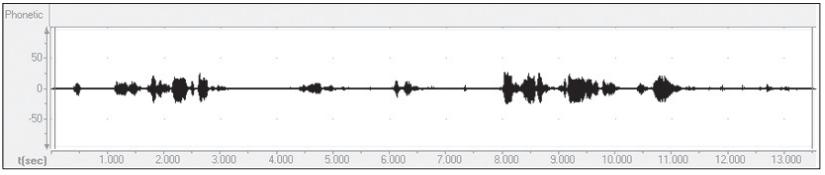


Рис. 15

– pitch (уровень тона)

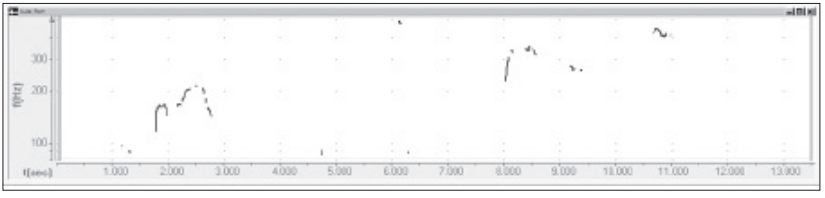


Рис. 16

— текстовом: если на входе в программу подается текст в виде xml, то xml-разметка должна сохраняться внутри программы, но быть невидимой и не доступной редактированию, чтобы при экспорте разметки из программы в результирующую таблицу Excel она

могла перейти туда в первоначальном виде. Кроме того, при экспорте данных из программы в результирующую таблицу Excel у пользователя должна быть возможность снять разметку xml и получить на выходе текст в формате txt.

7) Разметка текстового, видео- и аудиоряда происходит посредством привязки с помощью градуального движка к шкале времени, размеченной в миллисекундах. Таким образом, после проведенной разметки каждое слово (слог, часть слога, пауза), каждый элемент жеста, каждый элемент интонационного контура оказывается привязанным к одной и той же, общей для всех компонентов шкале времени (при этом жестикуляция разбивается на несколько линий — направление взгляда, закрытие/открытие глаз, движение головы, направление движение правой руки, левой руки, обеих рук, конфигурация ладони, ориентация, степень напряжения руки, движение корпуса; можно выбрать отдельную линию или две линии одновременно). Эта привязка к одной и той же шкале позволяет при выгрузке комбинировать данные: текст может автоматически накладываться на waveform, на уровень тона, на жестикуляционные линии, waveform и уровень тона могут комбинироваться с разметкой жестов.

8) Выгрузка разметки.

8.1) Основная идея состоит в том, что размеченный клип/фрагмент клипа должен в один клик выгружаться в таблицу Excel и формировать тем самым заготовку для базы данных, с которой затем пользователь будет работать. Пользователь выбирает, какая пара данных его интересует (например, waveform и текст, или направление движения руки, уровень тона и текст, или текст и направление взгляда, и т. д.), программа соединяет соответствующие данные и для пар, включающих в себя waveform или уровень тона, формирует графический объект (неподвижный или анимированный) под именем исходного видео- или аудиофайла с соответствующим уточнением в расширении. Для пар, включающих в себя текст и разметку жестов, формируется текстовый файл с обозначением жестового ряда, например, в верхнем регистре и/или в скобках — также под именем исходного видео-аудиофайла с соответствующим поясняющим расширением. При этом у пользователя должна быть возможность сохранить или убрать исходную xml-разметку, если она была в загружаемом файле.

8.2) Далее формируется таблица Excel в качестве заготовки для будущей базы данных: первая ячейка – гиперссылка на исходный видео- аудиофайл, вторая и далее ячейки – гиперссылки на файлы из п. 8.1, последняя ячейка – текстовый ряд файла в формате txt

9) Пожелание. Инструкция к программе не должна занимать более пяти страниц.

5.2.2. Мультимедийный параллельный русский корпус (МультиПАРК)

В конце 2014 г. открыт новый модуль в составе МУРКО, ориентированный на сопоставительные исследования.

Как известно, одним из самых серьезных ограничений при изучении устной речи является ее невоспроизводимость, а именно, невозможность получить одно и то же высказывание в одной и той же ситуации от разных говорящих. Это ограничение не соблюдается только для этикетных формул и иных фиксированных стандартизованных социальных реакций, которые в одной и той же ситуации у разных говорящих в значительной степени сходны. Все устные высказывания более сложной структуры уникальны в том смысле, что осуществляются только один раз и, не будучи зафиксированы на те или иные носители, не могут быть воспроизведены вместе с той ситуацией, которая их породила.

При этом чрезвычайно интересным и непростым является вопрос — что в данном устном высказывании является обязательным для любого говорящего, а что колеблется от говорящего к говорящему. Единственным способом приблизиться хотя бы вчерне к ответу на этот вопрос является создание для говорящего возможности произнести данную фразу в одной и той же ситуации. Понятно, что в реальной жизни — за рамками возможных искусственно организованных экспериментов — таких возможностей чрезвычайно мало (если мы, опять же, выходим за пределы этикетных формул). Однако такую возможность нам предоставляет сфера искусства.

Для исследования способов осуществления одного и того же высказывания разными говорящими в одной и той же ситуации (пусть не естественной, а смоделированной) и предназначен МультиПАРК. Предполагается, что МультиПАРК будет состоять из двух зон, направленных на выполнение двух разных, хотя и идеологически сходных задач.

5.2.2.1. Русскоязычный МультиПАРК.

Для проведения сопоставительного анализа русской устной речи предполагается использовать экранизации и театральные постановки одной и той же пьесы. Так, например, наличие трех экранизаций, порядка семи (легко доступных) театральных постановок и нескольких радиопостановок пьесы Н. В. Гоголя «Ревизор» дает нам уникальную возможность выровнять между собой и сопоставить более десятка вариантов произнесения начальной реплики пьесы «Господа, я собрал вас, чтобы сообщить пренеприятное известие: к нам едет ревизор». И таким образом может быть «размножена» не одна только эта фраза, а вся пьеса Гоголя. И не только Гоголя, но и, например, Чехова, Вампилова, Розова, Островского, и вообще любого другого автора, драматургия которого достаточно популярна для того, чтобы а) быть воспроизведенной хотя бы два раза, б) быть записанной на электронные носители, позволяющие воспроизводить данную постановку современными электронными средствами. Сопоставление разных произнесений одной и той же фразы в одной и той же ситуации профессиональными актерами, перед которыми поставлена задача произнести данную фразу максимально естественно, предоставляет нам возможность определить, какие интонационные, темповые, структурные (паузы), фонетические, жестовые особенности этой фразы являются обязательными, т. е. воспроизводимыми всеми говорящими, какие — уникальными или случайными. Кроме того, такое сопоставление дает нам возможность анализировать устный синтаксис с совершенно уникальных позиций, поскольку появляется реальный шанс определить, какие регулярные изменения в порядке слов происходят при переходе от письменного текста исходного драматического произведения к реальному звучанию на сцене или на экране. Естественно, на возможные выводы накладываются ограничения, связанные с искусственностью говорения в театральных и кинематографических условиях, но, тем не менее, определенные выводы, существенные как для понимания устной русской речи, так и русского языка в целом, мы сделать сможем.

5.2.2.2. Англо-русский МультиПАРК.

За три года существования Мультимедийного русского корпуса по его материалам было проведено много исследований, вполне

новаторских, результаты которых были доложены в интернациональных аудиториях. Практически всегда возникал вопрос — зафиксировано ли данное явление, характерное для устной русской речи (фонетическое или жестикуляционное), в других языках, прежде всего в английском? До сих пор никакого обоснованного ответа на вопросы такого рода дать было невозможно, поскольку не существовало корпусов, способных предоставить искомый материал в необходимом количестве.

Как представляется, ответом на данный запрос научного сообщества может стать англо-русский подкорпус МультиПАРКа. Как известно, параллельные подкорпуса и методика их подготовки уже хорошо освоены в Национальном корпусе русского языка. Однако все они ориентированы на письменную речь. Представляется, что МультиПАРК предоставит нам возможность проводить сопоставительные исследования и на устном материале.

Общая архитектоника англо-русского МультиПАРКа. В корпус будут включены расшифровки 1) сериалов и фильмов на английском языке, 2) расшифровки соответствующего русского дубляжа. Каждая видеодорожка (английская и русская) будет разрезана на небольшие фрагменты, одинаковые по содержанию. Каждый видеофрагмент будет выровнен с соответствующим фрагментом английского или русского транскрипта. Текстовые транскрипты при этом будут размечены по методике, принятой в НКРЯ и МУРКО (морфологическая и семантическая разметка для английской и русской зоны, акцентологическая, орфоэпическая разметка и разметка вокалической структуры слова — для русской зоны). Таким образом, за запрос пользователя будут выдаваться две пары «клип+текст» (на английском и русском языках), в которых будут выровнены между собой как видео-, так и текстовый ряд.

Такая подача материала позволит вести сопоставительные исследования в области интонации и фонетики, в области лексики, синтаксиса и семантики (в частности, как известно, английский язык является языком с твердым порядком слов,— даже в устной речи,— в отличие от русского, кроме того, имеет гораздо менее разветвленную систему частиц, которые в устной русской речи играют колоссальную роль; сопоставление устных текстов, полностью совпадающих по содержанию и конситуации, позволит проводить чрезвычайно плодотворные сопоставления столь разных языковых

систем). Несколько более сложным, однако, вполне реальным, будет проведение сопоставительных жестикуляционных исследований на данном корпусе. Во-первых, через поисковые формы корпуса исследователь получит прямой доступ к тем или иным лексемам, морфологическим характеристикам и синтаксическим конструкциям в английском тексте, которые (как и в русской жестикуляции) могут быть связаны с теми или иными жестами. Кроме того, получение жестикуляционных данных для английской жестикуляции будет возможно не напрямую, а через русский параллельный текст. Например, русская указательная частица *вот*, одно из самых распространенных слов в устной русской речи, 1) не имеет хорошего однословного перевода на английский язык, 2) регулярно сопровождается в русской жестикуляции указательными жестами строго определенного типа. Тем самым, получение английского жестикуляционного материала через обращение к русским контекстам с *вот* позволит определить, каким именно образом семантика вот передается в английском тексте и какие именно указательные жесты сопровождают соответствующий семантический компонент.

Литература

Grishina E. Multimodal Russian Corpus (MURCO): general structure and user interface. // NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25–27 November 2009. Proceedings. Tribun, 119–131

Grishina E. Multimodal Russian Corpus (MURCO): types of annotation and annotator's workbenches // Corpus Linguistics Conference CL2009, Universuty of Liverpool, UK, 20-23 July 2009b, http://ucrel.lancs.ac.uk/publications/cl2009/#papers,%20#165

Grishina E. Multimodal Russian Corpus (MURCO): First Steps // 7th Conference on Language Resources and Evaluation LREC'2010, Valetta, Malta/ http://www.lrec-conf.org/proceedings/lrec2010/pdf/143_Paper.pdf

Grishina E., Savchuk S., Sichinava D. Multimodal Russian Corpus (MURCO): Studying Emotions // 7th Conference on Language Resources and Evaluation LREC'2010, Valetta, Malta. Workshop Best Practice for Speech Corpora in Linguistic research/ http://www.corpora.uni-hamburg.de/lrec2012/Proceedings_Complete.pdf

Гришина Е. А. Национальный корпус русского языка как источник сведений об устной речи // Речевые технологии, № 3, 2008, с. 50–62

Гришина Е. А., Савчук С. О. Корпус звучащей русской речи в составе Национального корпуса русского языка // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7(14). По материалам международной конференции Диалог'2008. М., 2008. С. 125–132

Гришина Е. А., Савчук С. О. Корпус устных текстов в Национальном корпусе русского языка: состав и структура // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009. С. 129–149

Гришина Е. А. Мультимедийный русский корпус (МУРКО): проблемы аннотации // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009а. С. 150–174

Гришина Е. А. Корпус «История русского ударения» // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб., 2009б. С. 175–214

Гришина Е. А., Кудинов М. С. Инструменты полуавтоматической разметки для Мультимедийного русского корпуса (МУРКО) // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог 2009»). Вып. 8 (15), М., 2009. С. 248–261

Гришина Е. А. Мультимедийный русский корпус (МУРКО): современное состояние и перспективы развития // Труды международной конференции «Корпусная лингвистика – 2011», СПб., 27–30 июня 2011. С. 138-144

Захаров В. П. Корпуса русского языка // Труды Института русского языка им. В. В. Виноградова. Вып. 6. М., 2015. С. 20–64.

Elena A. Grishina
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
rudi2007@yandex.ru

THE MULTIMODAL MODULE AS PART OF THE RUSSIAN NATIONAL CORPUS

The article describes the Multimodal Russian Corpus (MURCO) as the new project in the framework of the Russian National Corpus (RNC). The pilot version of the MURCO was opened for general access in October, 2010 and since then MURCO has increased its volume up to 4.3 mln tokens and developed its structure. The corpus presently consists of the following subcorpora: 1) Movie speech, 2) Public Spoken Russian, 3) Private Spoken Russian 4) Theater speech 5) Written-to-be-Spoken Russian.

The MURCO is organized as a collection of clixts. A clixt is a pair of an audio- or video clip and the corresponding fragment of the text transcript. A user has the opportunity to download not only the text component of a clixt, but also its sound and video component, so after downloading a user may use any program to analyze it.

MURCO is marked up with different types of annotation. Some of them are standard for the RNC (metatextual, morphological, semantic annotation), some types are special for the spoken component of the RNC (sociological and accentological annotation), and some of the mark-up dimensions are specific only for the MURCO (orthoepic, annotation of vocal structure, speech act and gesture annotation). The speech act and gesture annotation is used in the smaller part of the MURCO, which is called the deeply annotated MURCO.

The article describes the main ways of retrieving information from the corpus and types of search queries. The prospects of further development of MURCO are as follows: increasing the number of texts, developing software for processing multimedia content, creating new modules of MURCO – multimodal parallel corpora.

Key words: Multimodal Russian Corpus, Russian National Corpus, Multimodal Parallel Corpus, annotation, searching instruments.

References

Grishina E. Multimodal Russian Corpus (MURCO): general structure and user interface. *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference,* Smolenice, Slovakia, 25–27 November 2009. Proceedings. Tribun, 119–131.

Grishina E. Multimodal Russian Corpus (MURCO): types of annotation and annotator's workbenches. *Corpus Linguistics Conference CL2009,* Universuty of Liverpool, UK, 20–23 July 2009. Available at URL: http://ucrel.lancs.ac.uk/publications/cl2009/#papers,%20#165 [accessed 3.05.2015].

Grishina E. Multimodal Russian Corpus (MURCO): First Steps. *7th Conference on Language Resources and Evaluation LREC'2010,* Valetta, Malta. Available at URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/143_Paper.pdf [accessed 3.05.2015].

Grishina E. Multimodal Russian Corpus (MURCO): Studying Emotions. Third International Workshop on emotion: Corpora for research on emotion and affect. *7th International conference on language resources and evaluation (LREC 2010)*. Valletta Malta. Available at: http://www.academia.edu/224975/Multimodal_Russian_Corpus_MURCO_Studying_Emotions (accessed 3.05.2015)

Grishina E., Savchuk S., Sichinava D. Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure. *8th Conference on Language Resources and Evaluation LREC'2012*, Istanbul, Turkey. Workshop Best Practices for Speech Corpora in Linguistic research. Available at URL: http://www.lrec-conf.org/proceedings/lrec2012/workshops/03.Speech%20Corpora%20Proceedings.pdf (accessed 3.05.2015).

Grishina E. A. [National corpus of the Russian language as a source of information on oral speech]. *Rechevye tehnologii*, 2008, no. 3, pp. 50–62. (In Russ.)

Grishina E. A., Savchuk S. O. [Speech Corpus within the Russian National Corpus]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Vypusk 7(14). Po materialam mezhdunarodnoi konferentsii Dialog'2008* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2008)]. Issue 7(14). Moscow, RSUH Publ., 2008, pp. 125–132. (In Russ.)

Grishina E. A., Savchuk S. O. [Corpus of spoken texts in the Russian National Corpus: the composition and structure]. *Natsional'nyi korpus*

russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy [The National Corpus of the Russian Language: 2006–2008. New results and perspectives]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 129–149. (In Russ.)

Grishina E.A. [Multimodal Russian Corpus (MURCO): Problems of Annotation]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2006-2008. New results and perspectives]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 150–174. (In Russ.)

Grishina E. A. [Corpus "The History of Russian accent"]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [The National Corpus of the Russian Language: 2006–2008. New results and perspectives]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 175–214. (In Russ.)

Grishina E. A., Kudinov M. S. [Tools of semi-automatic markup for Multimodal Russian corpus (MURCO)]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii (po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog 2009»* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference «Dialogue» (2009)]. Issue. 8 (15). Moscow, RSUH Publ., 2009, pp. 248–261. (In Russ.)

Grishina E. A. [Multimodal Russian Corpus (MURCO): current state and perspectives]. *Trudy mezhdunarodnoi konferentsii* "*Korpusnaya lingvistika — 2011*" [Proceedings of International Conference "Corpus Linguistics — 2011"]. St. Petersburg, St. Petersburg Univ. Publ., pp. 138–144. (In Russ.)

Zakharov V. P. [Corpora of the Russian language]. *Trudy instituta russkogo yazyka im. V. V. Vinogradova* [Proceedings of V. V. Vinogradov Russian language Institute], no. 6. Moscow, 2015, pp. 20–64. (In Russ.)

А. М. Молдован
Институт русского языка им. В.В.Виноградова РАН
(Москва, Россия)
alexandr.moldovan@gmail.com

ПАМЯТНИКИ ДРЕВНЕРУССКОЙ ПИСЬМЕННОСТИ В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА

В статье рассматриваются различные типы корпусного представления славянских рукописных памятников и их возможности при изучении языка древнерусской письменности.

Ключевые слова: корпусная лингвистика, древнерусская письменность, церковнославянский язык, источниковедение, лингвистическая документация.

Корпусное представление памятников древнерусской письменности — одна из форм их лингвистической документации, необходимой для того, чтобы памятники можно было дифференцировать по лингвистическим параметрам. Чем детальней документация, тем больше шансов получить данные для региональной и хронологической стратификации памятников, анализируя представленные в них различные функциональные комбинации генетически разнородных и разновременных лингвистических черт. Предпосылки для лингвистической документации создает источниковедческий анализ текстов, в ходе которого памятник письменности должен быть датирован, прочтен и переведен на язык грамматических определений, при этом в нем должны быть выявлены разнообразные искажения, описки, пропуски и вставки, определены его взаимоотношения с родственными текстами, объяснены непонятные слова и выражения и т. д. Это чрезвычайно трудоемкое дело, поэтому, хотя древние славянские памятники изучаются давно, объем неизученного все еще весьма велик.

Помимо источниковедческого исследования рукописных источников, при создании корпуса требуется решение специфических для «исторических» корпусов проблем — прежде всего кодировки нестандартных кириллических символов, релевантных для текстов разных типов; таксономии древнерусских текстов, параметров грамматической разметки, репрезентативности источников и т. д. [Молдован 2006]. Они активно обсуждаются, и много для их решения уже сделано[1], в частности, в последние версии Unicode включены почти все необходимые символы старой кириллицы. Работа в этом направлении ведется в славистических центрах России, Болгарии, США, Германии, Финляндии, Норвегии и некоторых других стран. К сожалению, достичь существенной координации в этой работе не удается, поскольку разнообразные задачи решаются в рамках грантов, налагающих на исполнителей различные ограничения и предполагающих короткие сроки выполнения. В итоге появилось множество разнообразных вариантов электронных изданий и программ компьютерной обработки славянских рукописных памятников.

Низшую ступень занимают электронные копии рукописей — сканированные изображения, выставленные в интернете или распространяемые в копиях. Их уже довольно много — и в интернете, и в частном обращении. Наиболее крупная коллекция такого рода — оцифрованные собрания Троице-Сергиевой лавры, Московской духовной академии и некоторые другие. Эта работа выполняется совместно Российской государственной библиотекой и Свято-Троицкой Сергиевой Лаврой. Благодаря ей, мы располагаем доступом к уникальным рукописям и ценным документам общим числом более пяти тысяч единиц (http://old.stsl.ru/manuscripts/). К сожалению, аналогичные работы в других хранилищах рукописей осуществляются лишь в ограниченном объеме.

Другой тип представления древних источников — это электронные книги, т. е. сканированные изображения печатных изданий памятников в форматах PDF, DjVu и т. п. Это наиболее многочисленная категория, включающая целые библиотеки (например, многотомную «Русскую историческую библиотеку» или Полное собрание русских летописей). Работа по оцифровке книг поддерживается

[1] См., в частности, материалы Международной комиссии по компьютерному изучению славянских памятников при Международном комитете славистов: http://www.obshtezhitie.net/

библиотеками, в ней заинтересованы все, поэтому можно полагать, что со временем все изданные памятники будут оцифрованы и доступны.

Третий тип представляют наборные цифровые издания текста отдельных памятников. В отличие от предыдущих двух, этот тип изданий позволяет осуществлять примитивный поиск по словам и буквосочетаниям (стрингам). Одним из первых опытов такого типа были цифровые издания, выполненные в рамках Хельсинкского проекта (Corpus Cyrillo-Methodianum Helsingiense: http://www.helsinki.fi/slaavilaiset/ccmh/). Благодаря ему специалисты впервые получили доступ к электронным текстовым версиям пяти важнейших источников из старославянского «канона» (евангелия — Ассеманиево, Зографское, Мариинское и Саввина книга, а также Супрасльская рукопись), а также к житиям Кирилла и Мефодия.

По этому пути пошли в 2002–2005 гг. киевские историки, которые создали большой сайт «Літопис» (http://litopys.org.ua/), поместив на нем наборные издания Лаврентьевской, Ипатьевской и Новгородской первой летописей, а также большое количество старо-украинских источников.

Простой набор представляет и болгарско-норвежский проект Corpus of Old Slavic Texts from the XIth Century (http://www.hf.ntnu.no/SofiaTrondheimCorpus/), включающий значительную часть славянских памятников XI–XII вв. К сожалению, в отличие от хельсинкского и киевского, на этом сайте поиск невозможен, видимые на экране тексты можно только читать и копировать, если на вашем компьютере есть соответствующий (ныне устаревший) шрифт.

Существуют разнообразные электронные издания памятников, соединяющие традицию бумажных изданий с некоторыми возможностями цифрового (в частности, они позволяют копировать фрагменты). Таково, например, издание Повести временных лет по нескольким спискам, подготовленное Дэвидом Бирнбаумом (http://clover.slavic.pitt.edu/pvl/ost1.html), в котором, наведя курсор на ту или иную строчку, можно получить все разночтения, представленные в пяти списках и в трех основных изданиях ПВЛ.

Свои преимущества имеет перенесенное в цифровой вид Дэвидом Бирнбаумом двухтомное издание Супрасльской рукописи (http://suprasliensis.obdurodon.org/) [Eckhoff, Birnbaum, Miltenova, Dimitrova 2012] и выполненное под его руководством издание Бдин-

ского сборника (http://bdinski.obdurodon.org/) [Sels, Stern 2012]. Удобство таких изданий в том, что здесь можно проверить сомнительные написания по фотокопии. Но по информативному потенциалу они остаются на уровне бумажного издания, отличаясь от него только тем, что фрагменты славянского или греческого текста можно скопировать.

Принципиально иначе подходят к представлению текстов создатели сайта «Манускрипт» в Ижевске (http://mns.udsu.ru/). На сайте представлена большая коллекция древнейших и средневековых славянских и русских текстов, подготовленных авторами или полученных от русских и болгарских коллег. Каждый текст на этом сайте представляет собой самостоятельную базу данных, допускающую некоторое преобразование исходного текста: можно выбрать опцию «текст оригинальный» и получить наборный текст без разделения на слова; можно выбрать «текст преобразованный» и получить текст с разделением на слова; можно по выбранному отрезку заказать выдачу индекса словоформ — прямого, обратного и частотного. Наведя курсор на адрес, можно выйти на соответствующее место в тексте — это весьма удобно. Возможен поиск по буквосочетаниям [Гнутиков и др. 2005].

Стремясь расширить информационные возможности системы, создатели сайта приняли участие в разработке совместно с Институтом русского языка им. В. В. Виноградова РАН автоматической системы морфологического анализа древнерусских словоформ (морфологического анализатора). Созданная программа опирается на накопленный в Институте русского языка фонд грамматически размеченных словоформ, представленных в ряде древнерусских памятников (http://mns.udsu.ru/mns/slov.prost_poisk) и дает неплохие результаты в пределах этого фонда. Но если мы выйдем за его пределы и зададим запрос, например, на слово **досѧщи**, которое содержится в Пандектах Антиоха (этот источник имеется в ижевском корпусе, но он не входит в корпус размеченных текстов ИРЯ РАН), программа выдаст отказ. Эта программа, кроме того, не умеет отождествлять разные орфограммы, поэтому она не узнаёт слово, если оно написано не в «канонической» орфографии — т. е. не в той нормализованной орфографии, в которой эти словоформы записаны в программе. Нечленная форма слова **гръдъ** в южнославянском написании получает правильный разбор. Но если мы напишем

это слово иначе, например, **гърдъ**, или зададим членную форму **гръдъи**, программа эти написания не опознает. Не справляется программа и с омонимией словоизменительных форм, чрезвычайно характерной для языка славянской письменности. Результат поиска нередко содержит слишком много альтернативных грамматических определений, чтобы его можно было считать удовлетворительным. Например, словоформа бързости имеет *десять* вариантов определений (*ед. род., ед. дат., ед. местн., ед. зват., мн. им., мн. зват., мн. вин., дв. им., дв. вин., дв. зват*.).

Подобные результаты получены и в других работах над автоматическим морфологическим анализатором — в Регенсбургском диахроническом корпусе (http://rhssl1.uni-regensburg.de/SlavKo/korpus/rrudi-new/) и в работе над проектом PROEL в Осло (https://nestor.uit.no/). Степень точности таких программ повышается для данного типа текстов при внесении в них дополнительной (в том числе синтаксической) информации. Однако объем необходимых для этого усилий пока еще вполне сопоставим с тем, который требуется для «ручной» разметки словоформ — заведомо более точной.

На протяжении многих лет в Институте русского языка им. В. В. Виноградова РАН создавался и на сегодня в значительной степени создан корпус древнерусских текстов домонгольской поры, включающий почти все восточнославянские переводы с греческого («История Иудейской войны», «Житие Андрея Юродивого», «Пчела», «Александрия», Изборник 1076 г. и другие памятники). Указатели к ним делались «вручную». С самого начала в корпус были заложены такие важные для исследователя параметры, как собственно тексты рукописей, грамматическая информация о каждом слове, начальная форма слов, греческие соответствия, текстологические варианты. Система построена так, что позволяет получить словарный индекс и церковнославянско-греческий словарь для одного или нескольких текстов. Наличие такой системы обеспечивает создание в полуавтоматическом режиме грамматических указателей к другим текстам — исследователю остается только выбирать из тех подсказок, которые предлагает ему программа.

Сначала этот корпус (точнее, отдельные базы данных) строился на основе Microsoft Access 2.0 [Зобнин, Пичхадзе 2005]. Со временем базы были конвертированы в новый формат и размещены в ин-

тернете. Базы устроены так, что можно, наведя курсор на любое слово в тексте, получить грамматическую информацию об этом слове. Перейдя на страницу поиска, можно получить выборку по искомой словоформе. При печатании запроса появляется подсказка с перечнем подходящих слов. Выбрав любое слово в выборке, можно перейти к его контексту. Кроме того — и это самое ценное — система позволяет вести поиск не только по словам и словоформам, но и по грамматическим категориям.

В этот формат переведены и представлены в интернете Повесть временных лет, Киевская, Галицкая и Волынская летописи по Ипатьевскому списку и Суздальская летопись по Лаврентьевскому списку (http://www.lrc-lib.ru/index.php%3fid=5). Единственным недостатком такого представления до недавнего времени было то, что нельзя было вести поиск сразу по всем памятникам. Это препятствие удалось снять при включении баз в Национальный корпус русского языка (НКРЯ), располагающий соответствующими возможностями.

Расширение хронологического диапазона было закономерным для развития НКРЯ, поскольку корпус, претендующий на полноту представления языка, не может обойтись без документации всей письменной истории русского языка, начиная с первых памятников XI в.

В настоящее время в составе Национального корпуса есть четыре подкорпуса, представляющих источники XI–XVII вв.: древнерусский, корпус берестяных грамот, старорусский и церковнославянский корпус. Эти условные названия присвоены им временно, условна и сама стратификация этих корпусов, связанная с их историей, качественными характеристиками и особенностями их наполнения.

Древнерусский подкорпус (http://ruscorpora.ru/search-old_rus.html) включает три типа памятников: оригинальные древнерусские произведения, выполненные на Руси переводы с греческого и памятники южнославянского происхождения, переписанные на Руси. Из оригинальных произведений в подкорпус входят все летописи, созданные в древнерусский период: Повесть временных лет, Киевская, Галицкая, Волынская летописи (все по Ипатьевскому списку), Новгородская I летопись (по Синодальному списку) и Суздальская летопись (по Лаврентьевскому списку). Кроме того, представлены некоторые сочинения Кирилла Туровского: «Сказание о черноризе-

ском чине», «Притча о душе и теле», «Повесть о беспечном царе и мудром советнике».

Из переведенных в Древней Руси памятников в подкорпусе сейчас доступны «История Иудейской войны» Иосифа Флавия, «Пчела», «Александрия» и цикл из восьми «Чудес» Николая Мирликийского, шесть из которых переведены на Руси или представляют собой переработку перевода, а два возникли у южных славян.

Русские списки древнеболгарских переводов представлены в подкорпусе Изборником 1076 г.

Поиск в древнерусском корпусе возможен по грамматическим признакам, перечисленным в таблице грамматических параметров НКРЯ. В результатах запроса можно получить краткие сведения о памятнике и указание на издание или рукопись, по которым выполнена разметка. При наведении курсора на любую словоформу в выборке появляется всплывающее окно с грамматической характеристикой этой словоформы, а в строке «Доп. признаки» указан адрес — страница издания памятника или лист рукописи и номер строки. У переводных памятников в этом окне приводится греческое соответствие данной словоформы. Например, к слову **безъславие** из «Пчелы», имеющему форму винительного падежа, указана греческая параллель ἀδοξίας в форме аблятива.

Леммы в этом корпусе имеют древнерусскую форму, т. е. отражают состояние языка до падения и прояснения редуцированных.

Общий объем словоформ древнерусского корпуса составляет 443 тысячи словоупотреблений. Это немного в сравнении с другими корпусами, особенно с основным корпусом НКРЯ. Но необходимо подчеркнуть, что, в отличие от большинства других, этот корпус получен не путем автоматической разметки, а в результате кропотливого филологического труда. Поэтому в нем нет неснятой омонимии — все омонимы разведены грамматическими определениями.

Аналогичным образом устроен подкорпус берестяных грамот, в котором грамматическая разметка является результатом выполненного А. А. Зализняком тщательного палеографического и филологического анализа.

Основу церковнославянского корпуса (http://ruscorpora.ru/search-orthlib.html) составляют современные богослужебные тексты (XIX–XX век) — около 60% объема корпуса. Кроме того, в корпусе содержатся более ранние тексты, относящиеся к XVII–XVIII

векам и представляющие другие жанры: Писание, святоотеческие сочинения и др. Общий объем корпуса — около пяти млн словоупотреблений. Поскольку новоцерковнославянский язык обладает известной грамматической и орфографической упорядоченностью, для этого корпуса была создана программа грамматического разбора словоформ, которая позволяет искать слова по лемме и грамматическим признакам, причем можно искать и по сочетанию слов и признаков. Например, можно задать запрос такого вида: «найти сочетания прилагательных с существительным *слава* в родительном падеже с расстоянием между ними до пяти слов». Результатом будут не только сочетания **сла́вы вре́менныя, сла́вы небе́сныя, сла́вы бж҃ія, сла́вѣ ва́шей** и т. п., но и более сложные определительные конструкции: **ра́дуйся, сла́во ц҃ркве всели́нскія, сла́вою ѹкраша́ются бж҃е́ственнагѡ причáстія, сла́вѣ бж҃ственнагѡ ꙁра́ка** и т. п.

Это дает очень широкие возможности для поиска лингвистического материала. Разумеется, нужно учитывать, что автоматическая разметка оставляет нераспознанными грамматические омонимы, и при запросе информации о той или иной словоформе система выдает все варианты характеристик. Например, для действительного причастия прошедшего времени, множественного числа **ца́рствовавшихъ** будут предложены два варианта определений падежа: родительный и местный. Снимать эту омонимию нужно вручную.

Старорусский корпус (http://ruscorpora.ru/search-mid_rus.html) представляет прежде всего тексты, создававшиеся в XV–XVII вв. Однако специфика древнерусской книжности состоит в том, что лингвистические инновации и развитие репертуара текстов не препятствовали сохранению рукописной традиции древних текстов, так что развитие языка текстов разного типа и назначения образует непрерывный континуум от XI-го до XVII-го века. Это значит, что в дальнейшем корпус должен будет объединить все источники XI–XVII вв. В настоящее время в корпусе содержится 4461 документ общим объемом около 6 миллионов словоупотреблений. В него включены доступные в цифровом виде научные публикации текстов разных жанров: летописи, памятники бытовой и деловой письменности, религиозная литература и др. Разметка в этом корпусе пока только метатекстовая (название текста, дата создания, дата списка, автор, жанр и т. п.), поэтому лингвистический поиск

возможен только по словам или частям слов. Однако метатекстовая разметка позволяет формировать запросный подкорпус по определенным параметрам (автор, жанр, дата создания и др.). Например, в таблице жанров можно выбрать разные виды грамот и вести поиск только в пределах этого жанра. Это создает условия для изучения типологии текстов.

Такой тип корпуса имеет свои прецеденты. Сходным образом организован созданный в Петербурге под руководством А. С. Герда корпус агиографических текстов (СКАТ: http://project.phil.spbu.ru/scat/page.php?page=project), который включает 15 памятников житийной литературы XV–XVII вв. Поиск в нем осуществляется по словам и буквосочетаниям. Поэтому, например, на запрос быти можно получить все адреса этой словоформы в имеющихся в программе житиях, а также адреса слов бытие, пребыти, избыти и т. п., но сведения о есмь, еси, соуть и т. п. нужно искать отдельно. По адресу словоформы можно выйти на соответствующий, довольно большой, фрагмент текста (в котором искомые словоформы, к сожалению, не выделены).

Аналогично устроен поиск в программе, созданной во Фрайбурге по тексту Апостола в Великих Минеях Четьих (http://www.vmc.uni-freiburg.de/Mens/) [Rabus, Savić, Waldenfels 2012].

В дальнейшем «исторические» корпусы НКРЯ, очевидно, нужно будет переструктурировать, сделав центральным корпус источников XI–XVII вв. Древнерусский корпус и корпус берестяных грамот, сохраняя автономное существование, могли бы одновременно стать той частью корпуса XI–XVII вв., в которой снята грамматическая омонимия.

Ближайшей задачей для корпуса XI–XVII вв. является работа по созданию программ автоматической лемматизации словоформ и их морфологической разметки. Эта задача представляет большие трудности из-за присущей древнерусскому словоизменению высокой степени омонимии, а также из-за разнообразия графико-орфографических систем, находящихся в непрямой зависимости от фонетических изменений и многого другого.

Представляется, однако, что решение этой задачи можно приблизить, разделяя корпус на однородные фрагменты (например, деловые памятники, вести-куранты XVII вв. и т. п.).

Литература

Гнутиков Р. М., Баранов В. А., Вотинцев А. А., Миронов А. Н. Редактор OldEd: Руководство пользователя (версия 1.1.0.2) — Электрон. издание. — Ижевск, 2005. — 21 с.

Зобнин А. И., Пичхадзе А. А. Корпус древнерусских переводов XI–XII вв.: результаты и перспективы // Научно-техническая информация. Серия 2. Информационные процессы и системы. М., 2005. № 3. С. 44–47.

Молдован А. М. Компьютерное воспроизведение славянских рукописных памятников и его задачи // «Мовознавство». 2006, № 2–3. С. 26–30.

Eckhoff H., Birnbaum D., Miltenova A., Dimitrova Tsv. The Tenth-Century Cyrillic Manuscript Codex Suprasliensis: the Creation of an Electronic Corpus // Преоткриване: Супрасълски сборник, старобългарски паметник от X век. Rediscovery: Bulgarian Codex Suprasliensis of 10th century. Sofia, 2012, p. 369–378.

Rabus A., Savić S., Waldenfels R. Towards an Electronic Corpus of the Velikie Minei Čet'i // Преоткриване: Супрасълски сборник, старобългарски паметник от X век. Rediscovery: Bulgarian Codex Suprasliensis of 10th century. Sofia, 2012, p. 343–354.

Sels L., Stern D. Preparing a Digital Edition of the *Bdinski Sbornik* // Преоткриване: Супрасълски сборник, старобългарски паметник от X век. Rediscovery: Bulgarian Codex Suprasliensis of 10th century. Sofia, 2012, p. 355–368.

A. M. Moldovan
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
alexandr.moldovan@gmail.com

OLD RUSSIAN MANUSCRIPTS IN THE RUSSIAN NATIONAL CORPUS

The article deals with different types of the corpus representation of Old Russian manuscripts in their ability to solve linguistics problems. Collections of digital copies and e-books in PDF and DjVu formats are

the simplest and widely spread type of electronic resources, containing old Russian manuscripts. Theoretically more interesting, though much more difficult, is the representation of digitized editions as textual corpora and databases. The main attention in the article is paid to the historical subcorpus within the Russian National Corpus representing the sources of the XI-XVII centuries and consisting of four parts: the Old Russian corpus, the Corpus of birchbark letters, the Middle Russian and the Church-Slavonic corpora.

Key words: corpus linguistics, old Russian literature, Church-Slavonic language, source study, linguistic documentation

References

Eckhoff H., Birnbaum D., Miltenova A., Dimitrova Tsv. The Tenth-Century Cyrillic Manuscript Codex Suprasliensis: the Creation of an Electronic Corpus. Preotkrivane: Suprasàlski sbornik, starobàlgarski pametnik ot X vek. Rediscovery: Bulgarian Codex Suprasliensis of 10th century. Sofia, 2012, pp. 369–378.

Gnutikov R. M., Baranov V.A., Votincev A.A., Mironov A. N. Redaktor OldEd: *Rukovodstvo pol'zovatelya (versiya 1.1.0.2)*. [Editor OldEd: user Manual (version 1.1.0.2)]. Electronic publ. Izhevsk, 2005. 21 p. (In Russ.)

Moldovan A. M. [Computer reproduction of Slavic manuscripts and its objectives]. *Movoznavstvo*, 2006, no. 2–3, pp. 26–30. (In Russ.)

Rabus A., Savić S., Waldenfels R. Towards an Electronic Corpus of the Velikie Minei Čet'i. *Preotkrivane: Suprasàlski sbornik, starobàlgarski pametnik ot X vek*. Rediscovery: Bulgarian Codex Suprasliensis of 10th century. Sofia, 2012, pp. 343–354.

Sels L., Stern D. Preparing a Digital Edition of the *Bdinski Sbornik. Preotkrivane: Suprasàlski sbornik, starobàlgarski pametnik ot X vek.* Rediscovery: Bulgarian Codex Suprasliensis of 10th century. Sofia, 2012, pp. 355–368.

Zobnin A. I., Pichhadze A.A. [The Corpus of the old Russian translations of the XI–XII centuries: results and prospects]. *Nauchno-tekhnicheskaya informatsiya. Seriya 2. Informatsionnye processy i sistemy* [Scientific and technical information. Series 2. Information processes and systems]. Moscow, 2005, no. 3, pp. 44–47. (In Russ.)

[1]Е. А. Мишина, [2]А. А. Пичхадзе
[1,2]Институт русского языка им. В.В. Виноградова РАН
(Москва, Россия)
[1]kmishina@mail.ru, [2]rusyaz@yandex.ru

ДРЕВНЕРУССКИЙ ПОДКОРПУС НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА[1]

Статья посвящена древнерусскому подкорпусу Национального корпуса русского языка, включающему оригинальные древнерусские памятники (летописи, жития и др.) и древнерусские переводы с греческого. Корпус содержит около полумиллиона словоупотреблений. В статье описывается грамматическая разметка подкорпуса и способы построения запросов с учетом специфики древнерусских текстов. Обращается внимание на возможность построения синтаксических запросов в подкорпусе. Приводятся и анализируются данные некоторых конкретных запросов, демонстрирующих: 1) употребление конструкции «личная форма глагола **бꙑти** + причастие» (типа **бѣ ловы дѣꙗ**), 2) отсутствие форм действительного причастия настоящего времени и имперфекта от перфективных глаголов **дати** и **поꙋстити** при отсутствии отрицания, 3) порядок слов в группе «личная глагольная форма — прямое дополнение — косвенное адресатное дополнение, выраженное местоимением **и** (**ємоꙋ**, **єи**, **имъ**)».

Ключевые слова: текстовые корпуса, древнерусский язык, грамматическая разметка, морфологические и синтаксические запросы.

Древнерусский подкорпус в составе Национального корпуса рус-

[1] Работа над статьей и над корпусом осуществлялась при поддержке Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» и РГНФ (проекты №14-04-12035 и № 15-04-12050).

ского языка (НКРЯ) изначально задумывался как ресурс по древнеславянским книжным[1] текстам восточнославянского происхождения. Проблема отбора памятников, которые нужно включать в ресурс, предназначенный для того, чтобы давать представление о древнерусском языке, всегда стояла остро и для составителей словарей языка древнерусского и старорусского периода. В отечественной лексикографии она решалась по-разному: И. И. Срезневский и составители «Словаря русского языка XI–XVII вв.», во-первых, считали допустимым использовать любые источники, даже южнославянского происхождения, если они были переписаны на Руси, а во-вторых, использовали памятники, сохранившиеся в поздних списках; составители «Словаря древнерусского языка (XI–XIV вв.)» пытались ограничить круг памятников теми, где встречаются лексические русизмы и которые сохранились в списках не позже рубежа XIV–XV вв. В результате в одни словари попала южнославянская лексика, не усвоенная древнерусским языком даже в его церковнославянском варианте, а в другие словари не попала лексика, представленная в памятниках, заведомо возникших на Руси в ранний период, но сохранившихся в поздних списках. При создании древнерусского подкорпуса НКРЯ был выбран средний путь: привлечение памятников с восточнославянскими элементами в лексике, возникших в период XI–XIV вв., даже если они сохранились только в поздних списках[2].

Разумеется, в подкорпус включались прежде всего оригинальные древнерусские тексты: на данный момент туда вошли древнерусские летописи и сочинения Кирилла Туровского. Сам по себе отбор оригинальных памятников не создает проблем, все они по возможности должны быть отражены в подкорпусе. Однако при отборе конкретных текстов возникают вопросы аутентичности и оценки источников, как при всяком обращении к средневековым памятникам. Выбор составителей подкорпуса опирался на новейшие источниковедческие исследования. С учетом работ последних лет было принято решение вводить в подкорпус «Повесть временных лет» не

[1] Материал берестяных грамот представлен в особом подкорпусе НКРЯ, поскольку их язык обладает специфическими особенностями, которые обусловливают иной формат разметки и поиска.

[2] О каждом памятнике дается краткая информация относительно времени создания или перевода. Она появляется в окне, всплывающем при выделении мышью названия памятника в результатах запроса.

по древнейшему Лаврентьевскому списку 1377 г., который сравнительно хорошо изучен и издан со словоуказателем, а по Ипатьевскому списку начала XV в., не менее важному для восстановления первоначального текста памятника (см. [Гиппиус 2002; 2007–2008]). Это, впрочем, не исключает возможности в дальнейшем присоединить и Лаврентьевский список — открытость подкорпуса позволяет объединять в его составе разные версии одного и того же текста. В ходе работы приходилось предпринимать специальные текстологические и лингвистические исследования, на основе которых, например, произведено разграничение Галицкой и Волынской летописей в составе Галицко-Волынского свода [Юрьева 2013], выполнена текстологическая группировка списков повествовательных сочинений Кирилла Туровского и отобраны наиболее представительные рукописи [Баранкова, Макеева 2013].

Что касается переводных церковнославянских памятников, то они отбирались на основании специального исследования переводов с восточнославянскими элементами в лексике [Пичхадзе 2011], которое позволило отделить небольшую группу переводных памятников, переведенную древнерусскими переводчиками, от более обширной группы, которая была переведена, по всей вероятности, южнославянскими переводчиками, работавшими на Руси. На настоящем этапе в подкорпус включена только первая небольшая группа текстов, принадлежащих восточнославянским переводчикам. В перспективе, при расширении подкорпуса, в него могут быть введены и тексты второй группы.

Единственный памятник южнославянского происхождения, представленный на сегодняшний день в подкорпусе, — Изборник 1076 г. Его присутствие в подкорпусе оправдано в силу того, что он подвергся ощутимому редактированию на русской почве и в некоторых отношениях (например, по количеству форм с восточнославянским полногласием) является уникальным для своего времени источником данных по истории именно древнерусского языка [Мушинская, Мишина, Голышенко 2009]. Представляется правильным и впредь вводить в древнерусский подкорпус памятники, отредактированные на Руси и впитавшие на восточнославянской почве заметное число языковых русизмов, прежде всего лексических.

На сегодняшний день древнерусский подкорпус, содержащий около полумиллиона словоупотреблений, представляет пользова-

телю книжные тексты XI–XIV вв., в наименьшей степени затронутые южнославянским влиянием и максимально отражающие восточнославянские языковые особенности. Грамматическая разметка текстов производилась вручную (с использованием прецедентных разборов) и затем редактировалась, что обеспечивает довольно высокую степень её надежности.

Тексты, входящие в древнерусский подкорпус, приводятся с сохранением древней оригинальной орфографии рукописей (в том числе и те, которые цитируются по изданиям), в некоторых из памятников (например, в «Истории Иудейской войны Иосифа Флавия», «Александрии») при разметке сохранены знаки акцентуации. В связи с этим для удобства пользователя в древнерусском подкорпусе НКРЯ разработана трёхступенчатая система поиска («точная», «упрощенная» и «модернизированная»), позволяющая в соответствии с той или иной исследовательской задачей учитывать разные орфографические режимы. В режиме «точной» орфографии можно искать словоформу или фрагмент словоформы с учетом оригинальной орфографии (буква в букву). В перспективе такой режим должен обеспечить возможность поиска нетривиальных орфографических приемов (например, написания «и десятеричного» без точки — **ı** — в заимствованиях). В режиме «упрощенной» орфографии можно строить запросы с учетом орфографических вариантов в одной и той же позиции: на виртуальной клавиатуре выделяются, например, узкие варианты букв (**ѥ**, **ѳ**) и **и** восьмеричное, и отдельно — их орфографические варианты, которые могут писаться в определенных позициях (например, **є=ѥ**, **ѡ=ѽ=ѿ** после букв гласных или в начале слова, **і=ї=ı** на конце строки и т. п.). Наконец, можно строить запрос в «модернизированной» орфографии, в которой приравниваются не только орфографические варианты, но и некоторые буквы, обозначающие разные фонемы: **ѥ** (и все его орфографические варианты) и **ѣ**, **и** (и все его орфографические варианты) и **ѵ**. Этот режим удобен для пользователей, ориентирующихся на современный облик слова, а также для запросов по словам, орфографический образ которых может варьироваться (например, слов с приставками **прѥ-**/**прѣ-**).

Для поиска по леммам наиболее удобными являются режимы, учитывающие упрощенную и модернизированную орфографию, поскольку леммы в общем для всех памятников словаре (всплываю-

щем при наборе леммы) представлены в унифицированной древнерусской орфографии[1].

Обновление и редактирование пополняемого сводного словаря текстов, включенных в древнерусский подкорпус НКРЯ, а также организация эффективного поиска по леммам, учитывающего, например, такие параметры, как разграничение по омонимам, представляет собой отдельную непростую задачу. При создании древнерусского подкорпуса была предпринята попытка её решить. Дополнительные трудности заключаются в том, что, во-первых, в некоторых памятниках представлен только один из омонимов, а во-вторых, в том, что оформление омонимов в разных текстах может отличаться (омонимы в словарях к отдельным памятникам могут получить отличающиеся толкования и т. д.); тем не менее все расхождения по отдельным текстам должны в конечном итоге быть унифицированы в сводном словаре. Создание эффективных алгоритмов автоматизации этого процесса — во многом задача будущего. На данный момент для разграничения омонимов при поиске конкретной леммы целесообразно указывать её частеречную характеристику (например, **и** «союз» и **и** «местоимение»). Если омонимы имеют одинаковые грамматические характеристики (например, **лоукъ** ‘оружие’ и **лоукъ** ‘лука седла’), то в словаре, который всплывает при наборе леммы, помимо недифференцированной леммы **лоукъ**, появляются еще две леммы **лоукъ**, при которых указаны их значения. Соответственно, можно осуществлять поиск по каждому омониму отдельно. Однако для некоторых памятников, где омонимы не разграничены (например, «История Иудейской войны»), поиск оказывается возможным только по недифференцированной лемме; в этом случае пользователю придется внимательно просматривать все найденные контексты и разграничивать омонимы самостоятельно. Подсказкой при определении значения омонимов (как и остальных слов) является греческое соответствие, если памятник переводной: оно указано в окне, всплывающем при выделении мышью соответствующей словоформы в результатах запроса.

Разбор древнерусского текста — работа, не вполне поддающаяся стандартизации. Разметка в определенной мере отражает взгляды исследователя, который её осуществляет, на предмет исследования,

[1] Она отражает состояние до падения и прояснения редуцированных.

на то, какой набор признаков необходим, полезен или актуален для научного анализа. Научная трактовка грамматических явлений, отраженных в древнерусских текстах, постоянно меняется, иногда (об этом ниже) непосредственно в процессе разметки. Так, в последнее время появились работы [Петрухин, Сичинава 2006; 2008], посвященные так называемому «русскому плюсквамперфекту», в которых было показано, что эта аналитическая форма, состоящая из *л*-причастия и перфекта в роли связки, представляет собой сверхсложное прошедшее со значением «аннулированного результата», т. е. имеет свою семантику, отличную от значений других перфектных форм, и обозначает действие, результат которого был отменен последующим ходом событий. Помета «сверхсложное прошедшее» была использована при разметке текстов подкорпуса, и теперь эта форма может быть легко найдена при помощи запроса.

Поскольку потребность добавлять релевантные пометы возникает при изучении древних текстов постоянно, система разметки разрабатывалась таким образом, чтобы исследователь мог легко вводить новые признаки по собственному усмотрению. В результате древнерусский подкорпус несет отпечаток индивидуальных подходов к анализу текста. Разные памятники, входящие в подкорпус, могут быть размечены как с разной степенью подробности, так и с разным наполнением одной и той же категории. Разумеется, это не касается базовых признаков и категорий, интерпретация которых однозначна.

Расхождения в грамматической разметке разных памятников проявляются, например, в разборе имен собственных. Как и все остальные леммы, они получают частеречные характеристики: личные имена — характеристику «личное», топонимы — «топоним», отчества и этнонимы определяются как «существительные», а обозначения жены или вдовы по мужу — как «прилагательные». В ряде памятников частеречные пометы отчеств, этнонимов и обозначений жены или вдовы по мужу исчерпываются базовыми («существительное» или «прилагательное»), в других памятниках одновременно с одной из этих обязательных помет дополнительно указывается другая, более детальная: «отчество», «этноним», «по мужу». Это означает, что при запросе на «существительное» во всех текстах подкорпуса будут найдены отчества и этнонимы, а при запросе «прилагательное» — наименования по мужу, но для некоторых памятников

есть возможность построить и более точный запрос на «отчество», «этноним», «по мужу». Расхождения в разметке вызваны прежде всего тематикой памятников: во многих церковнославянских текстах не встречаются ни отчества, ни наименования жены по мужу, поэтому эти пометы используются при разборе летописей и отсутствуют среди грамматических помет остальных памятников. Однако в случае с этнонимами расхождение объясняется исключительно предпочтениями исследователей: одни из них считают целесообразным выделять этнонимы специальной пометой, а другие — нет. В подкорпусе этнонимы размечены в летописях, «Александрии» и сочинениях Кирилла Туровского, а в остальных памятниках они имеют только помету «существительное».

Есть объективные причины, которые приводят к расхождениям в грамматической разметке. Они связаны с неоднородностью явлений, подпадающих в традиционной грамматике под одну категорию. Например, категория презенса по-разному трактуется исследователями: одни включают в нее формы с референцией к настоящему и будущему (так размечена Новгородская I летопись), другие разделяют её на две категории — настоящего и настоящего-будущего времени. Это относится и к аналитическим формам так называемого будущего I и будущего II, принадлежащих к числу недостаточно изученных и неоднозначно трактуемых категорий в древнерусском (ср. в последнее время новые интерпретации в [Юрьева 2010; Козлов 2014; Пенькова 2014]). Применительно к будущему I, состоящему из личной формы глаголов **начати**, **почати**, **имѣти**, **моци**, **хотѣти** в настоящем времени и инфинитива, во многих случаях проблематично уже само объединение этих двух глагольных форм: они могут разбираться и как несвязанные формы в свободной синтаксической конструкции. При разборе переводных памятников исследователь может опираться на данные оригинала: в греческом форма будущего времени образуется синтетически, и если в соответствии с ней в переводе употребляется сочетание **начати**, **почати**, **имѣти**, **моци**, **хотѣти** с инфинитивом, есть основания рассматривать сочетание как аналитическую форму будущего I. Однако при разборе оригинальных текстов исследователь вынужден исходить только из своих собственных представлений о будущем I и о семантике конкретного контекста, и четкие критерии здесь отсутствуют. В результате сочетания **начати**, **почати**, **имѣти**, **моци**, **хотѣти** с инфинитивом

в сходных контекстах могут быть размечены, в зависимости от точки зрения исследователя, либо как «будущее I», либо как свободные словосочетания. Трудность разбора таких конструкций связана еще и с тем, что в них входят глаголы разной семантики: сочетания с **начати** и **почати** легче трактовать как временные формы, чем сочетания с модальными глаголами, а сочетания с разными модальными глаголами могут развивать разную семантику, как это было показано применительно к конструкциям **хотѣти** и **имѣти** с инфинитивом [Козлов 2014]. Хотя на данный момент в подкорпусе используется традиционная помета «будущее I», не исключено, что в ближайшее время новые исследования позволят заменить её на несколько помет, более адекватно характеризующих явления, обобщенные под этой рубрикой.

Что же касается будущего II, представляющего собой сочетание *л*-причастия с презенсом от основы *буд-*, то дискуссия по поводу семантики и грамматического статуса этой формы продолжается уже полтора столетия: одни исследователи считают её временной формой, другие — формой наклонения (см. обзор мнений в [Пенькова 2014]). Работа над подкорпусом неожиданно подвела черту в этом споре: при разметке «Повести об Акире Премудром» обнаружилась форма повелительного наклонения **боуди вѕѧлъ** ‘твоя взяла, ты победил (букв.: будь взявший)’, что исключает трактовку будущего II как наклонения, хотя допускает модальные оттенки у этой формы, вообще присущие многим аналитическим перфективным формам. С другой стороны, при разметке «Жития Андрея Юродивого» были отмечены формы будущего II с референцией к прошедшему времени [Там же]. Эти факты показывают, что форма будущего II обладала очень разнообразной семантикой (подробнее см. указанную работу Я. А. Пеньковой), и объединение их под одной пометой «будущее II» выглядит достаточно условным. Можно предположить, что в ближайшее время формы, сейчас объединенные пометой «будущее II», по итогам дальнейших исследований получат какие-то дополнительные дифференцирующие характеристики.

Палеослависты обращаются к корпусам в первую очередь за лексической и морфологической информацией. Однако древнерусский подкорпус позволяет довольно успешно строить синтаксические запросы. В подкорпусе отмечается управление предлогов, принадлежность предиката целевому и косвенно-побудительному

придаточному предложению (глагольная форма получает помету «в значении долженствования»), а также специфическое синтаксическое употребление словоформ: имен прилагательных и причастий — в роли существительных или предикативов, некоторых падежных форм существительных — в роли наречий и т. п. Это дает возможность получать точные результаты в случаях, когда, например, необходимо построить запрос не только на существительные и местоимения, но и остальные субстантивы, т. е. прилагательные и причастия в роли существительных. В частеречных пометах предусмотрены две характеристики для местоимений: «местоимение-существительное» — для местоимений, имеющих субстантивный статус, и «местоимение-прилагательное» — для местоимений, выполняющих роль определения. Отмечаются некоторые предложно-падежные сочетания.

Некоторые синтаксические запросы не дают такого точного результата, как запросы по лексике и морфологии, и требуют ручного редактирования. Тем не менее они могут быть полезны для получения предварительного чернового материала, из которого нужно исключить не относящиеся к делу примеры. Так, неплохой результат дают запросы на дательный самостоятельный (запрос на причастие в дательном падеже).

Большое преимущество корпусов состоит в том, что они позволяют вести поиск по нескольким словам и/или формам одновременно (формы могут быть охарактеризованы только грамматически, без указания лемм). Это хорошее подспорье не только для поиска цитат и фразеологизмов, но и для синтаксического анализа — например, для получения сведений об управлении глаголов. В древнерусском подкорпусе, указав две соответствующие формы, нетрудно отыскать конструкции «**ꙗко** + инфинитив» со значением следствия, «личная форма глагола **бꙑти** + причастие», «3 л. ед. ч. глагола **бꙑти** + инфинитив» со значением долженствования, конструкции с отрицанием «**не** + глагольная форма» и др.

Результаты синтаксических запросов предоставляют иногда неожиданную информацию, которая ждет своего исследователя. Так, например, запрос на конструкцию «личная форма глагола **бꙑти** + причастие» (типа **бѣ бо ловы дѣꙗ ѡлегъ**) показывает 9 примеров из «Жития Андрея Юродивого», 5 примеров из «Александрии», 5 примеров из «Истории Иудейской войны», один пример из Избор-

ника 1076 г., более двух десятков примеров из Повести временных лет, 8 примеров из Киевской летописи, 8 примеров из совсем небольшой по объему Волынской летописи и 4 примера из Галицкой летописи. В «Пчеле», Новгородской I летописи и Суздальской летописи надежных примеров нет. Конструкция «личная форма глагола **быти** + причастие» считается специфически книжной, калькирующей соответствующую греческую конструкцию, распространенную в библейских текстах [Живов 2012: 235; Петрухин 2012: 239–243]. Она чрезвычайно частотна в русской агиографии — например в «Житии Феодосия Печерского» [Живов 2012: 236–238]. Между тем данные подкорпуса демонстрируют распределение, не коррелирующее напрямую со степенью книжности текста. Летописи используют указанную конструкцию интенсивнее, чем древнерусские церковнославянские переводные памятники: в одной только Повести временных лет она встречается чаще, чем во всех церковнославянских текстах подкорпуса вместе взятых, а из летописей её очень активно употребляет наименее книжная по языку Волынская летопись — в отличие от заметно ориентированной на церковнославянский стандарт Галицкой летописи. Этот факт требует специального исследования. Разумеется, летописи склонны к использованию клише, и рассматриваемая конструкция может принадлежать к их числу. Тем не менее её распространенность обращает на себя внимание. Возможно, конструкция «личная форма глагола **быти** + причастие» не была просто заимствованием и имела опору в восточнославянском языковом узусе. Если же признать за ней исключительно книжный статус, полученные данные, вероятно, можно интерпретировать как указание на независимость усвоения церковнославянских элементов летописями и церковнославянской древнерусской письменностью. Данные подкорпуса информативны еще в одном отношении, поскольку демонстрируют, что использование рассматриваемой конструкции обусловлено типом дискурса: она употребляется в нарративе (исторических повествованиях, агиографии) и почти полностью отсутствует в гномических текстах (Изборнике 1076 г., «Пчеле»).

Интересный материал дают и результаты запроса на «**ꙗко** + инфинитив». Наряду с ожидаемыми конструкциями следствия (**быша громи велици и блисканїи млънїа, ꙗко всемоу мироу подвиꙁатисѧ** ‘были громы великие и блистания молний, так что весь мир при-

шел в движение’), представленными как в церковнославянских, так и летописных источниках, запрос показывает множество примеров омонимичной конструкции при глаголах со значением ‘давать клятву, обещание; провозвещать; договориться’: **крестъ цѣловати, клѧтисѧ, съдоумати ꙗко/како** + инф., ср. в Новгородской I летописи: **сдоумаша <ꙗко> изгонити кн҃зѧ своего всѣволода** ‘договорились изгнать князя своего Всеволода’; **и тако оувѣчаста · ꙗко нѣ воюватι на ц(с)рьградъ** ‘и так договорились не воевать против Царьграда’; в Киевской летописи: **цѣлоуи кр҃тъ ꙗко имѣти братью въ любовъ** ‘целуй крест дружить с братьями’. Эта конструкция фиксируется в русских летописях, «Александрии» (83.9) и в «Истории Иудейской войны», вообще близкой по языку к летописям. Греческого аналога она не имеет и в церковнославянских текстах не употребляется: там используется либо инфинитивное дополнение без союза, либо придаточное предложение со сказуемым в финитной форме, вводимое союзом **ꙗко**. Судя по всему, **ꙗко** + инф. при глаголах клятвы и обещания — специфическая древнерусская книжная конструкция, генезис которой еще предстоит выяснить (см. о её функционировании [Власова 2014: 189–194]).

Иногда полезными оказываются запросы с исключением слов или грамматических характеристик. Так, например, одним из актуальных вопросов в исследовании формирования категории вида в древнерусском языке является способность перфективных глаголов образовывать формы имперфекта и причастий настоящего времени. Запрос на форму действительного причастия настоящего времени от глагола **дати** в подкорпусе выдает 19 примеров из 7 памятников, а на форму имперфекта от глагола **дати** — 18 примеров из 6 памятников. Однако эти же запросы с исключением отрицательной частицы **не** показывают, что без отрицания формы имперфекта и действительных причастий настоящего времени от глагола **дати** не встречаются ни разу. Аналогичный запрос на формы имперфектов от глагола **пустити** и его приставочных производных с исключением отрицательной частицы **не** показывает, что формы имперфекта с формантом **-стѫ-** (**поустѫхоу**, **допустѫхоу**, **припустѫхѹ**, **роспоустѧста**) — в отличие от форм имперфектов с формантом **-ща-** (**поущахоу**, **испоущаше**, **воспущаше**, **w̃пущаше**) — встречаются только с отрицанием (как и малочисленные имперфекты от других приставочных глаголов совершенного вида). Полученные дан-

ные могут свидетельствовать в пользу того, чтобы считать глаголы **дати** и **пустити** глаголами совершенного вида уже в ранний древнерусский период [Мишина 2012].

В известной мере подкорпус можно использовать для изучения порядка слов. Несмотря на неточность результатов, запросы дают в некоторых случаях картину, которая даже с учетом неточностей в целом позволяет составить адекватное представление о проблеме. В [Пичхадзе 2012: 185] на материале древнерусской «Пчелы» была отмечена тенденция, ставшая закономерностью в современном русском языке: дополнение, выраженное местоимением, предшествует дополнению, выраженному существительным, если только последнее не участвует в контрастном противопоставлении [Рогова 1962: 47]. Вот как по данным подкорпуса распределяются словопорядки в группе «личная глагольная форма (V) — прямое дополнение (O) — косвенное адресатное дополнение (A), выраженное местоимением **и** (**ємоу**, **єи**, **имъ**)» в русских летописях:

	VAO	OAV	VOA	OVA	AVO	AOV
Повесть временных лет	26	1	7	2	1	1
Киевская летопись	47	9	11	5	3	4
Новгородская I летопись	11	2	0	0	0	0
Галицкая летопись	12	3	0	3	2	0
Волынская летопись	14	1	3	3	2	0
Суздальская летопись	37	2	10	2	2	0

Приведенные в таблице данные подтверждают наличие тенденции к препозиции дополнения, выраженного местоимением, по отношению к дополнению, выраженному субстантивом, уже в древнерусском языке (сравнительно большие цифры третьей колонки объясняются частотностью в летописях конструкции **нарекоша имѧ ємоу**/**єи** X, которая употребляется с порядком VOA). Помимо этого, полученные данные лишний раз демонстрируют, насколько предпочтительнее для древнерусского книжного нарратива была препозиция глагольной формы по отношению к прямому и косвенному адресатному дополнениям по сравнению с постпозицией.

Древнерусский подкорпус моложе многих других подкорпусов НКРЯ, и практика его использования еще невелика. Он, безусловно,

нуждается в усовершенствованиях, в частности в унификации параметров разметки для всех текстов, в оптимизации словаря лемм, в расширении метаданных, в развитии поисковых возможностей. Ресурс для этого имеется, поскольку потенциал осуществленной разметки памятников шире того, что представлено на данный момент в древнерусском подкорпусе. Так, может быть добавлен поиск по греческим леммам и их славянским соответствиям; по альтернативным разборам словоформ, которые допускают неоднозначную трактовку; по ошибочным и испорченным словоформам и др. Тем не менее уже сейчас подкорпус позволяет получать разнообразную информацию, а главное, открывает неожиданные аспекты в изучении древнерусской письменности и ставит перед исследователями нетривиальные вопросы.

Литература

Баранкова Г. С., Макеева И. И. Повествовательные произведения Кирилла Туровского и проблемы русского литературного языка // Письменность, литература, фольклор славянских народов. История славистики. XV Международный съезд славистов. Минск, 20–27 августа, 2013 г. Доклады российской делегации. М., 2013. С. 52–80.

Власова Е. А. Инфинитив и сослагательное наклонение в косвенной речи в русских летописях XI–XVI вв. // Русский язык в научном освещении. 2014. № 1 (27). С. 185–205.

Гиппиус А. А. О критике текста и новом переводе-реконструкции Повести временных лет // Russian Linguistics. 2002. Vol. 26. № 1. С. 63–126.

Гиппиус А. А. К проблеме редакций Повести временных лет // Славяноведение. 2007. № 5. С. 20–44; 2008. № 2. С. 3–24.

Живов В. М. Об употреблении одной книжной конструкции в памятниках восточнославянской письменности: <глагол **бꙑти** с причастием наст. времени> // I. Podtergera (Hrsg.): Schnittpunkt Slavistik. Ost und West im wissenschaftlichen Dialog. Festgabe für Helmut Keipert zum 70. Geburtstag. Teil 3: Vom Wort zum Text. Göttingen, 2012. S. 235–249.

Козлов А. А. К грамматической семантике старославянских конструкций **хотѣти**/ **имѣти** с инфинитивом // Русский язык в научном освещении. 2014. № 1 (27). С. 122–149.

Мишина Е. А. «Ситуация напрасного ожидания» и отрицание // Русский язык в научном освещении. 2012. № 2 (24). С. 219–241.

Мушинская, Мишина, Голышенко 2009 — Изборник 1076 г. / Изд. подгот. М. М. Мушинская, Е. А. Мишина, В. С. Голышенко. Т. I–II. М., 2009.

Пенькова Я. А. К вопросу о семантике так называемого будущего сложного II в древнерусском языке (на материале «Жития Андрея Юродивого») // Русский язык в научном освещении. 2014. № 1 (27). С. 150–184.

Петрухин П. В., Сичинава Д. В. «Русский плюсквамперфект» в типологической перспективе // Вереница литер. К 60-летию В. М. Живова. М., 2006. С. 193–214.

Петрухин П. В., Сичинава Д. В. Еще раз о восточнославянском сверхсложном прошедшем, плюсквамперфекте и современных диалектных конструкциях // Русский язык в научном освещении. 2008. № 1 (15). С. 224–258.

Петрухин П. В. К проблеме реконструкции и перевода Повести временных лет // Русский язык в научном освещении. 2012. № 1 (23). С. 232–249.

Пичхадзе А. А. Переводческая деятельность в домонгольской Руси: лингвистический аспект. М., 2011.

Пичхадзе А. А. Порядок слов в сочетаниях «глагол + прямое и адресатное дополнение» в древнерусской «Пчеле» // I. Podtergera (Hrsg.): Schnittpunkt Slavistik. Ost und West im wissenschaftlichen Dialog. Festgabe für Helmut Keipert zum 70. Geburtstag. Teil 3: Vom Wort zum Text. Göttingen, 2012. С. 181–194.

Рогова К. А. Место прямого дополнения в простом повествовательном предложении // Ученые записки Ленинградского гос. университета. № 302. Серия филологических наук. Вып. 61. Исследования по грамматике русского языка. III. Л., 1962. С. 40–49.

Юрьева И. С. Особенности древнерусских инфинитивных сочетаний с глаголом **начати** // Русский язык в научном освещении. 2010. № 2 (20). С. 270–286.

Юрьева И. С. Некоторые особенности синтаксиса, морфологии и лексики так называемой Галицко-Волынской летописи // Лингвистическое источниковедение и история русского языка. <2012–2013>. М., 2013. С. 135–151.

[1]E. A. Mishina, [2]A. A. Pichkhadze
[1,2]Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Moscow, Russia)
[1]kmishina@mail.ru, [2]rusyaz@yandex.ru

OLD RUSSIAN SUBCORPUS OF THE RUSSIAN NATIONAL CORPUS

The paper deals with the Old Russian subcorpus of the Russian National Corpus that comprises original Old Russian texts (chronicles, hagiography etc.) and Old Russian translations from Greek. The subcorpus contains approximately 500 000 words. The paper describes the grammatical tagging used in the subcorpus and possible ways of constructing complex queries taking into account specific features of Old Russian texts and with special emphasis on possibility to construct syntactic queries. The results of some queries are presented and analyzed in the article: 1) the usage of the construction "finite form of the verb *byti* + participle" (e. g. *bě lovy děja*), 2) the lack of active praesens participle and imperfect forms of the perfective verbs *dati* and *pustiti* except the contexts with negation, 3) word orders in the group "finite verb + direct object + indirect object represented by the pronoun **и** (**ємѹ**, **ѥи**, **имъ**)".

Key words: text corpora, Old Russian, grammatical tagging, morphological and syntactic queries.

References

Barankova G. S., Makeeva I. I. [Narrative works of Cyril of Turov and problems of the Russian literary language]. *Pis'mennost', literatura, fol'klor slavyanskikh narodov. Istoriya slavistiki. XV Mezhdunarodnyi s"ezd slavistov. Minsk, 2013. Doklady rossiiskoi delegatsii* [Writing, literature, and folklore of the Slavic peoples. The history of Slavic studies. XV International Congress of Slavists. Minsk, 2013 Reports of the Russian delegation]. Moscow, 2013, pp. 52–80. (In Russ.)

Vlasova E. A. [The use of infinitive and subjunctive forms in modal patterns of indirect speech in Russian chronicles of the XI—XVI

centuries]. *Russkii yazyk v nauchnom osveshchenii,* 2014, no. 1 (27), pp. 185–205. (In Russ.)

Gippius A. A. [On text criticism and a new reconstructive translation of the Primary Chronicle]. *Russian Linguistics,* 2002, vol. 26, no. 1, pp. 63–126. (In Russ)

Gippius A. A. [On the problem of redactions of the Primary Chronicle]. *Slavyanovedenie*, 2007, no. 5, pp. 20–44; 2008, no. 2, pp. 3–24. (In Russ.)

Zhivov V. M. [About the use of one bookish construction in East Slavic manuscripts: <verb byti with present participle>]. *I. Podtergera (Hrsg.): Schnittpunkt Slavistik. Ost und West im wissenschaftlichen Dialog. Festgabe für Helmut Keipert zum 70. Geburtstag. Teil 3: Vom Wort zum Text. Göttingen,* 2012, pp. 235–249. (In Russ)

Kozlov A. A. [On the semantics of old Church Slavonic *xotěti*/ *iměti* + infinitive constructions]. *Russkii yazyk v nauchnom osveshchenii*, 2014, no. 1 (27), pp. 122–149. (In Russ)

Mishina E. A. ["Situation of futile expectations" and negation]. *Russkii yazyk v nauchnom osveshchenii*, 2012, no. 2 (24), pp. 219–241. (In Russ)

Mushinskaya, Mishina, Golyshenko 2009 — Izbornik 1076 g. M. M. Mushinskaya, E. A. Mishina, V. S. Golyshenko (eds.). V. I–II. Moscow, 2009.

Pen'kova Ya. A. [Towards the problem of the so called «compound Future II» semantics in the Old Russian language (on the material of The Life of St. Andrew the Fool)]. *Russkij jazyk v nauchnom osveshhenii,* 2014, no. 1 (27), pp. 150–184. (In Russ)

Petrukhin P. V., Sichinava D. V. ["Russian pluperfect" in typological perspective]. *Verenitsa liter. K 60-letiyu V. M. Zhivova*. [Collection of articles for the 60th anniversary of V. M. Zhivov]. Moscow, 2006, pp. 193–214. (In Russ.)

Petrukhin P. V., Sichinava D. V. [Once again about East Slavic complicated past, the pluperfect and modern dialectal constructions]. *Russkii yazyk v nauchnom osveshchenii*, 2008, no. 1 (15), pp. 224–258. (In Russ)

Petrukhin P. V. [On the reconstruction and translation of the Russian Primary Chronicle]. *Russkii yazyk v nauchnom osveshchenii,* 2012, no. 1 (23), pp. 232–249. (In Russ)

Pichkhadze A. A. *Perevodcheskaya deyatel'nost' v domongol'skoi*

Rusi: lingvisticheskii aspekt [Translational activity in pre-Mongol Rus': the linguistic aspect]. Moscow, 2011.

Pichkhadze A. A. [Word order in "verb + direct and indirect object" combination in Old Russian "Pchela"]. I. Podtergera (Hrsg.): *Schnittpunkt Slavistik. Ost und West im wissenschaftlichen Dialog. Festgabe für Helmut Keipert zum 70*. Geburtstag. Teil 3: Vom Wort zum Text. Göttingen, 2012, pp. 181–194. (In Russ)

Rogova K. A. [The place of direct object in a simple declarative sentence]. *Uchenye zapiski Leningradskogo gos. universiteta. № 302. Seriya filologicheskikh nauk. Vyp. 61. Issledovaniya po grammatike russkogo yazyka. III* [Scientific notes of the Leningrad State University. No. 302. Series in Philology. Vol. 61. Studies on the grammar of the Russian language. III]. Leningrad, 1962, pp. 40–49. (In Russ)

Yur'eva I. S. [The peculiarities of Old Russian nachati + infinitive constructions]. *Russkii yazyk v nauchnom osveshchenii*, 2010 no. 2 (20), pp. 270–286. (In Russ)

Yur'eva I. S. [Some peculiarities of syntax, morphology and vocabulary of the so-called Galician–Volhynian Chronicle]. *Lingvisticheskoe istochnikovedenie i istoriya russkogo yazyka*, <2012–2013>, Moscow, 2013, pp. 135–151. (In Russ.)

[1]Е. Р. Добрушина, [2]А. Г. Кравецкий, [3]А. Е. Поляков
[1]Православный Свято-Тихоновский Гуманитарный Университет,
edobrush@gmail.com
[2]Институт русского языка им. В.В. Виноградова РАН,
krav62@mail.ru
[3]Научная педагогическая библиотека К.Д. Ушинского
pollex@mail.ru
[1, 2, 3](Россия, Москва)

КОРПУС И ЧАСТОТНЫЙ ГРАММАТИЧЕСКИЙ КОРПУСНЫЙ СЛОВАРЬ ЦЕРКОВНОСЛАВЯНСКОГО ЯЗЫКА В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА

Статья посвящена описанию современного состояния работы над церковнославяянским подкорпусом Национального корпуса русского языка. Корпус отличается от собрания текстов наличием специальной разметки (грамматической, структурной, метатекстовой) и возможностью поиска по этой разметке. Рассмотрен состав, охарактеризованы основные жанровые рубрики, по которым распределяются включенные в корпус тексты, описаны принципы метаразметки, которые несколько отличаются от используемых в других корпусам, входящих в Национальный корпус русского языка. В необходимых случаях приводится историческая информация, на основе которой было принято то или иное решение. Поскольку для церковнославянских текстов ввод поисковых запросов с клавиатуры представляет определенные трудности, предлагается несколько вариантов упрощенной орфографической передачи, что позволяет ввести запрос, ограничиваясь возможностями стандартной клавиатуры. Наконец, в статье

содержится описание созданного в рамках работы над проектом частотного грамматического словаря.

Ключевые слова: церковнославянский язык, корпусная лингвистика, орфография, словоизменение, морфология, грамматический словарь

Корпус церковнославянского языка в рамках НКРЯ функционирует с мая 2012 года[1], а с декабря 2013 стал доступен созданный на его базе частотный грамматический словарь[2]. Здесь будет описано состояние Корпуса и Словаря на 2015 год[3].

Из специализированных корпусов НКРЯ, объединенных заголовком «Исторические», церковнославянский стал первым и самым объемным. Он включает тексты, созданные или отредактированные в XVII–XX вв.

Статус церковнославянского корпуса по отношению к диахронному и синхронному взгляду на язык и включение этого корпуса в блок исторических нуждается в некоторых комментариях. Дело в том, что вошедшие в корпус тексты, большая часть которых используются в современном богослужении Русской и некоторых славянских православных церквей, с равным успехом могут рассматриваться и как принадлежащие истории, и как относящиеся к современной языковой практике. Основанием для объединения с такими корпусами, как древнерусский, старорусский и берестяных грамот, может служить то, что речь

[1] Работа над статьей и над корпусом осуществлялась при поддержке РГНФ (проект 15-04-12050 «Развитие Исторических модулей НКРЯ»).

[2] Основным создателем церковнославянских Корпуса и Словаря является А.Е. Поляков; существенный вклад внесен также Е.Р. Добрушиной и А.Г. Кравецким. Кроме того, в создании церковнославянского корпуса на разных этапах его разработки в качестве создателей концепции, разработчиков и исполнителей технических задач, а также консультантов и вдохновителей работы принимали участие А.И. Зобнин (Яндекс), К.В. Литвинцева (ПСТГУ), Т.Ю. Иванова-Алленова (ПСТГУ), В.А. Плунгян (ИРЯ РАН), А.В. Жирова (ПСТГУ), А.А. Плетнева (ИРЯ РАН), Л.И. Маршева (ПСТГУ), свящ. Ф.Б. Людоговский (ИСл РАН), Р.Н. Кривко (ИРЯ РАН), И.В. Сегалович (Яндекс), свящ. К.О. Польсков (ПСТГУ), а также группа из 10 студентов ПСТГУ специальности «Прикладная филология» 2011 года набора и несколько студентов ОТИПЛа МГУ.

[3] О Церковнославянском корпусе на более ранней стадии разработки см. статью [Добрушина, Поляков 2013].

идет о языке определенного корпуса текстов, большая часть которого была окончательно сформирована не позже, чем к середине XVIII века. Однако при функциональном подходе, то есть, определяя церковнославянский как богослужебный язык, используемый в настоящее время в русской, болгарской, сербской и некоторых других Православных Церквях, следует обратиться к современной языковой ситуации.

До недавнего времени общепринятой оставалась точка зрения, согласно которой начиная с XVIII века церковнославянская книжность консервируется и уже не развивается (см., например: [Mathiesen 1984]). Лишь в последние десятилетия это представление было поколеблено, с одной стороны, благодаря изучению истории создания и редактирования имеющегося набора текстов, входящих в основной круг богослужебных книг, а с другой — благодаря исследованиям, показавшим, что лексика и метафорика богослужебных текстов продолжала и продолжает меняться и развиваться (подробнее см. [Кравецкий 2013]). Так что, помещая церковнославянский корпус в блок исторических, мы следуем традиции и отдаем себе отчет в том, что здесь возможны и иные решения.

***

Описывая **тексты, входящие в состав церковнославянского корпуса**, следует специально оговорить, что в него включены лишь церковнославянские тексты, относящиеся к периоду книгопечатания. Подготовка старославянского корпуса, а также корпусов различных изводов церковнославянского языка — это важные самостоятельные задачи, которые, мы надеемся, будут решены в рамках других проектов.

Церковнославянский корпус объединяет 1250 текстов, включающих около 4,6 миллиона словоупотреблений и около 150 тыс. различных словоформ. Для того чтобы наглядно представить, каков объем Церковнославянского корпуса, сравним его с объемом достаточно толстого книжного тома: сейчас его объем приблизительно аналогичен 25-ти подобным томам.

***

Особенностью корпуса является незначительное **упрощение графического облика** текстов. Найденные примеры на экране

выглядят близко к тому, как смотрелись бы в классической печатной церковнославянской книге, но с некоторыми упрощениями, а именно: отсутствуют знаки придыхания, букво-титла заменяются на буквы с надстрочным знаком, «юс малый» и «а-йотированное» заменяются на «я», «оу» и «ук» заменяются на простое «у» и др. Примеры со страницы выдачи результатов выглядят следующим образом:

Іи҃се, бу́рныхъ ѡти́шіе: іи҃се бж҃е, воздви́гни мя̀ па́дшаго. [Акафист Сладчайшему Господу нашему Иисусу Христу]

Тѣ́мже де́нь и но́щь къ тебѣ̀ да у́треннюю, де́нь и но́щь да тебѐ жа́жду, ѿ бж҃е мо́й! [Алфавит Духовный. Стихословия третия, рачительнаго к Богу вопиения]

Такой нетрадиционный облик текста разработчики избрали для того, чтобы пользователь мог использовать найденные примеры, не устанавливая для этого никаких дополнительных шрифтов. При использовании старых версий шрифтов типа Times часть символов будет отображаться как пустые квадраты, но в целом текст останется читаемым, а для комфортного просмотра церковнославянского текста нужно установить на компьютер любой достаточно полный юникодный шрифт, например, Arial Unicode MS. В будущем появится опция, дающая возможность копировать примеры в полноценном графическом оформлении тем, у кого нужные шрифты установлены.

***

Поиск организован так же, как в основном корпусе НКРЯ: по словарной форме (лемме) и грамматическим характеристикам. При проблемах с написанием леммы, можно воспользоваться грамматическим словарем, о котором см. ниже. Можно искать определенные части речи, например, в корпусе можно найти более 600 примеров с употреблениями междометий, а именно следующих трех: **ѡса́нна**, **ѿле**, **увы̀**. Можно искать по грамматической форме, например, можно получить примеры на все имеющиеся в Корпусе формы императива (более 99 000 форм) или на все формы аориста (более 98 000 форм). Можно получить примеры для любого выбранного слова (или нескольких слов) в заданной форме. Так, при запросе глагола *быти* в формах импефекта Корпус выдает почти 1500 употреблений.

Точно так же, как и во всем Национальном корпусе, можно искать сочетания слов с учетом грамматических характеристик. Например по запросу «наречие + *быти*» находится около 3000 сочетаний типа **при́снѡ сы́й**, **бли́ꙁъ су́щіи**, **гдѣ́ є́сть**, **свы́шє є́сть** и др.

Поскольку часть грамматической разметки создавалась автоматическими методами и эта разметка еще не выверена до конца, ошибки в грамматических трактовках встречаются, но, несомненно, постепенно будут устранены. Грамматическая омонимия в Церковнославянском корпусе не снималась.

***

На данном этапе Церковнославянский корпус по техническим причинам не дает **доступа к полным текстам**, в качестве примеров выдаются лишь фрагменты размера абзаца, что, несомненно, иногда создает неудобства для пользователя. Возможно, в дальнейшем опция перехода к полному тексту будет создана, пока же почти все тексты, помещенные в Корпус, можно найти в полном виде в электронной библиотеке «Библиотека святоотеческой литературы» (http://orthlib.ru).

***

Состав корпуса: тексты, заимствованные из церковнославянских книг, изданных типографским способом. Тексты, существовавшие лишь в рукописной традиции, в него не включены. Не вошли сюда и издания, в которых церковнославянские тексты были изданы как памятники языка или литературы. При описании жанровой структуры церковнославянского корпуса можно выделить несколько рубрик.

Книги, используемые в общественном и частном богослужении: Служебник, Требник, Минеи, Триоди, Октоих, Ирмологий и т. д. Именно богослужебные книги представляют собой ядро корпуса. В эту же рубрику следует, по всей видимости, включать Типикон, представляющий собою своеобразную инструкцию по использованию богослужебных книг. Богослужебные книги, в свою очередь, содержат значительное количество ремарок и комментариев, заимствованных из Типикона.

Священное Писание. Следует иметь в виду, что Священное Писание существует в двух вариантах — служебном (более архаич-

ном в языковом отношении) и четьем (см. об этом [Людоговский 2006 и 2010]). Богослужебная и четья редакции Писания отличались друг от друга еще в византийской традиции. Эти различия унаследовала и церковнославянская книжность. В XVIII веке в триодных паремьях богослужебная редакция была заменена четьим текстом, в то время как в минейных паремьях оставалась прежняя, более архаичная редакция. Таким образом, в современной богослужебной практике сосуществуют обе редакции Священного Писания, имеющие заметное количество лексических разночтений.

Акафисты. Среди церковнославянских текстов акафисты занимают особое место. Как известно, по степени русификации и неопределенности языковой нормы акафисты превосходят все другие виды церковнославянской книжности. При этом, согласно представлениям читателей (слушателей) акафистов, они написаны не на русском языке, а на церковнославянском. Среди церковнославянских текстов, написанных в XIX–XXI веке, акафисты явно преобладают. Эти тексты дают исследователям весьма ценный материал, поскольку они отчетливо демонстрируют представления широкой массы верующих о церковнославянском языке. Однако норма здесь оказывается весьма размытой. См. подробнее [Попов 2013, Людоговский и Плякин 2010, Людоговский 2010 а, Людоговский и Плякин 2013].

Источники канонического права. К этой рубрике можно отнести два текста: так называемую Никоновскую редакцию Кормчей книги (первое издание вышло в 1653, последнее — в 1834) и Книгу правил (об этой книге см. [Белякова 2004: 25–27]), которая с 1839 года заменяет Кормчую. Язык Кормчей достаточно архаичен, в то время как Книга правил сильно русифицирована.

Святоотеческая литература и агиография. Благодаря деятельности русских переводчиков второй половины XIX века, основной массив святоотеческих текстов существует на русском языке. В результате те миряне, которые имеют обыкновение читать дома святоотеческие тексты, как правило читают их по-русски. Церковнославянские переводы святоотеческих творений если переиздаются, то в качестве памятников, а не книг, востребованных читателями. Например, русский перевод «Добротолюбия» в конце XX — начале XXI века издавался многократно, в то время как славянский перевод воспроизводился репринтно только один раз

(это издание представлено в корпусе)[1]. При этом богословские сочинения, изначально написанные по-славянски, редко переводятся на русский язык. По-славянски читаются Иосиф Волоцкий[2], Нил Сорский, Паисий Величковский, Димитрий Ростовский[3] и др. Из текстов такого рода в Корпусе имеется Алфавит Духовный — сборник библейских толкований и молитвословных стихословий обычно приписывается святителю Димитрию Ростовскому, хотя на самом деле, его автором является Киевский митрополит Исаия Копинский[4]. Язык этих произведений имеет особенности, выходящие за пределы церковнославянской нормы.

Научная и художественная литература. В послепетровскую эпоху роль церковнославянского языка как языка науки и литературы постоянно сокращалась. Однако существует значительное количество церковнославянских изданий (в число которых входят, например, грамматика Мелетия Смотрицкого), включение которых в корпус кажется необходимым.

Основным источником текстов для корпуса явилась «Библиотека святоотеческой литературы» (http://orthlib.ru). Данный ресурс представляет собой результат огромной работы по оцифровке и переводу в текстовый формат основных церковнославянских книг[5], но тексты в том виде, как они представлены в библиотеке, были не вполне пригодны для Корпуса. Во-первых, они записаны в нестандартной кодировке HIP, которая удобна для набора на клавиатуре, но неудобна для чтения и обработки. Во-вторых, тексты не имеют грамматической разметки, что делает невозможным лексический поиск. Поэтому все тексты, взятые из библиотеки, были существенно переработаны и дополнены для нужд корпуса. В текстах Корпуса изредка встречаются разного типа ошибки и опечатки (например, ошибки сканирования, к счастью редкие, и вкрапле-

[1] Укажем также на славянское «Добротолюбие», изданное как литературный памятник: [Zamfiresco 1990].

[2] «Просветитель» Иосифа Волоцкого издается и в оригинале, и в русском переводе.

[3] Богословские сочинения Димитрия Ростовского не переводятся, в то время как агиографически чаще читаются в русской версии, чем в славянской.

[4] См. [Православная энциклопедия, т. 15: 17].

[5] Разработчики Церковнославянского корпуса рады возможности выразить признательность создателям Библиотеки священнику Владимиру Шину и М. Ю. Шин за их трудоемкую и бескорыстную работу.

ния русскоязычных комментариев к текстам), которые устраняются разработчиками по мере обнаружения.

В настоящее время основу Корпуса церковнославянских текстов составляют современные, то есть используемые в современной богослужебной практике богослужебные тексты: их доля в Корпусе примерно 60%. Кроме того, представлен и более ранний период — тексты XVII–XVIII веков, а также тексты небогослужебного назначения: Священное писание, святоотеческие сочинения, памятники канонического права. В дальнейшем планируется расширить Корпус за счет текстов из других источников. В таблице 1 представлен список основных книг и источников с указанием количества слов и доли от объема корпуса:

Таблица 1

№	Тип текста	Место, год издания	Слов	Доля в корпусе
1	Библия с параллельными местами	СПБ., 1900 (репринт: 1993)	650 253	14.36%
2	Евангелие	М., 1984	88 246	1.95%
3	Апостол	М., 1989	105 236	2.32%
4	Ирмологий	М., 1913 (репринт: 1995)	46 641	1.03%
5	Канонник	М., 1986	1 284	0.03%
6	Минея	Киев, 1893 (репринт: 1996–1997)	1 429 651	31.57%
7	Минея общая	М., 2002	102 493	2.26%
8	Минея праздничная	М., 1914 (репринт: 1993)	188 819	4.17%
9	Минейные службы (из разных источников)	–	13 923	0.31%
10	Молитвы, читаемые на молебнах	Пг., 1915 (репринт: 1993)	146 598	3.24%
11	Молитвы (из разных источников)	–	15 249	0.34%

12	Октоих	М., 1981	254 655	5.62%
13	Псалтирь следованная. Часть I	М., 1978	71 440	1.58%
14	Служебник	М., 1896	65 385	1.44%
15	Требник	М., 1906	108 325	2.39%
16	Триодь постная	М., 1992	217 295	4.80%
17	Триодь цветная	М., 1992	131 010	2.89%
18	Часослов	М., 1980	29 434	0.65%
19	Типикон	М., 1900	246 784	5.45%
20	Алфавит духовный	М., 1837 (репринт: 1997)	38 098	0.84%
21	Добротолюбие	М., 1902 (репринт: 2000)	294 909	6.51%
22	Ифика иерополитика, или Философия нравоучительная	СПБ., 1764	39 937	0.88%
23	Книга Правил святых апостолов, святых соборов и святых отец	М., 1893 (репринт: 1992)	75 513	1.67%
24	Пролог	М., 1643 (перенабор: 1915)	120 009	2.65%
25	Акафисты	(из разных источников)	47 597	1.05%

***

Метаразметка текстов корпуса организована по жанрам, с одной стороны, и по типу нормы с точки зрения временного периода — с другой. Типы текстов обозначены «корпусными ярлыками», то есть короткими условными обозначениями рубрик классификации, по которым тексты можно сортировать при создании пользовательского подкорпуса. Специфика метатекстовой разметки церковнославянского корпуса определяется тем, что тексты, входящие в состав богослужебных книг, имеют синтетический характер. Для большинства из них принципиально невозможно указать ни жанр, ни дату создания или перевода. Действительно, в состав богослу-

жебных последований входят поэтические тексты (каноны, стихиры), адресованные совершителю богослужения инструкции, нередки и чтения из Священного Писания. Эти тексты имеют различные жанровые характеристики и различную историю. Так, например, в минейном или триодном последовании содержатся гимнографические тексты разных жанров. Сюда входят тропари, подразделяющиеся на богородичны, троичны, крестобогородичны и т. д., стихиры, возгласы, чтения из Священного писания, уставные замечания и др. Точно так же в Типиконе содержится значительное число вкраплений из текстов, звучащих за богослужением. Поскольку метатекстовая разметка характеризует богослужебное последование целиком, а не каждый законченный фрагмент, в ней не может указываться тип текста даже с точностью до того, проза это или поэзия.

Поэтому в качестве характеристики типов текстов используются весьма общие ярлыки: (1) «Писание» (это Библия, Служебное евангелие и подборки паримий в богослужебных книгах, если они выделены в отдельную рубрику); (2) «святоотеческий»; (3) «служба» (это все богослужебные чины и службы, а также подборки богослужебных текстов (богородичны, кондаки и т. д.) в составе разных сборников); (4) «типикон»; (5) «акафист»; (6) «право» (один текст: «Книга правил»); (7) «научный» (один текст: «Ифика Иерополитика»).

***

Доступен **отбор текстов по типу языковой нормы**, связанной с периодом создания. Функционируют следующие ярлыки: (1) «архаичный тип» (например, «Добротолюбие»); (2) «гибридный тип» (например, «Алфавит Духовный»); (3) «стандартный тип» (это все тексты основных богослужебных книг за исключением текстов XX-го века); (4) «XX век» (это службы и акафисты, написанные в XX веке).

Вопрос о принадлежности текста к тому или иному временному периоду или к тому или иному типу языковой нормы очень запутан. Мы утверждаем, что Корпус объединяет тексты разных эпох в редакциях, возникших в эпоху книгопечатания. Все тексты подвергались нормализаторской правке в XVII — начале XX в., и с точки зрения языковых особенностей существенным является не время их создания или перевода, а время, когда текст был отредактирован и приобрел тот вид, в котором существует в печатной традиции.

Учитывая особенности истории церковнославянской книжности, создатели корпуса пришли к выводу, что в большинстве случаев рубрика «дата создания» является неактуальной. Между тем, в корпусе должна быть заложена возможность поиска на основе хронологических характеристик. В качестве компромисса мы предлагаем вместо даты создания текста, указывать время его последнего редактирования, ведь именно в процессе редактирования, касавшегося преимущественно лингвистических характеристик церковославянских текстов, формировалась языковая норма. При таком подходе выделяются следующие периоды:

1) XVII — сер. XIX в., характеризуется относительной свободой принципов редактирования;

2) сер. XIX — 1917 г. — период, для которого характерно стремления к созданию единой церковнославянской нормы;

3) вторая пол. XX — XXI в. — период, допускающий определенную нестабильность грамматической и орфографической нормы.

***

Описанная выше классификация нуждается в некотором **историческом комментарии**. С середины XIX по 1917 г. функционировала достаточно отлаженная система контроля за грамматической правильностью текста. Определяющим здесь было то, что государственные законы запрещали издавать церковную богослужебную литературу светским издательствам. На практике это означало, что почти все книги церковной печати издавались в трех синодальных типографиях. Здесь имелся штат редакторов (справщиков), обязанностью которых было наблюдение за языковой правильностью выпускаемых текстов и постепенная унификация выпускаемых изданий. Позже эти тексты воспроизводились без существенных изменений (см. подробнее [Кравецкий и Плетнева 2001: 17–24]).

Механизмы типографского контроля над грамматическим обликом церковнославянских книг формировались постепенно. Поэтому книги, изданные ранее, допускают куда большую грамматическую и орфографическую вариативность, что дает основания выделить их в отдельную группу. Для демонстрации подобной вариативности приведем несколько примеров из московского издания Библии 1663 года. В Московской Библии наряду со стандартным употреблением причастий встречаются причастия в пре-

дикативной функции. Употребление причастия в именительном падеже (из которых произошли деепричастные формы) в качестве сказуемого было характерной чертой русского извода церковнославянского языка (см. об этом [Алексеев 1987], [Горшкова, Хабургаев 1981: 336–337]). Однако в печатных книгах эта синтаксическая особенность воспринимается как нарушение нормы и исчезает. В Елизаветинской и позднейших изданиях Библии причастия в функции сказуемого уже не встречаются.

Библия. М., 1663	Библия. СПб., 1900
Быт. 19.23-24. в҆зы́деже сл҃нце над҆ зе́млю. лѡ́тъже вни́де въ сиго́ръ. гдⷭ҇ь же надожди́въ содо́мъ, и҆ гомо́ръ, ка́мы горѧ́щїи ѻ҆́гнь ѿ гдⷭ҇а съ нб҃сѐ.	Со́лнце взы́де над҆ зе́млю, лѡ́тъ же вни́де въ сигѡ́ръ. И҆ гдⷭ҇ь ѡ҆дождѝ на содо́мъ и҆ гомо́рръ жꙋ́пелъ, и҆ ѻ҆́гнь ѿ гдⷭ҇а съ небесѐ.
Быт. 19.35-37 Оу҆пои́стаже и҆ вино́мъ, в҆ше́дши ю҆нѣ́йшаѧ бы́сть со ѻ҆це́мъ своим. и҆ неразꙋмѣ̀ є҆гда̀ и҆стрезви́сѧ и҆́же бы́въ с҆ не́ю, и҆ воста̀. И҆ заче́нши ѻ҆́бѣ дще́ри лѡ́товѣ ѿ ѻ҆ц҃а своегѡ̀, И҆ родѝ старѣ́йшаѧ сн҃а, и҆ прозва̀ и҆́мѧ є҆мꙋ̀ мѡа́въ	Оу҆пои́ша же и҆ въ тꙋ̀ но́щь ѻ҆тца̀ своегѡ̀ вїно́мъ: и҆ вше́дши ю҆нѣ́йшаѧ преспа̀ со ѻ҆тце́мъ свои́мъ: и҆ не поразꙋмѣ̀ ѻ҆́нъ, є҆гда̀ преспа̀ и҆ є҆гда̀ воста̀. И҆ зача́ша ѻ҆́бѣ дще́ри лѡ́тѡвы ѿ ѻ҆тца̀ своегѡ̀: и҆ родѝ старѣ́йшаѧ сы́на и҆ наречѐ и҆́мѧ є҆мꙋ̀ мѡа́въ

Московская Библия сохраняет многие орфографические особенности, характерные для рукописной книжности, но исчезающие в период книгопечатания. Здесь иногда встречаются написания предлогов с паерком (Числ. 22.12. И҆ речѐ гдⷭ҇ь к҆ валаа́мꙋ) в Петербургском издании 1900: И҆ речѐ бг҃ъ къ валаа́мꙋ, слитное написание предлогов (Числ. 24.15–17 и҆ воста́нетъ чл҃къ ѿі҆и҃лѧ) в Петербургском издании 1900: и҆ воста́нетъ человѣ́къ ѿ і҆и҃лѧ. В Московской Библии встречаются выносные буквы без титла (Быт 11:1–9: и҆ и҆ спече́м ѧ҆̀ ѻ҆гне́мъ), в то время как в позднейших изданиях Библии выносные буквы употребляются только под взметом.

В отдельную группу выделяются тексты, вошедшие в богослужебную практику в советский период. Между 1918–1946 гг. на территории СССР богослужебные книги практически не переиздавались[1]. Новые тексты распространялись в машинописном виде

[1] Исключение составляли несколько частных изданий молитвословов, а также обновленческие издания, не оказавшие влияния на интересующие нас тексты.

(см. [Кравецкий 2012; Кравецкий, Плетнева 2014: 29–32]). Такой способ распространения делал невозможным какой бы то ни было контроль за орфографической правильностью текстов. Во второй половине XX века, когда появилась техническая возможность издавать богослужебную литературу, эти тексты постепенно стали входить в общецерковное употребление.

В некотором противоречии с тремя названными выше группами оказывается рубрика «Гибридные тексты». Сюда включены некоторые тексты, написанные на гибридном церковнославянском языке, которые впервые были изданы без участия синодальных типографий, а затем перепечатывались уже этими типографиями, однако радикальной языковой стандартизации для них проведено не было. Сейчас в Корпус включено два таких текста — «Алфавит духовный» и славянское «Добротолюбие»». В качестве иллюстрации приведем фрагмент «Увещания Антония Великого о нравах человеческих и благом житии»: «**Не дивно оубо, аще онъ къ простѣйшей бесѣдѣ, и древнешбразной и небрегомой, слогъ рѣчи склонилъ: но то дивно, акш ѿ толикїя простоты толикое спсенїе и польза чтущымъ бываетъ. Колико бо въ оныхъ ѡбрѣтаетсѧ оубѣренїе! колика ѿ оныхъ истекаетъ сладость! и вообще, колико источаетсѧ благїй нравъ и совершенство єѵльскаго жительства!**». Даже при беглом чтении бросается в глаза, что синтаксическая структура этого текста ближе к русскому литературному языку первой половины XVIII века. К тому же обращает на себя внимание использование восклицательного знака (церковнославянская пунктуация этого знака не знает) и т. д.

***

Основной **особенностью поиска в Корпусе** является возможность задавать искомое слово в нескольких **орфографических вариантах**: точном, упрощенном и модернизированном. Это удобно потому, что пользователь сможет найти нужное ему слово, даже если не уверен в его орфографии или затрудняется в введении со своей клавиатуры сложной буквы типа «ѱ». Например, если пользователь не уверен, как пишется нужное ему слово — через «е» или через «ѣ», то он может, не тратя время на выяснение этого, просто ввести слово в модернизированном варианте орфографии, при котором эти буквы не различаются. Результаты поиска всегда

будут выданы в одном орфографическом варианте — классическом церковнославянском.

В точном варианте орфографии используются все буквы, а именно 45 (плюс «титло»). В упрощенном — 39 букв, так как некоторые буквы совмещены ([«і» и «ї»], [«ѵ» и «ѷ»], [«ѱ» и «оѱ»], [«я» и «ѧ»] и др.), но сохраняются основные лексически значимые оппозиции («є» и «ѣ», «ꙁ» и «ѕ» и др.). В модернизированном варианте используются только 32 буквы, существующие в современном русском алфавите, например, совмещены в одной клавише [«є», «ѥ», «ѣ»] или [«и», «і», «ї», «ѵ», «ѷ»]. Различия в типах орфографии можно увидеть, просто вызвав виртуальные клавиатуры упрощенного и модернизированного типов запросов и изучив надписи на клавишах, совмещающих буквы. Впрочем, для успешного пользования корпусом нет необходимости в этом разбираться. Варианты орфографии нужны исключительно для упрощения составления запроса.

***

Как упоминалось выше, самое важное свойство лингвистического корпуса состоит в том, что любая словоформа в тексте снабжена грамматической информацией, по которой можно осуществлять поиск. Главная **особенность такого подхода к описанию грамматики** в Корпусе заключается в том, что, в отличие от традиционных грамматик, парадигмы не задаются априорно, а выводятся эмпирически на основе анализа множества словоформ, имеющих однотипное соотношение между грамматическими формами. Таким образом, номенклатура парадигм получается значительно более детальной и, будучи законченной, должна в эксплицитной форме включить тонкости словоизменения, которые, как правило, не учитываются или описываются в примечаниях. Например, традиционное первое склонение (рабъ) на самом деле распадается на 14 подтипов в зависимости от конечного согласного (парный твердый–мягкий, велярный, шипящий, йот), наличия беглого гласного и других особенностей. Все вариантные словоформы должны быть или включены в парадигму, или трактоваться как ошибки.

Грамматический словарь и грамматическая модель Корпуса создавались итеративно — методом последовательных приближений. Сначала из базы текстов был сгенерирован полный список слово-

форм и проведена его первичная проверка и чистка (исправлены явные ошибки). Затем для наиболее частотных слов была сделана ручная лемматизация, определены типичные словоизменительные шаблоны и таким образом построена первичная модель словоизменения. Параллельно с работой над словарем разрабатывалась грамматическая модель церковнославянского словоизменения: грамматические таблицы, номенклатура парадигм, состав грамматических признаков, правила порождения словоформы по лемме, список чередований и др. В процессе разработки словарь и модель постоянно корректируются и согласуются между собой.

***

Последним достижением в усовершенствовании церковнославянского корпуса, также имеющим самостоятельную ценность, стало создание **Частотного грамматического словаря церковнославянского языка**. Во избежание путаницы подчеркнем, что это именно грамматический словарь, а не двуязычный или толковый. Помимо него разработчики Церковнославянского корпуса участвуют в работе проектов по созданию еще двух словарей: в ИРЯ РАН разрабатывается двуязычный словаря современного церковнославянского языка (см. о нем [Добровольский и др. 2013]), в ПСТГУ — толковый словарь современной религиозной лексики, включающий подробное описание многозначности тех лексем церковнославянского происхождения, которые реально функционируют в современном русском христианском дискурсе (об этом словаре см. [Добрушина 2012] и [Добрушина, Польсков и др. 2011]).

Грамматический церковнославянский словарь в данный момент размещен в интернете по адресу http://feb-web.ru/febupd/slavonic/dicgram/, а в дальнейшем будет присоединен к Церковнославянскому корпусу НКРЯ. Цель присоединения к корпусу — упростить работу с ним, предоставив пользователю возможность при желании не самому набирать форму или лемму, что может быть затруднительно из-за графических или орфографических проблем, а выбирать ее из алфавитного списка или всплывающего на основе первых набранных букв списка лемм. На настоящий момент из НКРЯ перехода к словарю нет, но из самого словаря переход к списку примеров в НКРЯ работает.

Грамматический словарь церковнославянского языка — это описание грамматики языка, полностью определяемое имеющейся базой текстов, состоящее из словника и грамматических таблиц и отражающее около 35000 лемм и 60000 словоформ. Словник представляет собой список лемм, размещенных в прямом или обратном алфавитном порядках: пользователь может выбирать более удобный для его нужд порядок. Каждая из лемм снабжена следующей информацией: грамматические пометы; индекс, обозначающий тип словоизменения; перечень форм, найденных в текстах корпуса с указанием их частоты. При нажатии на грамматический индекс осуществляется переход к соответствующей грамматической таблице. При нажатии на лемму или форму слова осуществляется переход ко всем примерам корпуса, включающим эту лемму или форму. В дальнейшем будет возможность от грамматических таблиц при нажатии на индекс переходить к списку всех лемм, формоизменение которых соответствует этому индексу.

Словарные статьи имеют следующую структуру: (1) лемма, при нажатии на которую можно перейти ко всем примерам корпуса, содержащим ее, например «жре́ти» (см. образец словарной статьи целиком ниже); (2) грамматические пометы, описывающие лемму, например «*V, ipf, tr*» (то есть «глагол, несовершенного вида, переходный»); (3) число в скобках, обозначающее количество употреблений этой леммы во всех ее формах в текста корпуса, в данном случае (115); (4) грамматический индекс, отсылающий к определенной грамматической таблице, при нажатии на который происходит переход к соответствующей таблице; в данном примере индекс «V15er» отсылает к таблице, описывающей парадигму глаголов с основами на сонорный, а именно на «р» типа *умрети*; (5) краткий, не учитывающий многозначности перевод на русский — для некоторых проблемных лемм, например, «*приносить жертвоприношение*» для «жре́ти»; (6) список всех встретившихся в корпусе форм этой леммы с указанием количества употреблений в скобках: при нажаии на каждую происходит переход к списку включающих ее примеров: жре́мую (1), жре́мъ (2), жре́нъ (5), жре́ти (39), жре́тъ (5), жретѐ (1), жрѹ̀ (3), жрѹ́тъ (9), жрѹ́щаго (1), жрѹ́ще (5), жрѹ́щему (3), жрѹ́щими (1), жрѹ́щихъ (1), жрѹ́щымъ (2), жрѹ́щыя (1), жрѹ́щіи (1), жря̀ (5), жря́й (5), жря́ху (19), жря́ше (6).

Образцы словарных статей

жре́ти *V,ipf,tr* (115)

V15er *приносить жертвоприношение*

жре́мую (1), жре́мъ (2), жре́нъ (5), жре́ти (39), жре́тъ (5), жретѐ (1), жрѝ (3), жру́тъ (9), жру́щаго (1), жру́ще (5), жру́щему (3), жру́щими (1), жру́щихъ (1), жру́щымъ (2), жру́щыя (1), жру́щіи (1), жря̀ (5), жря́й (5), жря́ху (19), жря́ше (6)

ла́стовица *N,m,anim* (40)

N3c

ласточка

ла́стовица (22), ла́стовице (9), ла́стовицу (1), ла́стовицъ (1), ла́стовицѣ (2), ла́стwвицы (3), ла̂стовицы (2)

мину́ти *V,pf,intr*

V13a (53) *миновать*

мине́мъ (1), мине́тъ (7), минѝ (2), минѝ (4), мину́вшагw (2), мину́вшей (1), мину́вши (1), мину́вшимъ (1), мину́вшихъ (2), мину́вшу (6), мину́вшымъ (5), мину́въ (8), мину́ла (1), мину́ти (7), мину́хомъ (1), мину́ша (4)

ну́дити *V,ipf,tr* (42)

V21t

ну́ди (2), ну́дими (2), ну́дите (1), ну́дити (4), ну́дитъ (11), ну́диши (3), ну́дя (1), ну́дятъ (2), ну́дящаго (3), ну́дящагw (1), ну́дяще (1), ну́дящи (1), ну́дящихъ (1), ну́дящую (1), ну́дящымъ (3), ну́ждаста (3), нужда́ху (1), нужда́ше (1)

сапфі́ръ *N,m,inan* (37)

N1t

сапфи́ру (1), сапфі́ра (19), сапфі́ровѣ (1), сапфі́ромъ (3), сапфі́рою (2), сапфі́ру (1), сапфі́ръ (5), сапфі́ры (1), сапфѵ́ра (1), сапфѵ́ру (1), сапфѵ́ръ (1), сапфѵ́рѣ (1)

чревоволше́бникъ *N,m,anim* (5)

N1k

чревоволше́бники (1), чревоволше́бникwвъ (4)

шу́ій A (69)

A1a

шу́е (3), шу́ее (5), шу́ей (1), шу́ими (4), шу́ихъ (6), шу́ию (7), шу́ягw (8), шу́яя (2), шу́іи (2), шу́іими (1), шу́іихъ (4), шу́ія (13), шу́іягw (10), шу̂яя (2), шуЇимъ (1)

Образцы примеров к леммам и словоформам. При нажатии на лемму или словоформу пользователь переходит к списку примеров всех употреблений этой леммы или словоформы в обрабатываемой базе. Так, лемма «чревоволшебникъ» в корпусе использована только 5 раз. Примеры для нее выглядят следующим образом:

• Библия. Исайя [омонимия не снята] Все примеры (3)

• 19 **И а́ще реку́тъ къ ва́мъ: изыщи́те чревоволше́бникѡвъ и ѿ землѝ возглаша́ющихъ, тщесло́вующихъ, ѝже ѿ чре́ва глаша́ютъ, не язы́къ ли къ бг҃у своемꙋ҆ взы́щетъ; что̀ испытꙋ́ютъ ме́ртвыя ѡ живы́хъ;** [Библия. Исайя] [омонимия не снята] ←…→

• 3 **И возмяте́тся дꙋ́хъ еги́птянъ въ ни́хъ, и совѣ́тъ и́хъ разсы́плю, и вопро́сятъ богѡ́въ свои́хъ и кумі̑рѡвъ свои́хъ, и глася́щихъ ѿ землѝ и чревоволше́бникѡвъ.** [Библия. Исайя] [омонимия не снята] ←…→

• 25 **Кто̀ и́нъ разсы́плетъ зна́менія чревоволше́бникѡвъ и волшвы^ ѿ се́рдца, ѿвраща́яй мꙋ^дрыя вспя́тъ и совѣ́ты и́хъ ѡбꙋя́яй,** [Библия. Исайя] [омонимия не снята] ←…→

• Библия. 2 Паралипоменон [омонимия не снята] Все примеры (1)

• 6 **То́й же проведѐ ча^да своя^ чрезъ҆ ѻ́гнь во ꙋдо́ли веенно́мли и вража́ше (со жре́біи), и послѣ́доваше гада́ніемъ, и волхова́ніемъ служа́ше, и имѣ́яше чревоволше́бники и ѡбая́нники, и мно́га лꙋка^вая сотворѝ предъ҆ гдⷭ҇емъ, да прогнѣ́ваетъ его̀:** [Библия. 2 Паралипоменон] [омонимия не снята] ←…→

• Триодь Постная. Третья седмица (с понедельника по пяток) [омонимия не снята] Все примеры (1)

• **И а́ще реку́тъ къ ва́мъ: изыщи́те чревоволше́бникѡвъ, и ѿ землѝ возглаша́ющихъ, тщесло́вующихъ, ѝже ѿ чре́ва глаша́ютъ: не язы́къ ли къ бг҃у своемꙋ҆ взы́щетъ, что̀ испытꙋ́ютъ ме́ртвыя ѡ живы́хъ; зако́нъ бо въ по́мощь дадѐ, да реку́тъ не я́коже сло́во сіѐ, зане́ же не лѣ́ть да́ры дая́ти.** [Триодь Постная. Третья седмица (с понедельника по пяток)] [омонимия не снята]

Грамматические таблицы

Созданы следующие грамматические таблицы:

(1) Существительные (43 варианта словоизменения в зависимости от грамматических характеристик и типов основ).

(2) Прилагательные (8 вариантов словоизменения в зависимости от типов основ).

(3) Местоимения-прилагательные (9 вариантов словоизменения в зависимости от типов основ).

(4) Местоимения-существительные (7 вариантов словоизменения в зависимости от типов основ).

(5) Глаголы (50 вариантов словоизменения в зависимости от типов основ). Приведем образец одной из таблиц для глаголов:

Глаголы на -нути

Парадигма	**V13a**	**V13t**	**V13g**
Примеры	**мин-ути**	**кис-нути**	**двиг-нути, мерк-нути**
из.наст.ед.1	мин-у	кис-ну	двиг-ну
из.наст.ед.2	мин-еши	кис-неши	двиг-неши
из.наст.ед.3	мин-етъ	кис-нетъ	двиг-нетъ
из.наст.мн.1	мин-емъ	кис-немъ	двиг-немъ
из.наст.мн.2	мин-ете	кис-нете	двиг-нете
из.наст.мн.3	мин-утъ	кис-нутъ	двиг-нутъ
из.наст.дв.1	мин-ева(-ѣ)	кис-нева(-ѣ)	двиг-нева(-ѣ)
из.наст.дв.2/3	мин-ета(-ѣ)	кис-нета(-ѣ)	двиг-нета(-ѣ)
пов.ед.2/3	мин-и	кис-ни	двиг-ни
пов.мн.2	мин-ите	кис-ните	двиг-ните
пов.мн.1	мин-емъ	кис-немъ	двиг-немъ
пов.дв.1	мин-ева(-ѣ)	кис-нева(-ѣ)	двиг-нева(-ѣ)
пов.дв.2	мин-ита(-ѣ)	кис-нита(-ѣ)	двиг-нита(-ѣ)
из.имперф.ед.1	мин-яхъ	кис-няхъ	двиг-няхъ
из.имперф.ед.2/3	мин-яше	кис-няше	двиг-няше
из.имперф.мн.1	мин-яхомъ	кис-няхомъ	двиг-няхомъ
из.имперф.мн.2	мин-ясте	кис-нясте	двиг-нясте
из.имперф.мн.3	мин-яху	кис-няху	двиг-няху
из.имперф.дв.1	мин-яхова(-ѣ)	кис-няхова(-ѣ)	двиг-няхова(-ѣ)
из.имперф.дв.2/3	мин-яста(-ѣ)	кис-няста(-ѣ)	двиг-няста(-ѣ)
из.аор.ед.1	мин-ухъ	кис-нухъ/охъ	двиг-нухъ/охъ
из.аор.ед.2/3	мин-у	кис-ну/е	двиг-ну/движ-е
из.аор.мн.1	мин-ухомъ	кис-нухомъ/охомъ	двиг-нухомъ/охомъ
из.аор.мн.2	мин-усте	кис-нусте/осте	двиг-нусте/осте
из.аор.мн.3	мин-ша	кис-нуша/оша	двиг-нуша/оша

из.аор.дв.1	мин-ухова(-ѣ)	кис-нухова(-ѣ)/охова(-ѣ)	двиг-нухова(-ѣ)/охова(-ѣ)
из.аор.дв.2/3	мин-уста(-ѣ)	кис-нуста(-ѣ)/оста(-ѣ)	двиг-нуста(-ѣ)/оста(-ѣ)
инф.	мин-ути	кис-нути	двиг-нути
прич.перф.ед.м. прич.наст.	мин-ул-ъ	кис-нул-ъ	двиг-л-ъ/нул-ъ
действ.ед.м./ср.	мин-ый	кис-ный	двиг-ный
прич.наст.действ.	мин-ущ-и	кис-нущ-и	двиг-нущ-и
прич.наст.страд.		-	-
прич.прош.действ.	мин-ув-ъ		двиг-ъ/нув-ъ
прич.прош.страд.	мин-ут-ъ		двиг-нут-ъ

Условные обозначения. Грамматические признаки в словаре обозначены латиницей. Это связано с тем, что словарь создавался методами, принятыми в Национальном корпусе русского языка. Примеры помет: N — существительное, A — прилагательное, ADV — наречие, anim — одушевленное, inan — неодушевленное, poss — притяжательное, persn — личное и др. Возможно, для филологов-славистов, которые будут работать со словарем, такой вид помет не очень привычен, поэтому в дальнейшем планируется дублировать пометы на русском языке в стиле Грамматического словаря А. А. Зализняка.

На основе грамматического частотного словаря хорошо видна специфика церковнославянского языка: во-первых, вариативность, во-вторых, ограниченное количество существующих словоформ по отношению к потенциально возможным. Это следствие того, что церковнославянский, с одной стороны, малокодифицирован, с другой — существует на основе ограниченного набора текстов. Разумеется, в корпусе отражены не все тексты и по мере расширения корпуса в словарь будут добавляться новые леммы и новые словоформы, но опыт показывает, что далеко не всякий новый текст содержит новую лексику, в целом имеющийся объем весьма представителен. Специфика церковнославянского языка, в отличие от обычных языков, в частности в том, что набор текстов, представляющих его, ограничен. Создание новых текстов хотя и имеет место, но происходит медленно и в очень большой степени опирается на уже имеющиеся образцы. Поэтому, если для обычного языка для каждой

леммы с неущербной парадигмой, как правило, находится в письменных или устных текстах любая потенциально возможная форма, то в церковнославянском языке для многих лемм реально живут одна-две-три формы, как для слова «чревоволшéбникъ», словарная статья которого приведена выше. У более частотных слов форм, реализованных в текстах, больше, но как правило, заметно меньше, чем в их парадигме. Так описанная в приведенном выше образце таблицы парадигма глагола «**минути**» содержит 33 формы (без учета изменения причастий по падежам, родам и числам), а реальный глагол, как видно из образца словарной статьи, функционирует всего в 16, и вероятность появления текстов — новосозданных или существующих, но еще не включенных в корпус, с другими формами этого глагола, невелика. Для сравнения отметим, что из 24 потенциально возможных формы современного русского глагола *миновать*, 22 формы, включая настоящее страдательное причастие и повелительные формы обоих чисел, реально находятся в корпусе, не существует лишь форм страдательного причастия прошедшего времени и формы повелительного наклонения совместного действия. Поэтому картина языка, полученная корпусным методом и запечатленная в частотном словаре церковнославянского языка, отражает гораздо более объективную картину, чем просто грамматика или словарь, в которых указаны потенциально возможные, но не всегда реально существующие формы.

Литература

Алексеев А.А. Participium activi в русской летописи: особенности функционирования // Russian Linguistics. 1987. Vol. 11. № 2/3. С. 187–200;

Алипий (Гаманович). Грамматика церковно-славянскаго языка. Jordanville (N. Y.), 1964.

Белякова Е. В. Церковный суд и проблемы церковной жизни. М., 2004.

Горшкова К. В., Хабургаев Г. А. Историческая грамматика русского языка. М., 1981.

Добровольский И. С., Калужнина Н. В., Кравецкий А. Г., свящ. Федор Людоговский, Плетнева А. А., Хитров А. Н. Из опыта работы

над словарем современного церковнославянского языка. // Вестник ПСТГУ. Серия III: Филология. 2013. Вып. 4 (34). С. 40–57.

Добрушина Е. Р. Словарь христианской лексики: состав словника // Вестник ПСТГУ. Серия III: Филология. 2012. Вып. 3 (29). С. 105–113.

Добрушина Е. Р., Поляков А. Е. Корпус церковнославянского языка: возможности, методы создания, перспективы. // Вестник ПСТГУ. Серия III: Филология. 2013. Вып. 1 (31). С. 32–44.

Добрушина Е. Р., свящ. Константин Польсков, Литвинцева К. В., Хангиреев И. А. От «аббата» до «аналоя»: фрагмент «Лингво-энциклопедического словаря христианской лексики» // Вестник ПСТГУ. Серия III: Филология. 2011. Вып. 4 (22). С. 119–148.

Кравецкий А. Г. Церковнославянский язык как один из современных славянских языков // Славянское языкознание: XV Международный съезд славистов. Минск, 2013 г. Доклады российской делегации. М.: «Индрик», 2013. С. 571–589.

Кравецкий А. Г. Литургический самиздат XX века: языковые особенности и проблемы рецепции // Latopisy Akademii Supraskiej. Vol. 3. Język naszej modlitvy — dawnej I dziś. Pod redakcją Urszuli Pawluczuk. Białystok 2012. P. 85–94.

Кравецкий А. Г., Плетнева А. А. История церковнославянского языка в России (конец XIX–XX в.). М.: «Языки русской культуры», 2001.

Кравецкий А. Г., Плетнева А. А. Минейные службы Нового времени: история, поэтика, семантика // Минеи: образец гимнографической литературы и средство формирования мировоззрения православных / Под редакцией Е. Потехиной и А. Кравецкого. Olsztyn: Centrum Badań Europy Wschodniej. 2013. С. 15–90.

Людоговский Ф. Б. Функционирование и эволюция служебного и четьего вариантов церковнославянского Евангелия в эпоху книгопечатания: постановка проблемы // Лингвистическое источниковедение и история русского языка. 2004–2005. М.: «Древлехранилище», 2006. С. 400–418.

Людоговский Ф. Б. Ветхозаветные паремии в составе современного церковнославянского минейного корпуса // Лингвистическое источниковедение и история русского языка. 2006–2009. М.: «Древлехранилище», 2010. С. 523–545.

Людоговский Ф. Б. Церковнославянский акафист: структуры

и константы // Славянские языки: единицы, категории, ценностные константы: сб. науч. статей. Волгоград, 2010. С. 47–68.

Людоговский Ф. Б., Плякин М. Хвалебные гимны подвижникам // Журнал Московской Патриархии. 2010. № 6. С. 70–76..

Людоговский Ф. Б. Структура и поэтика церковнославянских акафистов. М., Институт славяноведения РАН, 2015.

Людоговский Ф. Б., Плякин М. Акафисты ангелам: опытный фрагмент каталога церковнославянских акафистов // Лингвистическое источниковедение и история русского языка <2012–2013>. М.: «Древлехранилище», 2013. С. 260–280.

Попов А. В. Православные русские акафисты. М.: Издательство Московской Патриархии, 2013.

Православная энциклопедия. Том 1–38. М., 2000–2015 (издание продолжается).

Mathiesen R. The Church Slavonic Language Question: An Overview (IX–XX Centeries) // Aspects of the Slavic Language Question. Ed. By R. Picchio and G. Goldblatt. New Haven, 1984. Vol. I: 61–64.

Zamfiresco Dan. La philocalie slavonne de Païssy Velichkovsky. Bucarest Roza vînturilor 1990.

[1]E. R. Dobrushina, [2]A. G. Kravetskii, [3]A. E. Polyakov
[1]St. Tikhon's Orthodox University
edobrush@gmail.com
[2]Vinogradov Russian Language Institute
of the Russian Academy of Sciences
krav62@mail.ru
[3]Ushinsky Scientific Pedagogical Library
pollex@mail.ru
[1,2,3](Moscow, Russia)

CORPUS AND FREQUENCY GRAMMAR CORPUS VOCABULARY OF THE CHURCH SLAVONIC LANGUAGE AS PART OF THE RUSSIAN NATIONAL CORPUS

The article describes the current state of work on the Corpus of Church Slavonic within the Russian National Corpus. The Corpus is

different from the collection of texts by the presence of a special markup (grammatical, structural, metatextual) and the possibility to search by this markup. The article considers the content, describes the main genre headings, according to which the texts included in the Corpus are distributed, describes the principles of the metamarkup that differ from those used in other Corpuses within the Russian National Corpus. Where necessary, the historical information, on the basis of which a certain decision was made, is provided. Since typing search queries on the keyboard presents certain difficulties for Church Slavonic texts, we offer several options for simplified spelling transmission, which enable a person to enter a query with limited capabilities of a standard keyboard. Finally, the article describes the frequency grammar vocabulary created during the work on the project.

Key words: Church Slavonic language, corpus linguistics, spelling, inflexion, morphology, grammatical dictionary

References

Alekseev A.A. [Participium activi in Russian chronicles: features o functioning]. *Russian Linguistics*, 1987, Vol. 11, no. 2/3, pp. 187–200. (In Russ.)

Alipii (Gamanovich). *Grammatika tserkovno-slavyanskago yazyka* [Grammar of the Church Slavonic language.]. Jordanville (N.Y.), 1964.

Belyakova E.V. *Tserkovnyi sud i problemy tserkovnoi zhizni* [Ecclesiastical court and the problems of church life]. Moscow, 2004.

Gorshkova K.V., Khaburgaev G.A. *Istoricheskaya grammatika russkogo yazyka* [Historical Grammar of the Russian Language]. Moscow, 1981, pp. 336–337

Dobrovol'skij I.S., Kaluzhnina N.V., Kravetskii A.G., svyashch. Fedor Lyudogovskii, Pletneva A.A., Khitrov A.N. [Working on the Dictionary of Contemporary Church Slavonic]. *Vestnik PSTGU. Seriya III: Filologiya* [Saint Tikhon Orthodox University Herald. Series III. Philology]. 2013, Vol. 4 (34), pp. 40–57. (In Russ.)

Dobrushina E.R. [Dictionary of Christian vocabulary: composition of glossary]. *Vestnik PSTGU. Seriya III: Filologiya*. [Saint Tikhon Orthodox University Herald. Series III. Philology]. 2012, Vol. 3 (29), pp. 105–113. (In Russ.)

Dobrushina E.R., Polyakov A.E. [Corpus of the Church

Slavonic language: possible methods of creation, prospects.]. *Vestnik PSTGU. Seriya III: Filologiya* [Saint Tikhon Orthodox University Herald. Series III. Philology]. 2013, Vol. 1 (31), pp. 32–44. (In Russ.)

Dobrushina E. R., svjashh. Konstantin Pol'skov, Litvinceva K. V., Hangireev I. A. [From "abbot" to "lectern": extract from "Linguistic and Encyclopedic Dictionary of the Christian vocabulary"]. *Vestnik PSTGU. Seriya III: Filologiya* [Saint Tikhon Orthodox University Herald. Series III. Philology]. 2011, Vol. 4 (22), pp. 119–148. (In Russ)

Kravetskii A. G. [Church Slavonic language as one of the modern Slavic languages]. *Slavyanskoe yazykoznanie: XV Mezhdunarodnyi s"ezd slavistov. Minsk, 2013. Doklady rossiiskoi delegatsii* [Slavic Linguistics: XV International Congress of Slavists. Minsk, 2013. Report of the Russian delegation]. Moscow, Indrik Publ., 2013, pp. 571–589. (In Russ.)

Kravetskii A. G. [Liturgic self-printed issues of the XXth century: language peculiarities and the perception problems]. *Latopisy Akademii Supraskiej*. Vol. 3. Język naszej modlitvy — dawnej I dziś. Pod redakcją Urszuli Pawluczuk, Białystok, 2012, pp. 85–94.

Kravetskii A. G., Pletneva A. A. *Istoriya tserkovnoslavyanskogo yazyka v Rossii (konets XIX–XX v.)* [History of the Church Slavonic language in Russia (end of XIX–XX c.)]. Moscow, Jazyki russkoj kul'tury Publ., 2001.

Kravetskii A. G., Pletneva A. A. [Menaia Services during the Early Modern and Modern Periods: History, Poetics and Semantics]. *Minei: obrazets gimnograficheskoi literatury i sredstvo formirovaniya mirovozzreniya pravoslavnykh* [Minaion: an example of hymnographyc literature and a tool to shape the Orthodox worldview]. Pociechina, Helena, Kravetsky, Alexander (eds.). Olsztyn, Centrum Badań Europy Wschodniej Publ., 2013, pp. 15–90. (In Russ.)

Lyudogovskii F. B. [Functioning and evolution of the service and menology variants of the Church Slavonic Gospel in the era of printing: the problem]. *Lingvisticheskoe istochnikovedenie i istoriya russkogo yazyka. 2004–2005* [Linguistic Source Study and History of the Russian language. 2004–2005]. Moscow, Drevlekhranilishche Publ., 2006, pp. 400–418. (In Russ.)

Lyudogovskii F. B. [Testament paremias as part of the modern Church Slavonic menological corpus]. *Lingvisticheskoe istochniko vedenie i istoriya russkogo yazyka. 2006–2009* [Linguistic Source

Study and History of the Russian language. 2006–2009]. Moscow, Drevlekhranilishche Publ., 2010, pp. 523–545. (In Russ.)

Lyudogovskii F. B. [Church Slavonic akathist: structures and constants]. *Slavyanskie yazyki: edinitsy, kategorii, tsennostnye konstanty: sb. nauch. statei* [Slavic languages: units, categories, value constants: Coll. of scientific articles.]. Volgograd, 2010, pp. 47–68. (In Russ.)

Lyudogovskii F. B. *Struktura i poetika tserkovnoslavyanskikh akafistov* [Structure and Poetics of Church Slavonic Akathists.]. Moscow, The Institue of Slavic Studies of the Russian Academy of Sciences, 2015.

Lyudogovskii F. B., Pljakin M. [Hymns of praise to hermits]. *Zhurnal Moskovskoi Patriarkhii*, 2010, no. 6, pp. 70–76. (In Russ.)

Lyudogovskii F. B., Plyakin M. [Akathists to angels: a practiced fragment of the Church Slavonic akathists directory]. *Lingvisticheskoe istochnikovedenie i istoriya russkogo yazyka <2012–2013>* [Linguistic Source Study and History of the Russian language <2012–2013>.]. Moscow, Drevlekhranilishche Publ., 2013, pp. 260–280. (In Russ.)

Mathiesen R. The Church Slavonic Language Question: An Overview (IX–XX Centeries). *Aspects of the Slavic Language Question.* Ed. By R. Picchio and G. Goldblatt. New Haven, 1984. Vol. I: 61–64.

Popov A. V. *Pravoslavnye russkie akafisty* [Russian Orthodox akathists]. Moscow, Izdatel'stvo Moskovskoi Patriarkhii, 2013.

Pravoslavnaya entsiklopediya [Orthodox Encyclopedia]. Vol. 1–38. Moscow, 2000–2015.

Zamfiresco Dan. *La philocalie slavonne de Païssy Velichkovsky.* Bucarest Roza vînturilor 1990.

[1]И. Б. Качинская, [2]Д. В. Сичинава
[1]Московский государственный университет им. М. В. Ломоносова
[2]Институт русского языка им. В. В. Виноградова РАН
[1,2](Россия, Москва)
[1]kacza@yandex.ru, [2]mitrius@gmail.com

ДИАЛЕКТНЫЙ ПОДКОРПУС СЕГОДНЯ[1]

В статье представлено современное состояние Корпуса диалектных текстов, являющегося подкорпусом НКРЯ. В 2005–2009 гг. проводилась работа по разметке пилотного диалектного корпуса. В последние годы проведена работа по усовершенствованию разметки диалектных текстов на различных уровнях (метатекстовая информация о времени и месте записи, жанре и других особенностях текста; фонетика, морфология, семантика, элементы синтаксиса), а также разработаны программные средства для разметки корпуса — прежде всего пользовательский интерфейс «Рабочее место диалектолога». Ведётся работа по сбору текстов в различных регионах России, а также по разметке ранее собранных текстов, как опубликованных, так и входящих в архивы исследователей. Разработана технология перевода транскрибированных текстов в орфографизированную форму; эта последняя версия затем подвергается полуавтоматической морфологической разметке. Предусмотрена возможность также включения в корпус текстов в орфографизированной записи. Морфологическая разметка сопровождается пометами о тех или иных нормативных особенностях на разных уровнях (основа, словообразование, словоизменение). Предусмотрена возможность

[1] Работа над Диалектным корпусом НКРЯ поддержана грантом РГНФ № 14-04-12012, проект «Корпус диалектных текстов Национального корпуса русского языка: пополнение и разметка», рук. Д. В. Сичинава.

включения также мультимедийной информации (видео- и аудиозапись). В статье приводятся сведения также о других существующих диалектных корпусах, разработчики многих из которых сотрудничают с командой диалектного корпуса НКРЯ.

Ключевые слова: Диалектный корпус, диалектология, морфологическая разметка, метатекстовая разметка, рабочее место

1. Концепция Корпуса диалектных текстов

Корпус диалектных текстов на сайте НКРЯ существует с 2005 года: http://www.ruscorpora.ru/search-dialect.html. Уже с самого начала отмечались особенности диалектной грамматики, приводился «паспорт» записи (место, время, данные об информанте и лице, предоставившем текст), имелись указания на жанрово-тематическую отнесенность текстов, на «диалектность» значений отдельных слов [Летучий 2005; 2009]. Вместе с тем тексты подавались в орфографизированном виде, не включалось толкование значений лексем.

За истекший период концепция подкорпуса существенно изменилась [Качинская 2009; 2011; Сичинава, Качинская 2014]. Возникла идея нового стандарта подачи как самих текстов, так и степени лингвистической обработанности каждой лексемы.

1.1. В настоящее время пользователю предоставлена возможность работать не только с фрагментами текстов (как во многих других подкорпусах, материалами для которых послужили тексты художественной литературы или СМИ), но и, если необходимо, с цельными текстами.

Корпус диалектных текстов НКРЯ предполагает включение любых диалектных текстов на русском языке, записанных как на территории исконного проживания русского населения (Европейская часть России), так и на территориях раннего заселения (Русский Север), позднего заселения (Сибирь, Дальний Восток, Дон, Нижнее Поволжье) и миграций (говоры старообрядцев/протестантов Латгалии, Азербайджана, Румынии, Австралии, Канады, Америки). Туда входят полевые записи, аудиорасшифровки, тексты из малотиражных сборников и изданий, войдут также видеорасшифровки и тексты из хрестоматий. Мы надеемся, что со временем этот корпус станет репрезентативным собранием диалектных текстов и будет

одним из самых посещаемых и востребованных пользователями. По материалам подкорпуса уже ведутся научные исследования [Пурицкая 2012].

1.2. Текст предоставляется пользователю в том виде, в котором он был первоначально записан, в том числе в фонетической транскрипции с сохранением ударений. Это особенно важно, т. к. диалектные тексты, особенно в транскрипционной записи, выпускаются в основном региональными университетами небольшими тиражами и малодоступны даже в своей «книжной» ипостаси.

1.3. Все тексты представлены с так называемой «снятой омонимией»: разметка осуществляется как на уровне грамматики (в том числе с указанием диалектных особенностей на уровне слова), так и на уровне метаразметки с указанием особенностей жанра, тематики и фонетических особенностей (в самых общих чертах).

1.4. Поиск может осуществляться по самым разным критериям.

На уровне **метаразметки** определяются:

1) география записи (с учетом области и района);

2) тематика;

3) жанр;

4) год записи;

5) период истории, время описываемых событий;

6) некоторые фонетические особенности;

7) сведения об информантах (пол, возраст, образование, конфессия и нек. др.).

На уровне отдельной лексемы:

8) грамматика (с учетом диалектных особенностей);

9) семантика (не везде);

10) возможен поиск лексемы, включая ее словообразовательные варианты и синонимы (не везде).

2. Среда «Рабочее место диалектолога»

Для ответов на запросы по грамматике, лексике, семантике, тематике и проч. текст должен быть первоначально размечен. Для этого создан особый системный продукт: среда «Рабочее место диалектолога» (РМ, автор Т. А. Архангельский). В РМ осуществляется разметка как на уровне слова, так и на уровне всего текста (метаразметка).

Метаразметка предлагается на трех уровнях:

1) Адрес-сопровождение.

2) Фонетическая.

3) Диалектная текстовая.

Каждый из уровней представляет собой развертку в таблицу.

2.1. Адрес-сопровождение

В этой таблице отмечаются:

• название текста;

• лицо, предоставившее текст (ФИО, информация для связи, научное звание, должность, место работы);

• кем производилась запись;

• место записи (область, район, населенный пункт);

• год записи;

• сведения об информанте (ФИО, возраст, год рождения, место рождения, образование, профессия/род занятий);

• где публиковался текст (если публиковался);

• место хранения записи.

2.2. Фонетическая метаразметка

2.2.1. Пока что специально отмечаются лишь немногие фонетические диалектные особенности:

• в области ударного вокализма: позиционные чередования гласных после мягких согласных на уровне «старого ятя» и <а>;

• в области безударного вокализма: оканье и аканье (включая указания на неполное оканье или диссимилятивное аканье);

• в области консонантизма: отмечаются <г> взрывной или фрикативный и цоканье.

2.2.2. Пользователю представляется текст в фонетической транскрипции в двух вариантах: Текст-1 (Т-1) соответствует первоначальной записи, т. е. подается так, как он был представлен держателями текстов (в фонетической транскрипции или в орфографизированном виде); Текст-2 (Т-2) — в «облегченной», унифицированной транскрипции. Т-1 и Т-2 подаются в шрифтах Unicode. Если первоначальный текст существовал в других шрифтах, он переводится в Unicode.

Т-2 создается автоматически из Т-1 с помощью специально подготовленной программы Детранскриптор-1 и практически не требует ручной доработки.

Соответствия Т-1 и Т-2:

Т-1	Т-2
в’и́т’ер, н’а́н’к’а, н’осу́ кон’, з’ем’е́л’к’е jéc’т’, пajду́, см’ijóца, ружйó	ви́тер, ня́нькя, нёсу́ конь, земе́льке йёсьть, пайду́, смиёца, ружьё

Некоторая сложность возникает при использовании в Т-1 знака *Ь*, обозначающего редуцированный гласный переднего ряда. Выход из положения видится следующий: в Т-1 текст остается с использованием знака *Ь* в качестве гласного. В Т-2 гласный [ь] заменяется на знак [и], а знак *Ь* используется как «мягкий знак» (так же, как в принятой орфографии).

Т-1: *в’ьч’ьре́лъ;* Т-2: *вичире́лъ*

Т-1 и Т-2 выступают как параллельные тексты.

2.2.3. В банке текстов, переданных для выставления на сайт Национального корпуса, многие тексты оказались поданы вовсе не в транскрипции, а в орфографии, поэтому в «фонетической» метаразметке существует помета «Орфографизированная ли запись?» Помета об «орфографизированной записи» будет постоянно присутствовать и на сайте, чтобы пользователь не пытался делать выводы о фонетических особенностях говора на основе текстов, которые первоначально представлены не в транскрипции.

Таким образом, если нет никаких отметок, то запись подана в транскрипции. Если стоит отметка *Орфографизированная запись,* значит, запись подана либо в орфографии, либо в орфографизированном виде — т. е. могут отмечаться какие-то отдельные диалектные особенности в области произношения на фоне общей установки на орфографию. Например, может быть отмечена утрата затвора у <ч>, так называемое «шоканье» (*девощка* = девочка = [дéвъщка]) при отсутствии отображения аканья в акающем говоре; могут быть отмечены какие-то особенности грамматики (*они стоя́ть* = они стоят = [ани стая́ть]).

Если запись первоначально была орфографической, то Т-1 и Т-2 совпадут.

Подача текстов в орфографизированном виде связана с достаточно устойчивой традицией публикации диалектных текстов. Иногда это делалось из-за недостатков полиграфической базы, сложности печати, но порой — для лучшего понимания таких текстов чита-

телями. Вместе с тем даже фольклористы, публикуя тексты в «орфографии», часто не могут устоять от соблазна показать особенно яркие черты фонетики (*Ванькя клюшник злой разлушник*…).

2.3. **Диалектная текстовая метаразметка** включает в себя 3 подуровня:

- жанр (тип) текста;
- место и время описываемых событий;
- тематика текста.

Каждый подуровень имеет свое деление.

2.3.1. **Жанр (тип) текста** делится на 3 категории: 1.1. Устные нефольклорные тексты; 1.2. Письменные нефольклорные тексты; 1.3. Фольклорные тексты. При выборе типа текста возникает возможность выбора конкретного жанра. Например, в **1.1. Устные нефольклорные тексты** включены следующие жанры: диалог; рассказ; пересказ (сна / разговора / фильма / телепередачи); ответы по вопроснику; спор. **1.2. Письменные нефольклорные тексты** включают жанры: автобиография; воспоминания/мемуары; личное письмо; поздравление; личный дневник / записная книжка; служебная / докладная записка; просьба, прошение; справка; заявление; характеристика; художественный текст, созданный носителем диалекта (в том числе поэтический).

Пока предпочтение при включении в корпус отдается устным нефольклорным текстам, хотя и там могут содержаться элементы фольклора, жанры которых отмечаются не в пределах встречаемости, а в пределах всего текста в целом (в устном рассказе естественно оказываются пословицы и поговорки, загадки, частушки и колыбельные песни и проч.). В то же время в банке диалектных текстов уже есть как письменные фольклорные тексты («песенники» или заговоры, записанные самими носителями), так и письменные нефольклорные (письма, мемуары, дневники, записные книжки, стихи и проч.).

2.3.2. **Место и время описываемых событий** во многом повторяет подобную таблицу в других корпусах НКРЯ и включает в себя следующие разделы:

- Доисторический период
- За пределами России
- Ирреальный (фантастический) мир
- Россия

Выделяются следующие исторические периоды, значимые как для русской деревни в целом, так и для наших информантов, их личной биографии (с XX века):

- до XVIII века
- XIX век
- 1900–1914
- 1914–1920
- 1920–1930-е (коллективизация)
- от коллективизации до войны 1941–1945
- война 1941–1945
- послевоенное восстановление 1945–1953
- 1950–1980-е
- постсоветский период

3. Особенности грамматической разметки

3.1.1. Грамматическая разметка диалектных текстов также осуществляется в РМ. Так как тексты в Диалектном подкорпусе поданы либо в транскрипции, либо в орфографии с сохранением некоторых фонетических и грамматических особенностей, возникла проблема их грамматического распознавания. Было принято решение к каждому тексту делать орфографический «подстрочник», создавать так наз. Текст-3 — орфографизированный вариант диалектного текста, по которому сможет работать стандартный грамматический анализатор, применяемый во всех корпусах НКРЯ. Для этой цели был создан **детранскриптор-2**, переводящий Т-2 (или Т-1) в Текст-3 (Т-3). Орфографический «подстрочник» требуется доводить до уровня орфографии вручную. Детранскриптор-2 основан на серии последовательных замен некоторых буквенных сочетаний, отличающих русскую фонетику от орфографии, отдельной строкой в него могут быть вписаны любые условия («*пробел — ана — пробел*» заменяется на *она*). Детранскриптор можно постоянно унифицировать, вписывая условия для наиболее часто повторяющихся замен. Унификация детранскриптора-2 увеличит степень «орфографизированности» текста, но не отменит ручной доводки. В РМ в Т-3 специальной программой осуществляется проверка стандартной орфографии, цветной меткой помечаются слова, не найденные в орфографическом словаре. Это или ошибки разметчика, которые легко исправить, или

диалектные слова (словоформы), отсутствующие в основном словаре.

Тексты 1, 2 и 3 выровнены по предложениям в формате XML, как в Параллельном корпусе НКРЯ.

Пример выровненного размеченного текста:

```
<para>
<se format="1">/ на / дифчёнкъ // а радить будить/ а рибяти́шки будут с красными въласами и ех замуш нихто́ ни вазьмёть // ну / тёть Шуръ Къралёвъ үъварит / ну куды ш паедим?</se>
<se format="2">/ <w><ana lex="на" gr="PART">
</ana><ana lex="на" gr="PR"><ana>на<w> / <w><ana
lex="девчонка" gr="S,anim,f,sg,nom,1decl">
</ana>дифчёнкъ</w> // <w><ana lex="а"
gr="INTJ"></ana><ana lex="а" gr="S,0">
</ana><ana lex="а" gr="CONJ"><ana>а<w> <w><ana
lex="родить"
gr="V,pf,inf,act,2conj"><ana><ana lex="родить"
gr="V,pf,tran,inf,act,2conj"></ana><ana
lex="родить" gr="V,ipf,tran,inf,act,2conj">
</ana>радить</w> <w><ana lex="будить"
gr="V,ipf,tran,inf,act,2conj"></ana>будить
</w>/ <w><ana lex="а" gr="INTJ"><ana>
<ana lex="а" gr="S,0"></ana><ana lex="а"
gr="CONJ"></ana>а</w> <w><ana lex="ребятишки"
gr="S,anim,pl,nom,pltantum"></ana>рибяти́шки
</w> <w><ana lex="быть" gr="V,pf,intr,in
dic,act,fut,3p,pl,1conj"></ana>будут</w>
<w><ana lex="с" gr="PR"></ana><ana lex="с"
gr="S,0"></ana>с</w> <w><ana lex="красное"
gr="S,inan,n,pl,ins"></ana><ana lex="красный"
gr="S,anim,m,pl,ins,adjdecl"></ana><ana
lex="красный" gr="A,pl,ins,plen"><ana>
красными</w> <w><ana lex="волосы"
gr="S,inan,pl,ins"></ana><ana lex="волос"
gr="S,inan,m,pl,ins,2decl"></ana>въласами</w>
```

```
<w><ana lex="и" gr="INTJ"></ana><ana lex="и"
gr="PART"></ana><ana lex="и" gr="S,0"><ana>
<ana lex="и" gr="CONJ"></ana>и</w> <w><ana
lex="их" gr="APRO"></ana><ana lex="они"
gr="SPRO,anim,pl,gen"></ana><ana lex="они"
gr="SPRO,anim,pl,acc"></ana>ех</w> <w><ana
lex="замуж" gr="ADV"></ana>замуш</w> <w><ana
lex="никто" gr="SPRO,anim,m,sg,nom"></
ana>нихто´</w> <w><ana lex="ни" gr="PART"></
ana><ana lex="ни" gr="CONJ"></ana>ни
</w> <w><ana lex="взять" gr="V,pf,tran,indic
,act,fut,3p,sg,1conj"></ana>вазьмёть</w> //
<w><ana lex="ну" gr="INTJ"></ana><ana lex="ну"
gr="PART"></ana>ну</w> / <w><ana lex="тетя"
gr="S,anim,f,pl,gen,1decl"></ana><ana lex="тетя"
gr="S,anim,f,pl,acc,1decl"></ana>тёть</w>
<w><ana lex="шура" gr="S,persn,anim,pl,gen,1de
cl"></ana><ana lex="шура" gr="S,persn,anim,pl,
acc,1decl"></ana>Шуръ</w> <w><ana lex="королев"
gr="S,famn,anim,m,sg,gen,famdeclm"></ana><ana
lex="королев" gr="S,famn,anim,m,sg,acc,famde
clm"></ana><ana lex="королева" gr="S,famn,an
im,f,sg,nom,1decl"></ana><ana lex="королева"
gr="S,anim,f,sg,nom,1decl"></ana><ana
lex="королев" gr="S,inan,m,sg,gen,famdeclm">
</ana>Къралёвъ</w> <w><ana lex="говорить" gr="V,
ipf,tran,indic,act,praes,3p,sg,2conj">
</ana>ӳъварит</w> / <w><ana lex="ну"
gr="INTJ"></ana><ana lex="ну" gr="PART">
</ana>ну</w> <w><ana lex="куда" gr="ADVPRO">
</ana>куды</w> <w><ana lex="ш" gr="S,0">
</ana>ш</w> <w><ana lex="поехать" gr="V,pf
,intr,indic,act,fut,1p,pl,1conj"></ana><ana
lex="поехать" gr="V,pf,intr,imper,act,1p,pl,
1conj"></ana>паедим</w>?</se>
<se format="3">/ <w><ana lex="она"
gr="SPRO,anim,f,sg,nom"></ana>на</w> / <w><ana
lex="девчонка" gr="S,anim,f,sg,nom,1decl">
```

```
</ana>девчонка</w> // <w><ana lex="а"
gr="CONJ"></ana>а</w> <w><ana lex="родить" gr
="V,pf,tran,indic,act,fut,3p,sg,2conj,dial
flex"></ana>родит</w> <w><ana lex="быть" gr
="V,pf,intr,indic,act,fut,3p,sg,1conj,di
alflex"></ana>будет</w>/ <w><ana lex="а"
gr="CONJ"></ana>а</w> <w><ana lex="ребятишки"
gr="S,anim,pl,nom,pltantum"></ana>ребятишки
</w> <w><ana lex="быть" gr="V,pf,intr,indic,act,
fut,3p,pl,1conj"></ana>будут</w> <w><ana lex="с"
gr="PR"></ana>с</w> <w><ana lex="красный"
gr="A,pl,ins,plen" lex_ref="рыжий"></
ana>красными</w> <w><ana lex="волосы"
gr="S,inan,pl,ins"></ana>волосами
</w> <w><ana lex="и" gr="CONJ"></ana>и</w>
<w><ana lex="они" gr="SPRO,anim,pl,acc,dialst
em,dialform"></ana>их</w> <w><ana lex="замуж"
gr="ADV"></ana>замуж</w> <w><ana lex="никто"
gr="SPRO,anim,m,sg,nom"></ana>никто</
w> <w><ana lex="не" gr="PART"></ana>не</w>
<w><ana lex="взять" gr="V,pf,tran,indic,act,f
ut,3p,sg,1conj,dialflex"></ana>возьмёт</w> //
<w><ana lex="ну" gr="INTJ"></ana><ana lex="ну"
gr="PART"></ana>ну</w> / <w><ana lex="тетя"
gr="S,anim,f,sg,voc,1decl"></ana>тёть</w>
<w><ana lex="шура" gr="S,persn,anim,sg,nom,1
decl"></ana>Шура</w> <w><ana lex="королева"
gr="S,famn,anim,f,sg,nom"></ana>Королёва</
w> <w><ana lex="говорить" gr="V,ipf,tran,in
dic,act,praes,3p,sg,2conj"></ana>говорит</
w> / <w><ana lex="ну" gr="PART"></ana>ну</w>
<w><ana lex="куды" gr="ADVPRO" lex_ref="куда"></
ana>куда</w> <w><ana lex="ж" gr="PART"></ana>ж</
w> <w><ana lex="поехать" gr="V,pf,intr,indic,act
,fut,1p,pl,1conj"></ana>поедем</w>?</se>
</para>
```

3.1.2. После создания орфографического «подстрочника» осу-

ществляется **грамматическая разметка** Текста-3 стандартным грамматическим анализатором. Дальше текст обрабатывается вручную:

1) снимается грамматическая **омонимия** (как в Основном корпусе),

2) отмечаются **диалектные грамматические особенности** тех лексем, где эти особенности встретились.

Таким образом, поверх стандартной грамматической разметки в диалектных текстах предусмотрена возможность отмечать диалектные грамматические особенности лексемы. Для этого в РМ внедрены грамматические таблицы по каждой из 5 изменяемых частей речи (глагол, существительное, прилагательное, местоимение, числительное).

3.2. Так как грамматические особенности русских говоров достаточно хорошо изучены, возникла идея использовать таблицы с реальными диалектными аффиксами, формами и проч. Однако при работе с конкретными текстами каждый раз выяснялось, что списки введенных аффиксов оказывались неполными. Чтобы не множить сущности, было принято решение вернуться к обобщенным указаниям на диалектные черты отдельных грамматических категорий: рода, числа, падежа, склонения, вида, переходности, возвратности и проч.

Для каждой изменяемой части речи предусмотрена возможность указывать диалектные особенности на многих уровнях. Приведем некоторые примеры.

3.3. Диалектные особенности у **существительных** могут быть связаны с категориями рода, числа, типа склонения, одушевленности; они могут иметь иные нежели в ЛЯ, падежные окончания.

3.3.1. Категория **рода**: *берлог* (= берлога), *пуза* (= пузо), *литра* (= литр), *девчонко* (= девчонка), *яблок* (= яблоко), *зайко*, *дедушко* и т. п. При указании на диалектные особенности рода принимаются в расчет два критерия: морфологический — по типу склонения, синтаксический — по согласованию с адъективами или с глаголами в прош. вр. (*плачь у ней была такая*), в том числе в случаях рассогласования, т. е. когда синтаксический и морфологический критерии не совпадают (*така хороша девчёшко, парна́я мълако*). Если особенности рода сопровождаются особенностями в типе склонения, то отмечается одновременно и «диалектный род», и «диалектное склонение» (*такую колечку*).

3.3.2. Изменение типа **склонения** (или особенности склонения) могут быть связаны как с распадением категории среднего рода — при ориентации сущ. среднего рода на женский (*мою колечку, в девятую царству*) или мужской (*вкусный яблок*), так и с ориентацией 3 скл. на 1-е: *на пече*, *ночей*, *осеньей*. В этом случае как диалектные особенности отмечаются «диалектное склонение» и «флексия» (dialflex).

3.3.3. **Диалектная флексия (dialflex)** у имен существительных может указывать на разные вещи. Так, у существительных 1 скл. в части говоров могут совпадать окончания Д=П и Р: *к козы́, в реки́, в Москвы́, на войны́*. Есть говоры, в которых Р совпадает с Д=П (окончание <*е*> из «ятя»): *у жене́*, *без сестре́*, в безударной позиции в акающих говорах: *у баби* (<*у бабе*>), *блис школи* (<*близ школе*>). У существительных 2 скл. особенности окончаний могут встречаться в Р. (говоры с расширенной сферой окончания *-у*: *без чаю*, *без мужику*, *из городу*) или в П. (говоры, в которых окончания П. *-е* и *-у* распределены иначе, чем в ЛЯ: *в лесе*, *в глазе*, *в шкафе*, *на береге* или *на коню*; в П. встречается грамматикализованное окончание *-и*: *на кони*). Особые окончания могут встретиться у так называемых разносклоняемых существительных (существительных с нерегулярным склонением): слова на *-мя*, *мать*, *дочь*, *свекровь*, *день*, *путь* и некоторых других, а также во множественном числе (*без бабов*, *месяцей*, *лопаткима*). Если исследователя интересует диалектные грамматические особенности определенной лексемы или определенного падежа, достаточно задать в поиске параметры падежа, числа, типа склонения + диалектные особенности.

3.4. Диалектные особенности **глагола** могут быть связаны с категориями вида, переходности, возвратности, наклонения, типов спряжения; глаголы в русских народных говорах могут иметь специфические окончания в личных формах, особые основы, иные формы инфинитива, императива, причастий, деепричастий, употребляться в конструкциях, связанных с иным временем (перфект, плюсквамперфект). Во всплывающей таблице «диалектные особенности» для глагола отмечаются как «диалектное»: **основа, суффикс, флексия, форма, вид, переходность, возвратность, время**.

3.4.1. Иная, чем в ЛЯ, **основа** глагола (**dialstem**) может быть связана с принадлежностью глагола к регулярному или нерегулярному классу, т. е. с иным соотношением основ настоящего и прошедшего

времени (инфинитива): наст. время может ориентироваться на прошедшее (*торговать ~ торгует*, *ездил ~ ездию*, *мыл ~ мыет*); прошедшее — ориентироваться на настоящее (*жму ~ жмал*, *трёт ~ трать*, *умрёт ~ умрал*). Это явление может считаться «архаическим» или «новаторским» (*гоню ~ гонить* / *гнать*, *мяукать* / *мяучить ~ мяукает* / *мяучит*, *чихать ~ чишу* / *чихаю*, *полоскать ~ полощу* / *полоскаю*). Диалектная основа может быть связана с процессом выравнивания (унификации) основ личных форм глагола в настоящем — простом буд. времени: *просю*, *любю*, *бегёшь*, *бежат*, *пекёт*, *пекот* (= *печет*).

3.4.2. Диалектный **суффикс (dialsfx)** отмечается при особом **императиве**: *поста́ви* (= поставь), *удари* (= ударь), *посоль* (= посоли) — т. е. учитывается и нулевой суффикс. Если форма императива возникает от иной, нежели в ЛЯ, основы (*доедь*, *ехай*, *пахай*), то отмечается не dialsfx, а dialform (диалектная форма). Таким образом, исследователю для того, чтобы получить сведения об особом императиве, достаточно задать в поиске императив + диалектные особенности.

3.4.3. Именно суффиксом отличаются в разных говорах формы **инфинитива**: «архаический» *-ти* безударный (*гуляти*, *смотрети*); «новаторский» *-ть* с утратой гласного (*потрясть* = потрясти, *подместь* = подмести); «новаторский» *-чи* в глаголах с основой на заднеязычный (*пекчи* = печь, *стерегчи* = стеречь). В этих случаях отмечается диалектный суффикс (dialsfx). В случаях, когда встречается унификация суффикса инфинитива (*идти́ть*, *прийти́ть*, *пойди́ть*), отмечается не dialsfx, а dialform (диалектная форма). Если особая форма инфинитива связана с переходом / непереходом глагола в регулярный класс, т. е. с изменением соотношения основ непрошедшего / прошедшего времени: *жмать* = жать (руку), *жнить* = жать (траву), *бости* = бодать, *ноять* = ныть, *бечь* = бежать, то одновременно отмечается dialform (диалектная форма) и dialsteam (диалектная основа).

3.4.4. Из четырех видов **причастий** в говорах, как правило, встречаются лишь страдательные причастия прошедшего времени (чаще в краткой форме), которые образуются с помощью тех же суффиксов, что и в ЛЯ (*-енн-*, *-нн-*, *-т-*), однако в разных лексемах эти суффиксы могут быть распределены иначе, чем в ЛЯ (*убратый* = убранный, *датый* = данный, *расколонный* = расколотый). В этих

случаях отмечается диалектный суффикс (dialsfx). Если причастия образованы от иных основ (*даденный*), в таком случае отмечается dialsteam (диалектная основа) и/или dialform (диалектная форма).

3.4.5. Как правило, в говорах встречается иной набор суффиксов **деепричастий**: *-вши*, *-лши*, *-мши*, *-тши*, *-чи* (*выпивши*, *выпимши*, *выпилши*, *пришёдци*). В этом случае отмечаются диалектный суффикс (dialsfx) и диалектная форма (dialform). Часто такие деепричастия используются в функции предиката и имеют значение перфекта (*Он пришоццы. Корыто было пропалшы*) — тогда кроме диалектного суффикса и диалектной форма должно быть отмечено диалектное время (dialtense). Таким образом, если пользователь хочет найти деепричастия в функции перфекта, он задает запрос на деепричастия + диалектное время.

3.4.6. В графу **диалектная возвратность (dialrefl)** попали самые разные случаи, характеризующие особенности возвратных глаголов. Так, в говорах на месте возвратного глагола (в ЛЯ) может оказаться невозвратный (*смеять*), а на месте возвратного — невозвратный (*дрожаться*, *дежуриться*). Возможно присоединение постфикса *-ся* не только к основе на согласный, но и на гласный (архаические формы: *умывалася*, *посмотрелися*). В постфиксе может присутствовать как мягкий *с*, так и твердый (*умывалас*, *умывалса*). Гласный постфикса может быть не только *а/я*, но и *и/ы*, *е/э*, *ё/о* (*умывался*, *умывалсе*, *умывалси*, *умывалсё*; *умывалса*, *умывалсэ*, *умывалсы*, *умывалсо*). Если пользователь хочет найти невозвратные глаголы на месте возвратных, он задает запрос по диалектной возвратности, при этом отмечает отсутствие *-ся* в конце слова. Ассимилятивные процессы на стыке основы и постфикса (*смеёсся*, *смеёша*) являются фонетическими и в разметке особенностей грамматики не учитываются.

3.4.7. **Диалектные окончания (dialflex)** характеризуют глаголы 3 л. ед. и мн. ч., сюда также попадают самые разные случаи: ударные окончания без перехода *е* > *о* (*идóшь*), конечное *-ть* в 3 л. ед. и мн. ч. (*растёть*, *растуть*), формы без *-т* (*идё*), общее спряжение (*смотрют*), так называемое 3-е спряжение (*играт*) и проч. Думается, что специалисту в этом достаточно легко разобраться, особенно при возможности учитывать географические фильтры. Например, если исследователь хочет найти особенности спряжения и ищет «общее спряжение» (любют), он задает в поиске «наст. вр. 3 л. мн. ч.

+ dialflex+ окончание *-ут*, *-ют*». Если его интересует 3-е спряжение, характерное для северных и среднерусских говоров (*он гулят*), лучше включать географические фильтры и указывать 2 и 3 л. ед. и 1, 2 и 3 л. мн. ч. наст. вр. + диал. флексия.

3.4.8. Кнопка **форма** зарезервирована для случаев, когда в ЛЯ нет соответствий или трудно / невозможно разделить основу и флексию: типа *jo* или *jy* как формы местоимения *её* (= она), 3 л. ж. р. В.-Р. ед., или *ихной*, *ихний*, *ихой* (= их), *тэй* / *тый* парень (= тот), *оне*, *оны* (= они), сравн. ст. прилаг. / наречий — *первеющий* и проч.

4. Развитие корпусной диалектологии на текущем этапе

4.1. Создание диалектных корпусов — дело во многом новое, особенно это касается общедоступных корпусов. Назовем известные нам продукты, доступные через интернет, в которых отражены сведения о русских говорах или представлены диалектные тексты:

• сайт «Школьный диалектологический атлас "Язык русской деревни"» (включающий карты и подробные комментарии к ним): http://gramota.ru/book/village (ИРЯ РАН, Москва).

• Фонетика русских диалектов: http://dialect.philol.msu.ru/index.php (МГУ имени М. В. Ломоносова, Москва).

• «Электронная библиотека русских народных говоров» (собраны материалы экспедиций по различным говорам Европейской части России): http://dialekt.rx5.ru/index.html (Казанский [Приволжский] федеральный университет, Казань).

• «Лингвогеографическая система «Диалект»»: http://lgw2.udsu.ru:9001/(Удмуртский университет, Ижевск).

• «Региональная этнолингвистика» (материалы по Кубанским говорам): http://www.ethnolex.ru/(Кубанский государственный университет, Славянск-на-Кубани).

• Тексты из Шатурского р-на Московской обл. и Харовского р-на Вологодской обл. в рамках проекта «Электронные базы данных по русским народным говорам»: http://starling.rinet.ru/cgi-bin/main.cgi?root=ruscorpora&encoding=utf-rus (ИРЯ РАН, Москва, авторы С. А. Крылов и А. В. Тер-Аванесова)

• тексты и база по говору д. Пушкино Устьянского р-на Архангельской обл. (Ustja River Basin Corpus Query interface): http://www.

slavist.de/Pushkino/login.php (Национальный исследовательский университет «Высшая школа экономики», Москва, и институт славистики Бернского университета, Берн, Швейцария; Научно-учебная группа «Корпусное исследование вариативности в речи носителей говора Устьянского района Архангельской области», руководители М. А. Даниэль и Р. фон Вальденфельс).

4.2. Во многих научных центрах (как правило, организованных на базе университетов) ведется активная работа по созданию и совершенствованию диалектных корпусов (пока еще без выхода в Интернет). «Саратовский диалектологический корпус» (http://sarteorlingv.narod.ru/dialekt/kru4kova-goldin.html) создан по материалам трех русских говоров — двух южных и одного северного (Центр изучения народно-речевой культуры Саратовского государственного университета им. Н. Г. Чернышевского, руководители проф. В. Е. Гольдин и проф. О. Ю. Крючкова; http://www.sgu.ru/structure/philological/narrech). Материал более чем из ста говоров Архангельской области содержится в Корпусе «Электронная картотека "Архангельского областного словаря"» (МГУ имени М. В. Ломоносова, Москва) — как видно из названия, этот Корпус имеет жесткую лексикографическую направленность. Вышло несколько выпусков «Тамбовской фонохрестоматии» (Тамбовский университет), в которой расшифрованные тексты даны в сопровождении аудиоматериалов, имеется карта области, разделенная на районы, включена система поиска, т. е. по сути эта фонохрестоматия является корпусом. Ведется активная работа по созданию корпусов по русским народным говорам в Томске, Тюмени, Челябинске, Смоленске и других научных центрах [Русская устная речь 2011; Юрина 2011].

4.3. Сбор диалектных текстов для выставления их на сайте НКРЯ не должен и не может препятствовать созданию диалектных корпусов, так сказать, «местного масштаба». С одной стороны, мы хотели бы задать некоторый стандарт подачи диалектного текста — речь идет прежде всего о текстах, которые будут специально расшифровываться для Корпуса диалектных текстов НКРЯ. С другой стороны, нельзя игнорировать уже имеющиеся образцы записи народных говоров, которые оказались далеки от нашего «стандарта». Главное отличие Диалектного подкорпуса НКРЯ от других русских диалектных корпусов видится нам в установке на **сплошную грамматическую разметку** текстов, что соответствует общей стратегии всего Нацио-

нального корпуса в целом, тогда как региональные корпуса, скорее всего, будут основываться на жанрово-тематической разметке.

В Корпусе диалектных текстов НКРЯ предполагается давать ссылки на существующие корпуса, имеющие сайты в интернете.

4.4. Если грамматическая разметка дается в Диалектном подкорпусе для каждого слова, то семантика слова указывается лишь в том случае, если тексты сопровождаются словарем или есть ссылка на имеющиеся словари. В некоторых случаях семантика уточняется за счет указанных синонимов (в поле lexref). Конечно, этого недостаточно даже для лингвиста-недиалектолога, не говоря уже о пользователях-нелингвистах, не знакомых с говором. Выход видится в том, чтобы в Диалектном подкорпусе ввести ссылки на диалектные словари, например на «Словарь русских народных говоров» http://iling.spb.ru/vocabula/srng/srng.html (ИЛИ РАН, Санкт-Петербург); «Архангельский областной словарь»: http://www.philol.msu.ru/~dialectology/dictionary/ (МГУ имени М. В. Ломоносова, Москва). К сожалению, подавляющее количество диалектных словарей не имеет в интернете официальных адресов (сайтов).

4.5. В Диалектном подкорпусе возможно включать любого рода комментарии к текстам: фонетические, грамматические, семантические, экстралингвистические (предоставить этнографическую, этнокультурную информацию), возможно сопровождение текстов фотоматериалами. Некоторые тексты предполагается сопровождать звуко- и видеорядом (в случае, когда тексты явились расшифровками аудио- и видеозаписей). В последующем планируется создание серии интерактивных карт с указанием точки на карте, соответствующей данному пункту с демонстрацией запрашиваемого явления на карте в масштабе области / Европейской части РФ / всей России.

4.6. Свободное предоставление в интернете текстов, записанных в русских народных говорах, их грамматическая, семантическая и метатекстовая характеристика позволит специалистам-диалектологам, другим лингвистам и нелингвистам, филологам, историкам, культурологам, этнографам и всем, кто интересуется народным русским словом, обращаться к подкорпусу в самых разных целях: примеры из текстов и сами тексты могут выступать в качестве справочного материала, материала для научной и педагогической работы, демонстрации этнографических, этнокультурных традиций, особенностей русского менталитета.

Литература

Качинская И. Б. Корпус Диалектных Текстов в Национальном корпусе русского языка: состояние и перспективы // Лексический атлас русских народных говоров (Материалы и исследования). 2009. СПб. С. 57–68. (http://www.philol.msu.ru/~ruslang/pdfs/kachinskaya.i.b/19.pdf от 01.03.2014)

Качинская И. Б. Диалектный Подкорпус НКРЯ. Новый стандарт подачи. Новое рабочее место // Русская устная речь. Материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения». Межвузовское совещание «Проблемы создания и использования диалектных корпусов». Саратов, СГУ, 15–17 ноября 2010 г.. Саратов, СГУ, 2011. С. 245–255.

Летучий А. Б. Корпус диалектных текстов: задачи и проблемы // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005. С. 215-232.

Летучий А. Б. Диалектный корпус: состав и особенности разметки // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 114–128.

Сичинава Д. В., Качинская И. Б. Корпус диалектных тестов в национальном корпусе русского языка: сегодняшнее состояние и перспективы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). М.: Изд-во РГГУ, 2014. С. 593–600.

Пурицкая Е. В. Диалектный подкорпус национального корпуса русского языка как источник изучения лексической динамики диалекта // Северно-русские говоры. Вып. 12: межвуз. сб. / отв. ред. А.С. Герд. – СПб.: Изд-во С.-Петерб. ун-та, 2012. С. 14–22.

Русская устная речь. Материалы международной научной конференции «Баранниковские чтения. Устная речь: русская диалектная и разговорно-просторечная культура общения». Межвузовское совещание «Проблемы создания и использования диалектных корпусов». Саратов, СГУ, 15–17 ноября 2010 г.. Саратов, СГУ, 2011.

Юрина Е. А. Томский диалектный корпус: в начале пути // Вестник Томского гос. ун-та. Филология. № 2, 2011. С. 58-63 (http://cyberleninka.ru/article/n/tomskiy-dialektnyy-korpus-v-nachale-puti от 01.09.2014).

[1]Irina B. Kachinskaya, [2]Dmitri V. Sitchinava
[1]Moscow State University
[2]Vinogradov Russian Language Institute
of the Russian Academy of Sciences
[1,2](Russia, Moscow)
[1]kacza@yandex.ru, [2]mitrius@gmail.com

TODAY'S STATE OF THE DIALECTAL SUBCORPUS[1]

The paper presents the current state of the Corpus of dialectal texts which is a subcorpus of the RNC. Between 2005 and 2009, a pilot dialectal corpus was marked up. Since then a considerable work has been done to improve the markup of the dialectal texts on different levels (metatextual information on the time and place of collection, genre and other properties of the texts; phonetics, morphology, semantics, elements of syntax). Programming means for corpus markup, including the graphic user interface, are created. Texts are collected in different regions of Russia; previously collected texts, both published and taken from archives, are digitalized and marked up. A technology converting the transcribed texts into an orthographized form is applied. An orthographized version is then marked up semi-automatically. It is also possible to include the texts that exist only in orthographized version. The morphological annotation includes tags on dialectal properties on different levels (stem, inflection, and derivation). Multimodal information (video and audio) can be also included. The paper offers information on other dialectal corpora that collaborate with the RNC team.

Key words: Dialectal corpus, dialectology, morphological markup, metatextual markup, user interface

References

Kachinskaya I. B. [Corpus of dialectal texts within the RNC: current state and perspectives]. *Leksicheskii atlas russkikh narodnykh govorov (Materialy i issledovaniya)* [Lexical atlas of Russian rural dialects: Materials and studies]. St. Petersburg, 2009, pp. 57–68. (Available

[1] The work on the Dialectal Corpus is supported by the RFH grant № 14-04-12012; name of the project "Corpus of dialectal texts within the RNC", head of the project Dmitri V. Sitchinava.

at http://www.philol.msu.ru/~ruslang/pdfs/kachinskaya.i.b/19.pdf Accessed on 01.03.2014). (In Russ.)

Kachinskaya I. B. [Corpus of dialectal texts within the RNC: new standards and new GUI]. *Russkaya ustnaya rech'. Materialy mezhdunarodnoi nauchnoi konferentsii «Barannikovskie chteniya. Ustnaya rech': russkaya dialektnaya i razgovorno-prostorechnaya kul'tura obshcheniya». Saratov, SGU, 15–17 noyabrya 2010 g*. [Russian oral speech. Proc. of an International Barannikov conference "Oral speech: Russian dialectal and colloquial culture of speech". Saratov State University, November 15–17 2010]. Saratov, Saratov University, 2011, pp. 245-255. (In Russ.)

Letuchii A. B. [Corpus of dialectal texts: tasks and perspectives]. *Natsional'nyi korpus russkogo yazyka: 2003–2005* [Russian National Corpus 2003–2005. Results and perspectives]. Moscow, Indrik Publ., 2005, pp. 215–232. (In Russ.)

Letuchii A. B. [Dialectal corpus: composition and properties of tagging]. *Natsional'nyi korpus russkogo yazyka: 2006-2008. Novye rezul'taty i perspektivy* [Russian National Corpus 2006–2008: New Results and perspectives]. St. Petersburg, Nestor-Istoria Publ., 2009, pp. 114–128. (In Russ.)

Sichinava D. V., Kachinskaya I. B. [Corpus of dialectal texts within the RNC: current state and perspectives]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog» (Bekasovo, 4–8 iyunya 2014 g.). Vyp. 13 (20)*. [Computer linguistics and intellectual technologies: Proc. of the annual international Dialogue conference, Bekasovo, June 4–6, 2014, Vol. 13 (20)]. Moscow, RSUH Publ., 2014, pp. 593–600. (In Russ.)

Puritskaya E. V. [Dialectal subcorpus of the RNC as a source for studying the dynamics of a dialectal lexicon]. *Severno-russkie govory. Vyp. 12* [Northern Russian dialects: Volume 12]. St. Petersburg, Saint Petersburg university, 2012, pp. 14–22. (In Russ.)

Russkaya ustnaya rech'. Materialy mezhdunarodnoi nauchnoi konferentsii «Barannikovskie chteniya. Ustnaya rech': russkaya dialektnaya i razgovorno-prostorechnaya kul'tura obshcheniya». Saratov, SGU, 15–17 noyabrya 2010 g. [Russian oral speech. Proc. of an International Barannikov conference "Oral speech: Russian dialectal and colloquial culture of speech". Saratov State University, November 15–17, 2010] Saratov, Saratov University, 2011.

Yurina E. A. [Tomsk dialectal corpus: beginnings]. *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya*, 2011, no. 2, pp. 58-63 (Available at http://cyberleninka.ru/article/n/tomskiy-dialektnyy-korpus-v-nachale-puti Accessed on 01.09.2014). (In Russ.)

С. О. Савчук
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
savsvetlana@mail.ru

КОРПУС РЕГИОНАЛЬНЫХ ГАЗЕТ РОССИИ И ЗАРУБЕЖЬЯ[1]

В статье описывается опыт создания нового модуля в составе НКРЯ — корпуса региональной и зарубежной прессы. Главную задачу этого корпуса мы видим в том, чтобы предоставить исследователям инструмент для изучения региональной вариативности русского языка. В состав корпуса в настоящий момент входят: 1) корпус газет гродненского региона (7 изданий, 1,9 млн словоупотреблений), подготовленный совместно со специалистами Гродненского университета; 2) подкорпус российских региональных газет (13 изданий, около 2 млн словоупотреблений); 3) корпус региональных российских газет 1990–2000-х гг. (40 газет, 2,6 млн словоупотреблений); 4) региональные выпуски газеты «Комсомольская правда» (6,5 млн словоупотреблений). Тексты корпуса снабжены морфологической, семантической аннотацией и подробной метаразметкой. В статье описываются принципы отбора текстов, стандарт и программные средства обработки текстов, организация поиска по корпусу, перспективы развития состава и корпусного инструментария.

Ключевые слова: Национальный корпус русского языка, газетный корпус, корпус региональной прессы, состав и структура, аннотация текстов.

[1] Работа выполнена при поддержке РГНФ (гранты № 13-24-01004 и № 15-04-12018) и РФФИ (грант № 15-06-04334).

I. Введение

Корпус региональной и зарубежной прессы представляет собой самостоятельный модуль в составе Национального корпуса русского языка. Он открыт для свободного доступа с октября 2014 года по адресу http://www.ruscorpora.ru/search-regional.html. Найти его в составе НКРЯ несложно: он объединен с большим корпусом СМИ[1] в один газетный блок, вход осуществляется с боковой панели главной страницы Корпуса.

Главную задачу этого корпуса мы видим в том, чтобы предоставить исследователям инструмент для изучения региональной вариативности русского языка. Вероятность возникновения вариантов узуса повышается в том случае, если язык функционирует в иноязычном окружении, в условиях языковых контактов, выступает как язык межнационального или международного общения. Для таких языков, как английский, французский, немецкий, установлен состав национальных языковых вариантов (в частности британский, американский, канадский, австралийский, новозеландский варианты английского языка; варианты французского языка Франции, Бельгии, Швейцарии и Канады; немецкий, австрийский, швейцарский варианты немецкого языка), описаны особенности их словаря и грамматики. Материал о национальных вариантах включается в состав учебников и пособий для изучающих эти языки, создаются корпуса вариантов того или иного языка, которые используются не только в качестве исследовательских ресурсов, но имеют большую популярность в педагогической практике.

Что касается русского языка, то изучение его как средства межнационального общения на большой территории, населенной разными народами, было начато еще в СССР. Значительный материал о взаимодействии русского и родного языка был накоплен специалистами в области преподавания русского языка как иностранного. В последние десятилетия в изменившихся геополитических условиях для России стала актуальной тема вариативности русского языка в ближнем и дальнем зарубежье (в Белоруссии, Казахстане, Киргизии, Молдавии, Узбекистане, Америке, Финляндии, Австралии, Израиле) и в национальных республиках Российской

[1] Принципы формирования, состав и структура и первый опыт использования Корпуса СМИ 2000-х годов (газетного корпуса) см. в [Савчук 2011].

Федерации. Приводит ли новая языковая ситуация в государствах постсоветского пространства к усилению процессов дивергенции региональных вариантов и кодифицированного языка метрополии, к формированию национальных вариантов русского языка за пределами России? Для того чтобы ответить на эти вопросы, нужна репрезентативная экспериментальная база и продуманная методика анализа. Учитывая то обстоятельство, что изменения в русском языковом узусе того или иного региона могут накапливаться постепенно и быть незаметными невооруженным глазом, эффективным в данном случае может оказаться использование лингвостатистических методов, применяемых на большом массиве текстов. Все это делает создание такого массива настоятельно необходимым и актуальным.

Работы по формированию корпуса региональных СМИ начались в рамках международного проекта, выполняемого коллективами исследователей ИРЯ им. В. В. Виноградова РАН и Гродненского государственного университета им. Я. Купалы[1]. Целью проекта было изучение лексико-семантических и культурно-специфических особенностей русской речи на территории Гродненской области Республики Беларусь[2] на материале областных и районных газет.

Были согласованы общие принципы и методика организации корпусов, программное обеспечение и технология подготовки и описания текстов, используемые в НКРЯ, были модернизированы и настроены под новые задачи. В течение 2013 года белорусской стороной был подготовлен корпус русскоязычных газет Гродненского региона, а российской стороной — корпус российских региональных СМИ 2010-х годов. Пилотные версии этих корпусов были размещены на сайте ruscorpora.ru и стали ядром создаваемого корпуса региональных СМИ. Рассмотрим его состав более подробно.

---

[1] При поддержке РГНФ (грант № 13-24-01004, руководитель А.Я. Шайкевич) и БРФФИ (договор № Г13Р-050, руководитель Л.В. Рычкова).

[2] Ситуация в Гродненском регионе представляет особый интерес в том плане, что, во-первых, русский язык функционирует в нем, как и на всей территории Белоруссии, в качестве государственного, то есть обслуживает все сферы речевой коммуникации, а во-вторых, существует в контакте не только с белорусским, но и с польским и литовским языками.

Состав корпуса

Русскоязычные газеты Гродненского региона

• «Вечерний Гродно» (http://www.vgr.by), городская вечерняя газета г. Гродно, издается с 2000 г., выходит 1 раз в неделю, тираж около 23000–25000 экз., в корпус вошло 696 статей за 2012 г.

• «Перспектива» (http://www.perspektiva-info.by), региональная газета, г. Гродно и Гродненский район, издается с 1939 г. (первоначально — «Свободная Беларусь», затем — «За свободную Беларусь», «Сельская новь» и «Сельская навіна»), выходит 2 раза в неделю, тираж около 11000 экз., в корпусе 750 статей за 2012 г.

• «Островецкая правда»/«Астравецкая праўда» (http://www.ostrovets.by) районная газета г. п. Островец, издается с 1941 г. (первоначально — «Бальшавіцкі арганізатар»), выходит 2 раза в неделю, в корпусе 915 статей за 2012 г.

• «Берестовицкая газета»/«Бераставіцкая газета» (http://www.beresta.by), районная газета г. п. Б. Берестовица, издается с 1944 г. («Знамя Советов», с 1966 г. «За камунізм», с 1991 г. «Бераставіцкая газета»), выходит 2 раза в неделю, тираж около 3000 экземпляров, в корпусе 1235 статей за 2012 г.

• «Ивьевский край»/«Іўеўскі край» (http://ivyenews.by), общественно-политическая районная газета Ивьевского района, г. Ивье, выходит 2 раза в неделю, тираж более 5600 экз., в корпусе 198 статей за 2012 г.

• «Праца» (http://www.praca.zelva.by), районная газета г. п. Зельва, под таким названием издается с 1967 года, выходит 2 раза в неделю, тираж более 4000 экз., в корпусе 926 статей за 2012 г.

• «Свислочская газета»/«Свіслацкая газета» (http://www.svisloch.info), районная газета г. Свислочь, издается с 1943 г., выходит 2 раза в неделю, тираж более 4000 экз., в корпусе 903 статьи за 2012 г.

В таблице 1 дается характеристика газет по типу аудитории и приводится объем текстов каждого издания, выраженный в количестве словоупотреблений.

Таблица 1

Название	**Тип газеты**	**Год издания**	**Количество с/у**
«Вечерний Гродно»	региональная вечерняя	2012	182 504
«Перспектива»	региональная	2012	184 970
«Бераставіцкая газета»	районная	2012	414 545
«Іўеўскі край»	районная	2012	133 580
«Астравецкая праўда»	районная	2012	467 517
«Праца»	районная	2012	352 597
«Свіслацкая газета»	районная	2012	330 007
ИТОГО			2 074 720

Отбор источников производился на основе анализа СМИ Гродненского региона, который был проведен белорусскими коллегами [см. Рычкова и др. 2012]. С учетом опыта формирования газетных корпусов в составе НКРЯ были разработаны критерии отбора. Их можно разделить на технические и содержательные. К техническим критериям относятся наличие сайта газеты и качество его технической поддержки. Техническая поддержка обеспечивает существование архива газетных номеров, его регулярные пополнения, логическую организацию сайта, аннотацию текстов, формат их представления. От этих показателей зависит скорость и качество обработки текстов соответствующих газет для размещения в корпусе. Поскольку время, отпущенное на формирование корпуса, ограничено рамками договора по проекту, скорость и качество подготовки корпуса как экспериментальной базы исследования приобретает решающее значение.

К содержательным критериям отбора относится разнообразие источников, обеспечивающее представительность корпуса. Источники должны варьироваться по географии, по типу газет, типу аудитории, политической принадлежности и др. В рамках конкретного проекта важно найти баланс между максимальными содержательными запросами и ограниченными техническими возможностями. Как представляется, сформированный пилотный вариант газетного корпуса СМИ Гродненщины вполне отвечает требованию представительности.

Рис. 1.

Во-первых, обеспечено географическое разнообразие: в корпус включены газеты шести районов Гродненской области, расположенных на севере (Островецкий район), западе (Гродненский, Берестовицкий районы), юго-западе (Свислочский, Зельвенский районы) и на востоке области (Ивьевский район).

Во-вторых, представлены газеты разных типов: две региональные («Перспектива», «Вечерний Гродно») и пять районных газет («Берестовицкая газета», «Островецкая правда», «Ивьевский край», «Свислочская газета», «Праца» (г. п. Зельва).

В-третьих, газеты различаются по типу аудитории: районные ориентированы на сельских жителей, они освещают в основном местные новости и проблемы локального уровня, в то время как региональные газеты ориентируются на жителей города Гродно и области, наряду с местными новостями уделяют внимание проблемам регионального, государственного и международного уровня. Кроме того, две гродненские газеты представляют другое типологическое различие: *«Перспектива»* относится к типу общественно-политических газет, с большой долей деловой и социально значимой информации, а *«Вечерний Гродно»* — к типу вечерних городских газет, предназначенных для досуга и семейного чтения.

В процессе подготовки текстов обнаружилось, что часть текстов районных газет, озаглавленных по-русски, оказалась на белорусском языке[1]. Это не только комментарии читателей, но и корреспондентские материалы. В отдельных номерах доля белорусскоязычных текстов достигала 40%. Было принято решение включить эти тексты в состав корпуса, введя еще один признак метаописания — «язык текста» — и снабдив их соответствующей пометой. Статьи на белорусском языке (общим объемом около 500000 словоупотреблений) доступны для точного поиска в составе общего корпуса, но их мож-

[1] Использование русского языка во всех заголовках электронной версии газеты, в том числе и в статьях на белорусском языке, обусловлено требованиями поисковой системы.

но выделить в отдельный подкорпус, выбрав значение опции «язык текста» — белорусский. Подкорпус газетных текстов на белорусском языке в русскоязычных газетах пяти районов, представленных в корпусе, свидетельствует, с одной стороны, о реальном двуязычии жителей этих районов, а с другой стороны, о стремлении газет быть ближе к читателям, ориентироваться на разные типы аудитории, в том числе и в языковом плане.

Ориентации газет на читателя, развитию их интерактивности способствует электронная форма изданий: во многих газетах бывшие отделы писем преобразованы в отделы обратной связи в виде комментариев, форумов и под. Тексты, помещенные в этих разделах, принадлежат непрофессиональным авторам и стилистически отличаются от основных публикаций: в отличие от них, комментарии читателей ближе к разговорной речи, чем к публицистике. Они представляют не меньший интерес для исследования региональных особенностей русского языка, чем основные газетные жанры, так как могут обнаружить еще один аспект регионального варьирования: если основной корпус газетных текстов мы используем для изучения вариантов нормы, то в комментариях мы будем иметь дело с вариантами разговорного узуса. Комментарии читателей к публикациям помещены в корпус, оформлены как самостоятельные тексты, каждый из которых включает все отклики на статью (их может быть от одного до нескольких десятков в зависимости от актуальности темы). Они имеют в составе заголовка название основной публикации (например, *У озера Свитязь появится хозяин / Комментарии, Лидские автобусы изменят дизайн / Комментарии*) и аннотированы как «электронная коммуникация».

Корпус-эталон российских газет

Согласно основной гипотезе проекта, региональные особенности в текстах гродненских газет могут быть выявлены при сравнении с корпусом-эталоном российских СМИ. Приступая к его формированию, мы учитывали, что этот корпус должен быть сопоставимым по объему и по временному периоду с гродненским корпусом. Что касается конкретного состава газет, здесь открывался веер возможностей. Можно было бы выбрать какую-нибудь область, например, Новгородскую или Смоленскую, и собрать местные газеты разных

типов; можно было действовать так же на уровне какого-либо административного образования, например, собрать газеты субъектов РФ, входящих в Северо-Западный административный округ (Ненецкий АО, Республика Коми, Республика Карелия, Архангельская, Мурманская, Вологодская, Ленинградская, Новгородская, Псковская, Калининградская области). В этом случае, вероятно, следовало бы ожидать значительного присутствия в корпусе-эталоне российских СМИ текстов с местными региональными особенностями. Мы решили нивелировать зависимость от регионального фактора. Сделать это можно было бы либо путем случайной выборки газет, либо путем отбора в корпус достаточно большого количества изданий из разных регионов (с учетом технических критериев, о которых говорилось выше). Для пилотной версии корпуса был выбран второй способ.

В настоящее время в корпус-эталон российских СМИ включены 13 газет, представляющих семь регионов России: Центральный, Северо-Западный, Южный, Приволжский, Уральский, Сибирский, Дальневосточный. Время выхода газет — в основном 2012–2013 гг. Общий объем корпуса — около 2 млн словоупотреблений.

• «Амурская правда» (http://www.ampravda.ru) — ежедневная региональная общественно-политическая газета Амурской области, г. Благовещенск. Издается с 24 февраля 1918 года. В советскую эпоху — печатный орган Амурского областного комитета КПСС и областного Совета народных депутатов. Тираж — 8000 экз., в корпус включено 174 текста за 2012–2013 г.

• «Байкальские Вести» (http://www.baikvesti.ru) — общественно-политическая еженедельная газета Иркутской области, г. Иркутск, издается с 2001 г. Первоначально газета называлась «Восточно-Сибирские вести», а в декабре 2003 года переименована в «Байкальские вести», причем коллектив остался прежним. С 1 января 2010 года газета выходила дважды в неделю, с июля 2012 года выходит 1 раз в неделю по понедельникам, тираж 10000 экз., в корпусе 490 текстов в основном за 2012 г.

• «Богатей» (http://bogatej.ru) — еженедельная независимая деловая газета, г. Саратов, издается с 1996 г., тираж 4500 экз. Читателями газеты, как сказано на сайте газеты, "являются бизнесмены, чиновники, политики, люди интеллектуальных профессий, которые уча-

ствуют в управлении хозяйством и обществом". В корпусе 270 текстов в основном за 2013 г.

• «Вечерний Орел» (http://vechorel.ru) — первая орловская электронная газета, общественно-политическое издание, организовано в 2010 г. как городская еженедельная вечерняя газета, тираж составлял 8000 экз. В корпус включено 138 текстов за 2012–2013 г.

• «Вятский край» (http://vk.ru) — ежедневная общественно-политическая газета, одна из четырёх областных государственных газет Кировской области. Газета с таким названием выходила еще в дореволюционной Вятке, но сегодняшний «Вятский край» свой отсчет ведет с 5 октября 1990 г., тираж составляет 60000 еженедельно, в корпусе 272 текста за 2012–2013 г.

• «Ижевская газета» (http://www.izhtime.ru/izhgazeta) — городская газета, г. Ижевск, интернет-версия издавалась с 2006 по 2010 г. В корпусе 192 текста за 2010 г.

• «Красное знамя» (http://komikz.ru/) — старейшая газета Республики Коми. Ведет своё начало от газеты «Зырянская жизнь» (вышла 10 июня 1918 года в г. Усть-Сысольске Вологодской губернии). До 1955 года выпускалась также под названиями «В Зырянском краю», «Удж» (труд), «Югыд туй» (светлый путь), «За новый Север». До 1991 года являлась официальным органом Коми обкома КПСС, Верховного Совета и Совета Министров Коми АССР. В настоящее время — независимая частная газета, выходит 1 раз в неделю, тираж около 10000 экз., регион распространения — Республика Коми, г. Сыктывкар; в корпусе 248 текстов за 2012–2013 г.

• «Новая газета Кубани» (http://ngkub.ru/) является региональным партнером федерального печатного общественно-политического издания «Новая газета». Первый номер «Новой газеты Кубани» вышел в свет 20 мая 2004 г. Целевая аудитория газеты — политики, предприниматели, бизнесмены, работники культуры и искусства. Издатель — ООО «Центр политического мониторинга». Газета выходит 2 раза в неделю. Тираж издания — 15000 экз. Регион распространения — Южный федеральный округ. Соотношение региональных и федеральных материалов составляет 40% : 60%. В корпусе 389 текстов за 2009–2013 г.

• «Новости Саратовской губернии» (http://sarnovosti.ru/) — региональное сетевое общественно-политическое издание, выходит с 2000 г., г. Саратов. Учредитель — министерство информации и пе-

чати Саратовской области. Тематика издания — освещение происходящих в области событий, деятельности органов власти всех уровней, общественных организаций. 8000–10000 посещений в сутки.

• «Рабочий путь» (http://www.rabochy-put.ru) — еженедельная общественно-политическая региональная газета, издается с марта 1917 года. Как сказано на сайте газеты, «Это не только летописец, но и участник славной истории Смоленщины и России. Газета не прекращала издаваться и во время оккупации территории области фашистскими войсками — «РП» печатался в Москве и распространялся с воздуха. «Рабочий путь» — политический тяжеловес. Он был и остается любимой газетой четырех поколений смолян». Тираж газеты 25000 экз. В корпусе 290 статей за 2008–2013 г.

• «Уральский рабочий» (http://газета-уральский-рабочий.рф) — ежедневная общественно-политическая газета, г. Екатеринбург. Главная и старейшая газета Среднего Урала издается с 14 февраля 1907 года. В советскую эпоху — орган Свердловского обкома КПСС и областного Совета депутатов трудящихся. Тираж около 20000 экземпляров, выходит пять раз в неделю. В корпус включено 112 текстов в основном за 2012–2013 г.

• VL.ru (Владивосток) — информационно-справочный портал города Владивостока, существует с 1997 г. Имеет разделы: Новости Владивостока (http://www.newsvl.ru) и справочные разделы (Афиша, Недвижимость, Транспорт, Базы отдыха, Работа, Форумы и пр.). В корпусе 229 текстов за 2011–2013 г.

• «Новгородские ведомости» (http://novved.ru) — единственная областная газета Новгородской области, издается с 29 декабря 1990 г. Распространяется во всех районах области, в Великом Новгороде и Северо-западном регионе. Выходит 3 раза в неделю. Номер в среду посвящён публикации официальных документов (областные законы, постановления и распоряжения администрации Новгородской области), в номерах во вторник и субботу публикуются материалы о политике, общественной жизни, культуре, спорте и др. Тираж газеты более 21000 экз. В корпус включено 849 текстов за 2012–2013 г.

Количественный вклад каждого издания в корпус-эталон российских СМИ 2010-х годов отражен в таблице 2.

Таблица 2

Название	**Регион**	**Тип газеты**	**Год издания**	**Кол- во с/у**
«Амурская правда» (Благовещенск)	Дальневосточный	региональная	2012-2013	114460
VL.ru (Владивосток)	Дальневосточный	городское электронное издание	2011-2013	44238
«Байкальские Вести» (Иркутск)	Сибирский	региональная	2010-2013	541580
«Уральский рабочий» (Екатеринбург)	Уральский	региональная	2011-2013	66215
«Богатей» (Саратов)	Приволжский	региональная	2013	136438
«Новости Саратовской губернии»	Приволжский	региональное электронное издание	2013	13912
«Ижевская газета» (Ижевск)	Приволжский	городская	2010	90516
«Вятский край» (Киров)	Приволжский	региональная	2012-2013	111826
«Вечерний Орел» (Орел)	Центральный	городское электронное издание	2012-2013	33587
«Рабочий путь» (Смоленск)	Центральный	региональная	2008-2013	112885
«Красное знамя» (Республика Коми)	Северо-Западный	региональная	2012-2013	158071
«Новгородские ведомости» (В. Новгород)	Северо-Западный	региональная	2012-2013	351679
«Новая газета Кубани» (Краснодар)	Южный	региональная/ федеральная	2009-2013	251323
ИТОГО				2026730

Рис. 2.

Сравнение гродненского корпуса и корпуса-эталона российских СМИ показывает, что они отличаются по составу газет: гродненский корпус на 80% состоит из текстов районных газет (причем районов с сельскохозяйственной специализацией), в российском корпусе собраны 8 региональных (областных) газет, 3 городские и 1 федеральная газета представлена своим региональным партнером. В гродненский корпус вошли электронные версии газет, которые выпускаются также и в печатной форме. В российском корпусе таких газет 10, а 3 издания представляют собой сетевые (электронные) СМИ. Издания гродненского корпуса локализованы в одном регионе Беларуси, география российского корпуса — максимально широка. Для контрастивных исследований такое отличие корпусов — скорее плюс, но для проведения сопоставительных исследований необходимо развитие экспериментальной базы в нескольких направлениях: в первую очередь включение районных газет в российский корпус, увеличение доли текстов, представляющих собой комментарии читателей. В долгосрочной перспективе предстоит пополнение обоих корпусов новыми газетами — в гродненский корпус войдут газеты других районов Гродненщины, студенческие газеты и малотиражные издания, в российский корпус будут включены газеты новых регионов, а также другие типы изданий, в том числе электронные.

Корпус региональных российских газет 1990–2000-х гг.

Гродненский корпус и корпус российских газет 2010-х годов формируют ядро регионального корпуса, но не единственную его составляющую. Концепция корпуса, который планируется создавать в течение нескольких лет, а затем постоянно пополнять, состоит в следующем. Корпус планируется сделать принципиально открытым, в том числе для сотрудничества с разными коллективами исследователей и заинтересованными организациями. Его можно будет расширять, присоединяя к нему новые модули, например, русскоязычные газеты Казахстана или Прибалтики, региональные газеты советского периода или начала XX в. и под. В настоящее время разработан общий дизайн корпуса, параметры описания текстов и стандарты их обработки. Внутри каждого модуля может быть своя специфика, которая определяется особенностями материала, но единственным общим требованием является соблюдение стандартов аннотации и обработки текстов, которое обеспечит возможность их интеграции в составе целого.

Одним из таких модулей, включенных в состав корпуса, является корпус региональных СМИ 1990–2000-х годов. В нем представлены тексты 40 изданий (в том числе районных и городских) из 7 федеральных округов России, и по объему он сопоставим с корпусом российских газет 2010-х годов. Состав этого подкорпуса приведен в таблице 3.

Таблица 3

Название	**Регион**	**Тип газеты**	**Год издания**	**Кол- во с/у**
«Амурский Меридиан» (Хабаровск)	Дальневосточный	региональная	2004	21937
«Биробиджанер Штерн» (Биробиджан)	Дальневосточный	региональная	2004	15194
«Владивосток» (Владивосток)	Дальневосточный	региональная	2003	32906
«Ежедневные новости» (Владивосток)	Дальневосточный	региональная	2003	13080

Название	Регион	Тип газеты	Год издания	Кол- во с/у
«Республика Саха» (Якутск)	Дальневосточный	региональная	1996	8484
«Рыбак Приморья» (Владивосток)	Дальневосточный	деловая региональная	2003	60103
«Биржа плюс свой дом» (Н. Новгород)	Приволжский	региональная	2002	152914
«Богатей» (Саратов)	Приволжский	региональная	2003	268266
«Вечерняя Казань» (Казань)	Приволжский	региональная	2003	52109
«Дело» (Самара)	Приволжский	деловой региональный журн.	2001-2002	421307
«Марийская правда» (Йошкар-Ола)	Приволжский	региональная	2003	46071
«Московский комсомолец» в Нижнем Новгороде	Приволжский	федеральная региональный выпуск	2004	2387
«Московский комсомолец» в Саранске	Приволжский	федеральная региональный выпуск	2004	2659
«Нефтяник» (Пермь)	Приволжский	региональная	2003	59741
«Нижегородские губернские ведомости» (Н. Новгород)	Приволжский	региональная	1998, 2003	23795
«Оренбуржье» (Оренбург)	Приволжский	региональная	1997	26884
«Пермский строитель» (Пермь)	Приволжский	региональная	2003-2004	196941
«Петербургский Час пик» (С.-Петербург)	Северо-Западный	региональная	2003	130765

Название	**Регион**	**Тип газеты**	**Год издания**	**Кол- во с/у**
«Санкт-Петербургские ведомости» (С.-Петербург)	Северо-Западный	региональная	2003	63790
«Смена» (С.-Петербург)	Северо-Западный	региональная	2003	1173
«Калининградская правда» (Калининград)	Северо-Западный	региональная	2003	11813
«Калининградские Новые колеса» (Калининград)	Северо-Западный	региональная	2004	37695
«Московский комсомолец» в Сыктывкаре	Северо-Западный	федеральная региональный выпуск	2003	1954
«Восточно-Сибирская правда» (Иркутск)	Сибирский	региональная	2003	194482
«Континент Сибирь» (Новосибирск)	Сибирский	региональная	2004	50873
«Красноярский рабочий» (Красноярск)	Сибирский	региональная	2003	77306
«Свободный курс» (Барнаул)	Сибирский	региональная	1997	25837
«Деловой квартал» (Екатеринбург)	Уральский	городской еженедельный журнал	2003	12671
«Вечерний Екатеринбург»	Уральский	вечерняя	2004	12554
«Сургутская трибуна» (Сургут)	Уральский	региональная	2001	9431
«Уральский автомобиль» (Миасс)	Уральский	районная	2004	62006
«Весть» (Калуга)	Центральный	региональная	2002	108438

Название	Регион	Тип газеты	Год издания	Кол- во с/у
«Воронежские вести» (Воронеж)	Центральный	региональная	2003	21903
«Встреча» (Дубна)	Центральный	районная	2003	301243
«Наша жизнь» (с. Перемышль, Калужская обл.)	Центральный	районная	2003	6674
«Северный край» (Ярославль)	Центральный	региональная	1997	5605
«Дагестанская правда» (Махачкала)	Северо-Кавказ-ский	региональная	2003-2005	49126
«Краевые новости» (Краснодар)	Южный	региональная	2003	6383
«Новороссийский рабочий» (Новороссийск)	Южный	городская	2003	9654
«Приазовский край» (Ростов-на-Дону)	Южный	региональная	2004	17769
«Сочи» (Сочи)	Южный	городская	2002	2502
ИТОГО				2626425

Таким образом, два корпуса газет — 1990–2000-х и 2010-х годов — представляют российскую региональную прессу в двух временных срезах и дают возможность проводить сравнительные микродиахронические исследования, отслеживать изменения в тематическом наполнении, в жанровом составе и в самих жанровых формах, стилистике взаимодействия с читателем и пр.

Региональные выпуски газеты «Комсомольская правда»

Другой модуль, включенный в состав текущей версии регионального корпуса,— это региональные выпуски КП. Вкладки с местными новостями издаются более чем в 40 регионах России (Абакан, Барнаул, Белгород, Благовещенск, Брянск и т. д.) и в зарубежных странах (Беларусь, Кыргызстан, Молдова, Северная Европа, Бал-

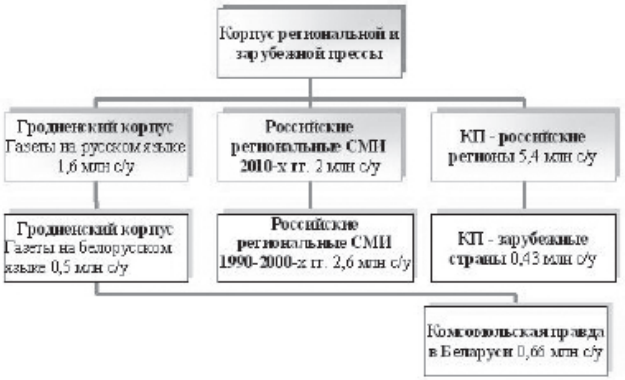
каны, Кипр). Материалы пишутся сотрудниками на местах, потому эти выпуски КП можно рассматривать как региональные издания. Однако будучи самостоятельными, редакции и издатели этих выпусков не могут не придерживаться общей политики центрального издания, принципов освещения и приемов подачи материала и др. Поэтому имеет смысл рассматривать местные выпуски КП и других федеральных газет как самостоятельный тип изданий. Состав подкорпуса приведен в таблице 4.

Таблица 4

Название	**Регион**	**Кол- во с/у**
«Комсомольская правда» (российские регионы)	40 субъектов РФ	5 422 196
«Комсомольская правда» в Беларуси	Беларусь	662 451
«Комсомольская правда» – Северная Европа	Эстония, Латвия, Литва	65 535
«Комсомольская правда» в Кыргызстане	Кыргызстан	67 193
«Комсомольская правда» в Молдове	Молдова	300 119
ИТОГО		6 517 494

Таким образом, в текущей версии регионального газетного корпуса выделяются четыре относительно самостоятельные коллекции (см. Рис. 3): 1) тексты русскоязычных газет Гродненщины на русском и белорусском языке (Гродненский корпус), две коллекции региональных СМИ России с дистанцией в 10 лет — 2) издания 1990–2000-х годов и 3) СМИ 2010-х годов и 4) коллекция региональных выпусков «Комсомольской правды». С ними можно работать как с единым массивом, так и с каждой коллекцией в отдельности, а также использовать коллекции в различных комбинациях. В частности, можно объединить в один подкорпус газеты Гродненщины и белорусские выпуски «Комсомольской правды». Эти и многие другие возможности обеспечиваются поиском по корпусу.

Рис. 3.



II. Подготовка корпуса. Стандарт обработки текстов

Превращение текста со страницы интернет-издания в элемент корпуса — долгий процесс преобразований, на отдельных этапах которого текст должен быть освобожден от одной аннотации, встраивающей его в электронный газетный гипертекст, и снабжен новой метатекстовой и лингвистической аннотацией, обеспечивающей поиск в корпусе. При подготовке текстов мы опирались на опыт разработки большого газетного корпуса, которая осуществлялась с использованием программных средств. Полностью отказаться от ручной обработки текстов не удалось (о причинах будет сказано ниже), но как представляется, был достигнут разумный компромисс в сочетании автоматических средств и ручного труда.

Первый и самый важный этап переформатирования текстов выполнялся с помощью специальной программы пакетной обработки файлов (автор Л. Д. Алексеевский). В программу встроен базовый шаблон преобразования текста из исходного формата в выходной формат, что позволяет обрабатывать файлы, имеющие одинаковый формат (например, тексты одного издания), в сколь угодно большом количестве. Однако поскольку в разных интернет-изданиях используются разная программная поддержка, для каждого издания приходится модифицировать шаблон замены с помощью ручной настройки. Изменения в шаблоне касаются в основном извлечения

метаинформации, поэтому справиться с задачей ручной настройки можно даже не владея специальными навыками программирования. После отладки нового модифицированного шаблона на нескольких текстах запускается автоматический режим работы. Шаблоны рекомендуется сохранять, поскольку они могут пригодиться для обработки новых текстов того же издания, а также подойти для текстов какого-то другого издания (в нашей практике такое случилось дважды).

Программная обработка файлов осуществляется в два приема: первый запуск программы обеспечивает извлечение нужной текстовой и метатекстовой информации (автор, название текста, дата, рубрика, если есть, адрес страницы, если есть) и очистку от лишней html-разметки. В результате второго запуска программы текст приводится к правильному xml-формату. В данном случае используется единый шаблон, поэтому производить эту операцию можно не по отдельным изданиям, а со всей обрабатываемой коллекцией. В полученном файле имеются две зоны — текст с размеченным заголовком и абзацами и зона метаданных, оформленная специальными тегами.

```
<?xml version="1.0" encoding="UTF-8"?><html><head>
<title>До первой звезды</title>
<meta content="До первой звезды" name="title">
</meta>
<meta content="15.03.2013" name="date"></meta>
<meta content="Старая Русса" name="topic"></meta>
<meta content="Новгородские ведомости"
name="source"></meta>
<meta content="http://novved.ru/staraya-russa/1419-
do-pervoj-zvezdy.html" name="url"></meta>
</head>
<body>

<p class="h1">До первой звезды</p>
<p>Наши ветераны — сильнее всех </p>
<p>В начале марта в Валдае прошел традиционный
турнир по мини-футболу, посвященный памяти одного
из основателей команды «Юпитер» Александра
Белоусова. В играх участвовала и команда
```

```
старорусских ветеранов, возраст которых перешагнул
50-летнюю отметку. </p>
<p>Как сообщил вратарь старорусских спортсменов, а
по совместительству и директор МАУ «Спорткомплекс»
Борис Васильев, наши на турнире были лучшими. </p>
<p>В финале мы встречались с командой хозяев и
выиграли со счетом 5:4. В баталиях участвовали
команды из Валдая, Старой Руссы, Парфина и
Крестец.</p>
</body>
</html>
```

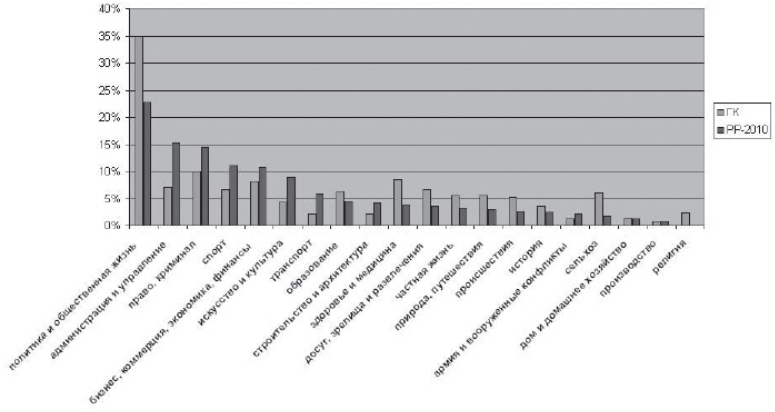
Следующий этап обработки — метатекстовая аннотация — выполняется программными средствами с последующим редактированием. Для этого используется программа QTMetabaseEditor (автор Л. Д. Алексеевский), которая считывает метаданные из файлов в соответствующие поля базы данных. Дальнейшее заполнение полей (прежде всего определение текстовых характеристик) производится вручную, причем удобнее это делать с помощью той же программы, которая по каждому полю предлагает выбор из списка значений, что, несомненно, сокращает количество формальных ошибок при заполнении. Экспортируется содержимое базы в формат таблицы. csv, которая сопровождает каталоги с текстовыми файлами и хранит всю метаинформацию о текстах.

При разработке формата таблицы метаданных, в отличие от большого газетного корпуса, мы решили не отказываться от использования ручного редактирования при заполнении полей по нескольким причинам. Во-первых, ядро корпуса — это белорусско-русский проект, который предполагал подробную аннотацию текстов с целью изучения лексико-семантических и культурно-специфических особенностей газетных текстов, в том числе и в жанровом отношении. Во-вторых, корпус сравнительно невелик по объему, и такой объем работы оказался по силам даже небольшому коллективу. Кроме того, в большинстве изданий удалось извлечь информацию о рубрике, что облегчило определение типа текста и тематики. Всего лишь облегчило, хотя первоначальный расчет был на то, что наличие рубрики позволит автоматически заполнить поле тематики. Однако на практике оказалось, что организация текстов в рубрики не кор-

релирует с тематическим членением текстов. Действительно, чаще всего рубрика в самом деле соответствует тематическому содержанию публикации (см., например, такие рубрики, как «Политика», «Общество», «Бизнес», «Происшествия», «Здоровье», «Спорт»). Реже она дает информацию о жанре и помогает определить тип текста (например, в рубриках «Новости», «Комментарии» или «Обзор прессы» размещаются соответственно информационные сообщения, комментарии и обзоры; в рубрике «Персона» — в основном интервью с известными людьми, «Наш репортаж» — репортажи). Иногда название рубрики отражает логику внутренней организации сайта, но не несет нужной нам информации о жанре и тематике текста (например, такие рубрики, как «Старая Русса» в «Новгородских ведомостях» или «Киров и область» в «Вятском крае», или «Без рубрики» в «Вечернем Гродно»). Кроме того, в исходных файлах некоторых изданий информация о рубрике отсутствовала. Поэтому полностью отказаться от ручного редактирования метатекстовой аннотации оказалось невозможным. Наконец, в-третьих, качественная ручная разметка пилотного корпуса позволит в дальнейшем использовать его в качестве эталона для обучения программ автоматического определения жанров и тематики текстов.

Тематическое распределение текстов в Гродненском и эталонном корпусах представлено на Рис. 4.

Рис. 4.



Учитывая, что в последнее время исследования по автоматической атрибуции жанровой и тематической принадлежности текстов на больших массивах данных ведутся с большой интенсивностью, в будущем мы надеемся автоматически разметить тексты «Комсомольской правды» и освоить автоматический способ разметки тематики и типов текста в пополнениях корпуса.

III. Поиск по корпусу

При разработке стандартов аннотации и организации поиска по корпусу был учтен опыт создания основного и газетного корпусов НКРЯ. Метатекстовая аннотация состоит в приписывании каждому тексту атрибутов, по которым в дальнейшем будет осуществляться поиск в корпусе. В метатекстовой разметке регионального корпуса есть значительные отличия как от основного, так и от газетного корпуса. Набор метатекстовых признаков и основных значений приведен ниже.

0. Название файла и путь к нему

I. Информация об авторе

1. **Имя автора**
2. **Пол автора**
3. **Возраст автора**

II. Информация о тексте

4. **Название текста**
5. **Дата создания текста**
6. **Сфера функционирования текста:** публицистика, к которой относятся основная масса газетных текстов, реклама, деловая, художественная.
7. **Тема текста**, или предметная область*
8. **Хронотоп**, или место и время описываемых событий: только для художественных текстов
9. **Тип текста***
10. **Жанр художественной литературы**
11. Стиль текста

III. Информация об аудитории

12. Возраст аудитории: н-возраст, если это незначимый фактор, детский, подростковый, взрослый
13. Уровень образования аудитории: для газет как средства массовой информации, в основном н-уровень

14. **Размер аудитории**: национальная (для федеральных газет), региональная (для областных газет), районная (для районных газет, а также городских, аудитория которых ограничена конкретным языковым поселением).

IV. Библиографическое описание текста

15. Источник текста — адрес страницы в интернете

16. **Название издания** — название газеты или электронного издания

17. Тип источника — газета, журнал, электронный текст

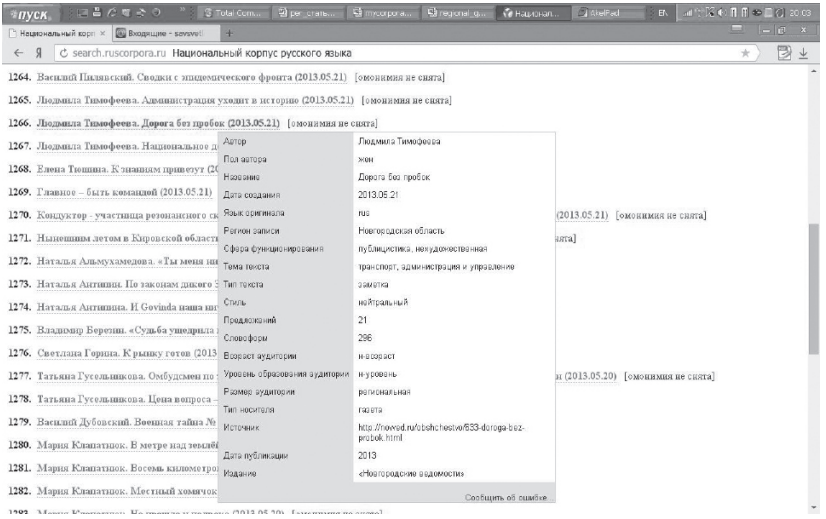
18. **Страна**

19. **Регион**

20. **Язык**

Тексты регионального корпуса, за исключением региональных выпусков КП, имеют полный набор метапризнаков, который приводится в «паспорте» текста, доступном при нажатии на активную строку с названием текста на странице выдачи подкорпуса или выдачи контекстов (см. Рис. 5). В текстах КП пока нет разметки по тематике и типу текста (эти признаки в приведенном выше списке отмечены звездочкой).

Рис. 5.

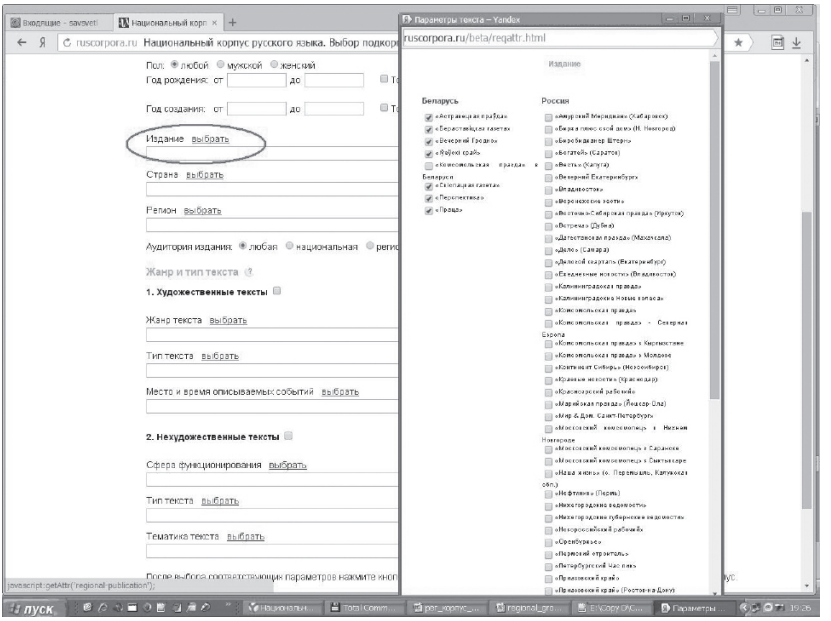


Почти все метаатрибуты (они выделены в списке жирным шрифтом) представлены на странице http://ruscorpora.ru/mycorpora-regional.html, и по ним организован выбор подкорпуса. При выборе рабочего подкорпуса мы имеем возможность задавать комбинации признаков и конструировать корпус в соответствии с поставленными задачами. Спектр возможностей широк — можно работать как со всем массивом текстов, так и с текстами отдельной газеты. Удобный и понятный интерфейс позволяет делать это с минимальными усилиями — установка большинства опций предполагает выбор из списков.

Выбирая значение «Язык» = «белорусский», мы задаем небольшой подкорпус текстов на белорусском языке в гродненских газетах. Выбор значения «Язык» = «русский» позволяет нам работать со всем 13-миллионным корпусом текстов на русском языке.

Для выбора Гродненского корпуса текстов на русском языке можно использовать следующий набор атрибутов: «Язык» = «русский», «Страна» = «Беларусь», «Регион» = «Гродненская область». Другой путь к тому же корпусу — через атрибут «Издание»: в открывшемся

Рис. 6.

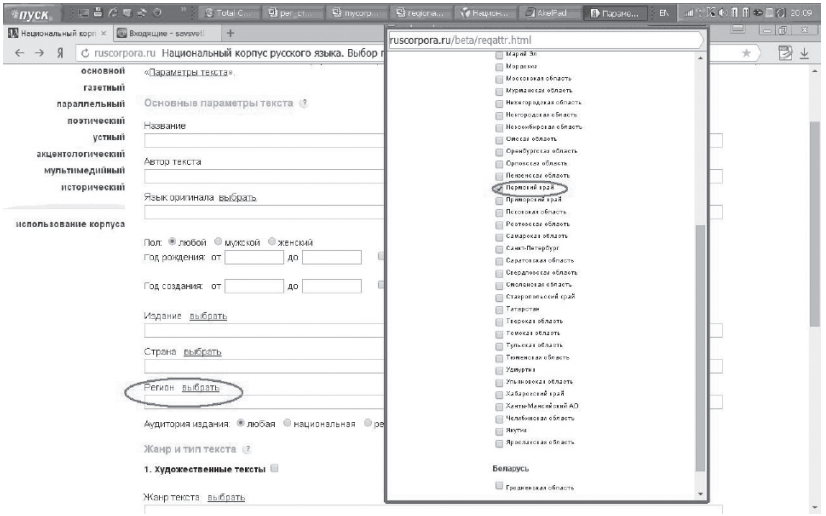


списке нужно отметить галочками все белорусские газеты (кроме «Комсомольской правды» в Беларуси) или выбрать нужные.

Коллекция российских СМИ отбирается аналогично: «Страна» = «Россия» (язык «русский» можно не устанавливать), газеты выбираются из списка «Издание». Если нужен весь эталонный корпус российских СМИ, используем следующие параметры: «Страна» = «Россия», «Год создания» от 2009 до 2013, из списка газет исключаем «Комсомольскую правду». Если нужен ранний подкорпус российских газет, устанавливаем дату 1996–2004 (выпуски «Комсомольской правды» в него не попадают, поскольку они датированы 2011–2012 годами). Если предстоит работать только с текстами «Комсомольской правды», подкорпус формируется из списка «Издание».

Корпус предоставляет возможность работы с газетами какого-либо региона (для Российской федерации можно выбрать уровень отдельного субъекта федерации — области, края, республики, или административного округа, например Сибирского, Уральского, Южного). Для этого нужно использовать параметр «Регион», выбрав нужное значение или несколько значений из списка.

Рис. 7.



Например, Пермский край представлен четырьмя газетами — «Пермский строитель», «Нефтяник», «Уральский автомобиль»

и местные выпуски «Комсомольской правды», всего в подкорпусе 925 текстов, почти 495 тыс. словоупотреблений. Доля газет Иркутской области («Байкальские Вести», «Восточно-Сибирская правда» и региональные выпуски КП) составляет 1102 текста, более 845 тыс. словоупотреблений, Краснодарского края — 713 текстов (около 456 тыс. словоупотреблений) из пяти газет («Краевые новости», «Новая газета Кубани», «Новороссийский рабочий», «Сочи», региональные выпуски КП). По мере пополнения корпуса новыми газетами доля газет отдельного региона будет также расти, и это даст возможность исследовать с помощью корпуса особенности функционирования русского языка в отдельных российских регионах и странах ближнего и дальнего зарубежья.

IV. Перспективы развития корпуса

В настоящее время по адресу http://ruscorpora.ru/search-regional.html доступна рабочая версия регионального корпуса. В нем представлены газеты нескольких уровней — региональные выпуски центральных газет, газеты регионального уровня и местные издания — районные и городские. Временные рамки текстов 1996–2013 годы. География печатных изданий широка и охватывает все федеральные округа России, а также страны ближнего зарубежья (Беларусь, Молдова, Кыргызстан), страны Балтии. В корпус включены издания, разнообразные во многих отношениях: общественно-политические и отраслевые, деловые и вечерние, предназначенные для досуга, ежедневные и еженедельные, печатные издания и электронные СМИ. Помимо корреспондентских материалов в корпусе присутствуют отклики и комментарии непрофессиональных авторов, в которых реализуется обратная связь газеты с читателем. Корпус используется в качестве экспериментальной базы исследования региональных особенностей русской речи, что отражено в публикациях [Кустова, Савчук 2013; Рычкова 2014; Рычкова, Станкевич 2014; Шайкевич, Савчук 2014 и др.].

Корпус можно рассматривать как пилотный проект с заданными параметрами развития будущего большого корпуса региональных СМИ. Основное направление в развитии корпуса — формирование полноценных газетных подкорпусов для разных регионов, так чтобы все регионы были в нем равномерно представлены. Ориентиром

могут служить параметры Гродненского корпуса, то есть объем для каждого региона должен быть не менее 1,5–2 млн словоупотреблений, и в нем должны присутствовать газеты разных уровней. Исходя из этого задачей ближайшего будущего является увеличение доли текстов районных газет и читательских комментариев в корпусе российских СМИ. Другой первоочередной задачей является пополнение корпуса за счет изданий тех регионов, которые пока слабо в нем представлены. Русскоязычные газеты ближнего и дальнего зарубежья пополнят блок зарубежных СМИ. Гродненский корпус, в случае продолжения совместного проекта, будет пополняться газетами других районов области.

Другим вектором развития корпуса, наряду с территориальным, является временной. В настоящее время корпус содержит тексты только современных газет, но в него можно включать коллекции региональных газет 1960–1980-х годов или более ранних периодов. Однако старые газеты менее доступны и подготовка их потребует больших усилий, чем работа с современными электронными изданиями.

Требование разнообразия обусловливает поиск и включение в корпус газет с разными целевыми аудиториями, которые различаются по величине, возрасту, профессиональной и политической ориентации, — разумеется, с учетом реального положения дел в сфере региональных СМИ. Таким образом, целью дальнейшего развития регионального корпуса является создание сбалансированного ресурса, в котором разные регионы были бы представлены с достаточной полнотой, необходимой для проведения лексико-семантических и грамматических исследований.

Предполагается также дальнейшее развитие корпусного инструментария: организация поиска по n-граммам и графическое представление статистической информации — распределения частот употребления словоформ по разным осям — по годам или по регионам. В целях совершенствования интерфейса планируется разработать справочные страницы с информацией о регионах, включая карты, с кратким описанием периодических изданий — все это упростит и облегчит выбор подкорпуса и сделает работу с ним более занимательной. Справочный раздел корпуса можно будет расширять и путем размещения на сайте работ с результатами исследований, проведенных с использованием корпуса.

Мы надеемся на сотрудничество с коллективами региональных исследовательских центров — объединение усилий поможет осуществить проект.

Литература

Кустова Г. И., Савчук С. О. Изучение лексико-семантической и социокультурной специфики русской речи на территории Республики Беларусь (на материале текстов СМИ) // Труды Международной конференции «Корпусная лингвистика — 2013». Санкт-Петербург, 2013. С. 344–352.

Рычкова, Л. В., Станкевич, А. Ю., Бодак Ю. А. Принципы создания газетного корпуса СМИ Гродненщины // Компьютерная лингвистика: научное направление и учебная дисциплина [Текст] сборник научных статей. Вып. 2. / Отв. ред. В. И. Коваль [и др.] М-во образования РБ, ГГУ им. Ф. Скорины; Научно-методический центр русистики. Гомель: ГГУ им. Ф. Скорины, 2012. С. 88–92.

Рычкова Л. В. Специальная лексика в языке региональных СМИ // Терминология и знание. Вып. IV: материалы IV Междунар. симпозиума, Москва, 6–8 июня 2014 г. / Под ред. С. Д. Шелова. М., 2014. С. 157–172.

Рычкова Л. В., Станкевич А. Ю. «Диалог» языков на страницах СМИ Гродненщины // Актуальные проблемы теории дискурса: материалы Междунар. электрон. конф., Актюбинск, 30 мая 2014 г. / «Actual problems of the theory of discourse» proceedings of the international scientific e-conference, May 30, 2014. Актобе: Актюбинский регион. гос. ун-т им. К. Журбанова. С. 51–61.

Рычкова Л. В. Изучение социально-коммуникативных особенностей русского языка на основе использования лингвистических корпусов региональных средств массовой информации (на примере корпуса Гродненщины). // Universitas Catholica Rosenbergensis Studia Russico-Slovaca. Verbum. Ruzomberok, 2014. С. 48–58.

Савчук С. О. Корпус современной русской прессы: из опыта создания и использования // Труды Международной конференции «Корпусная лингвистика — 2011». СПб: Санкт-Петербургский государственный университет, 2011. С. 149–154

Савчук С. О. Корпусное исследование вариантов родовой при-

надлежности имен существительных в русском языке // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». М.: РГГУ, 2011. С. 562–580

Станкевич А. Ю. Технология сбора и систематизации электронного контента для корпуса русскоязычных СМИ Гродненщины // Карповские научные чтения: сб. науч. ст. Вып. 8: в 2 ч. / редкол.: А. И. Головня (отв. ред.) [и др.]. Минск: Белорус. Дом печати, 2014. Ч. 2. С. 6–10.

Шайкевич А. Я., Савчук С. О. Анализ лексико-семантических особенностей региональной прессы (на примере газет Гродненского региона Беларуси) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). М.: Изд-во РГГУ, 2014. С. 585–598.

S. O. Savchuk
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
savsvetlana@mail.ru

THE CORPUS OF REGIONAL RUSSIAN-LANGUAGE NEWSPAPERS IN RUSSIA AND ABROAD

The article describes the corpus of regional and foreign press as a new module within the RNC. Unlike newspaper corpus including only materials of the national press, the new corpus focuses on texts from regional publications. The corpus presently consists of the following subcorpora: 1) newspapers of the Grodno region (7 editions, 1.9 million tokens in total), prepared jointly with the specialists of the Grodno University, 2) Russian regional press of 2010s (13 newspapers, about 2 million tokens), 3) Russian regional press of 1990-2000 (40 newspapers, 2.6 million tokens), 4) regional releases of the national newspaper "Komsomolskaya Pravda" (6.5 million tokens). Our approach to selecting texts, the standards and software for word processing, organization of the corpus inter-

face are described in the article as well as the prospects of the development of corpus composition and searching instruments.

The texts of the regional corpus are provided with morphological, semantic and detailed metatextual annotation. The corpus is freely available at the address http://www.ruscorpora.ru/search-regional.html. The main purpose of this corpus as we see it is to provide researchers with a tool of studying the regional variation of the Russian language.

Key words: newspaper corpus, regional press, composition and structure of corpus, text annotation

References

Kustova G. I., Savchuk S. O. [The study of lexical-semantic and socio-cultural specifics of the Russian language on the territory of Belarus (on the material of mass media texts)]. *Trudy Mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2013"* [Proceedings of International Conference "Corpus Linguistics – 2013"]. St. Petersburg, 2013. pp. 344–352. (In Russ.)

Rychkova, L. V., Stankevich, A. Yu., Bodak Yu. A. [The principles of creating newspaper corpus of the Grodno region]. *Komp'yuternaya lingvistika: nauchnoe napravlenie i uchebnaya distsiplina.* [Computational linguistics: research area and academic discipline]. Vol. 2. V. I. Koval' et al. (eds.). Gomel, GSU them. F. Skorina, 2012, pp. 88–92. (In Russ.)

Rychkova, L. V. [Special vocabulary in regional media]. *Terminologiya i znanie. Vyp. IV: materialy IV Mezhdunar. simpoziuma* [Terminology and knowledge. Vol. IV: proceedings of the IV Intern. Symposium]. S. D. Shelov (ed.). Moscow, 2014, pp. 157–172. (In Russ.)

Rychkova L. V. [Studying the socio-communicative features of the Russian language at the base of regional mass-media text corpora: The case of the Corpus of Grodno region media]. *Universitas Catholica Rosenbergensis Studia Russico-Slovaca. Verbum.* Ruzomberok, 2014, pp. 48–58. (In Russ.)

Rychkova L. V., Stankevich A. Yu. ["Dialogue" of languages in the media of Grodno region]. *Aktual'nye problemy teorii diskursa: materialy Mezhdunar. elektron. konf.* [Actual problems of the theory of discourse: proceedings of the international scientific e-conference, May 30, 2014]. Aktobe, Aktobe region. state Univ. K. Gurbanov, pp. 51–61. (In Russ.)

Savchuk S. O. [The Corpus of Modern Russian Press: Compilation and Use]. *Trudy Mezhdunarodnoi konferentsii "Korpusnaya lingvistika – 2011"* [Proceedings of International Conference "Corpus Linguistics – 2011"]. St. Petersburg, St. Petersburg State Univ. Publ., 2011, pp. 149–154. (In Russ.)

Savchuk S. O. [Corpus-based study of morphological variability: Variation in gender forms of Russian nouns]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog – 2011"]. Moscow, RSUH Publ., 2011, pp. 562–580. (In Russ.)

Stankevich A. Yu. [Technology of collecting and organizing electronic content for Russian-language mass-media of Grodno region]. *Karpovskie nauchnye chteniya: sb. nauch. st.* [Karpovsky readings: collection of scientific articles]. Vol. 8: in 2 parts. A. I. Golovnya (ed.). Minsk, Belarusian printing house, 2014, p. 2, pp. 6–10. (In Russ.)

Shajkevich A. Ya., Savchuk S. O. [Distributional-statistical analysis of regional press (newspapers of Grodno region of Belarus)]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog – 2014"]. Moscow, RSUH Publ., 2014, pp. 585–598. (In Russ.)

Д. В. Сичинава
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
mitrius@gmail.com

ПАРАЛЛЕЛЬНЫЕ ТЕКСТЫ В СОСТАВЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: НОВЫЕ НАПРАВЛЕНИЯ РАЗВИТИЯ И РЕЗУЛЬТАТЫ[1]

В статье представлено текущее состояние параллельных корпусов в составе НКРЯ и работа, проведенная над этими корпусами в период 2009–2015 гг. Параллельные корпуса НКРЯ включают следующие параллельные двуязычные (с русским) корпуса: английский, армянский, белорусский, болгарский, испанский, итальянский, латышский, немецкий, польский, украинский, французский, эстонский. Практически для всех из этих языков представлены обе симметричные языковые пары. В соответствии с мировым опытом построения параллельных корпусов, теперь входит многоязычный корпус, состоящий из 9 текстов и задействующий более 20 языков (в основном славянских). Как поливариантный с 2012 г. разрабатывается русско-французский корпус, включающий до 4 вариантов перевода для некоторых текстов. Поливариантные тексты включены и в многоязычный корпус. В параллельные корпуса теперь включаются как художественные, так и нехудожественные тексты с той же классификацией, что и в основном русском корпусе (публицистика, производственно-технические, учебно-научные, церковные, юридические тексты). Разработан инструментарий разметки несоответствий и вольностей при переводе (пропуск части предложения,

[1] Работа выполнена в рамках Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» и при поддержке РГНФ (грант № 15-04-12018).

вставка, значимая неадекватная замена, выполненная переводчиком), значительная часть текстов размечена с учетом этого инструментария. Тексты на большинстве языков получают морфологическую разметку. Рассмотрены конкретные примеры использования корпусов для исследования лексики и грамматики.

Ключевые слова: Параллельный корпус, разметка, многоязычный корпус, лексическая типология, грамматическая типология, лингвоспецифичная лексика, перфект

За время, прошедшее с выхода сборника «Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы», параллельные корпуса в составе НКРЯ (http://ruscorpora.ru/search-para.html) пережили очень важный этап развития как с технологической, так и с содержательной стороны. Теперь это один из самых объёмных (более 70 миллионов словоупотреблений) и важных проектов НКРЯ, задействующий несколько команд специалистов из разных стран (ср. публикацию [Sitchinava 2012]). Это направление получило поддержку Программы Президиума РАН «Корпусная лингвистика» (проект «Создание и развитие параллельных русско-иноязычных корпусов в Национальном корпусе русского языка», руководитель Д. О. Добровольский). Ряд параллельных корпусов (например, русско-французский мультиязычный, белорусский, польский) развивался в эти годы и как самостоятельные подпроекты, поддержанные отдельными грантами. Параллельные корпуса — едва ли не самый «многолюдный» проект в составе Национального корпуса; как и диктует специфика задачи, его команда — это целый ряд групп разработчиков, представляющих разные страны.

Назовём **языковой парой L1— L2** параллельный (под)корпус, состоящий из текстов, созданных в оригинале на языке L1 и переведённых на язык L2. В дальнейшем мы для обозначения двух симметричных пар «русский-L корпус» и «L-русский корпус» будем также использовать более краткое выражение **L-й параллельный корпус** (например, «армянский параллельный корпус» = «армянско-русская и русско-армянская языковые пары»).

В 2008 году Национальный корпус русского языка включал только три языковые пары: англо-русский, русско-английский и немецко-русский корпус. Эти корпуса включали в себя только художественные тексты.

За прошедшие шесть лет параллельные корпуса получили развитие в следующих направлениях:

1. Новые языковые пары. Сейчас параллельные корпуса включают следующие параллельные двуязычные (с русским) корпуса: английский, армянский, белорусский, болгарский, испанский, итальянский, латышский, немецкий, польский, украинский, французский, эстонский. Практически для всех из этих языков представлены обе симметричные пары, например, армянско-русский и русско-армянский (под)корпуса. В настоящее время нет болгарско-русских текстов (в НКРЯ включен готовый русско-болгарский корпус, составленный в Тырновском университете). Русско-немецкая языковая пара, отсутствовавшая в 2008 году, появилась и активно развивается. Часть из этих корпусов представляет собой небольшие по объёму пилотные проекты из нескольких текстов (особенно испанский и в какой-то мере латышский), часть — вполне полноценные корпуса (крупнейшими по объёму, помимо английского и немецкого, являются также украинский и польский). Существуют проекты создания и других двуязычных пар (например, венгерского, литовского, шведского корпусов).

2. Многоязычный (мультиязычный) корпус. В НКРЯ, в соответствии с мировым опытом построения параллельных корпусов, теперь входит многоязычный корпус, состоящий из 9 текстов и задействующий более 20 языков (в основном славянских).

3. Поливариантный корпус, включающий более одного перевода с языка L1 на язык L2. Как поливариантный с 2012 г. разрабатывается русско-французский корпус, включающий до 4 вариантов перевода для некоторых текстов. Поливариантные тексты включены и в многоязычный корпус. Имеются также русско-итальянские поливариантные тексты. Есть планы аналогичного развития также иных языковых пар (английские, немецкие тексты).

4. Жанровое разнообразие. В параллельные корпуса теперь включаются как художественные, так и нехудожественные тексты на основе той же классификации, что и в основном русском корпусе (публицистика, производственно-технические, учебно-научные, церковные, юридические тексты).

5. Средства разметки и выравнивания. Разработан инструментарий разметки несоответствий и вольностей при переводе (пропуск части предложения, вставка, значимая неадекватная за-

мена, выполненная переводчиком), значительная часть текстов размечена с учетом этого инструментария. Тексты на большинстве языков получают морфологическую разметку. Польские тексты размечены с автоматическим снятием омонимии, армянские — с рядом облегчающих поиск особенностей (перевод лемм на английский язык, версия в латинице). Для всех русских текстов в составе параллельных корпусов доступен такой же семантический поиск, что и в основном русском корпусе. Расширена и развита метаразметка текстов: сейчас метаинформация параллельного корпуса включает сведения о сфере функционирования текста, дате создания перевода, предусматривает указания названия текста и имени автора на нескольких языках. Разработана специальная программа-оболочка «Евклид» (автор Т. А. Архангельский), в которую включено автоматическое алгоритмическое выравнивание, осуществляемое программой HunAlign; в программе предусмотрена удобная корректировка ошибок автоматического выравнивания, разметка неточности в переводе (ср. выше), а также модуль заполнения таблицы метаразметки и конвертация текста в XML в формате UTF-8. При помощи этой программы один разметчик может самостоятельно собирать тексты, выравнивать и метаразмечать их, получая на выходе файлы XML и метатаблицы в формате, готовом для размещения в параллельном корпусе.

Ниже рассмотрим все эти пять направлений развития параллельных корпусов НКРЯ. Кроме того, будут рассмотрены конкретные примеры использования корпусов для исследования лексики (раздел 6) и грамматики (раздел 7).

1. Новые языковые пары и развитие старых

1.1. Английский корпус

Английский и немецкий параллельный корпуса, как и на момент публикаций в сборнике НКРЯ 2006–2008, разрабатываются под руководством А. А. Кретова и Д. О. Добровольского сотрудниками Воронежского университета (следует назвать вносящих особо активный вклад И. А. Меркулову, Ю. П. Плешкову, Е. Н. Подтележникову, Ю. А. Суворову, Кс. М. Шилихину).

В 2008–2014 гг. в английский параллельный корпус ежегодно включались новые художественные тексты писателей XIX — XXI вв., в том числе т. н. «классические» переводы, представляющие большое культурно-историческое значение, сопоставимое с оригинальными текстами на данном языке (например, переводы русских классиков, выполненные К. Гарнет, или «Над пропастью во ржи» Сэлинджера в переводе Р. Райт). Кроме того, в корпус введены предоставленные компанией ABBYY (проект ABBYY Lingvo) переведённые с английского нехудожественные тексты: технические (инструкции, документы и официальные отчеты организаций), учебно-научные, публицистические. В 2015 г. в корпус также включено представительное собрание сочинений В. В. Набокова как русского, так и американского периода, в оригиналах и переводах соответственно на английский и русский языки, в том числе авторских и авторизованных (тексты собраны и выровнены Я. А. Жеребцовой). Английский корпус остаётся самым большим из параллельных корпусов НКРЯ (18,1 млн словоупотреблений в англо-русской и 6,5 млн в русско-английской части), на его материале активно проводятся исследования лексики и грамматики (см. разделы 6 и 7), в том числе диахронические.

1.2. Армянский корпус

Создан армянский паралельный корпус размером 2,1 млн словоупотреблений, большая часть которых приходится на русско-армянские тексты. В основу корпуса положены переводы с русского языка на армянский (в основном классической художественной литературы), входящие в Восточноармянский национальный корпус (EANC; http://eanc.net). При отборе текстов обращалось внимание на репрезентативность с точки зрения включения максимального количества разных авторов и переводчиков. В корпус вошли оригиналы и переводы на армянский язык прозаических произведений М. Булгакова, И. Бунина, Н. Гоголя, И. Гончарова, Ф. Достоевского, В. Ерофеева, А. Куприна, О. Мандельштама, А. Н. Островского, Н. А. Островского, А. Пушкина, А. Рыбакова, М. Салтыкова-Щедрина, Л. Толстого, И. Тургенева, А. Фадеева, А. Чехова. В армянско-русскую языковую пару включены оригиналы и переводы на русский язык произведений Х. Абовяна, А. Бакунца и Д. Демирчяна.

Все армянские тексты снабжены морфологической разметкой с неснятой омонимией, основанной на морфологическом модуле EANC. Эта разметка включает в себя переводы армянских лемм на английский язык. Кроме того, учитывая особенности армянской графики, труднодоступной неспециалисту, все армянские тексты продублированы в латинской транслитерации (включая морфологическую разметку лемм). Три слоя — русский, армянский в графике и армянский в латинице — автоматически выровнены предложение в предложение.

Выравниванием текстов занималась Т. О. Шаврина, технической поддержкой нестандартной конвертации разметки и транслитерации — Т. А. Архангельский.

1.3. Белорусский корпус

Белорусско-русская языковая пара — в своей основе плод отдельного совместного проекта РГНФ и БРФФИ (№ 11–24–01004а/Bel «Корпусные сопоставительные исследования русского и белорусского языков и разработка параллельных электронных корпусов»). Корпус пополняется и после окончания этого гранта. Беларусь — страна, где в последние годы активно развиваются корпусные проекты, в том числе и с участием российских специалистов; в этом ряду параллельный корпус стал одним из первых (по крайней мере из доступных в Интернете). Руководство проектом с белорусской стороны взял на себя один из главных энтузиастов корпусной лингвистики в этой стране В. А. Кощенко (Национальная академия наук Беларуси), в разработке корпуса на разных этапах активно участвовали также И. В. Глинник, А. В. Зубов, И. Л. Копылов, О. В. Мицкевич, Е. Н. Скопинова, Ю. А. Стасевич (Коровко).

Именно в ходе работы над этим корпусом по предложениям работавших над ним белорусских специалистов была создана и существенно усовершенствована программа выравнивания «Евклид», появилась технология разметки неточностей в переводе (связанная со спецификой художественного перевода с «языков народов СССР»).

Белорусский корпус включает новостные тексты (материалы двуязычных новостных агентств «БелаПАН» и «Телеграф»), а также тексты существующих в двух языковых версиях белорусских

законов 1990–2000-х гг. Язык оригинала для этих текстов указан как белорусский условно, поскольку реальное направление перевода не всегда можно установить. Русско-белорусский компонент также недавно пополнился новым переводом библейских новозаветных книг, выполняемых библейской комиссией Белорусского экзархата РПЦ. В качестве оригинала с долей условности указан русский синодальный перевод Библии, на решения которого белорусский перевод в известной степени ориентируется, хотя делается с греческого оригинала. Текст белорусских новозаветных книг, включенных в корпус, частично снабжён знаками ударения.

В основу морфологической разметки белорусского корпуса положен модуль морфологии, разработанный И. В. Совпелем и усовершенствованный компанией «Яндекс» (в частности, этот модуль теперь частично обрабатывает так называемую тарашкевицу, или белорусское классическое правописание, с передачей предконсонантной мягкости: *сьвет, прывітаньне*). Ведется работа над усовершествованием этой разметки (в частности, включением в словарь таких слов, как, например, местоимений *іхні* и *ягоны*, частотных в текстах, но ограниченно представленных в словарных стандартах, ориентированных на советские нормы).

Общий объём белорусско-русского параллельного корпуса — 4,9 млн словоупотреблений в белорусско-русской и 1,8 млн в русско-белорусской части.

1.4. Болгарский корпус

Русско-болгарский компонент основывается на текстах, любезно предоставленных Великотырновским университетом (группой под руководством проф. Гочо Гочева, http://rbcorpus.com/ekip_rus.php). Тексты совокупным объёмом 3,6 млн словоупотреблений представляют собой оригиналы и переводы на болгарский язык произведений XX — XXI вв., причём включают не только художественные тексты, но и публицистику (например, фельетоны М. Кольцова). Все тексты в рамках российского проекта переконвертированы в формат XML и снабжены морфологической разметкой с неснятой омонимией на основе болгарского морфологического модуля компании «Яндекс».

1.5. Испанский и итальянский корпуса

В пилотной стадии в настоящее время находятся новые романские корпуса НКРЯ. В испанский и итальянский корпус входят как переводы с русского (например, «Анна Каренина» Толстого, «Град обреченный» Стругацких), так и переводы на русский (например, «Улей» Камило Хосе Селы, «Имя розы» Умберто Эко). Все эти тексты относятся к художественной литературе, задача жанровой репрезентативности пока не ставилась. Тексты получают морфологическую разметку на базе анализаторов компании «Яндекс», которая ещё нуждается в дополнительном усовершенствовании.

Над выравниванием испанских текстов работали В. С. Люсина и С. Ю. Бочавер, существует также проект сотрудничества в этом направлении с университетом Гранады (Р. Гусман Тирадо). Продолжается работа над развитием итальянского корпуса с участием Католического университета Милана (А. Бонола и В. Ноземда) и Болонского университета (Ф. Бьяджини, Кс. Д. Балакина), в частности, корпус включает теперь поливариантные итальянско-русские тексты. К моменту выхода этой статьи из печати объём итальянского корпуса превысит 4 миллиона словоупотреблений (русская классическая литература и философия XIX–XX веков, итальянские пьесы, новеллы и повести XVIII–XIX веков).

1.6. Латышский корпус

В корпус впервые включены выровненные латышско-русские тексты, над которыми работала Н. В. Перкова. Переводы с русского на латышский пока немногочисленны — «Белая гвардия» М. А. Булгакова и рассказы Чехова; среди латышских текстов — как классическая проза (Я. Райнис, Р. Блауманис), так и литература 1970–1980-х годов, переводившаяся в советское время. Совокупный объём латышско-русского и русско-латышского компонентов — 700 тыс. словоупотреблений. Латышские тексты при их размещении были не снабжены морфологической разметкой и доступны только для текстового поиска. Отметим, что даже в Латвии задача разработки латышского морфологического анализатора до сих пор далека от удачного решения.

1.7. Немецкий корпус

В последние годы благодаря группе под руководством Д. О. Добровольского и А. А. Кретова (о которой см. раздел 1.1) в корпусе появилась русско-немецкая языковая пара, в основном состоящая из культурно значимых переводов русской классики («Капитанская дочка», «Герой нашего времени», «Война и мир» полностью, романы Достоевского и др.). Таким образом, имевшаяся в предшествующие годы асимметрия (значительный по объему англо-русский корпус, относительно небольшие русско-английский и немецко-русский корпуса при полном отсутствии русско-немецкого компонента) постепенно выравнивается. Пополнялась новыми текстами и существовавшая ранее немецко-русская языковая пара.

В корпус включены предоставленные компанией ABBYY нехудожественные тексты, переведённые с немецкого и на немецкий: производственно-технические (инструкции), учебные (санитарно-просветительские тексты), публицистика. Объем немецкого корпуса равен 7,6 млн словоформ.

Предусмотрено развитие поливариантного русско-немецкого корпуса; о существующем опыте в этом направлении см. вошедшую в сборник НКРЯ 2006–2008 статью Д. О. Добровольского. Развивается и отдельный подкорпус «Русская классика в немецких переводах» при финансовой поддержке Фонда содействия развитию интернета «Фонд поддержки интернет».

1.8. Польский корпус

Польский корпус разрабатывался в ходе совместного проекта с Варшавским университетом (руководитель с польской стороны — М. А. Лазиньский; в работе над корпусом и его программным обеспечением принимали участие П. и Н. Годлевские, Е. В. Гурина, В. И. Демидова, Е. Д. Ильина, М. Куратчик, С. О. Минлос, Г. А. Мороз, Б. В. Орехов, Е. Слободян (Выналек) и другие). Это одна из самых больших языковых пар — свыше 6,3 миллионов словоупотреблений.

Здесь хорошо представлена польская художественная проза разных периодов, от Сенкевича и Пруса до Хмелевской и Сапковского. В корпус входят также газетные статьи, законодательные тексты.

Русско-польский параллельный корпус пока не вполне репрезентативен, но в него входит, в частности, такой нестандартный для параллельных корпусов материал, как проза русского Серебряного века (Брюсов и Сологуб) в современных польских переводах.

Польский корпус размечен при помощи морфологического анализатора TaKiPi с автоматическим снятием омонимии.

1.9. Украинский корпус

Украинский корпус (работу над ним ведет в основном М. А. Шведова; в выравнивании участвовали также А. Л. Кривенко и О. А. Тищенко-Монастырская) — одна из крупнейших языковых пар в НКРЯ, вторая по объёму после английского корпуса (см. публикацию [Січінава, Тищенко-Монастирська, Шведова 2011]). Его совокупный объём составляет теперь 9,3 млн словоупотреблений. В корпусе достаточно репрезентативно собрана украинская проза, переводившаяся в разные эпохи на русский язык (от Котляревского до Жадана), представлены также выполненные в разные эпохи и с разными установками переводы с русского языка на украинский. Особо укажем на такой жанр, как украинская (и русская, переводившаяся на украинский) фантастика и приключенческая проза XX в. В украинский корпусактивно включаются нехудожественные тексты: научные, правовые, публицистические, религиозные. Контекст создания украинского корпуса несколько напоминает ситуацию с белорусским: в ситуации отсутствия большого общедоступного одноязычного корпуса украинского языка параллельный корпус в известной степени на первых порах перенимает его функции. В настоящее время вне рамок НКРЯ развиваются также украинско-польский (Н. Коцыба) и украинско-болгарский (Е. Б. Сирук, И. А. Держанский) параллельные корпуса.

1.10. Французский корпус

В формировании русско-французского компонента параллельного корпуса определяющую роль играет его поливариантный состав (подробнее см. ниже, раздел 3). Поливариантный русско-французский корпус поддержан специальным грантом (РФФИ № 12–06–33038 «Контрастивные корпусные исследования русских и фран-

цузских глагольных категорий в поливариантных параллельных текстах», руководитель Д. В. Сичинава). Основным вдохновителем и автором концепции корпуса является И. М. Зацман (ИПИ РАН), в работе над ним участвуют Н. В. Бунтман, М. Г. Кружков, Е. Ю. Лощилова, О. А. Петрушкина, Е. А. Роганова, Анна А. Зализняк и другие. На базе поливариантного корпуса построена уникальная французско-русская база данных соответствий грамматических категорий (см. подробнее ниже и публикации Loiseau et al. 2013, Бунтман и др. 2014).

В русско-французский корпус включается несколько альтернативных переводов русской классики (Гоголь, Чехов, Гончаров, Толстой) на французский язык; поливариантный состав позволяет исследовать вариативность в переводе лексики и грамматики в идентичных контекстах. Повести Гоголя «Нос» и «Шинель» в настоящее время включены в корпус в четырёх разных переводах. Наряду с поливариантными текстами в корпус включаются также и традиционные моновариантные пары (в том числе субтитры художественных фильмов). Эти моновариантные тексты выровнены Венсаном Бене (Париж, INALCO). В проекте с французской стороны участвует также Сильвен Луазо (университет Paris-XIII).

Французско-русский корпус включает моновариантные (с одним переводом) тексты, как художественные (Мопассан, Бальзак, Флобер, Госинни, Модиано), так и нехудожественные, включая предоставленные компанией ABBYY официально-деловые и технические тексты.

Совокупный объём французского компонента — 2,3 млн словоупотреблений.

1.11. Эстонский корпус

С 2015 г. в корпус входят также параллельные эстонско-русские и русско-эстонские тексты. Особая сложность для подготовки корпуса связана с плохой представленностью в Сети и других источниках электронных версий эстонской прозы (оригинальной и переводной); велась работа по сканированию и распознаванию бумажных источников. Особое влияние планируется уделить переводам «эстонского текста» русской литературы (Довлатов, Аксенов и др.) Над корпусом работали М. В. Боровикова и Е. С. Фомина (Тартуский университет).

В морфологической разметке эстонских текстов используется анализатор vabamorph, разработанный в Тартуском университете Х. Калепом и К. Муйшнек. Объем эстонского корпуса — более 400 тыс. словоупотреблений.

2. Многоязычный корпус

Популярным направлением развития многоязычных корпусов является создание так называемых «массовых параллельных текстов» [Cysouw, Wälchli 2007], включающих большое количество (не менее 10–20, а для таких массово распространяемых текстов, как Библия, и сотни) выровненных переводов некоторого текста на различные языки. Такие корпуса включают религиозные, политические, классические художественные, правовые тексты (ср. известный корпус Acquis communautaire — переводящихся на языки стран-членов ЕС правовых документов Европы) и т. п.

НКРЯ постоянно сотрудничает с построенными на основе схожей методологии (в том числе на принципе общедоступности материалов для научной работы) параллельными корпусами славянских языков, такими, как ParaSOL (ранее Regensburg Parallel Corpus, http://www-korpus.uni-r.de/ParaSol/, см. также [von Waldenfels 2006]), ASPAC (Amsterdam Slavic Parallel Aligned Corpus, авторский проект А. Барентсена, http://home.medewerker.uva.nl/a.a.barentsen/), InterCorp в рамках Чешского национального корпуса (http://www.korpus.cz/intercorp/).

В настоящее время в состав НКРЯ входит многоязычный корпус из 9 текстов («Алиса в стране чудес» и «Алиса в Зазеркалье» Л. Кэррола, «Собака Баскервилей» А. Конан Дойля, «Винни-Пух» А. А. Милна, «Код да Винчи» Д. Брауна, «Маленький принц» А. де Сент-Экзюпери, «Пиноккио» К. Коллоди, «Алхимик» П. Коэльо, «Мастер и Маргарита» М. Булгакова). Каждый из этих текстов представлен в оригиналах и в переводах (от 6 до 25 переводов, в основном на славянские языки). В некоторых случаях включено несколько переводов на один язык.

В многоязычном корпусе представлены следующие славянские языки: русский, украинский, белорусский, польский, чешский, словацкий, верхнелужицкий, словенский, хорватский, сербский, болгарский, македонский. Из неславянских языков в него входят тексты

на английском, немецком, нидерландском, шведском, итальянском, португальском, румынском, французском, греческом, латышском и литовском языках. Для части из этих языков НКРЯ поддерживает морфологический поиск (основывающийся на морфологических словарях компании «Яндекс»).

В основе многоязычного корпуса НКРЯ лежат тексты корпуса ASPAC, любезно предоставленные А. Барентсеном, и затем пополненные некоторыми текстами (например, переводами «Винни-Пуха» на балтийские языки). Кроме того, в проекте ASPAC выравнивание текстов проведено поабзацно; для нужд НКРЯ оно полуавтоматически заменено на выравнивание по предложениям.

На базе многоязычных корпусов, в частности, возможно массовое типологическое сравнение грамматических параметров языков [Dahl 2014], [von Waldenfels 2014], см. о перфекте ниже, раздел 7.

3. Поливариантный корпус

Со второй половины 2012 года команда НКРЯ совместно с исследователями Института проблем информатики РАН и французскими лингвистами (университет Paris-13, Институт восточных языков INALCO) разрабатывает поливариантный параллельный русско-французский корпус (см. подробнее [Loiseau et al. 2013, Бунтман и др. 2014]). Одновременно на его материале строится база данных поливариантных соответствий (полиэквиваленций) глагольных аспектуальных и временных форм в русском и французском языках (http://a179.ipi.ac.ru/corpora_dynasty/main.aspx).

Поливариантные параллельные корпуса — сравнительно слабо развитое направление в рамках корпусной лингвистики, и тем меньше существует собственно сопоставительных грамматических исследований, выполненных на их базе. Существует поливариантный русско-немецкий корпус на материале выполненных в разное время переводов романов Достоевского, созданный в Австрийской академии наук в Вене [Добровольский 2009], но он не находится в открытом доступе. Он используется прежде всего для исследования лексики и эволюции переводческих техник. По нескольку вариантов перевода ряда текстов на один и тот же язык (например, «Алиса в стране чудес» — шесть переводов на русский и четыре на польский) находится в ряде многоязычных параллельных корпусов, на-

пример, в Амстердамском корпусе параллельных текстов (ASPAC) и Регенсбургско-Бернском параллельном корпусе.

Для поливариантного корпуса выбираются тексты, существующие не менее, чем в двух переводах на французский язык, причём эти переводы должны быть созданы, как правило, не ранее середины XX века. Более ранние французские переводы русской литературы, особенно выполненные в XIX веке, устарели в языковом отношении, а главное, содержат много ошибок и сокращений исходного текста.

Для многих произведений русской литературы (Н. В. Гоголь, И. А. Гончаров, Л. Н. Толстой) существует несколько новых переводов, выполненных на высоком уровне. Они демонстрируют как различные подходы к художественному переводу, так и различный выбор языковых выражений, в том числе грамматических.

Пример выравнивания оригинального текста с двумя переводами («Обломов» И. А. Гончарова):

Цвет лица у Ильи Ильича не был ни румяный, ни смуглый, ни положительно бледный, а безразличный или казался таким, может быть, потому, что Обломов как-то обрюзг не по летам: от недостатка ли движения или воздуха, а может быть, того и другого.	Le teint d'Ilia Ilitch n'était ni rose, ni mat, pas véritablement pâle non plus; neutre plutôt, ou du moins ayant cette apparence, et cela peut-être parce qu'Oblomov était prématurément un peu flasque, faute d'exercice, ou faute d'air, ou faute d'exercice et faute d'air à la fois.	Le teint d'Ilia Ilitch n'était ni rose, ni hâlé, ni carrément pâle, mais indifférent ou, du moins, il le paraissait. Peut-être parce que la chair d'Oblomov était prématurément flasque: faute d'exercice ou manque d'air, peut-être l'un et l'autre.

На базе поливариантного корпуса построена контрастивная русско-французская база данных по соответствиям грамматических категорий. В базе данных эквивалентных глагольных форм объектом анализа являются «соответствия», т. е. пары функционально эквивалентных форм русского и французского языков. На материале русско-французского параллельного корпуса речь может идти о соответствиях двух типов. Это:

• «модели перевода» — множество переводов {Fn… Fn+m} для русской формы R (например, НСВ переводится как present, imparfait или в известных ограничительных контекстах passé simple);

• «стимулы перевода» — множество «стимулов» {Rn… Rn+m}, «реакцией» на которые является французская форма F (например, passé antérieur может появиться для русских слов «немедленно» или «внезапно»).

Соответствия первого типа позволяют получить новые сведения об устройстве русского языка, второго — об устройстве французского. Поскольку объектом нашего интереса является русский язык, в рамках данного проекта рассматриваются лишь соответствия в направлении «русский => французский», т. е. «модели перевода».

Второе ограничение состоит в том, что в создаваемую базу данных функционально эквивалентных лексико-грамматических глагольных форм русского и французского языков включаются только те глагольные формы русского языка, которые содержат спрягаемый глагол — verbum finitum (т. е. исключаются слова категории состояния, перифразы с глаголом «быть» и некоторые другие типы предикативных единиц).

Когда ведение базы будет отработано на тестовом материале, она будет расширена. Последующее ее расширение возможно также в направлении включения данных, полученных из параллельного французско-русского корпуса; они позволят ответить на вопрос «На какие французские «стимулы» появляется в переводе в качестве «реакции» данная русская форма?».

В рамках данного проекта рассматривается кластер глагольных категорий, который в типологической литературе получил обозначение TAM (Tense-Aspect-Mood; ср. также типологические исследования языков Европы в рамках комплексного проекта EUROTYP).

Вводится понятие «лексико-грамматической формы», которое представляет собой комбинацию значений трех категорий для русского языка (Tense, Aspect, Mood) и двух (Tense, Mood) — для французского (см. [Loiseau et al. 2001]). Понятие «лексико-грамматической формы» (ЛГФ) близко по своему содержанию к понятию «конструкции» — в том значении, которое придается этому термину в Грамматике конструкций [Goldberg 1995, 2006], т. е. это может быть как набор значений грамматических и семантических категорий, задающий класс удовлетворяющих данным требованиям реальных форм (т. е. «грамматическая конструкция» в традиционном понимании), так и единичная лексема или словоформа. Понятие лексико-грамматической формы представляется

наиболее адекватным инструментом контрастивного анализа поливариантного параллельного корпуса.

4. Расширение жанрового многообразия корпуса

В параллельный корпус в настоящее время активно включаются нехудожественные тексты. Преимущество их использования — в меньшей степени переводческой вольности на всех уровнях передачи текста по сравнению с художественным переводом, в лучшей представленности в текстах современной лексики, терминологии. К сложностям следует отнести неопределенность в ряде случаев языка оригинала и языка перевода (например, в двуязычных инструкциях к лекарствам, текстах Интернет-СМИ и т. п. о направлении перевода можно строить только более или менее правдоподобные предположения), а также сомнительную лингвистическую корректность использования текстов, возникших в результате вручную отредактированного машинного перевода (это не редкость, например, в двуязычных украинских СМИ; однозначной диагностике такие случаи не поддаются). Существенное подспорье Корпусу оказал проект ABBYY Lingvo, с любезного согласия которого в английский, немецкий и французский корпуса были включены нехудожественные тексты разных жанров.

Сейчас в параллельных корпусах 3,2 млн словоупотреблений нехудожественных текстов (4,6% совокупного объёма). Уже представлены те же основные сферы функционирования текстов, что и в основном одноязычном русском корпусе: официально-деловая, обиходно-бытовая, производственно-техническая, публицистика, религиозная, учебно-научная. В различных языковых парах они представлены неравномерно; в «малых» парах (армянском, испанском, латышском, эстонском корпусах) нехудожественных текстов пока нет. Больше всего удельный вес нехудожественных текстов в белорусском (9%) и в украинском (11%) корпусах.

Здесь же можно упомянуть об использовании «неканоничной» художественной литературы: литературно обработанного фольклора (народных сказок), массовых жанров вроде детектива или фантастики. Художественные тексты таких жанров пополняют также прежде всего украинский и белорусский корпуса.

5. Средства разметки

5.1. Выравнивание

Для параллельных корпусов НКРЯ были разработаны специальные программные средства выравнивания и разметки. Команда, работающая над английским и немецким корпусом, использует программу «ПарТекс» (идея А. А. Кретова, руководство И. Е. Ворониной, программирование Д. Спесивцева), имеющую на входе два параллельных текста (оригинал и перевод). Выравнивание осуществляется на уровне предложений. Программа выдает на выходе синтезированный текст, в котором последовательно за каждым предложением оригинала следует соответствующее предложение перевода. В таком синтезированном тексте поиск интересующих пользователя слов может осуществляться штатными средствами обычных текстовых редакторов. В программе «ПарТекс» для поиска может быть «подстрока в строке» и на выход подается текстовый файл, в который входят все пары предложений оригинала и перевода, содержащие заданную последовательность символов. Возможен поиск как английских или немецких, так и русских слов или словосочетаний.

Одна из наиболее существенных трудностей выравнивания заключается в том, что авторское членение текста на предложения и абзацы не всегда выдерживается в тексте перевода. Кроме того, в разных языках (а иногда и разных изданиях) приняты различные способы графического оформления, что иногда затрудняет определение границ предложения в автоматическом режиме. Ср., например, различные способы оформления переходов от прямой речи персонажей к авторским ремаркам. В таких случаях результаты автоматического выравнивания нуждаются в коррекции, осуществляемой вручную. Программа позволяет обнаружить в параллельных текстах асимметрию такого рода. Результаты работы программы ПарТекс далее просматриваются и дополнительно корректируются вручную (Д. О. Добровольским и Д. В. Сичинавой, в последние годы также М. А. Хохловой из СПбГУ / ИЛИ РАН).

Для выравнивания остальных параллельных корпусов (кроме английского и немецкого) используется программа «Евклид» (идея Д. В. Сичинавы, программист Т. А. Архангельский, учитывались за-

мечания и предложения разметчиков из разных стран, прежде всего В. А. Кощенко [Беларусь] и М. А. Шведовой [Украина]). Эта программа является графическим интерфейсом пользователя (GUI) к доступной в открытом режиме программе выравнивания текстов HunAlign [Varga et al. 2005].

Входящий стандарт программы — файл в текстовом формате (ANSI или Юникод) с любым числом абзацев, разбиений строки, пробелов в начале строк и т. д. По запросу пользователя загружается несколько текстов — оригинал и перевод(ы), после чего они проходят автоматическую предобработку: удаляются пустые строки, лишние пробелы, каждое предложение переносится на новую строку, при этом учитываются основные сочетания знаков препинания и типы границы предложения.

Затем запускается программа HunAlign, при помощи статистического механизма (информация Гейла-Чёрча о длине предложений и автоматическое составление словаря на её базе) автоматически выравнивающая тексты попарно полностью, и выход её загружается в графический интерфейс программы «Евклид» для ручного постредактирования.

Программа HunAlign, основываясь на ряде статистических весов (длина предложения, совпадение символов, структура знаков препинания и т. п.) приписывает каждой выровненной ею паре предложений определённый коэффициент вероятности выравнивания. Если он ниже нуля, значит, выравнивание данных отрезков текста маловероятно. Кроме того, она автоматически склеивает последовательности предложений на одном языке, соответствующие, с его точки зрения, друг другу. Места склейки в выходном формате при этом указываются.

Пары предложений с отрицательным коэффициентом выравнивания и все склеенные последовательности предложений считаются сомнительными отрезками, нуждающимися в первоочередной проверке вручную. Для графической чёткости коэффициенты в сомнительных отрезках выделяются в программе цветом. При помощи специальной опции в программе «Евклид» скрываются все предложения, кроме сомнительных отрезков. Они просматриваются редактором и в случае ошибки выравнивания (а также в случае сохранения в файле лишней информации — например, записанных в текстовом виде выходных данных и т. п.) редактируются вручную

при помощи быстрых механизмов программы. Предусмотрены следующие автоматизированные операции (в виде графических кнопок при каждой строке, некоторые из которых открывают отдельное диалоговое окно):

добавление пустой строки;

удаление пустой строки;

перенос предложения в соседнюю строку;

перенос части склеенного предложения — т. н. «обрезка» — в соседнюю строку.

При необходимости ближайший контекст сомнительных отрезков может быть раскрыт (двойным щелчком по скрытой строке) без раскрывания всего остального текста. Затем для просмотра и редактирования можно сделать доступным весь остальной текст (без сомнительных отрезков) и отредактировать его по тому же принципу.

5.2. Формат

Выровненный текст можно сохранить в формате XML (кодировка Юникод), как в окончательном корпусном, так и в промежуточном формате, доступном для дальнейшего редактирования. При сохранении выровненного текста следует указать язык оригинала и перевода. Они будут сохранены в XML в составе соответствующих тегов. В сохранённом XML пары предложений пронумерованы (им сопоставлены уникальные номера) для облегчения ссылок.

Программа «Евклид» позволяет также добавить к каждому сохранённому тексту метатекстовую информацию (метаразметку), которая записывается в качестве отдельной строки в электронную таблицу (документ, доступный для редактирования программами типа Excel). Пользователь заполняет в специальной форме название текста и имя автора в оригинале и переводе, дату создания текста; язык оригинала и перевода сохраняются автоматически исходя из ранее сохранённой информации.

Выровненные пары предложений объединяются при помощи XML-тега <para>. Тег имеет атрибут ID, в который записывается номер предложения.

Предложения объединяются при помощи XML-тега <se>. Тег имеет обязательный атрибут lang, где указан язык текста, например, <se lang=ru> для русского или <se lang=be> для белорусского.

5.3. Метаинформация

Метаинформация сохраняется в отдельном файле (электронной таблице) в формате CSV (значения ячеек разделены при помощи знака «точка с запятой»). Таблица заполняется при помощи соответствующей формы программы «Евклид». Таблица включает в себя следующие поля метаинформации:

1) название текста в оригинале;
2) год создания текста;
3) имя автора в оригинале;
4) год рождения автора;
5) название текста в переводе;
6) имя автора в переводе;
7) имя переводчика (на языке перевода);
8) дата перевода;
9) сфера функционирования текста (художественная, публицистика...);
10) язык оригинала;
11) язык перевода.

5.4. Маркирование вольности перевода

Важно отметить, что художественный перевод советского времени с русского языка на языки народов СССР и в обратном направлении (а именно переводы этой эпохи лежат в основе нынешнего состава украинско-русского и белорусско-русского параллельных корпусов) отмечен высокой степенью вольности на разных уровнях, и при составлении и анализе параллельного корпуса это постоянно приходится учитывать. Причины этого разнообразны — общая установка ряда переводчиков, роль автора, участвующего в творческом пересоздании авторизованного перевода (в частности, он может и сам выступать переводчиком), а нередко и цензурные причины. Большое значение имеет корпус для исследований по теории и практике перевода, например, по передаче заглавий, идиом, различного рода идеологем (так, анализ переводов советского времени с украинского на русский язык показывает, что в них систематически различными переводчиками устранялась религиозная фразеология и оскорбительная в современном языке [но далеко не всегда для

авторов XIX в.] этнонимия — вместо неба крестьяне апеллировали к справедливости, а *жид* и *москаль* достаточно регулярно представали, к примеру, *ростовщиком* и *солдатом*). Вольности и ошибки различного уровня фиксируются и в переводах с участием других языков, хотя и не в таких масштабах, но также систематически.

Для указания наличия той или иной вольности в переводе,— пропусков, добавлений, замены текста с изменением смысла,— в корпусе используется специальный уровень разметки — атрибут **loose**. Значения атрибута следующие:

Добавление: <se lang=rus loose=add>
Пропуск: <se lang=rus loose=omit>
Изменение: <se lang=rus loose=change>

Пример из украинского рассказа «Невольница» Марко Вовчок (автоперевод); выделены добавленные в переводе фрагменты:

<para id="84">
<se lang="uk">Ще гаразд сонечко не підбилося у небі, вже скрізь взброювалися як спроможність, коней виводили, сідлали,— в кого чого не було — біг той питаючи, шукаючи; козачки допомагали й кохали як мога; діти клопоталися, начеб їм на турка перед вести йшлося…</se>
<se lang="ru" loose="add">Еще солнышко хорошенько не взошло на небе, уж повсюду в городе вооружались чем кому бог послал, всюду коней выводили, седлали — кому чего недоставало, тот бежал, спрашивая, ища, отыскивая; козачки помогали, как только в силах были; *надо сказать, что* они и любили тогда, как только в силах были: *не раз вы, верно, деточки, замечали сами, что вас с большей нежностью гладили по головке, когда вы, по детской неосмотрительности, падали, например, в грязь — каково ж должно быть чувство к тому, кто идет честно на смертный бой?* Дети хлопотали и *суетились*, словно им приходилось самим в передовом отряде на турок идти…</se>
</para>

Пример из «Смерти Ивана Ильича» Л. Н. Толстого во французском переводе М. Оффмана; выделен опущенный в переводе фрагмент.

<para id="353">
<se lang="ru">*Было лучшее общество, и* Иван Ильич танцевал с княгинею Труфоновой, сестрою той, которая известна учреждением общества «Унеси ты мое горе».</se>
<se lang="fr" loose="omit">Ivan Illitch dansa avec la princesse Troufonova, la propre sœur de l'illustre fondatrice de la société « Foin de peines! ».</se>
</para>

Пример из «Похищения чародея» К. Булычева в белорусском переводе Р. Саматыи; выделены изменённые и добавленные фрагменты:

<para id="90">
<se lang="ru" variant_id="0" loose="change">*Речь шла* об опустении рек и лесов, о том, что некий купец еще до революции возил с холма камень в Полоцк, чем обкрадывал культурное наследие, о том, что население этих мест смешанное, потому что сюда все кому не лень ходили, о том, что каждой деревне нужен музей…</se>
<se lang="be" variant_id="1" loose="add, change">*Яго доўгі маналог* быў пра горкі лёс рэк і лясоў; пра тое, што якісьці купец яшчэ да рэвалюцыі вазіў з пагорка камень у Полацак, абкрадаючы *гэтым самым* культурную спадчыну; пра тое, што насельніцтва гэтых мясцін мяшанае, бо сюды прыходзілі *ваяваць усе, хто хацеў*; што кожнай вёсцы патрэбны музей…</se>
</para>

В рамках поливариантного корпуса разработана более тонкая разметка неточных соответствий перевода (см. подробнее Сичинава 2013, Loiseau et al. 2013).

6. Пример использования параллельного корпуса для количественного изучения лингвоспецифичной лексики[1]

Использование параллельных корпусов для сравнительных исследований — предмет уже обширной исследовательской литера-

[1] Данное исследование поддерживалось грантом РФФИ № 13-06-00403 «Контрастивное корпусное исследование специфических черт семантической системы русского языка» (руководитель Анна А. Зализняк).

туры, назовём, прежде всего, специальный номер журнала STUF под редакцией М. Сисоу и Б. Вельхли [Cysouw, Wälchli (eds.) 2007]. При переводе лексики, как и грамматических конструкций (чёткую границу между ними не всегда можно провести), существенную роль играют **модели перевода** и **стимулы перевода**. Обычно для конструкции А в языке В имеется (реально представлено в текстах) несколько неслучайно повторяющихся соответствий — моделей перевода (translation patterns; ср. сборник [Hasselgård, Oksefjell (eds.) 1999]. Практически можно утверждать, что, имея дело с переводными текстами и располагая сколько-либо большим корпусом, мы наверняка можем выделить несколько моделей перевода конструкции языка А в языке В. Стимулом перевода назовем конструкции в языке оригинальных текстов, соответствующие определённой конструкции в языке перевода («что стимулирует появление конструкции X в переводе на язык В?»).

Параллельный корпус может использоваться для точного определения и выявления так называемой лингвоспецифичной лексики (единиц, как нередко утверждается, т. н. «языковой картины мира»), которая «трудно» или особенно неоднозначно переводится на другие языки (ср. [Добровольский 2009]). Вокруг этого и схожих понятий, как известно, ведётся оживлённая дискуссия (ср. резко критикующие любые исследования «языковой картины мира» публикации, собранные в книге [Павлова (сост.) 2013], а также взвешенный обзор дискуссии в [Руссо 2014]). Представляется, что некоторая попытка точных формализуемых определений здесь была бы как минимум не лишней.

Возможных мер «разброса различных переводов» некоторой лексемы можно предложить несколько, в частности:

отношение абсолютной частоты самой частотной модели перевода (F (Mmax)) к количеству различных эквивалентов (NumM);

средняя частота вхождений на один эквивалент (F (O)/NumM, где F (O) — частотность данного слова в оригинале);

отношение абсолютной частоты самой частотной модели перевода к частоте второй (F (Mmax)/F (Msec));

отношение абсолютной частоты самой частотной модели перевода к общему количеству вхождений (F (Mmax)/F (O)).

Наконец, существуют несколько самых общих статистических мер разброса (diversity indexes), из которых мы используем две:

энтропию дискретного распределения относительной частотности каждой модели перевода ($-\sum (F (M_i)/F (O))\log2 (F (M_i)/F (O))$ и более простой индекс Герфиндаля (имеющий также ряд других названий), первоначально использовавшийся для оценки уровня монополизма в экономике ($\sum (F (M_i)/F (O))^2$; минимальное значение: 1/NumM, максимальное 1[1]). Высокое значение энтропии соответствует низкому индексу Герфиндаля и большому разбросу значений.

Для «лингвоспецифичного» слова (если таковые существуют) предполагается, что моделей перевода будет много, в среднем на каждую будет приходиться сравнительно немного контекстов, а частота самой частотной из них не будет сильно отличаться от остальных (и он будет занимать лишь небольшой процент от общего числа соответствий). Энтропия будет достаточно велика, а индекс Герфиндаля — низок.

Данные показатели имеет смысл вычислять также для стимулов перевода (в этом случае предполагаемая лингвоспецифичная лексика появляется при переводе; соответственно вместо буквы M выше выступает S, а вместо частотности в оригинале F (O) — частотность в переводе F (T))

Кроме того, имеет значение отношение частоты лексемы (на миллион слов) в оригинальных текстах на языке L (F (O)ipm) к частоте ее же (на миллион слов) в переводных текстах на этом же языке L (F (T)ipm). Ожидается, что лингвоспецифичная лексика скорее появится при создании оригинального текста, чем при переводе.

Данные показатели можно получить и сравнить для нескольких языков (например, для лексики русского языка использовать русско-английский, русско-украинский и другие корпуса).

Проверим эти предположения на материале нескольких лексем, отнесенным в [Зализняк, Левонтина, Шмелев 2012] к лингвоспецифичным, и нескольких лексем, которые такими обычно не считаются.

Однокоренные модели перевода обычно засчитываются как одна модель (bore, boredom, boring, bored); то же относится к случаям, когда в некоторых вхождениях имеются также другие слова, например, recklessness приравнивается к reckless jockeying. Повторяющиеся в одном тексте идиоматизмы (фразеологизмы типа *заливать то-*

[1] Существует также нормализованный индекс Герфиндаля, независимо от объёма множества принимающий значения от 0 до 1: $(\sum(F(Mi) /F(O)) — 1/NumM)/(1-1/NumM)$.

ску, прозвища типа *Смертная Тоска*), внутренние цитаты и т. п. явно зависимые вхождения считаются только один раз. Случаи пропуска слова в переводе или, напротив, привнесения переводчиком («отсебятины»: *Fredrica's cat watched her from the high window* / *Из окна под самой крышей дома за ней* ***с тоской в глазах*** *наблюдал кот Фредерики*; *He was propped behind the wheel of his shiny yellow Lincoln, talking about his grandmother* / *Он сидел за рулем своего сверкающего желтого «Линкольна» и,* ***уставившись куда-то в пространство остекленевшими глазами****, о чем-то разговаривал со* [sic! ср. about] *своей бабушкой*) не учитываются. Иногда неясно, какое слово соответствует какому или же перевод «распределен» по нескольким; например, во фразе *И как чудна она сама, эта дорога: ясный день, осенние листья, холодный воздух… покрепче в дорожную шинель, шапку на уши, тесней и уютней прижмемся к углу!* [Н. В. Гоголь. Мертвые души (1835–1852)] // *and how interesting for its own sake is a highway, should the day be a fine one (though chilly) in mellowing autumn, press closer your travelling cloak, and draw down your cap over your ears, and snuggle cosily, comfortably into a corner of the britchka.* [Nikolay Gogol. Dead Souls]. Все три корня *snug*, *cosy* и *comfort* выступают в других текстах для передачи слова *уют(ный)*.

Если сравнить лексемы *удаль*, *тоска* и *страсть* (последнее в значении «желание, стремление, чувство», а не омоним со значением «страх»), то оказывается, что у первых двух очень сильный разброс лексических соответствий в параллельных английских текстах. В оригинальных текстах слово *удаль* встретилось 4 раза и переведено каждый раз по-новому, соответственно, значения F (Mmax)/F (Msec) = F (O)/NumM = 1 (минимально возможное значение); F (Mmax)/NumM = F (Mmax)/F (O) = 0,25. В переводных текстах слово удаль встретилось 11 раз, 2 стимула встретились дважды, 9 по одному. Соответственно, это даёт F (Smax)/NumS = 0,22, F (Smax)/F (T) = 18%, F (Smax/Ssec) = 1, F (T)/NumS =1,2. В оригинальных текстах (на миллион слов) слово удаль встречается в 1,6 раз чаще, чем в переводных. Показатель энтропии H для моделей перевода — 2, для стимулов перевода — 3,4. Индекс Герфиндаля для моделей перевода — 0,25 (минимально возможный для 4 значений; нормализованный тем самым равен 0), для стимулов перевода — 0,14 (нормализованный 0,05).

Более частотная лексема *тоска* переводится на английский

22 различными способами и соответствует 66 различным английским корням в оригинале (различие связано, в частности, с тем, что англо-русский корпус в 4,6 раза больше русско-английского). Этот огромный список включает, в частности, новые заимствования: *ennui* из французского и *angst* из немецкого, что может уже косвенно указывать на специфику русского слова. При переводе на английский, благодаря индивидуальным предпочтениям переводчиков, выделяются два слова: *anguish* (его предпочитают переводчики Булгакова Пивир и Волохонская) и *misery* (его предпочитает переводчица Чехова К. Гарнетт; в частности, существует одноименный рассказ Чехова, где заглавное слово *тоска*/misery встречается семь раз). При том, что средняя частота соответствия, естественно, выше, чем у редкого слова *удаль* (3,2), выше и отношение частоты лидирующего варианта к количеству остальных (0,72), остальные показатели близки: F (Mmax)/F (Msec) = 1, F (Mmax)/F (O) = 23%. Для стимулов перевода, где лидерство — за основой long-, а на втором месте — предлагаемое А. Вежбицкой в качестве основного эквивалента yearning и его производные, разрыв между двумя местами больше, чем в русско-английском: F (Smax/Ssec) = 1,92, очень близкая средняя частота соответствия (3,02) и гораздо более низкий рейтинг лидирующего перевода — 13%. Показатели энтропии вдвое ниже, чем у *удаль* (для моделей перевода — 1,08, для стимулов — 1,6), зато ниже и индекс Герфиндаля: 0,13 и 0,04 (нормализованный 0,12 и 0,03) соответственно. Слово *тоска*, как и слово *удаль*, встречается в оригинальных текстах в 1,6 раз чаще, чем в переводных.

Совсем иная картина открывается при анализе профиля слова *страсть*, примерно столь же частотного в русском тексте (26 на миллион), как *тоска* (24 на миллион). В 85% случаев это слово переводится с русского как passion, разброс стимулов больше — на самый частотный стимул (разумеется, тоже passion) приходится 34% случаев. Дело в том, что слово *страсть* несколько чаще встречается в переводных текстах — в 1,3 раза; оно обслуживает, в частности, такую английскую лексику, как lust (‘~похоть’), obsession (‘~одержимость’) и другие. Отметим, что общие статистические показатели демонстрируют резкое различие между оригинальными и переводными текстами: показатель энтропии 0,31 для моделей, но достаточно высокий (0,99) для стимулов; индекс Герфиндаля 0,73 для моделей и 0,16 (нормализованный 0,14) для стимулов. Та-

ким образом, слово *страсть* ведёт себя (судя по моделям перевода в русском оригинале) не похоже на лингвоспецифичное слово, но (как показывают стимулы русского перевода) вместе с тем является общим термином (umbrella term) для целого ряда английских наименований чувств.

Интересный материал представляют родственные квазисинонимы *простор* и *пространство*, первый из которых нередко считается знаком русской «языковой картины мира», а второй — как правило, нет. Слово *простор* неоднократно привносится переводчиком как часть речения (сложного предлога?) *на просторах, в просторе*: in the ocean — *на просторах* океана; такие случаи, в соответствии с общей нашей методикой, при подсчётах не учитывались, но отметить их заметное количество необходимо. Обнаруживается значительный разброс моделей перевода слова *простор* (самый частотный перевод, space и его производные — 21% вхождений, средняя частота переводов — 1,5 слова, энтропия 0,92, нормализованный индекс Герфиндаля 0,07); еще сильнее это выражено для стимулов — самый частотный перевод 19%, средняя частота 2,6, энтропия 1,34, нормализованный индекс Герфиндаля 0,06. Слово *пространство* имеет существенно более низкий разброс соответствий: самый частотный перевод (тоже space) — 55% среди моделей и 66% среди стимулов; частота среднего перевода — 3,8 среди моделей и 4,6 среди стимулов; энтропия 0,72 среди моделей и 0,63 среди стимулов, нормализованный индекс Герфиндаля соответственно 0,32 и 0.44. При этом оба слова чаще встретились в переводных текстах, чем в оригинальных: *пространство* — примерно в два раза, *простор*, что отчасти предсказуемо, лишь в 1,3 раза.

Рассмотрим также еще два слова, часто считающиеся лингвоспецифичными: *уютный/уют* и *пошлый/пошлость*. Показатели слова *уют* напоминают показатели слова *страсть*: переводчики с русского используют резко ограниченный круг соответствий (cozy/cosy, comfort и snug — первое из них характерно для переводчицы Стругацких А. Буис и уже упомянутых Пивира и Волохонской, последнее — для К. Гарнет), в то время как переводчики с английского привносят слово *уют* как общий umbrella term для положительных характеристик «атмосферы» такого рода, в результате чего оно почти вдвое чаще встречается в переводных текстах. Лингвоспецифичность тут если и есть, то особого рода, и зависит от направления

перевода. Действительно, на невыразительные показатели для моделей перевода (37% для самой частотной модели, средняя частотность модели 7,5, энтропия 0,52, нормализованный индекс Герфиндаля 0,29), приходится заметно большая энтропия при переводе на русский (0,92, нормализованный индекс Герфиндаля 0,20), при том, что лидирующий стимул перевода даёт даже более высокий показатель, чем лидирующая модель (44%).

Слово *пошлый (пошлость)* обнаруживает ряд признаков этого же типа (ограниченный круг выбора в переводах с русского и более широкий разброс стимулов при переводах с английского), но всё же оно заметно ближе к «хорошим» лингвоспецифичным словам типа *удаль* и *простор*. Среди моделей перевода обнаруживается достаточно высокая энтропия (0,8), но нормализованный индекс Герфиндаля даёт 0,26 (почти в половине случаев используется слово vulgar). По стимулам перевода разброс, как обычно, выше: энтропия вырастает до 1,1, нормализованный индекс Герфиндаля падает до 0,06, а на vulgar приходится лишь пятая часть соответствий. Наконец, *пошлость*/*пошлый* существенно чаще встречается в оригинальных текстах (почти в пять раз), и это наиболее яркая его лингвоспецифичная черта.

Любопытно привлечь также материал близкородственного языка — украинского и сравнить меру специфичности украинских соответствий с английскими соответствиями; соответствующие показатели для 9-миллионного параллельного русско-украинского корпуса также вполне надёжны. Оказывается, что только два слова из этого перечня могут, исходя из статистических критериев, считаться лингвоспецифичными для русского по сравнению с украинским — *пошлый (пошлость)* и *удалой (удаль)*: как для моделей, так и для стимулов перевода обоих этих слов энтропия выше 0,65, а средний нормализованный индекс Герфиндаля — 0,18. Слово *тоска* даёт средние показатели разброса, несмотря на большое количество разных соответствий, а слова *уют(ный)* и *страсть* уже имеют четко выраженные ведущие переводные эквиваленты (*затишок/затишний* и *пристрасть* соответственно). Показательно, что слово *простор*, представавшее как «хорошее» лингвоспецифичное при сопоставлении с английским, при сопоставлении с украинским даёт, наоборот, самый низкий разброс соответствий: дело в том, что в украинском языке имеется его точное этимологическое соответ-

ствие *простір*, семантически отвечающее не только русскому *простор*, но и *пространство*. По моделям перевода слово *пространство* при этом имеет несколько больший разброс, чем *простор* (и даже *уют* и *страсть*), но по стимулам перевода выглядит примерно так же, как и *простор*.

Итак, применение статистических методов к выделению лингвоспецифичной лексики на параллельном корпусе кажется перспективным и выявляет, кроме того, ряд общих свойств переводных текстов. Например, разброс моделей перевода как общее правило ниже, чем стимулов. Выявлены слова, типа *уют* или *страсть*, для которых эта разница особенно сильна и которые используются именно при переводе как соответствие большому ряду английских эквивалентов. Применение этой методики к целому ряду романских, германских и славянских языков позволит выделить кластеры специфичной для тех или иных языков лексики.

7. Пример использования параллельного корпуса для грамматических исследований: перфект в параллельном корпусе

Параллельный корпус является хорошим материалом для исследования такой типологически нетривиальной и варьирующей даже в близкородственных языках категории, как перфект[1]. Об исследовании перфекта в «массовом» параллельном корпусе библейских переводов ср. также выполненную независимо от нашей работу [Dahl 2014]; подробнее материалы нашего исследования изложены в работе [Сичинава 2016].

На материале анализа переводов «Винни-Пуха» А. А. Милна на другие германские языки видно, что английский перфект почти всегда переводится при помощи германского аналитического перфекта (с точностью до выбора вспомогательного глагола в немецком и нидерландском), но при этом имеется и ряд отклонений, прежде всего, связанных с небазовыми значениями перфекта. (О семантике перфекта в типологическом освещении, и особенно в европейских языках, см., прежде всего, [McCoard 1978], [Dahl, Hedin 2000], [Lindstedt 2000]).

[1] Работа поддерживалась грантом № 10-04-00168 «Механизмы изучения перфекта в типологическом освещении», руководитель проекта Ф. А. Елоева.

Перфект пассива с результативным значением может переводиться презенсом (в данном примере — на немецкий):

EN: The flood-level has reached an unprecedented height [WP: IX]
DE: Der Pegelstand issst (PRAES) unverhältnisssmäßßig hoch.
'Уровень паводка достиг небывалой высоты'

Перфект со значением оценки ситуации («оказался») переводится на немецкий простым претеритом, на нидерландский — конструкцией без глагола («так глупо меня (можно) ввести в заблуждение»). Из германских языков только шведский, язык с «сильным перфектом», сохраняет форму. Романские языки и македонский сохраняют перфект, но в балтийских и болгарском выступают другие времена:

EN: «I have been Foolish and Deluded,» said he [WP: III]
DE: «Ich war (PRAET) ein verblendeter Narr», sagte er.
NL: «Stommerd om me zo op een dwaalspoor te laten brengen,» zei hij.
SV: — Jag har varit enfaldig och blivit lurad, sa han.
ES: «He sido Crédulo y Estúpido», dijo.
IT: «Sono stato Stupido e Ingenuo» disse.
LT: — Buvau (PRAET) žioplas Apsigavėlis,— pridūrė jis.
LV: — Es biju (PRAET) muļķis un piekrāpos,— viņš bēdājās.
MK: «Колку сум бил глупав и забудален» рече.
BG: — Аз съм (PRAES) Глупав и Измамен…
'— Я был Глуп и Сбит с толку,— сказал он'

Претеритом в немецком передан и перфект «расширенного настоящего», охватывающего точку отсчёта. Наиболее ярко это выражено в примере, где говорящий противопоставляет настоящее и перфект «расширенного настоящего» при помощи наречия *so far* 'до сих пор' (перевод презенсом, таким образом, невозможен). В немецком и нидерландском выбран претерит, при этом в шведском перфект сохранён:

EN: «Friends», he said, «including oddments, it is a great pleasure, or perhaps I had better say it has been a pleasure so far, to see you at my party». [WP: X]
NL: «Vrienden», zei hij, «vrienden en alles wat daarbij hoort, het is mij

een groot genoegen, of misschien moet ik zeggen, tot nu toe was (PRAET) het een groot genoegen om jullie allen op mijn partijtje te zien».

DE: «Freunde», sagt er, «und sonstiges herumwuselndes Kroppzeug eingeschlossen, es ist ein großes Vergnügen oder vielleicht sollte ich eher sagen, Es war (PRAET) bisher ein großes Vergnügen euch auf meiner Party zu sehen».

'Друзья,— начал он,— друзья мои… включая прочих! Для меня большая радость — во всяком случае, до настоящей минуты было большой радостью — видеть вас на моём Пиргорое'.

Отметим, что большинство отмеченных отклонений связано с бытийным глаголом, перфект которого в немецком (*ich bin gewesen*) вообще существенно более редок, чем претерит (*ich war*).

При обращении к англо-русскому параллельному корпусу оказывается, что в большинстве случаев английский перфект передаётся русским претеритом без каких-либо дополнительных лексических средств. Перфекту с обстоятельством ограниченного периода времени в русском соответствует глагол СВ длительно-ограничительного способа действия (пердуратив) с приставкой *про-*:

«Vicky», he said carefully. «I have driven fifteen hundred miles on turnpikes since we left Boston».

'— Вики, я проехал (// ?ехал) по шоссе пятнадцать тысяч миль <…>'. [Стивен Кинг. Дети кукурузы / Пер. не указан].

Обращает на себя внимание использование несовершенного вида прошедшего времени (выражающего одновременность ситуации) для передачи значения «результатив-в-прошедшем» (в составе инфинитива); СВ *осталось* в этом контексте менее вероятен. При этом фраза перестроена: вместо 'что я оставила' — *что у меня оставалось*. Если бы агентивная конструкция оригинала была сохранена (*что я оставила достаточно времени*), то здесь как раз невозможен был бы НСВ (**что я оставляла*…).

(26) I strolled for a bit, happy to have left enough time to get as lost as I was, and finally ducked into a deli for a cup of coffee. [Lauren Weisberger. The Devil Wears Prada (2003)]

'Я немного поблуждала, радуясь, что у меня еще оставалось на это

время, и наконец толкнулась в закусочную, чтобы выпить чашку кофе'. [Лорен Вайсбергер. Дьявол носит Прада / М. Маяков, Т. Шабаева].

Несовершенный вид выступает в случае, когда перфект сочетается с итеративным значением (фактически эта функция близка к таксисной):

Sometimes I try to call up old girl friends on the telephone late at night, after my wife has gone to bed. [Kurt Vonnegut. Slaughterhouse-Five Or The Children's Crusade (1969)]

'Иногда поздно ночью, когда жена уходит спать, я пытаюсь позвонить по телефону старым своим приятельницам' [Курт Воннегут. Бойня номер пять, или Крестовый поход детей / Р. Райт-Ковалёва]

Перфекту со значением «расширенного настоящего» (актуальной ситуации, охватывающей момент речи) в русском отвечает презенс:

And now for four or five lifetimes of men the Godkings have ruled all the four lands together, and made them an empire. [Ursula Le Guin. The Tombs of Atuan (1971)]

'И вот уже четыре или пять поколений Божественный Король правит всеми Четырьмя Странами, объединенными в Империю' [Урсула Ле Гуин. Гробницы Атуана / Пер. не указан]

Перфект прогрессива передаётся несовершенным видом (настоящего либо прошедшего времени) в актуально-длительном значении. Отмечено такое лексическое средство для передачи охватывающего точку отсчёта перфекта прогрессива, как глаголы типа *начать, решить*, указывающие на актуальность для текущего момента начального предела ситуации:

Perhaps people have been celebrating Bonfire Night early — it's not until next week, folks! [Joanne Kathleen Rowling. Harry Potter and the Sorcerer's Stone (1997)]

'Кажется, народ уже **начал** праздновать день Порохового Заговора — рановато, господа, он будет только на следующей неделе!' [Дж. К. Роулинг. Гарри Поттер и Волшебный камень / М. Спивак].

Well, it's a new thing the boss has been trying. [Stephen King. The Lawnmower Man (1975)]

— Это новая методика, которую **решил** испробовать наш босс. [Стивен Кинг. Газонокосильщик / Пер. не указан]

Перфект пассива в результативном значении передаётся стандартно выражающим «результатив-в-настоящем» русским презенсом пассива (ср. Князев 1983):

The golden rule of the crime scene, people, is don't touch anything until it has been studied, photographed and charted. [Michael Connelly. City Of Bones (2002)]

'Ребята, золотое правило осмотра места преступления заключается в том, что нельзя притрагиваться к находке, пока она не осмотрена, не сфотографирована и не отмечена на плане' [Майкл Коннели. Город костей / Д. Вознякевич].

Ср. не вполне обычно звучащий перевод системного сообщения: *Установка Windows была успешно завершена* в соответствии с *has been (*was) successfully completed*. Действительно, по-русски претерит пассива употребляется скорее в антирезультативных контекстах или по крайней мере в нарративе, чем в контексте «свежих новостей» (hot news) и достигнутого в настоящем результата (ср., впрочем, поэтическое: *Москва, спалённая пожаром, французу отдана*).

Отметим употребление презенса для передачи перфекта при глаголе (пересказываемой) речи. Такая сочетаемость в контекстах пересказа текста характерна для русского языка (*говорят, что…, ты пишешь, что*…, сюда же, вероятно, «вневременной презенс» типа *Аристотель делит формы государства на правильные и неправильные*). При этом перфект прогрессива (описывающий предшествующие моменту сообщения события) передан русским претеритом НСВ:

And finally, bird-watchers everywhere have reported that the nation's owls have been behaving very unusually today. [Joanne Kathleen Rowling. Harry Potter and the Sorcerer's Stone (1997)]

'И в завершение нашего выпуска, орнитологи страны сообщают, что сегодня повсеместно наблюдалось крайне странное пове-

дение сов'. [Дж. К. Роулинг. Гарри Поттер и Волшебный камень / М. Спивак]

Ср. аналогичную корреляцию презенса и перфекта в переводе с русского на английский:

Ведь целый ряд очень видных ученых полагает, что находки в зонах посещения способны изменить весь ход нашей истории. [А. Н. Стругацкий, Б. Н. Стругацкий. Пикник на обочине (1971)]

'After all, many very important scientists have proposed that the discoveries made in the Visitation Zones are capable of changing the entire course of our history' [Arkady Strugatsky, Boris Strugatsky. Roadside Picnic / Antonina W. Bouis]

Перфекту в экспериенциальном значении соответствует несовершенный вид в общефактическом значении:

Until recently antimatter has been created only in very small amounts (a few atoms at a time). [Dan Brown. Angels and Demons (2000)]

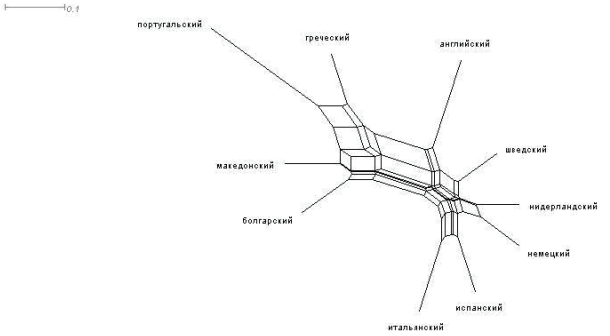
'До недавнего времени антивещество получали лишь в мизерных количествах (несколько атомов за один раз)' [Дэн Браун. Ангелы и демоны / Г. Косов]

"God damn it; who the hell has been tampering with this?" [Isaac Asimov. The Gods Themselves (1972)]

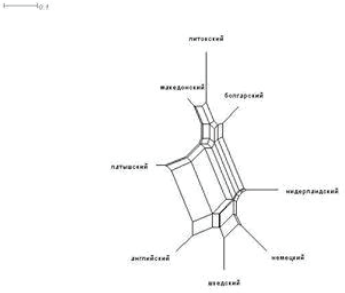
'— Черт подери! Какой сукин сын трогал эту колбу?' [Айзек Азимов. Сами боги / Н. Рыбакова]

Наконец, на материале многоязычного параллельного корпуса возможно типологическое сопоставление перфектов в языках Европы методом измерения расстояний (в программе NeighbourNet/SplitsTree, первоначально применявшейся в биологических исследованиях [Huson, Bryant 2006]; ср. также аналогичное исследование перфекта в [Dahl 2014] и исследование императива в славянских языках в работе [von Waldenfels 2014]). Строится матрица, строки которой соответствуют языкам, а столбцы — предикациям текста; единица означает, что в этой точке встретился перфект, ноль — что употреблено другое средство выражения. На базе этой матрицы программа порождает граф рас-

стояний между рядами данных. Оказывается, что получившаяся сеть расстояний «группирует», помимо объединений, примерно изоморфных генетическим и ареальным (итальянский + испанский, нидерландский + немецкий), также вместе балканские языки и португальский язык, в которых перфект сохраняется прежде всего в контексте экспериенциального значения (материал — переводы «Алисы в стране чудес»):



Аналогичная картинка по переводам «Винни-Пуха» (где во включенном в корпус материале отсутствовали романские языки и греческий, но присутствовали два балтийских) показывает явную типологическую близость малоупотребительного в этом переводе литовского перфекта не к германским языкам с «сильным перфектом», а к балканским, в то время как латышский перфект занимает скорее промежуточную позицию.



Таким образом, объём и проработанность параллельных корпусов НКРЯ уже достаточны для проведения достаточно сложных со-

поставительных и типологических исследований, как на лексическом, так и на грамматическом уровне.

Литература

Бунтман Н. В., Зализняк Анна А., Зацман И. М., Кружков М. Г., Лощилова Е. Ю., Сичинава Д. В. Информационные технологии корпусных исследований: принципы построения кросслингвистических баз данных, Информатика и её применения, 8:2 (2014), 98—110

Добровольский Д. О. Корпус параллельных текстов в исследовании культурно-специфичной лексики // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 383—401.

Зализняк Анна А., *Левонтина И. Б., Шмелев А. Д.* Константы и переменные русской языковой картины мира. М.: Языки славянских культур, 2012.

Павлова А. В. (сост.) От лингвистики к мифу: лингвистическая культурология в поисках «этнической ментальности». СПб.: Антология, 2013.

Руссо М. М. Неогумбольдтианская лингвистика и рамки «языковой картины мира» // Политическая лингвистика, 1 (47), 2014, с. 12—24 (http://journals.uspu.ru/attachments/article/622/%D0%9F%D0%BE%D0%BB%D0%B8%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F%20%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_2014_1_%D1%81%D1%82.%2001.pdf)

Сичинава Д. В. Поливариантные параллельные тексты в составе НКРЯ // Труды международной конференции «Корпусная лингвистика-2013», СПб., 2013 (http://corpora.phil.spbu.ru/Works2013/%D0%A1%D0%B8%D1%87%D0%B8%D0%BD%D0%B0%D0%B2%D0%B0.pdf)

Сичинава Д. В. Европейский перфект сквозь призму параллельного корпуса // Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН / Отв. ред. Н. Н. Казанский. Т. XII. Ч. 2. Исследования по теории грамматики. Выпуск 7: Типология перфекта / Отв. ред. Т. А. Майсак, В. А. Плунгян, Кс. П. Семенова. СПб.: Наука, 2016.

Січінава Д. В., Тищенко-Монастирська О. О., Шведова М. О. Паралельні українсько-російський та російсько-український корпуси //Лексикографічний бюлетень, 20. К., 2011: 35–38

Cysouw M., Wälchli B. (eds.). Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie & Universalienforschung STUF 60.2, 2007.

Dahl Ö., Hedin E. Current relevance and event reference. // Osten Dahl (ed.), Tense and Aspect in the Languages of Europe, 385—402. Berlin, New York: de Gruyter, 2000.

Dahl Ö. The perfect map: Investigating the cross-linguistic distribution of TAME categories in a parallel corpus // Szmrecsanyi, Benedikt & Walchli, Bernhard. (eds.) Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech. Linguae & litterae 28. Berlin: Walter de Gruyter, 2014: 268–289.

Goldberg A. Constructions: A Construction Grammar Approach to Argument structure. Chicago, 1995.

Goldberg A. Constructions at Work. The nature of generalization in grammar. Oxford, 2006.

Hasselgård H., Oksefjell S. (eds.). Out of Corpora: Studies in. Honour of Stig Johansson. Amsterdam — Atlanta, GA:Rodopi, 1999.

Huson D. H., Bryant D. Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23(2), 2006: 254—267.

Lindstedt J. The perfect — aspectual, temporal and evidential. // Östen Dahl (ed.), Tense and Aspect in the Languages of Europe, 365–384. Berlin, New York: de Gruyter, 2000.

Loiseau S., Sitchinava D. V., Zalizniak A. A., Zatsman I. M. Information technologies for creating the database of equivalent verbal forms in the Russian—French multivariant parallel corpus [Информационные технологии создания баз данных эквивалентных глагольных форм в русско-французском поливариантном параллельном корпусе] // Информатика и её применения, 7:2 (2013), 100–109.

McCoard R. W. The English Perfect: Tense-Choice and Pragmatic Inferences.Amsterdam: North-Holland, 1978.

Sitchinava D. Parallel corpora within the Russian National Corpus // Prace Filologiczne, LXIII, 2012, 271—278.

Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. Parallel

corpora for medium density languages // Proceedings of the RANLP 2005: 590-596.

v. Waldenfels R. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment // Beitrage der Europaischen Slavistischen Linguistik (POLYSLAV) 9. Munchen, 2006, S. 123─138 (доступно по адресу http://www-nw.uni-regensburg.de/%7E.war05297.slavistik.sprachlit.uni-regensburg.de/pub/WaldenfelsParallelCorpora2006.pdf)

v. Waldenfels R. Explorations into variation across Slavic: Taking a bottom-up approach // Szmrecsanyi, Benedikt&Walchli, Bernhard. (eds.) 2014. Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech. Linguae & Litterae 28. Berlin: Walter de Gruyter, 290-323.

Dmitri V. Sitchinava
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
mitrius@gmail.com

PARALLEL TEXTS WITHIN THE RUSSIAN NATIONAL CORPUS: NEW DIRECTIONS AND RESULTS

The paper presents the current states of the parallel corpora within the RNC and the updates of the last six years. The parallel corpora with the RNC include the following bilingual Russian—X corpora: Armenian, Belarusian, Bulgarian, English, Estonian, French, German, Italian, Latvian, Polish, Spanish, and Ukrainian. Virtually for all these language we have language pairs in both directions of translation. The RNC now includes a multilingual corpus that consists of nine texts and involve more than 20 languages, mostly Slavic. The Russian-French corpus has been developed, since 2012, as a polyvariant corpus with multiple translations of the same text into the same language. Polyvariant texts are also included also in the multilingual corpus. Parallel corpora now include texts of different genres, not only fiction, but also news, technical, scientific, religious and

legal texts, viz. all the main subdivisions represented in the main RNC. The discrepancies between the original text and the translation are now specially marked (arbitrary omission, adding, or change by the translator). Most languages represented in the corpus also have a morphological tagset and annotation. Some examples of corpus-based studies of lexicon and grammar are examined in more detail.

Key words: Parallel corpus, annotation, multilingual corpus, lexical typology, grammatical typology, language-specific lexicon, perfect

References

Buntman N. V., Zaliznyak Anna A., Zatsman I. M., Kruzhkov M. G., Loshchilova E. Yu., Sichinava D. V. [Informational technologies in corpus research: framework for building cross-linguistic databases]. *Informatika i ee primeneniya*, 8:2 (2014), pp. 98–110 (In Russ.)

Cysouw, M., Wälchli B. (Eds.). *Parallel Texts. Using Translational Equivalents in Linguistic Typology. Theme issue in Sprachtypologie & Universalienforschung STUF* 60.2, 2007.

Dahl Ö. The perfect map: Investigating the cross-linguistic distribution of TAME categories in a parallel corpus. Szmrecsanyi, Benedikt & Walchli, Bernhard. (eds.) *Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech.* Linguae & litterae 28. Berlin, Walter de Gruyter, 2014, pp. 268–289.

Dahl Ö., Hedin E. Current relevance and event reference. In: Osten Dahl (ed.), *Tense and Aspect in the Languages of Europe*, 385–402. Berlin, New York, de Gruyter, 2000.

Dobrovol'skij D. O. [Corpus of parallel texts in studying culture-specific lexicon]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus 2006–2008: New Results and perspectives]. St. Petersburg, Nestor-Istoria Publ., 2009, pp. 383–401. (In Russ.)

Goldberg A. *Constructions at Work. The nature of generalization in grammar.* Oxford, 2006.

Goldberg A. *Constructions: A Construction Grammar Approach to Argument structure.* Chicago, 1995.

Hasselgård, H., Oksefjell, S. (Eds.). *Out of Corpora: Studies in Honour of Stig Johansson*. Amsterdam – Atlanta, GA:Rodopi, 1999.

Huson D. H., Bryant D. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2), 2006, pp. 254–267.

Lindstedt J. The perfect — aspectual, temporal and evidential. In: Östen Dahl (ed.), *Tense and Aspect in the Languages of Europe*, 365–384. Berlin, New York, de Gruyter, 2000.

Loiseau S., Sitchinava D. V., Zalizniak A. A., Zatsman I. M. Information technologies for creating the database of equivalent verbal forms in the Russian–French multivariant parallel corpus. *Informatika i ee primeneniya*, 7:2 (2013), pp. 100–109.

McCoard R. W. *The English Perfect: Tense-Choice and Pragmatic Inferences*. Amsterdam, North-Holland, 1978.

Pavlova A. V. (Ed.) *Ot lingvistiki k mifu: lingvisticheskaya kul'turologiya v poiskakh «etnicheskoi mental'nosti»* [From linguistics to myth: linguistic cultural studies searching "ethnic mentality"]. St. Petersburg, Antologiya Publ., 2013.

Russo M. M. [Neo-Humboldtian linguistics and the boundaries of the "language image of the world"]. *Politicheskaya lingvistika*, 1 (47), 2014, pp. 12—24 (Available at: http://journals.uspu.ru/attachments/article/622/%D0%9F%D0%BE%D0%BB%D0%B8%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F%20%D0%BB%D0%B8%D0%BD%D0%B3%D0%B2%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0_2014_1_%D1%81%D1%82.%2001.pdf, accessed on 4.7.2015) (In Russ.)

Sichinava D. V. [Polyvariant parallel texts within the RNC]. *Trudy mezhdunarodnoi konferentsii «Korpusnaya lingvistika-2013»* [Proceedings of the international conference "Corpus linguistics 2013"]. St. Petersburg, Saint Petersburg University, 2013 (Avaliable at http://corpora.phil.spbu.ru/Works2013/%D0%A1%D0%B8%D1%87%D0%B8%D0%BD%D0%B0%D0%B2%D0%B0.pdf, accessed on 4.7.2015)

Sichinava D. V., Tyshchenko-Monastyrska O. O., Shvedova M. O. [Parallel Russian-English and English-Russian corpora]. *Leksikografichnyi byuleten'* 20, Kyiv, 2011, pp. 35–38. (In Ukrainian)

Sitchinava D. Parallel corpora within the Russian National Corpus. *Prace Filologiczne*, LXIII, 2012, pp. 271–278.

Sitchinava D. V. [European perfect within a parallel corpus] *Acta linguistica Petropolitana XII. Part 2. Studies in Grammar theory: Vol. 7: Typology of perfect.* St. Petersburg, Nauka Publ., 2016 [In Russ.].

Varga D., Németh L., Halácsy P., Kornai A., Trón V., Nagy V. Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, pp. 590–596.

v. Waldenfels R. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment. *Beitrage der Europaischen Slavistischen Linguistik* (POLYSLAV) 9. Munchen, 2006, pp. 123–138 (available at http://www-nw.uni-regensburg.de/%7E.war05297.slavistik.sprachlit.uni-regensburg.de/pub/WaldenfelsParallelCorpora2006.pdf)

v. Waldenfels R. Explorations into variation across Slavic: Taking a bottom-up approach. Szmrecsanyi, Benedikt&Walchli, Bernhard. (eds.) 2014. *Aggregating Dialectology, Typology, and Register Contents Analysis. Linguistic Variation in Text and Speech*. Linguae & Litterae 28. Berlin, Walter de Gruyter, 2014, pp. 290–323.

Zaliznyak Anna A., Levontina I. B., Shmelev A. D. *Konstanty i peremennye russkoi yazykovoi kartiny mira* [Constants and variables of the Russian language image of the world]. Moscow, Languages of the Slavic cultures Publ., 2012.

К. М. Корчагин
Институт русского языка им. В. В. Виноградова РАН
(Россия, Москва)
stivendedal@gmail.com

ПОЭЗИЯ XX ВЕКА В ПОЭТИЧЕСКОМ ПОДКОРПУСЕ НАЦИОНАЛЬНОГО КОРПУСА РУССКОГО ЯЗЫКА: ПРОБЛЕМА РЕПРЕЗЕНТАТИВНОСТИ[1]

В статье рассматриваются правила отбора поэтических текстов XX века для поэтического подкорпуса Национального корпуса русского языка. Перед создателями поэтического корпуса стояла задача создать инструмент, который одновременно упрощал работу исследователям метрики и поэтического языка и, в то же время, выступал бы как репрезентативный источник поэтических текстов. В XX веке картина русской поэзии существенно усложнилась, разделившись на несколько независимых друг от друга ветвей. При создании поэтического корпуса предполагалось, что корпус должен быть репрезентативен для каждой ветви русской поэзии XX века (для «официальной» и «неофициальной», эмигрантской и советской и т.д.). Создатели поэтического корпуса стремились сделать поэтический корпус полезным для двух групп пользователей — для исследователей метрики и исследователей поэзии. В связи с этим создатели корпуса предпочли отказаться от строгой иерархии авторов и текстов, свойственной большинству обобщающих работ о литературе XX века, и сделать акцент на восстановлении поэтического контекста эпохи. Таким образом, наравне с поэтами первого ряда в корпус включались «малые» поэты, которые не занимали столь значитель-

[1] Исследование выполнено в рамках Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» (2012—2014) и при поддержке РГНФ (грант № 15-04-12018).

ного места в литературе эпохи, однако при этом пользовались определенной известностью. При этом в корпус не включалась массовая стихотворная продукция, которая могла бы интересовать исследователей метрики, так как ее присутствие существенно уменьшало бы репрезентативность корпуса с точки зрения развития русской поэзии за последние сто лет.

Ключевые слова: метрика, стиховедение, поэтический корпус, НКРЯ, правила отбора, репрезентативность, поэтический контекст, русская поэзия XX века

На первоначальном этапе создания поэтического подкорпуса НКРЯ перед нами стояла задача сформулировать правила последовательного описания метрики поэтических текстов, в то время как вопрос о составе корпуса казался второстепенным (Гришина et al. 2009). При работе над текстами XVIII — XIX веков мы следовали тем правилам отбора, что были сформулированы М. Л. Гаспаровым в книге «Современный русский стих»: эти правила предписывали исследователю опираться на наиболее полные академические издания (прежде всего, на тома «Библиотеки поэта») и на антологии, состав которых не предполагал специального отбора по метрическим критериям (Гаспаров 1974: 41–42). Наша выборка поэтических текстов XVIII — XIX веков отличается от той выборки, что представлена в «Современном русском стихе», но составлена по тем же принципам (отличия вызваны сложностью автоматической обработки изданий XVIII — XIX веков, присутствующих в выборке М. Л. Гаспарова, а в некоторых случаях появлением новых, более полных собраний произведений отдельных поэтов).

При работе с текстами XVIII — XIX веков вопрос определения правил метрической разметки казался первоочередным: корректная разметка упорядочивала состав поэтического корпуса и для того относительно небольшого количества текстов, с которым мы первоначально имели дело, такое упорядочивание выглядело достаточным. Но ситуация поменялась как только создатели корпуса «вступили» в XX век — на территорию множества разнообразных поэтических практик, где отсутствовала единая иерархия литературных достижений. На этой территории стремление к максимальному охвату приводит к неоправданному увеличению объема корпуса и заставляет поставить вопрос о том, в какой мере такой

«максимально полный» корпус может дать представление о состоянии русской метрики соответствующего периода, не содержит ли он слишком большое количество «случайных» текстов — текстов, которые занимают не вполне определенное место в литературном контексте эпохи.

Уже в годы возникновения русского литературного модернизма происходит валовое увеличение поэтической продукции, на фоне которой, впрочем, еще можно выделить поэтов «первого ряда», занимающих центральное место в культурной жизни эпохи. Однако по мере приближения к нашему времени картина становится всё более сложной: некогда единая русская поэзия разделяется на несколько линий, ведущих параллельное существование: в довоенное время — на поэзию эмиграции и советскую поэзию, в послевоенное — на поэзию «официальную» и поэзию «неподцензурную» и т. д. Подобные изменения заставляют заново поставить вопрос о правилах формирования корпуса и критериях его репрезентативности.

Задача этой статьи — ответить на вопрос, какие именно поэты XX века попадают в поэтический корпус и в каком количестве. Для ответа на этот вопрос мы коснемся нескольких проблемных зон, с которыми вынуждены считаться все исследователи поэзии этого периода: соотношения истории стиха и истории поэзии, литературного контекста и поэтической иерархии, уникальности и массовости поэтических практик, а также разделения на «неофициальную» и «официальную» литературу. При составлении корпуса поэзии XX века мы старались найти компромисс между всеми этими оппозициями, и ниже мы поясним, каким именно образом этот компромисс достигался.

1. История стиха и история поэзии

Перед создателями поэтического корпуса стояла специфическая задача: создать инструмент, который одновременно упрощал работу исследователям метрики и поэтического языка и, в то же время, выступал бы как репрезентативный источник поэтических текстов. Поэтический корпус не задумывался как собрание «канонических» произведений, но должен был представлять собой такую выборку из всего массива русской поэзии, которая позво-

ляла бы реконструировать поэтический и метрический контекст того или иного времени. Однако решение такой задачи для поэзии XX века оказывается достаточно проблематичным: корпус должен быть репрезентативен для каждой ветви русской поэзии XX века (для «официальной» и «неофициальной», эмигрантской и советской и т.д.) при том, что для каждой из них стоит вопрос о полноте существующих изданий и общих принципах отбора текстов.

Кроме того, создатели поэтического корпуса искали компромисс между двумя группами пользователей — исследователями метрики и исследователями поэзии. Для первой группы более важной оказывается репрезентативность корпуса с точки зрения того, насколько тот или иной его хронологический срез может репрезентировать метрический репертуар русской поэзии — это в точности та задача, которую ставил перед собой М.Л. Гаспаров в своих классических работах (Гаспаров 1974; 1984/2002). Вторая группа более заинтересована в том, чтобы в таком срезе были представлены произведения тех авторов, чья деятельность была хотя бы в малейшей степени значима для литературной жизни эпохи (при этом представления о важности того или иного автора могут достаточно сильно варьироваться — особенно применительно к XX веку). Обобщая, можно сказать, что первых больше интересует стих, а вторых — поэзия.

И здесь встает вопрос полноты имеющегося корпуса: если мы вслед за М.Л. Гаспаровым ориентируемся на наиболее авторитетные издания поэтических текстов, то, обращаясь к материалу XX века, мы неизбежно столкнемся с тем, что некоторые сегменты русской поэзии изучены с исчерпывающей полнотой, в то время как другие лишь только ждут своего исследователя. Например, московское издательство «Водолей» в последние несколько лет выпускает академические издания «забытых» поэтов серебряного века (серия «Серебряный век. Паралипоменон»), среди которых собрания Юрия Верховского, Марии Веги, Георгия Голохвастова, Льва Гомолицкого, Владимира Корвин-Пиотровского и многих других. Все издания этой серии удовлетворяют тем критериям, которые сформулировал М.Л. Гаспаров, — это академические издания, снабженные справочным аппаратом, предполагающие серьезную текстологическую работу и т.д. Однако место большинства авторов этой серии в литературном контексте эпохи достаточно неод-

нозначно: они либо редко публиковали собственные произведения, либо не публиковали их вообще, а след, который был ими оставлен в истории отечественной поэзии, зачастую почти не заметен (это нередко подчеркивают и составители соответствующих томов). В отличие от поэтов группы Обэриу такие поэты пока не сформировали отдельную «область» на поэтической карте эпохи. По этой причине лишь отдельные издания серии становились предметом для метрического анализа и были включены в поэтический корпус (собрания Юрия Верховского, Сергея Соловьева, Моисея Цетлина, Марка Тарловского — тех поэтов, которые занимали достаточно заметное место в литературном ландшафте XX века, но по разным причинам систематически не издавались ни в советское ни в постсоветское время). Прочие поэты представлены в корпусе либо ограниченными по объему подборками (Алексей Лозина-Лозинский, Мария Вега), либо оставлены за его пределами.

Другая проблема связана с наличием более или менее представительных собраний произведений «неподцензурных» поэтов: несмотря на появление ряда важных изданий в последние годы (книги Яна Сатуновского, Всеволода Некрасова, Генриха Сапгира, Елены Шварц, Д.А. Пригова, Геннадия Айги), на настоящий момент значительная часть массива этой поэзии не издана удовлетворительным образом. Например, не существует собрания произведений одного из наиболее заметных поэтов «неподцензурного» Ленинграда — Виктора Кривулина (1944—2001), в отсутствии стихов которого соответствующий раздел поэтического корпуса будет заведомо неполон. При отсутствии авторитетных собраний создателям корпуса приходится прибегать к существующим изданиям, чаще всего не удовлетворяющим академическим эдиционным принципам.

В случае отсутствия обобщающих изданий «неподцензурных» поэтов мы опираемся на тот материал, что представлен в антологии «Неофициальная поэзия», подготовленной Иваном Ахметьевым и размещенной на Интернет-странице «Русской виртуальной библиотеки» (rvb.ru): в этом издании не только перечислены практически все «неподцензурные» поэты советского времени, но и указаны все издания их произведений, которые по возможности и используются при подготовке выборок для корпуса. Зачастую отдельные книги этих поэтов располагаются на другом ресурсе —

сайте «Вавилон» (vavilon.ru), создатель которого Дмитрий Кузьмин оказывал и оказывает нам помощь при работе с произведениями «неподцензурных» поэтов второй половины XX века.

Наконец, в силу того, что задачи репрезентативности получающихся выборок для нас оказываются важнее, чем текстологические критерии, мы отказались от идеи представлять в корпусе различные варианты одних и тех же текстов (тем более что их выявление, например, для «неподцензурных» поэтов представляет отдельную трудоемкую задачу) — практически во всех случаях тексты используются в том виде, в каком они размещены в соответствующих изданиях (источник текста указан в соответствующей графе метаразметки). Возможно, в силу этого поэтический корпус теряет в текстологической точности, но зато приобретает в репрезентативности.

2. Иерархия и контекст

Одна из наиболее очевидных проблем, с которой сталкиваются не только создатели поэтического корпуса, но и стиховеды, ставящие перед собой задачу описать метрический репертуар некоторой эпохи,— это проблема поэтической иерархии. Отказ от опоры на «образцовые» тексты — одна из предпосылок корпусной лингвистики, но в ориентированных на исследования литературы корпусах такой подход требует значительных оговорок. Причина этого в том, что такие корпуса предназначены для отслеживания не только языковых, но и культурных процессов, а для этого необходимо обращаться к уже заранее определенному полю тех текстов и авторов — репрезентативность в таком случае не может быть достигнута простым увеличением объема выборки.

В центре нашего внимания находятся наиболее значимые поэтические произведения XX века. Однако если исследователь поэзии ставит перед собой задачу описать поэтический язык или метрику эпохи, он не может удовольствоваться только «вершинными» явлениями — его внимание должно быть направлено на тех авторов, чье присутствие в литературе менее заметно, хотя они и вносят существенный вклад в формирование общего литературного пространства. Очевидным выходом из этого противоречия был следующий компромисс: вместо строгой иерархии авторов и текстов, свойствен-

ной в явном или неявном виде большинству обобщающих работ о литературе XX века, мы сосредоточили внимание на восстановлении поэтического контекста эпохи[1].

На практике это означает, что наравне с авторами «первого ряда», отсутствие которых в любом крупном собрании русской поэзии немыслимо (такими как Александр Блок, Велимир Хлебников, Марина Цветаева, Александр Введенский), мы обращали внимание на тех поэтов, которые не занимали столь значительного места в литературе эпохи, однако при этом пользовались определенной известностью (как, например, Георгий Шенгели или Владимир Шилейко), либо впоследствии становились предметом детального филологического изучения (Василий Комаровский или Павел Зальцман). Такой отбор помогает корпусу быть, прежде всего, аналитическим инструментом, позволяющим делать выводы о состоянии поэтического языка эпохи, а не собранием «образцовых» текстов, дополненным метрической разметкой[2].

В этой ситуации принципиальным оказывается вопрос — кто именно из поэтов «второго ряда» должен быть включен в поэтический корпус при условии, что последний не может расширяться бесконечно? Возможно, для ответа на этот вопрос могут применяться автоматизированные критерии (подсчет упоминаний в периодической печати, посвященных автору работ и т. п.), однако формулировка таких критериев составляет отдельную и довольно трудоемкую задачу, и нами она была решена лишь в первом приближении: при составлении поэтического корпуса учитывались критические и аналитические работы, посвященные тому или иному поэту, причем работы, опубликованные в значимых изданиях, принадлежащих как к миру литературы, так и к миру науки. Ориентирами здесь служат ведущие филологические издания,

[1] Стремление к полноте контекста наталкивается на одно важное ограничение — на проблему поэтического перевода. Очевидно, что переводные тексты (например, «Илиада» Гнедича) могут играть крайне важную роль в литературе эпохи, однако по техническим причинам в поэтическом корпусе такие переводы занимают довольно скромное место. Возможно, в будущем этот недостаток корпуса удастся устранить.

[2] Можно сказать, что формат поэтического корпуса противостоит формату поэтической *антологии*: последний опирается именно на «образцовые» тексты и потенциально допускает существенные лакуны. При этом работа по составлению антологии русской поэзии, снабженной метрической разметкой, представляет собой отдельную задачу: известный опыт такого рода антология В.Е. Холшевникова «Мысль, вооруженная рифмами».

чаще всего обращающиеся к русской поэзии XX века: «Russian Literature», «Wiener Slawistischer Almanach», «Новое литературное обозрение», «Toronto Slavic Quarterly» и другие — эти издания приложили больше всего усилий для выстраивания единой линии русской поэзии XX века. Кроме того особое значение для создателей поэтического корпуса имели наиболее авторитетные антологии русской поэзии XX века: антология «Строфы века» Евгения Евтушенко, «Самиздат века» (составитель поэтического раздела Генрих Сапгир) и ее электронная версия, представленная на ресурсе «Русская виртуальная библиотека» под названием «Неофициальная поэзия» (составитель Иван Ахметьев), «Освобожденный Улисс» (составитель Дмитрий Кузьмин), «Русские стихи 1950–2000-х годов» (составители: Иван Ахметьев, Герман Лукомников, Владимир Орлов, Андрей Урицкий) и некоторые другие. Кроме того, составление списка референтных фигур сопровождалось консультациями с ведущими специалистами по современной русской поэзии (Д. В. Кузьминым, Д. М. Давыдовым, Ю. Б. Орлицким).

Всё это позволило для каждого сегмента литературы XX века сформировать список поэтов, в котором бы присутствовали как фигуры, определяющие эпоху, так и те, кто во многом шел в их фарватере, но при этом пользовался достаточным вниманием со стороны читателей и исследователей поэзии.

Так, подборка поэтов парижской ветви первой эмиграции опирается на несколько центральных фигур, среди которых Георгий Иванов, Георгий Адамович, Борис Поплавский, Владислав Ходасевич и Владимир Набоков. Всё это поэты, которые имели громкую славу при жизни — их творчество было признано важной страницей в истории русской литературы. Но если в исследованиях поэзии первой эмиграции ограничиваться только этими поэтами, мы получим заведомо искаженную картину — как с точки зрения метрики, так и с точки зрения истории литературы. Говоря о значении этих поэтов, нужно помнить, что они существовали на фоне менее знаменитых авторов, таких как Анатолий Штейгер, Лидия Червинская, Юрий Терапиано, Довид Кнут, Борис Божнев, Юрий Одарченко, Анна Присманова и другие. Благодаря привлечению более широкого контекста становится видно, что те особенности метрики, которые можно было бы считать индивидуальными (например, преобладание 5-стопного ямба среди всех других силлабо-тонических разме-

ров или предпочтение дольника прочим формам «неклассического» стиха) может быть определяющей чертой для всего этого сегмента литературы — более того, у поэтов «первого ряда» такие тенденции выражены не так отчетливо, как у поэтов «второго». Таким образом, общие выводы относительно метрического репертуара эпохи и/или литературной формации возможны только при обращении к расширенному контексту, заведомо включающему не только «вершинные» произведения.

3. Уникальное и массовое

Исследователи поэзии могут изучать как поэтические новации, так и некий «усредненный» поэтический язык эпохи, и, следовательно, внимание потенциальных пользователей поэтического корпуса может быть направлено либо на уникальное и единичное, либо на массовое и повторимое. Оба этих подхода имеют право на существование: первый позволяет указать на те явления, которые радикальным образом меняют саму ткань поэзии (такие как ранние дольники Блока или акцентный стих Маяковского), второй — понять, на каком именно фоне эти явления возникали и что, напротив, представляло собой «среднее» стихотворение эпохи. Любое более или менее объемное исследование метрики требует компромисса между этими двумя подходами, и исследователи зачастую комбинируют их в разных пропорциях в зависимости от поставленных задач и доступного материала. Так, М. Л. Гаспаров в работах о метрике серебряного века обращал внимание по преимуществу на единичное, а в работах о советской метрике — на массовое (особенно показательны в этом отношении соответствующие главы «Современного русского стиха»).

Задача найти компромисс такого рода стояла и перед нами. При этом мы заведомо отказались от идеи включить в корпус все доступные источники (это потребовало бы, по меньшей мере, существенного увеличения состава рабочей группы), но при этом стремились сделать корпус репрезентативным ресурсом, демонстрирующим различные «синхронные» срезы литературного процесса, а не только произведения поэтов «первого ряда». Покажем, каким именно образом мы достигали этого компромисса, на двух примерах.

1. Предположим, что мы имеем дело с поэтом, произведения которого не публиковались (или почти не публиковались) при жизни, однако посмертно он получил внимание со стороны исследователей литературы, которые провели работу по (ре)контекстуализации его творчества. Для многих поэтов XX века характерна такая судьба: советская цензура исключала из литературной жизни множество авторов — некоторые из них могли найти относительный компромисс со временем, присутствуя в официальной литературе как детские поэты (обэриуты, позднее поэты Лианозовской школы) или как переводчики (Сергей Петров, Георгий Шенгели), но другие оставались почти полностью исключены из литературного мира (Владимир Щировский, Георгий Оболдуев, Павел Зальцман, Даниил Андреев). Такая исключенность была характерна, прежде всего, для поэтов 1930—1940-х годов: в эти два десятилетия не существовало литературных институций, которые были бы параллельны официальным и могли самостоятельно поддерживать «связность» литературной среды (само их существование в условиях массовых репрессий было, конечно, невозможно).

Несмотря на то, что такие поэты не могли принимать живое участие в литературной жизни эпохи, постфактум им приписывалась принципиальная роль в том общем поэтическом контексте (как метрическом, так и собственно литературном), с которым имеют дело создатели корпуса. Этими поэтами были предложены многие метрические новации (среди которых можно упомянуть хотя бы пятисложные стопы Даниила Андреева[1]) и хотя восприятие этих новаций было отсрочено, позднее они заняли важное место в работах исследователей поэзии и литературных критиков.

В качестве иллюстрации можно обратить внимание на двух поэтов, стихи которых стали известны совсем недавно — на Павла Зальцмана (1912—1985) и Геннадия Гора (1907—1981)[2]. Этих двух поэтов можно считать своего рода промежуточным звеном между поэтами довоенного авангарда (прежде всего, поэтами группы Обэриу) и неподцензурной литературой 1950-х годов, и

[1] См. например: Корчагин 2014.

[2] Недавние собрания стихотворений этих поэтов подготовлены И. Кукуем (2011) и П. Урбаном (2007). О месте этих поэтов в литературной жизни эпохи уже написано довольно много. См. например библиографию научных и критических работ о Павле Зальцмане и Геннадии Горе на посвященных им Интернет-ресурсах (http://pavelzaltsman.org и http://www.gsgor.ru/).

это касается всех особенностей их творчества — в том числе и метрики.

Для метрики Гора и Зальцмана характерны те же черты, что и для метрики обэриутов, но представлены они в несколько менее радикальной форме. Полное описание метрического репертуара этих поэтов выходит за пределы задач этой статьи, однако мы можем обратить внимание на одну характерную особенность «неклассических» размеров этих поэтов, которая роднит Зальцмана и Гора с Даниилом Хармсом и Александром Введенским.

Согласно нашим данным, среди дольников Хармса и Введенского преобладают вольные дольники (70% и 75% соответственно), при этом у первого постоянная анакруса присутствует только в 16% случаев, а у второго отсутствует вообще. У Гора обнаруживается 73% вольных дольников и 20% дольников с постоянной анакрусой; у Зальцмана — 47% вольных дольников и 28% дольников с постоянной анакрусой. На первый взгляд, кажется, что получающиеся числа сильно отличаются друг от друга (особенно это заметно, если принимать во внимание количество известных нам текстов этих поэтов: наследие Хармса и Зальцмана существенно превышает по объему наследие Введенского и Гора), но достаточно сравнить эти показатели с соответствующими показателями других поэтов эпохи, чтобы увидеть принципиальное отличие между обэриутами и их последователями и всей остальной поэтической продукцией второй половины 1920-х — 1930-х годов (стихи Зальцмана 1940-х годов также можно рассматривать на фоне поэзии 1930-х: с началом войны поэт был эвакуирован в Казахстан и был оторван от литературной жизни эпохи, сохранив, судя по всему, ту поэтическую манеру, которая была характерна для его стихов предыдущего десятилетия). Для контроля сделаем выборку из дольников Ольги Берггольц и Владимира Луговского, написанных в эти годы: доля вольных дольников у первой — 6%, у второго — 10% (постоянная анакруса — 70% и 49% соответственно). Это очень приблизительная оценка, но она показывает, что тяготение к определенным размерам (например, вольным дольникам) выделяло обэриутов и их последователей среди других поэтов, не важно, ориентировались ли последние на сталинскую неоклассику (как Берггольц) или на революционный авангард (как Луговской).

Случай Гора и Зальцмана достаточно очевиден: в их отсутствие

поэтический корпус будет заведомо не полон. Однако зачастую внимание исследователей направлено на тех авторов, чье место в истории отечественной словесности не столь очевидно. Хорошим примером такой работы можно считать недавнюю антологию А. Л. Соболева «Летейская библиотека» (Соболев 2014), посвященную тем поэтам 1910–1920-х годов, которые практически неизвестны даже специалистам. В этой антологии были предприняты значительные усилия по (ре)контекстуализации фигур многих «забытых» поэтов, однако в силу того, что в основном последние отмечены только в этом издании и почти не привлекали внимания других исследователей, нами было принято решение не включать их тексты в корпус (в случае если интерес к тому или иному поэту станет более широким, это решение, безусловно, будет пересмотрено). Разумеется, это касается не только этой конкретной антологии, но и всех происходящих в последнее время довольно часто «открытий» ранее неизвестных или практически неизвестных поэтов.

2. Другой случай можно назвать в известном смысле обратным. Предположим, что мы имеем дело с поэтом, который активно публиковался в значимых для эпохи изданиях, регулярно выпускал поэтические сборники и т.п., однако не привлекал или почти не привлекал к себе внимание исследователей, критиков, других поэтов и всех тех, чьими усилиями формируется литературный контекст эпохи. Использование сочинений таких поэтов в исследованиях метрики и поэтического языка не редкость — именно на них ориентировался М.Л. Гаспаров при написании послевоенной истории советского стиха. Среди поэтов, исследованных этим ученым, основное место занимали те, чей вклад в историю литературы и истории метрики не может не вызывать вопросов (Александр Коваленков, Михаил Матусовский, Николай Рыленков, Алексей Недогонов и другие): эти поэты в последние десятилетия практически не привлекают к себе внимания исследователей, и причину этого можно видеть также и в том, что с их именами не связываются никакие поэтические новации: они не оказали никакого влияния на поэтический язык, и все их заслуги перед метрикой сводятся к упрощению и тривиализации достижений предшественников (по крайней мере, именно так описывал стих этих поэтов М.Л. Гаспаров). Обращение к стиху этих поэтов в лучшем случае может показать общие особенности поэтического языка эпохи (и такая задача имеет смысл),

в худшем — замаскировать подлинные тенденции развития стиха. Последняя опасность представляется нам более значимой, в связи с чем эти поэты были оставлены за пределами поэтического корпуса[1].

В силу особенностей развития русской поэзии в XX веке многие значимые поэты были лишены возможности публиковаться в периодической печати и выпускать авторские сборники. При этом именно эти поэты зачастую оказывали решающее влияние на последующее развитие поэзии. Наиболее яркий пример здесь, возможно, Иосиф Бродский, который в значительной степени определял литературную ситуацию 1980–1990-х годов, несмотря на то, что его произведения не печатались в Советском Союзе. В то же время, неофициальные институции 1980-х годов — такие, как Премия Андрея Белого предлагают совсем иную картину литературы этого периода — картину, которая почти не находит точек пересечения с «официальной» и даже эмигрантской печатью. В связи с этим предпочтение той или иной картины («официальной», «неофициальной» и т. п.) оказывается методологически некорректным: во всех таких случаях мы имеем дело с принципиальным сужением литературного поля. Выборка, построенная на основании только одной из ветвей советской литературы, будет заведомо нерепрезентативна.

Но здесь «официальная» советская литература имеет преимущества: «официальные» поэты издавались большими тиражами и количество опубликованных ими произведений значительно превышает весь массив литературы «неподцензурной». При этом для тех поэтов, которые не привлекали к себе внимания исследователей и критиков в постсоветские десятилетия, мы не можем предложить ни одного критерия для включения в поэтический корпус: с точки зрения истории литературы и истории стиха произведения этих поэтов представляли собой однородную массу, внутри которой почти невозможно провести какое-либо структурирование. Количество публикаций в советской периодической печати, награждение «официозными» знаками отличия, очевидно, не подходит на роль структурного критерия в силу того, что основывается на идеологической конформности соответствующих текстов и авторов, а не на их реальном значении для литературной жизни эпохи.

[1] Эта проблема касается не только советских поэтов. См., например, недавнюю статью Олега Лекманова о массовой русской поэзии 1913 года (Лекманов 2013a; 2013b).

4. Официальная и неофициальная поэзия

Пожалуй, центральная проблема, с которой сталкиваются исследователи русской поэзии советского периода, — это проблема существования в рамках этой эпохи нескольких литературных миров, каждый из которых предполагал собственную иерархию текстов и авторов, собственный художественный и социальный контекст. Эти миры могли переплетаться друг с другом, образовывать сложные сочетания, но при этом они не формировали единого непрерывного пространства — такого, что было свойственно поэзии XVIII–XIX веков. Исследователи послевоенной советской поэзии обычно говорят о двух литературах — литературе «официальной» и «неофициальной», или «неподцензурной»[1]. «Неподцензурная» литература не была едина и появилась не сразу: ей предшествовало несколько попыток создания «альтернативных» литературных миров в довоенный период: к числу таких миров можно отнести литературу первой эмиграции, которая имела несколько региональных центров, и деятельность ряда поэтов Москвы и Ленинграда в 1930-е годы, которая была прервана репрессиями и затем войной[2] (здесь можно вспомнить группу Обэриу, московского поэта Георгия Оболдуева и некоторых других).

В силу разных причин именно «неподцензурная» литература (во всех своих вариантах) в последние двадцать лет привлекает к себе основное внимание исследователей поэзии. Именно в недрах этой литературы происходили знаменательные изменения поэтического языка и поэтической субъективности. Без «неподцензурной» поэзии невозможно представить поэзию 1990–2000-х годов, когда именно наследие неподцензурных поэтов оказывало решающее влияние на младшее поэтическое поколение, осознавшее себя продолжателями этой линии советской поэзии[3].

При этом метрика «неподцензурной» поэзии, несмотря на отдельные первопроходческие работы (например, Орлицкий 2002;

[1] Об истории этой дихотомии см.: Львовский 2012.

[2] «Предыстория» «неподцензурной» литературы до сих пор ждет своего исследователя, но довольно подробно описана в литературной критике. См. прежде всего: Юрьев 2013.

[3] См. например недавнее обобщение этого исследовательского опыта в монографии Житенев 2012.

Андреева, Орлицкий 2009), тем не менее, почти не исследована[1], в силу чего до сих пор не создано обобщающее описание метрики советского периода, которое невозможно без учета этой линии, тем более, что для развития литературы она оказалась определяющей. Все выводы, которые были сделаны М.Л. Гаспаровым о метрике советской поэзии, сделаны на материале «официальных» поэтов, и после ухода ученого работа по интегральному исследованию метрики советского периода так и не была продолжена. Поэтический корпус не может подменить такую работу, однако он может дать исследователю инструмент, который позволит ликвидировать лакуны и создать новое описание метрики советского периода, лишенное вынужденной предвзятости более ранних работ.

Это не означает, что создатели поэтического корпуса отказываются от «официальной» поэзии: напротив, в корпусе ей должно отводиться значительное место, однако для этой поэзии предполагаются более строгие (чем, например, в работах М.Л. Гаспарова) критерии отбора: наше внимание будут привлекать только те авторы, что были интегрированы в общий поэтический контекст, оставили значимый след в истории литературы и привлекали внимание исследователей и других поэтов. Диапазон таких авторов довольно широк: так, поэтический корпус был бы заведомо неполон без Александра Твардовского, Евгения Евтушенко или Николая Рубцова, которые несмотря на принадлежность к «официальной» ветви советской поэзии заняли важное, хотя и не всегда однозначное место в отечественной культуре XX века, но, в то же время, едва ли репрезентативность корпуса уменьшится, если в нем будут отсутствовать «официозные» поэты-лауреаты вроде Егора Исаева или Николая Грибачева (тем более что в современной филологической науке преобладает тенденция рассматривать творчество этих и подобных им поэтов с антропологических, а не с литературоведческих позиций — Гюнтер 1997).

При этом некоторые «официальные» поэты занимают крайне важное место в истории поэзии и стиха. Один из таких поэтов — Борис Слуцкий (1919–1986). Слуцкий был одним из наиболее ярких поэтов «фронтового» поколения, и хотя его стихи довольно часто

[1] Отдельное, но крайне значимое исключение здесь составляет метрика Иосифа Бродского, которой посвящен ряд работ, среди которых можно выделить: Семенов 1998, 2010; Андреева 2003; Власов 2005; Левашов, Ляпин 2012.

появлялись в советской печати, он оставил после себя большое количество неопубликованных текстов. Ученик Ильи Сельвинского, Слуцкий на протяжении всего творческого пути находился в живом диалоге с авангардной поэзией 1920-х годов, хотя понимал и воспринимал ее по-своему. Слуцкому посвящен ряд работ, с разных сторон проясняющих его место в общем контексте эпохи (прежде всего, Grimberg 2013), однако мы обратим внимание на значимость его метрики для русского стиха.

Рассмотрим 3-иктный дольник Слуцкого — размер, который согласно М. Л. Гаспарову, у послевоенных официальных советских поэтов проявлял значительную тенденцию к логаэдизации и упрощению метрических форм (Гаспаров 1974: 242–244). Среди прочего, для этого размера было характерно усиление постоянных анакрус, среди которых начинают преобладать анапестические, и упорядочивание внутренней структуры строки («равносложие в стихе уже не ограничивается метрическим рядом, а распространяется на весь стих»; там же: 235–236).

В собрании стихотворений Слуцкого 257 3-иктных дольника; среди них только 16 (6%) содержат переменную анакрусу, а в шестидесяти (23%) встречаются отдельные нарушения анакрусы. Таким образом, в 3-иктном дольнике Слуцкого наблюдается тенденция к удержанию постоянной анакрусы, и в 65% случаях эта анакруса — анапестическая. Эти выводы согласовываются с выводами М. Л. Гаспарова, однако упорядочивание анакрусы было лишь одним из элементов эволюции 3-иктного дольника — другой особенностью этой эволюции была логаэдизация, и именно в этом аспекте дольник Слуцкого кардинальным образом отличается от усредненного дольника эпохи.

3-иктные дольники Слуцкого демонстрируют следующее распределение форм (на материале выборки в 1000 строк):

I	II	III	IV	V	прочие	число стихов
37,7	20,3	30,6	3,2	7,4	0,8	1000

Эти данные можно сравнить с данными М.Л. Гаспарова (Гаспаров 1974: 239), который в своих подсчетах опирался на существенно более ограниченный корпус произведений поэта — лишь на те тексты, что были опубликованы в советской печати:

I	II	III	IV	V	прочие	число стихов
31,1	18,0	31,8	2,6	15,1	1,4	422

Легко заметить, что, несмотря на некоторое расхождение в показателях (особенно по I и V формам), в обоих случаях мы имеем дело с одним и тем же типом дольника, резко отличающимся от преобладающего в прочей «официальной» поэзии. Этот тип, в котором преобладают формы I — III, традиционно называется «есенинским». Согласно М. Л. Гаспарову, он был характерен для ранних стадий эволюции дольника: «Есенинский тип приблизительно соответствует общему облику дольника на начальной стадии его развития, цветаевский — на последней стадии; с помощью переходных типов мы можем даже построить схему постепенного перехода от есенинского и гумилевского типов дольника к цветаевскому» (Гаспаров 1974: 242). Таким образом, 3-иктный дольник Слуцкого ведет себя не так, как усредненный дольник эпохи.

Внимательное изучение подготовленных М. Л. Гаспаровым таблиц может показать, что усредненный, логаэдизированный дольник в 1950–1960-е годы был характерен для таких поэтов, как Михаил Матусовский, Алексей Недогонов, Александр Коваленков, Сергей Островой, Лев Озеров, Игорь Кобзев. Эти поэты занимали видное место в «официозном» литературном процессе, однако не привлекали к себе внимание исследователей и критиков поэзии в постсоветскую эпоху. Выше мы уже отмечали, что согласно нашим критериям эти поэты не попадают в поэтический корпус, и обращение к метрическим особенностям их стиха предоставляет еще один аргумент в пользу такого решения.

Тем не менее, между местом поэта в «официальной» литературной иерархии и значимостью его достижений нет прямого противоречия, и случай Бориса Слуцкого подтверждает это: Слуцкий публиковался в советской печати и был признанным «официальным» автором, однако уже особенности его стиха выделяют его на фоне усредненного стиля эпохи и заставляют внимательнее приглядеться к его поэтической практике и тому месту, которое она занимала в советской литературе.

Таким образом, работа по формированию поэтического корпуса каждый раз заставляет заново проводить работу по (ре)контекстуа-

лизации творчества того или иного поэта. Мы стремимся к такому отбору текстов и авторов, который учитывал бы место поэта в литературном процессе эпохи и степень внимания к нему со стороны следующих поколений. В силу этого принципы формирования корпуса русской поэзии XX века непосредственно зависят от компромисса между несколькими противопоставленными друг другу исследовательскими ориентирами. В этой статье мы описали, каким образом создатели поэтического подкорпуса НКРЯ пытаются достигнуть такого компромисса при работе с поэзией XX века.

Мы стремимся представить исследователю такую картину поэзии, которая позволяла бы делать выводы о состоянии метрики и поэтического языка для любого выбранного периода и объединяла бы в себе различные линии литературы XX века, несмотря на то, что они подчас довольно далеко отстояли друг от друга. Это требует переосмысления принципов отбора произведений и авторов: место поэта в поэтическом корпусе определяется в зависимости от его места в поэтическом контексте эпохи, в зависимости от того, привлекал ли он внимание исследователей и критиков поэзии, а также поэтов младших поколений. С нашей точки зрения, только опора на такие, учитывающие место поэта в литературном процессе, критерии позволяет поэтическому корпусу достигнуть репрезентативности, в отсутствии которой любое обобщающее исследование поэзии теряет смысл.

Литература

Андреева А. Н. Эволюция тонического стиха в поэзии Иосифа Бродского. Дисс. ... канд. филолог. наук. М., 2003.

Андреева А. Н., Орлицкий Ю. Б. Гетероморфный (неупорядоченный) стих в русской поэзии // Славянский стих VIII: Стих, язык, смысл. / Под ред. А.В. Прохорова, Т.В. Скулачевой. М.: Языки славянских культур, 2009. С. 365—389.

Власов К. А. Система стиха И.А. Бродского (Проблемы метра и ритма). Дисс. ... канд. филол. наук. Орел, 2005.

Гаспаров М. Л. Современный русский стих. Метрика и ритмика. М.: Наука, 1974.

Гаспаров М. Л. Очерк истории русского стиха. Метрика. Ритмика. Рифма. Строфика. Изд. 2 (доп.). М.: Фортуна Лимитед, 2002.

Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках Национального корпуса русского языка: общая структура и перспективы использования. // Национальный корпус русского языка: 2006—2008. Новые результаты и перспективы / Отв. ред. В.А. Плунгян. СПб.: Нестор-История, 2009. С. 71—114.

Гюнтер Х. Поющая родина (Советская массовая песня как выражение архетипа матери) // Вопросы литературы. 1997. № 4. С. 46—61.

Житенев А. А. Поэзия неомодернизма. СПб.: Инапресс, 2012.

Корчагин К. М. Метрическая амбивалентность в русском стихе («сверхкороткая» силлаботоника, силлабический стих, кольцовский пятисложник и гиперпеон) // Корпусный анализ русского стиха: Сборник научных статей. Вып. 2. Отв. ред. В.А. Плунгян, Л.Л. Шестакова. М.: Издательский центр «Азбуковник», 2014. С. 87—104.

Левашов А. М., Ляпин С. Е. Ритмико-синтаксическое строение «Прощальной оды»: к гексаметрической концепции шестииктного стиха Бродского // Иосиф Бродский: проблемы поэтики / Под ред. А.Г. Степанова, И.В. Фоменко, С.Ю. Артемова. М.: Новое литературное обозрение, 2012. С. 139—151.

Лекманов О. А. Русская поэзия в 1913 году (Часть первая) // Новое литературное обозрение. 2013. № 119. С. 140—174.

Лекманов О. А. Русская поэзия в 1913 году (Часть вторая) // Новое литературное обозрение. 2013. № 124. С. 180—211.

Львовский С. Две литературы: тридцать два года спустя // Русская поэзия от Пушкина до Бродского. Что дальше? / Под ред. К. Скандуры. Roma: Edizioni Nuova Cultura, 2012. С. 315—334.

Орлицкий Ю.Б. Стих и проза в русской литературе. М.: РГГУ, 2002.

Семенов В. Стихосложение Иосифа Бродского: Метрика; Строфика; Ритмика; Ритм и синтаксис. Дисс. ... Magister Artium. Tartu, 1998.

Семенов В. Полустишия в позднем неклассическом стихе И. Бродского: Квазицезура // Studia slavica: Сборник научных трудов молодых филологов: [Вып.] IX. Таллин: Tallinna Ülikooli Kirjastus, 2010. С. 199—222.

Соболев А.Л. Летейская библиотека: В 2 т. М.: Трутень, 2014.

Юрьев О. Заполненные зияния: Книга о русской поэзии. М.: Новое литературное обозрение, 2013.

Grinberg M. "A Am to Be Read Not from Left to Right, But in Jewish: from Right to Left": The Poetics of Boris Slutsky. Mass.: Academic Studies Press, 2013.

K. M. Korchagin
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
stivendedal@gmail.com

20TH-CENTURY POETRY IN THE RUSSIAN NATIONAL CORPUS: HOW TO CREATE A REPRESENTATIVE POETICAL CORPUS

The paper regards the principles of selection of poems for Russian National Corpus' poetical subcorpus. The developers of the corpus try to create an instrument that both simplifies studies in the field of metrics and analysis of poetical language and serves as a representative source of 20th-century poems. The structure of 20th-century Russian poetry is very complicated: the poetry was split between several areas that were almost independent of each other. Developing poetical subcorpus had to be representative for each of these areas (such as "official" poetry and "unofficial" one, *emigré* poetry, soviet poetry etc.). Poetical subcorpus has to be useful for two groups of scholars, for metricists and researchers of Russian poetry, and therefore corpus's developers have decided to reject a hierarchy of poets and poems that characterizes most of 20th-century poetry studies and concentrate on the reconstruction of 20th-century poetical context. Thereby we have included in the poetical subcorpus not only the most prominent or famous poems and poets, but also *poetae minores* that were crucial for their own epoch's poetical context nevertheless. Poetical subcorpus includes *poetae minores*, but doesn't include mass versification and/or *poésie naïve* that can still be interesting for metricists, but can reduce representativeness of the corpus in the same time.

Key words: metrics, theory of verse, poetic corpus, Russian National Corpus, representativeness, poetical context, 20th-century Russian poetry

References

Andreeva A. N. *Evolyutsiya tonicheskogo stikha v poezii Iosifa Brodskogo. Diss. ... kand. filolog. nauk*. [The evolution of tonic verse in the poetry of Joseph Brodsky]. Phd. diss. Moscow, 2003.

Andreeva A. N., Orlitskii Yu. B. [Heteromorphic (unregulated) verse in Russian poetry]. *Slavyanskii stikh VIII: Stikh, yazyk, smysl* [Slavic verse VIII: Verse, Language, Sense]. Ed. by A. V. Prokhorov, T. V. Skulacheva. Moscow, Yazyki slavyanskikh kul'tur Publ., 2009, pp. 365–389. (In Russ.)

Gasparov M. L. *Ocherk istorii russkogo stikha. Metrika. Ritmika. Rifma. Strofika* [Sketch of the history of Russian verse. Metrcis. Rhytmic. Strophic]. Moscow, Fortuna Limited, 2002.

Gasparov M. L. *Sovremennyi russkii stikh. Metrika i ritmika* [Modern Russian verse. Metrics and rhythmic]. Moscow, Nauka Publ., 1974.

Grinberg M. *"A Am to Be Read Not from Left to Right, But in Jewish: from Right to Left": The Poetics of Boris Slutsky*. Mass., Academic Studies Press, 2013.

Grishina E. A., Korchagin K. M., Plungyan V. A., Sichinava D. V. [Poetical corpus in National Russian corpus: structures and the ways of instrumentation]. *Natsional'nyi korpus russkogo yazyka: 2006—2008. Novye rezul'taty i perspektivy* [National Russian corpus: 2006—2008. New results and perspectives]. V. A. Plungyan (Ed.). St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 71–114.

Günter H. [The motherland singing: Soviet mass chants as expression of mother's archetype]. *Voprosy literatury*, 1997, no. 4, pp. 46–61. (In Russ.)

Yur'ev O. *Zapolnennye ziyaniya: Kniga o russkoi poezii* [Completed lacunae: The book about Russian poetry]. Moscow, Novoe literaturnoe obozrenie, 2013.

Korchagin K. M. [Metrical ambiguity in Russian verse ("ultrashort" accentual-syllabic verse, syllabic verse, Kolcov's pentameter and hyperpeon)]. *Korpusnyi analiz russkogo stikha: Sbornik nauchnykh statei* [The corpus analysis of Russian verse. Collection of articles].

V. A. Plungyan, L.L. Shestakova (Eds.), Moscow, Azbukovnik, 2014, pp. 87–104. (In Russ.)

Lekmanov O. A. [Russian poetry in 1913 (I)]. *Novoe literaturnoe obozrenie*, 2013, no. 119, pp. 140–174. (In Russ.)

Lekmanov O. A. [Russian poetry in 1913 (II)]. *Novoe literaturnoe obozrenie*, 2013, no. 124, pp. 180–211. (In Russ.)

Levashov A. M., Lyapin S. E. [Rhythmic and syntactic structure of Proshhal'naja oda: the problem of hexametrical origin of Brodsky's 6-icti verse]. *Iosif Brodskii: problemy poetiki* [Iosif Brodskii. Issues of poetics]. A. G. Stepanova, I. V. Fomenko, S. Yu. Artemova (Eds.), Moscow, Novoe literaturnoe obozrenie, 2012, pp. 139–151. (In Russ.)

L'vovskii S. [Two literatures: 32 years later]. *Russkaya poeziya ot Pushkina do Brodskogo. Chto dal'she?* [Russian poetry from Pushkin to Brodskii. What next?]. C. Scandura (Ed.). Roma, Edizioni Nuova Cultura, 2012, pp. 315–334. (In Russ.)

Orlitskii Yu.B. *Stikh i proza v russkoi literature* [Verse and prose in Russian literature]. Moscow, RSUH Publ., 2002.

Semenov V. [Hemistich in J. Brodsky's late non-classical verse: Quasicaesura]. *Studia slavica: Sbornik nauchnykh trudov molodykh filologov*, no. IX. Tallin, Tallinna Ülikooli Kirjastus, 2010, pp. 199–222. (In Russ.)

Semenov V. *Stikhoslozhenie Iosifa Brodskogo: Metrika; Strofika; Ritmika; Ritm i sintaksis* [Joseph Brodsky's versification: Metrics; Strophic; Rhythmic; Rhythm and Syntax]. Diss. ... Magister Artium. Tartu, 1998.

Sobolev A. L. *Leteiskaya biblioteka* [Lethe's library]. Moscow, Truten' Publ., 2014. (In Russ.)

Vlasov K. A. *Sistema stikha I.A. Brodskogo (Problemy metra i ritma). Diss. ... kand. filol. nauk*. [The system of Joseph Brodsky's verse (Problem of meter and rhythm)]. Phd. diss. Orel, 2005.

Zhitenev A. A. *Poeziya neomodernizma* [The poetry of neomodernism]. St. Petersburg, Inapress Publ., 2012.

[1]Е. А. Гришина, [2]Ю. Г. Зеленков, [3]Б. В. Орехов
[1]Институт русского языка им. В. В. Виноградова РАН, [2]Яндекс,
[3]Национальный исследовательский университет
«Высшая школа экономики»
[1,2,3](Россия, Москва)
[1]rudi2007@yandex.ru, [2]yuryz@yandex-team.ru,
[3]nevmenandr@yandex.ru

НАИВНАЯ ПОЭЗИЯ В АКЦЕНТОЛОГИЧЕСКОМ КОРПУСЕ

В статье рассматривается особенный материал, способный пополнить акцентологический корпус, специфический подкорпус в составе НКРЯ, отражающий место постановки ударений. Наивная поэзия – непрофессиональные стихи, созданные поэтами-любителями, их текстовая продукция не прошла редакторских фильтров и не была опубликована в авторитетных периодических изданиях и издательствах. Так как большинство этих текстов написаны в правильной силлабо-тонике, оказывается возможным автоматически предсказывать ударения и делать разметку для корпуса. Наивная поэзия была извлечена с сайта stihi.ru, самого старого сайта в России, публикующего подобного рода произведения поэтов-любителей. Несмотря на появление альтернативных площадок для публикации, сайт всё ещё остаётся популярным и количество публикаций на нём растёт. Для разметки текстов была применена специальная программа, предсказывающая расстановку ударений на основе машинного обучения. В тексте статьи приводится таблица, показывающая, насколько пополнение корпуса увеличило количество вхождений некоторых конкурирующих форм.

Ключевые слова: наивная поэзия, акцентология, корпусная лингвистика, лингвистические данные

Введенный в свое время фольклористами термин «наивная поэзия» (см. Неклюдов 2001; Минаева, Жигарина 2009) не имеет оценочной окраски. Речь идет о стихах, сочиняемых непрофессиональными литераторами. Определение непрофессиональной литературы наталкивается на известные методологические сложности, но в целом для нас это в первую очередь литература, не прошедшая редакторские фильтры, то есть не опубликованная в признанном сообществом периодическом издании или авторитетном издательстве [Bonch-Osmolovskaya, Orekhov 2014]). У такой текстовой продукции есть несколько содержательных и формальных особенностей, которые отличают ее от современной профессиональной поэзии. В частности, считается, что наивные поэты в большей степени ориентируются на воспроизведение усвоенных из школьной программы классических образцов, чем на создание принципиально новых произведений. Среди следствий этого принципа — доминирование силлабо-тоники в наивной поэзии. Действительно, так как традиционное представление о стихотворной речи предполагает, что поэтическое произведение должно быть написано одним из привычных двухсложных или трехсложных размеров, большинство непрофессиональных литераторов избегают неурегулированных размеров. Предварительные оценки, правда, свидетельствовали, что и современная авангардно-профессиональная поэзия также предпочитает метрическую организованность (хотя и в меньшей степени, чем непрофессиональные поэты) [Сонькин 2009].

Мы устраняемся от обсуждения художественных достоинств этих стихотворных произведений, оно должно быть делом литературных критиков. Для нас, как для составителей корпуса, наиболее важной является именно метрическая урегулированность большинства текстов наивных литераторов, позволяющая в автоматическом режиме предсказывать расстановку иктов в строках и, следовательно, ударений в словах, что представляется ценным при пополнении акцентологического подкорпуса НКРЯ. Другим ценным для составителя корпуса свойством наивной поэзии является ее доступность в последние годы. Непрофессиональные литераторы благодаря развитию Интернета получили широкие возможности для публикации своих произведений на специализированных сайтах, исповедующих принципы User Generated Content, то есть на таких, содержательное наполнение которых происходит за счет обычных пользователей, а

не нанятых администрацией ресурса редакторов или авторов. Одним из таких специализированных сайтов, на которых пользователь имеет простую возможность для публикации своего произведения, является сайт stihi.ru, — безусловно, старейший в Рунете ресурс, предоставляющий такие услуги. За время его работы пользователи stihi.ru успели объединиться в особую социокультурную среду (о некоторых принципах ее функционирования см. [Дианова 2009]). Даже с пришествием в сетевое пространство блогов и социальных сетей популярность stihi.ru не падает. На рис. 1 изображен график количества стихотворений, публикуемых в день на stihi.ru.

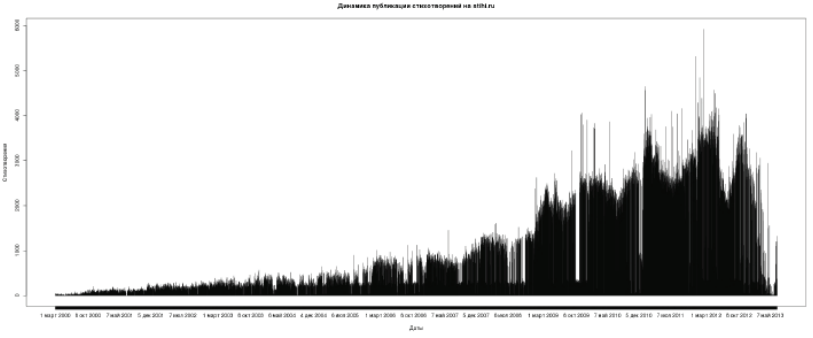


Рис. 1. Динамика публикации стихотворений на stihi.ru

В составе НКРЯ есть поэтический подкорпус, и размечаемые для него тексты автоматически попадают также и в акцентологический подкорпус, так как метрическая разметка предполагает расстановку ударений. Однако наивная поэзия не вписывается в концепцию пополнения поэтического подкорпуса, предполагающую включение культурно значимых и прошедших редакторские фильтры текстов (см. об этом статью К. М. Корчагина в настоящем сборнике). Тем не менее, наивные поэтические тексты представляют интерес для акцентологического корпуса как ценный и объемный материал для изучения актуального состояния русской акцентологии[1].

Локальная копия сайта stihi.ru по состоянию на декабрь 2013 года предоставлена разработчикам корпуса Поиском Mail.Ru, ко-

[1] По сообщению К. К. Богатырёва, А. А. Зализняк еще в доинтернетную эпоху говорил о пользе для акцентологии не принятых редакциями стихов.

торым мы приносим свою благодарность. Из общего объема более 300 млн словоупотреблений нами было отобрано два набора текстов общим объемом примерно 17 млн словоупотреблений. Первый набор текстов (приблизительно 10 млн словоупотреблений) был отобран случайным образом. В отборе второго (около 6,5 млн) учитывался процент совпадений биграммов текста с набором словесных биграммов, составленным по поэтическому корпусу НКРЯ (во второй набор текстов допускались только такие, у которых совпадение было не меньше 60%). Ожидалось, что высокий процент совпадения биграммов с поэтическим корпусом даст и большее число правильных силлабо-тонических текстов. Это соображение основывалось на приводившемся выше принципе, согласно которому наивная поэзия развивается в орбите существующих литературных стереотипов. БОльшая стереотипность в лексике и сочетаниях слов должна означать и бОльшую метрическую стереотипность.

На все отобранные тексты с помощью программы, разработанной Ю. Г. Зеленковым, была автоматически наложена метрическая разметка, то есть каждой строке присвоена помета, сообщающая, каким метром эта строка написана, а также учитывающая стопность и клаузулу.

Алгоритм автоматической расстановки ударений в использованной программе метрической разметки является открытым, т.е. обрабатывает любые русские словоформы бессловарным способом. Алгоритм имеет точность 0.96-0.97 для нарицательной лексики и 0.94-0.95 для имен собственных. В основе алгоритма лежит идея о существовании в русском языке высокой корреляции между буквенным составом концов слов и их грамматическими характеристиками, в частности, позицией ударения.

В качестве иллюстрации ниже, в таблице 1, приводится фрагмент большого (около 150 элементов) списка словоформ, упорядоченных в обратном лексикографическом порядке.

Таблица 1. Словоформы, упорядоченные в обратном лексикографическом порядке

*вольнослу`шательницами
свиде`тельницами
лжесвиде`тельницами

благоде`тельницами
избави`тельницами
прави`тельницами
…
состави`тельницами
вдохнови`тельницами
усынови`тельницами
покрови`тельницами
заяви`тельницами
победи`тельницами
…
руководи`тельницами
сопроводи`тельницами
роди`тельницами
распоряди`тельницами
жи`тельницами
скази`тельницами
…
проси`тельницами
искуси`тельницами
соврати`тельницами
посети`тельницами
похити`тельницами
укроти`тельницами
…
учи`тельницами
утеши`тельницами
прия`тельницами
настоя`тельницами
ва`фельницами
*отше`льницами

Из таблицы хорошо видно, что все без исключения словоформы (кроме самых крайних, которые показывают границы данного списка и сами входят в другие, аналогичные списки) имеют ударение в одной и той же позиции. В таком случае, наш список может быть сокращен всего до одной записи, т.е. примерно в 150 раз, как показано в таблице 2.

Таблица 2. Сокращенный вариант списка

*вольнослу`шательницами
свиде`тельницами
* отше`льницами

Другими словами, словоформа «свиде`тельницами» представляет целый класс словоформ, у которых правильная позиция ударения определяется на основе совпадения их конечных букв с конечными буквами представителя класса.

Этот подход был использован для машинного обучения, в результате которого, с помощью последовательных итераций, был построен достаточно большой список представителей «акцентологических классов», обеспечивающий вполне приемлемые показатели точности при расстановке ударения для любых русских слов.

Процесс анализа стихотворения включает три этапа. На первом этапе для каждой строки (стиха) выполняется расстановка ударений. Затем метрические ударения (икты) основных метров силлабо-тоники накладываются на словесные ударения и путем сравнения их позиций определяются наиболее вероятные метр, стопность и клаузула строки. При этом может распознаваться сверхсхемное ударение и корректироваться исходная расстановка ударения («и го`рьки мне`, горьки` твои` упрё`ки»).

На втором этапе анализируется структура строфы. Поскольку анализ отдельных строк может быть неоднозначным, число комбинаций, после объединения этих строк в строфы, часто оказывается больше одного. При снятии этой неоднозначности используется энтропия, вычисляемая для каждого варианта структуры строфы. В качестве основного выбирается вариант с минимальной энтропией.

Наконец, на третьем этапе, если структуры некоторых строф «выбиваются» из общей картины, используется метрическая инерция, т.е. делается попытка изменить метры отдельных строк на наиболее частотные в данном стихотворении. Здесь также возможна корректировка словесного ударения.

В заключение выполняется фонетическая транскрипция всех строф и определяются их схемы рифмовки.

Выяснилось, что для большей точности метрического анализа к текстам наивных поэтов требуются применять особые правила. В

частности, в случаях, которые программа могла бы трактовать двояко, при выборе между гетерометрией и нарушением стопности при сохранении метра, алгоритму следует предпочесть второй вариант: наивные авторы чаще не могут выдержать стопность, чем используют иной метр.

Все строки, для которых не был распознан силлабо-тонический размер, были помечены как «НУР» (неурегулированный размер). Стихотворения, в которых строк с НУР было больше 30%, были отброшены и в корпус в итоге не попали.

Благодаря автоматической метрической разметке была проверена гипотеза о корреляции между лексической и сочетаемостной традиционностью, с одной стороны, и метрической урегулированностью, с другой. В таблице 3 показано общее число строк и число строк в неурегулированных размерах в первом (отобранном случайно) и втором (отобранном неслучайно) наборах текстов.

Таблица 3. Наборы данных и неурегулированные размеры

Наборы данных	Строки	НУР
1	2 186 617	211 244
2	1 416 562	118 972

Была проведена статистическая проверка гипотезы с использованием критерия хи-квадрат и V Крамера: X-squared = 1644.823, df = 1, p-value < 2.2e-16, Cramer's V: 0.02136568.

Эти цифры свидетельствуют, что зависимость между наборами данных и «неурегулированными размерами» отсутствует.

Конечный результат составил пополнение акцентологического корпуса на 14,3 млн словоупотреблений (162,8 тыс. стихотворений).

Приведенная ниже таблица 4 позволяет оценить существенность пополнения акцентологического корпуса с точки зрения присутствия в текстовой коллекции некоторых «проблемных» случаев вариативности ударения.

Таблица 4. Вариативные словоформы в пополнении акцентологического корпуса

словоформа	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
балованный	0	0	0	0	0	1
баловать	0	6	7	13	2	16
балую	0	2	6	8	1	3
бедны	2	28	29	59	2	8
бледны	1	84	75	160	1	10
близки	6	133	79	218	25	137
бодры	0	13	7	20	0	2
важны	1	32	35	67	21	99
верны	5	54	59	118	3	57
видны	17	135	167	219	24	138
вкусны	1	14	4	19	0	4
влажны	2	11	4	17	0	11
вольно	6	319	17	342	22	103
вольны	2	44	23	69	3	28
воспринял	0	28	1	29	14	19
восприняла	0	0	0	0	0	1
восприняли	0	16	0	0	14	6
вредны	2	25	10	40	3	2
глупы	3	41	5	49	1	51
гнусны	1	14	0	15	0	0
годны	1	6	5	12	0	6
горды	2	63	47	112	4	17
горстей	0	0	3	3	0	2
горьки	3	44	6	53	0	37

слово-форма	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
грешно	4	2	123	129	1	44
грешны	0	11	10	21	2	20
грозны	4	59	5	68	0	3
грубы	3	81	15	99	1	15
грузны	0	3	4	7	0	2
грустны	2	24	15	41	0	35
грязны	2	8	3	13	0	3
дружны	1	38	16	55	1	13
душны	0	6	0	6	0	1
жадны	0	23	2	25	0	1
жирны	0	5	5	10	0	3
завит	0	2	8	10	1	0
залит	1	38	23	62	0	38
звучны	0	21	4	25	0	2
знатны	2	7	0	9	0	0
йогурт	0	0	0	0	0	6
квартал	2	2	64	68	6	12
кислы	0	3	0	3	0	1
колки	0	15	1	16	0	9
крепки	3	61	44	108	0	29
круглы	1	6	15	22	0	4
ловки	0	8	11	19	0	1
милы	2	206	31	239	2	68
мокры	2	9	10	21	0	8
мрачны	0	72	8	80	0	11
мудры	1	21	5	27	3	6
мутны	0	15	3	18	0	4
мягки	3	22	2	27	0	15
налит	1	22	11	34	0	3

слово-форма	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
нежны	3	138	47	188	0	84
низки	0	38	0	38	0	2
новы	2	142	13	157	0	17
нужны	13	194	901	1108	228	1209
областей	1	0	20	21	8	4
плотны	0	3	1	4	0	2
полно	60	745	289	1094	47	278
полны	13	391	415	819	2	303
предпринял	1	15	0	16	2	5
пресны	0	1	0	1	0	2
прожит	1	44	2	47	0	48
пройден	1	21	4	26	3	36
простыней	0	0	3	3	0	24
прочны	0	11	4	15	2	10
пышны	1	62	3	66	0	3
пьяны	5	76	31	112	4	19
редки	2	66	2	70	1	31
резвы	2	15	3	20	0	3
ровны	0	9	3	12	0	2
серьгам	0	0	3	3	0	0
серьгами	0	2	5	7	1	1
серьгах	1	7	9	17	0	2
сильны	10	76	73	159	10	64
скромны	1	16	10	27	0	9
скудны	2	18	1	21	0	5
скучны	1	38	16	55	1	9
слабо	10	207	24	241	25	77
слабы	4	90	16	110	2	33

слово-форма	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
сложны	0	3	7	10	2	17
слышны	12	181	63	256	3	168
смелы	0	45	11	56	0	9
смуглы	0	8	7	13	0	0
сочны	1	8	2	11	0	5
спелы	1	8	1	10	0	0
средам	0	4	1	5	0	7
страшны	9	154	87	250	2	233
строги	2	127	5	234	0	19
стройны	0	29	25	54	0	13
тверды	7	57	23	87	0	9
творог	2	8	16	26	6	7
теней	9	78	607	694	4	250
тесны	1	18	6	25	0	17
тихи	5	87	31	123	0	16
толсты	1	8	6	15	0	2
тонки	1	51	18	70	0	10
точны	0	14	6	20	2	7
трезвы	0	4	0	4	0	5
трудны	1	15	4	20	1	6
тусклы	0	21	4	25	0	2
тучны	0	22	3	25	0	0
узки	0	21	12	33	2	3
узко	1	41	0	42	5	5
хитро	3	60	59	122	8	59
хитры	0	9	11	20	2	6
храбры	4	22	5	31	0	1
черствы	1	1	4	6	0	3
честны	3	12	8	23	0	15

словоформа	До пополнения акцентологического корпуса					В пополнении с текстами stihi.ru
	Без уд	1 вар. уд	2 вар. уд	Всего	2000-е (1 785 992 слов)	
чисты	6	91	57	154	1	86
чужды	5	216	29	250	2	104
шумны	0	35	2	37	0	1
шустры	0	0	1	1	0	1
щедры	0	13	10	23	4	9
щелей	1	14	28	43	2	18
ярки	1	90	7	98	0	16
ясны	4	158	61	223	1	45

Как видим, пополнение акцентологического корпуса в «проблемных» точках — по сравнению с предшествующим состоянием данных, относящихся к XXI веку, — достаточно существенно (ячейки, которые демонстрируют особенно заметную разницу, выделены полужирным). Таким образом, очевидно, что пополнение акцентологического корпуса наивными поэтическими текстами дает специалистам по современной русской акцентологии принципиально новый материал, который позволяет изучать состояние и тенденции развития современной русской акцентологической системы на статистически значимом материале, который, кроме того, в значительной степени отражает живое русское произношение, а не рекомендации нормативных справочников, — что позволяет, очевидно, как строить предположения о тенденциях развития системы, так и более полно описывать реализацию тенденций, заложенных в языке предшествующими этапами его развития[1].

[1] В дальнейшем возможности программы Ю.Г. Зеленкова предполагается несколько расширить за счет автоматической разметки в поэтических текстах т.н. «зоны рифмовки», т.е. тех слов и словосочетаний, которые включают в себя последнюю клаузулу строки. Это позволит исследовать проблемы, связанные, в частности, с выбором между *е* и *ё* в ряде лексем и словоформ (например, не вполне очевидно, какое в настоящий момент предпочитается произнесение — *крестный/крёстный*, *Крез/Крёз* и под.). И такие данные можно получить либо из прозаического модуля акцентологического корпуса (где анализируемые лексемы встречаются довольно редко), либо из его поэтического модуля. И здесь данные о зоне рифмовки в наивных поэтических текстах могут оказаться совершенно бесценными.

Литература

Гришина Е. А. Микроизменения в акцентологической системе русских прилагательных по материалам Национального корпуса русского языка // Вопросы культуры речи. Вып. 11. М., 2012

Дианова Т. Б. «Эффект Сидоренко»: образ виртуальной личности в сети и манипуляция массовым сознанием // Интернет и фольклор. М.: Государственный республиканский центр русского фольклора, 2009. С. 32–43.

До и после литературы: тексты наивной словесности. / Сост. А. П. Минаева. Отв. ред. Е. Е. Жигарина. М., 2009.

«Наивная литература»: исследования и тексты / Сост. С. Ю. Неклюдов. М., 2001. [URL: http://www.ruthenia.ru/folklore/luriem43.pdf]

Савчук С. О. Активные процессы в системе русского словоизменения: опыт корпусного исследования акцентологических норм // Труды II Международной конференции «Русский язык и литература в международном образовательном пространстве: современное состояние и перспективы». Т. 2. Гранада, 2010. С. 1549–1554.

Сонькин В. Традиционный стих в контексте современной русской поэзии // Славянский стих. VIII: Стих, язык, смысл. М: Языки славянских культур, 2009. С. 390–393.

Bonch-Osmolovskaya A., Orekhov B. Distant reading of naïve poetry: corpora comparison as research methodology // Digital humanities Lausanne — Switzerland '14. URL: http://dharchive.org/paper/DH2014/Paper-777.xml

Larsson J. О некоторых изменениях тенденций акцентного развития прилагательных в современном русском языке // Russian Linguistics. 2006. Vol. 30. P. 235–262

[1]E. A. Grishina, [2]Yu. G. Zelenkov, [3]B. V. Orekhov
[1]Vinogradov Russian Language Institute
of the Russian Academy of Sciences,
[2]Yandex, [3]National Research University "Higher School of Economics"
[1,2,3](Russia, Moscow)
[1]rudi2007@yandex.ru, [2]yuryz@yandex-team.ru,
[3]nevmenandr@yandex.ru

NAÏVE POETRY IN ACCENTOLOGIC CORPUS

The article deals with specific material that is able to enlarge the Accentologic Corpus, the subcorpus in the RNC reflecting accent patterns in Russian words. Naïve poetry is the term for unprofessional poems written by amateur poets. Their textual products have not passed any editorial filters and have not been published in reputable periodicals and publishing houses. Since majority of these texts are written in the correct syllabic-tonic, it is possible to predict stress automatically and make the markup for the Corpus. Examples of naive poetry have been downloaded from the site stihi.ru, the oldest one in Russia that publishes such works by amateur poets. Despite the existence of alternative platforms for publication, the site is still popular and the number of publications is on the rise. A special program for the markup of texts was used. This program based on machine learning predicts the place of the accents. There is a table in the article that shows how the enlargement has increased the number of occurrences of some competing forms in Corpus.

Key words: naïve poetry, accentology, corpus linguistics, linguistic data.

References

Grishina E. A. [Micromutations in accentologic system of Russian adjectives based on Russian National Corpus]. *Voprosy kul'tury rechi* [Questions of speech culture]. Issue 11. Moscow, 2012. (In Russ.)

Dianova T. B. [“Sidorenko Effect”: an image of virtual personality on the web and mass conscience manipulation]. *Internet i fol'klor* [Internet and Folklore]. Moscow, Gosudarstvennyi respublikanskii tsentr russkogo fol'klora Publ., 2009, pp. 32–43. (In Russ.)

Do i posle literatury: teksty naivnoj slovesnosti [Before and after a

literature: naïve texts]. A. P. Minaeva (comp.), E.E. Zhigarina (ed.). Moscow, 2009.

«Naivnaja literatura»: issledovanija i teksty [Naïve literature: studies and texts]. S. Yu. Nekljudov (comp.). Moscow, 2001. Available at URL: http://www.ruthenia.ru/folklore/luriem43.pdf (accessed on 14.06.2015)

Savchuk S. O. [Active processes in the system of Russian inflections: a corpus-based study of accentologic norms] *Trudy II Mezhdunarodnoi konferentsii "Russkii yazyk i literatura v mezhdunarodnom obrazovatel'nom prostranstve: sovremennoe sostoyanie i perspektivy"* [Proceedings of the IInd International conference "Russian language and literature in the international educational space: current status and prospects"]. B. 2. Granada, 2010, pp. 1549-1554. (In Russ.)

Son'kin V. [Traditional verse in a context of contemporary Russian poetry]. *Slavjanskij stih. VIII: Stikh, jazyk, smysl* [Slavic verse VIII. Verse, language, meaning]. Moscow, Jazyki slavyanskikh kul'tur Publ., 2009, pp. 390–393. (In Russ.)

Bonch-Osmolovskaya A., Orekhov B. Distant reading of naïve poetry: corpora comparison as research methodology. *Digital humanities* Lausanne — Switzerland '14. Available at URL: http://dharchive.org/paper/DH2014/Paper-777.xml (accessed on 14.06.2015)

Larsson J. [About some mutations of tendencies of accentologic evolution of adjectives in contemporary Russian language]. *Russian Linguistics*, 2006, vol. 30, pp. 235–262.

[1]П. В. Дяченко, [2]Л. Л. Иомдин, [3]А. В. Лазурский, [4]Л. Г. Митюшин,
[5]О. Ю. Подлесская, [6]В. Г. Сизов, [7]Т. И. Фролова, [8]Л. Л. Цинман
[1,2,3,4,5,6,7,8]Институт проблем передачи информации
им. А.А. Харкевича РАН
[1,2,3,4,5,6,7,8](Россия, Москва)
[1]pavelvd@iitp.ru, [2]iomdin@iitp.ru, [3]lazur@iitp.ru, [4]mit@iitp.ru,
[5]olga@iitp.ru, [6]sizov@iitp.ru, [7]tfrolova@iitp.ru, [8]cinman@iitp.ru

СОВРЕМЕННОЕ СОСТОЯНИЕ ГЛУБОКО АННОТИРОВАННОГО КОРПУСА ТЕКСТОВ РУССКОГО ЯЗЫКА (СИНТАГРУС)[1]

В статье излагаются основные особенности, принципы создания и параметры синтаксически аннотированного корпуса русских текстов «СинТагРус». Помимо синтаксической разметки каждого предложения в виде деревьев зависимостей, корпус содержит информацию об аргументах и значениях лексических функций слов, входящих в предложение, а также сведения о лексических значениях слов. Рассматривается подкорпус предложений, содержащих различные виды эллипсиса. Обсуждаются возможности применения корпуса для решения научных и практических задач.

Ключевые слова: СинТагРус, синтаксически размеченный корпус русских текстов, грамматика зависимости, лексические функции, разрешение неоднозначности, эллипсис.

Вводные замечания

Глубоко аннотированный корпус русских текстов (СинТагРус) в течение последнего десятилетия разрабатывается Лабораторией компьютерной лингвистики ИППИ РАН. Он основан на идеологии

[1] Данная работа выполнена при частичной поддержке РФФИ (грант № 13-06-00756).

многоцелевого лингвистического процессора ЭТАП-3 (см., в частности, Apresjan et al. 2003) и в первую очередь синтаксического анализатора (парсера) русского языка, используемого в различных приложениях лингвистического процессора, в том числе в системе русско-английского машинного перевода. СинТагРус – это составная, но полностью автономная часть Национального корпуса русского языка. Материалы корпуса (за исключением информации о дополнительных типах разметки, см. ниже раздел 2) доступны на сайте НКРЯ по адресу http://ruscorpora.ru/search-syntax.html.

Корпус СинТагРус содержит русские тексты, снабженные полной морфосинтаксической разметкой со снятой омонимией. Общий объем корпуса на середину 2015 года составляет около 64000 предложений (свыше 925 000 словоупотреблений) и постоянно растет. Большой объем корпуса и детальность содержащегося в нем материала делают его весьма ценным и уникальным ресурсом, который может быть успешно использован в широком классе исследовательских и практических задач компьютерной лингвистики (см. об этом, например, Иомдин, Иомдин 2013).

Тексты, включенные в СинТагРус, относятся к художественному, научно-популярному, публицистическому, биографическому и новостному жанрам. Выбор текстов этих жанров позволяет отразить в синтаксически аннотированном корпусе современное состояние русского литературного языка и в то же время ограничить появление в корпусе материала, синтаксическая, а частично и лексическая кодификация которого затруднительна (разговорная речь, поэзия, диалектные тексты, техническая документация). Как следствие, материал корпуса оказывается синтаксически достаточно однородным, что приводит к повышению качества синтаксической разметки и достижению большего единообразия принимаемых лингвистических решений.

В последние годы создатели корпуса ведут активную работу над обогащением синтаксически аннотированного корпуса лексико-функциональной разметкой. Лексические функции понимаются в соответствии с теорией лексических функций И. А. Мельчука и с учетом корректив, внесенных в нее Ю. Д. Апресяном. Около 14 000 предложений размечены лексическими функциями, осуществляется поиск по аргументам и значениям лексических функций, и в ближайшем будущем он также будет доступен для исследовательских целей на сайте НКРЯ.

Этапы развития корпуса и различные аспекты, связанные с его созданием, поддержкой и использованием, неоднократно публиковались в литературе (см., в частности, Boguslavsky et al., 2000; Апресян и др., 2005; Богуславский и др., 2008а,б; Шеманаева, Фролова, 2010; Boguslavsky 2014).

Настоящая статья отражает современное состояние СинТагРус'а. Статья содержит пять разделов. В первом разделе излагаются общие принципы построения и синтаксического аннотирования корпуса. Второй раздел посвящен дополнительным видам разметки корпуса: сведениям о лексических функциях и информации о наличии в предложении эллипсиса разных типов. В третьем разделе говорится о применении корпуса СинТагРус для регрессионного тестирования синтаксического парсера ЭТАП-3, а также для построения на его основе гибридного парсера русского текста. Четвертый раздел коротко характеризует исследования и разработки, выполненные в других коллективах, использующих СинТагРус. Наконец, пятый раздел посвящен ближайшим перспективам развития корпуса, в частности, автоматическому пополнению корпуса за счет части основного корпуса НКРЯ, а именно, подкорпуса со снятой омонимией.

1. Построение корпуса и принципы синтаксического аннотирования

Построение глубоко аннотированных корпусов текстов разных языков является весьма актуальной и быстро развивающейся областью современной компьютерной лингвистики. Это обусловлено, в частности, тем, что при создании систем автоматической обработки текстов все более широко применяются методы машинного обучения, и размеченные корпусы представляют собой именно те массивы данных, на которых такое обучение проводится. СинТагРус тут не является исключением; об использовании СинТагРус'а в машинном обучении различных компьютерно-лингвистических систем см. Богуславский и др. 2008а,б, Nivre et al. 2008, Казенников 2010, Казенников, Куракин 2011, Boguslavsky et al. 2011, Boguslavsky 2014.

Важной особенностью данного корпуса является представление синтаксических структур предложений в виде деревьев зависимостей. При этом различается около 70 типов синтаксических связей между словами, что обеспечивает более полное и лингвистически

содержательное представление синтаксической структуры, чем в других существующих корпусах с синтаксической разметкой.

В настоящее время существует не менее 70 корпусов для различных языков, в которых аннотация достигает синтаксического уровня, так называемых treebanks (в том числе не менее 15 для английского языка). СинТагРус занимает здесь особое место, являясь единственным в мире большим стопроцентно отредактированным корпусом текстов на русском языке с аннотацией на морфосинтаксическом уровне.

Корпус СинТагРус создается следующим образом. Вначале непрепарированный текст обрабатывается программой сегментации текста, которая автоматически разбивает его на отдельные предложения. После этого каждое предложение пропускается через парсер многофункционального лингвистического процессора ЭТАП-3, в результате чего вырабатывается его синтаксическая структура (в виде дерева зависимостей). Парсер выполняет также лексико-семантическую и лексико-функциональную (см. раздел 2.1) разметку предложения.

Чтобы лучше представить себе арсенал синтаксических решений, используемых в корпусе СинТагРус, рассмотрим примеры предложений, синтаксические структуры которых не относятся к числу самых простых.

На рис. 1 представлена синтаксическая структура предложения

(1) *Те детали, без которых прошлое не складывается в гармоничную картину, нарушает покой и расстраивает сон*

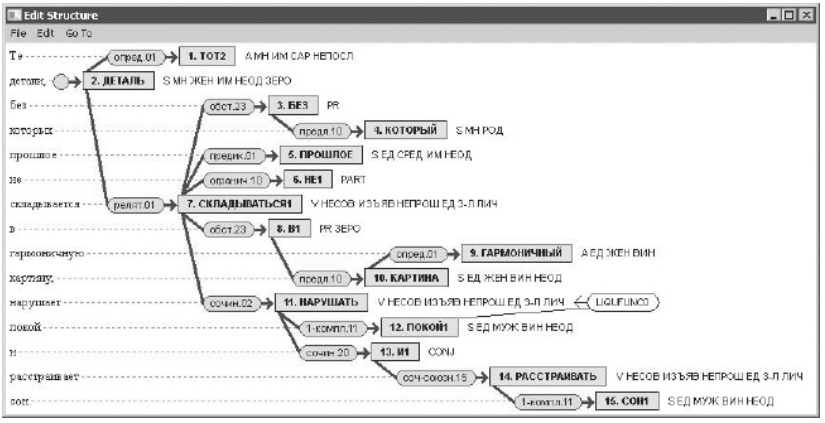


Рис. 1. Синтаксическая структура предложения (1).

Как видно из рисунка, в узлах дерева зависимостей стоят слова предложения, представленные именами лексем (в прямоугольниках) и цепочками грамматических характеристик (справа от прямоугольников), а ветви помечены именами синтаксических отношений (в овалах). Как явствует из синтаксической структуры, предложение (1) — номинативное, его вершиной является существительное *детали*. Справа от него стоит придаточное определительное, вершина которого — личный глагол *складывается* — присоединяется к существительному по релятивной синтаксической связи. Этот глагол открывает сочинительную цепочку, куда входят еще два глагола — *нарушает* и *расстраивает*.

Лексико-семантическая разметка заключается в том, что многозначным словам приписывается одно из значений, имеющихся в комбинаторном словаре процессора ЭТАП-3. Так, в (1) обнаруживается лексическая многозначность, которая разрешается парсером (или лингвистом-экспертом при редакции), и мы видим результаты разрешения этой многозначности: в предложении фигурирует а) (местоименное) прилагательное ТОТ2, а не местоименное существительное ТОТ1 (ср. *Жалок тот, в ком совесть нечиста*), глагол СКЛАДЫВАТЬСЯ1 в значении ‘происходить, завершаться’ (ср. *Всё сложилось хорошо*) в отличие от глагола СКЛАДЫВАТЬСЯ2 в значении ‘состоять’ (как в *Роман складывается из нескольких частей*), союз И1, а не частица И2 (как в *Я и не знал, что он в городе*), существительное ПОКОЙ1 в значении ‘спокойное состояние’, а не существительное ПОКОЙ2 в значении ‘комната’ (ср. *царские покои*) и существительное СОН1 в значении ‘состояние с минимальным уровнем мозговой деятельности’ в отличие от существительного СОН2 в значении «сновидение».

На следующем рисунке представлена СинтС предложения

(2) *Что говорить: корпус действительно нуждается в капитальной реконструкции (это вовсе не означает его полный снос)*

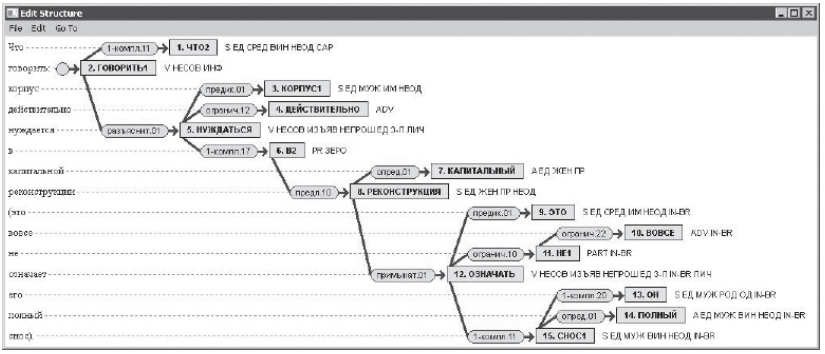


Рис. 2. Синтаксическая структура предложения (2).

Вершиной предложения (2) является глагол в инфинитиве *говорить*, лексическое значение которого ГОВОРИТЬ1 ‘сообщать’ (в отличие от ГОВОРИТЬ2 ‘разговаривать’, как в *говорить по-французски*). Этому глаголу подчиняется по 1-му комплетивному отношению местоименное существительное ЧТО2 в винительном падеже (ЧТО1 — это союз), тем самым констатируется, что *что* выступает в качестве прямого дополнения. Разумеется, это решение не отражает того факта, что *что говорить* — это своего рода синтаксическая фразема со значением ≈ ‘незачем говорить вследствие очевидности’, но на данном этапе синтаксически размеченный корпус такую информацию передать не может.

Отметим еще несколько деталей СинтС предложения (2): в нем фигурирует лексема КОРПУС2 ‘здание’ (а не КОРПУС2 ‘совокупность’ как в *корпус текстов*, КОРПУС3 ‘оболочка’ как *часы в золотом корпусе* и не КОРПУС3 ‘тело, туловище’, как в *подался всем корпусом*), лексема СНОС1 ‘разрушение’ (а не СНОС2 ‘смещение’, как в *снос корабля течением*); вводное предложение в скобках своей вершиной *означает* присоединяется к предшествующему существительному *реконструкции*.

(3) *«Теперь мы, пользуясь статистическими и компьютерными методами, будем изучать синтаксис и грамматическую структуру надписей из долины Инда», – говорит Раджеш Рао.*

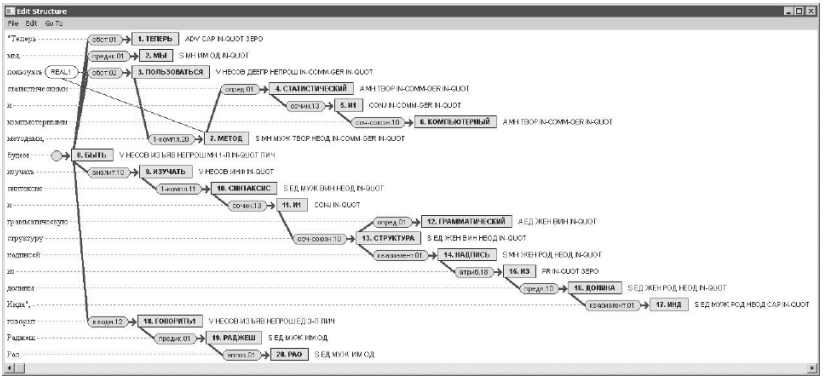


Рис. 3. Синтаксическая структура предложения (3).

Вершиной этого предложения является глагол *будем*, ему по вводному синтаксическому отношению подчиняется глагол со значением ГОВОРИТЬ1. Тут все дело в порядке слов: если бы глагол *говорит* предшествовал глаголу *быть*, то он был бы вершиной предложения и глагол быть подчинялся бы ему по 1-му комплетивному отношению (как в предложении *Раджеш Рао говорит: теперь мы будем изучать синтаксис*). В данной фразе мы видим также примеры двух сочинительных конструкций — сочинение прилагательных статистическими и компьютерными (методами) и существительных, синтаксис и (грамматическую) структуру.

Фактически почти все затраты труда при разработке корпуса приходятся на последний этап работы, выполняемый специально подготовленными аннотаторами-лингвистами. Хотя лингвистический процессор ЭТАП-3 насыщен весьма богатой и разнообразной лингвистической информацией, он не способен строить морфосинтаксические структуры со стопроцентной надежностью. Аннотаторы проверяют все элементы структур, созданных парсером процессора, и вносят необходимые коррективы с помощью специального программного комплекса «Редактор структур» (основная составная часть программной среды «Рабочее место аннотатора»).

Использование процессора ЭТАП-3 для разработки корпуса создает обратную связь, важную для самого процессора, поскольку при выполнении этой работы база русских грамматических правил и русский словарь процессора постоянно пополняются и совершен-

ствуются. Отметим, что процессор рассчитан на обработку русских и английских текстов без каких-либо тематических и лексико-грамматических ограничений. Русский морфологический словарь в настоящее время содержит свыше 130 000 входов, а комбинаторный словарь — более 105 000 входов.

2. Дополнительные виды разметки корпуса

2.1. Лексико-функциональная разметка.

В ходе лексико-функциональной разметки обнаруживаются и отмечаются словосочетания, допускающие интерпретацию в терминах лексических функций; при этом также используется информация, содержащаяся в комбинаторном словаре. Затем структуры, построенные системой ЭТАП-3, проверяются экспертами-лингвистами.

Главный тезис теории лексических функций звучит так: в языке можно выделить несколько десятков значений высокого уровня абстракции (‘высокая степень’, ‘начало’, ‘прекращение’, ‘каузация’, ‘ликвидация’, ‘манифестация’ и т.п.), каждое из которых выражается большим классом слов. При этом выбор конкретного слова W для выражения данного значения ‘S’ целиком зависит от того слова X (аргумента ЛФ), с которым оно сочетается в тексте. Он идиоматичен, т.е. семантически не полностью мотивирован.

Маркировка ЛФ непосредственно в текстах и удобные средства поиска обеспечивают возможность прямого наблюдения контекстов, в которых реализуются ЛФ. Эти данные также важны для компьютерных систем переработки текстов на естественных языках. Системы, использующие ЛФ как инструмент анализа и перевода текстов, обладают преимуществом по сравнению с системами, которые их не используют. Кодируя универсальные смыслы, ЛФ становятся своеобразным языком-посредником между разными языками и позволяют описывать значительную часть лексики каждого рабочего языка независимо от других языков. Между тем, в большинстве систем машинного перевода правила перевода носят принципиально двуязычный характер и требуют постоянной корректировки уже имеющихся словарных данных при включении в систему новых языков.

В СинТагРус’е лексико-функциональные связи отражаются отдельно от синтаксической структуры с помощью особых элементов

XML. В визуализаторе же для удобства пользователя имена ЛФ записываются прописными латинскими буквами в овалах, от которых протянуты стрелки к аргументу и значению функции.

За примером обратимся снова к предложению (1). Нетрудно убедиться в том, что в этом предложении имеется несвободное словосочетание *нарушает покой*, которое описывается в терминах лексических функций. Из рис. 1 видно, что это словосочетание оформляется с помощью ЛФ LIQUFUNC0 от аргумента ПОКОЙ1 со значением НАРУШАТЬ. Аналогичным образом, из рис. 3 видно, что в предложении (3) встречается несвободное словосочетание *пользуясь методами*, которое оформляется с помощью ЛФ REAL1 от аргумента МЕТОД со значением ПОЛЬЗОВАТЬСЯ.

В номинативном предложении

(4) … *Который ставит под серьезную угрозу принцип равенства перед законом людей с разной сексуальной ориентацией*

отмечены два лексико-функциональных сочетания: *серьезную угрозу*, где СЕРЬЕЗНЫЙ — значение ЛФ MAGN от аргумента УГРОЗА2, и *ставить под угрозу*, где СТАВИТЬ — значение ЛФ INCEPLABOR1–2 от того же аргумента УГРОЗА2, а предлог ПОД1 — обязательный контекст этой лексико-функциональной зависимости.

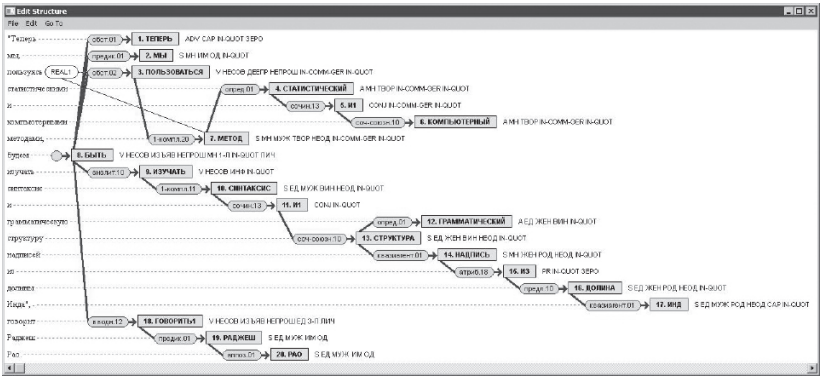


Рис. 4. Разметка предложения (4) несколькими лексическими функциями.

2.2. Обработка эллиптических конструкций

Одной из наиболее трудных задач, возникающих при автоматическом синтаксическом анализе, является обработка эллиптических конструкций.

Предложения с эллипсисом составляют 2–3 процента от общего числа предложений корпуса СинТагРус. Они представлены в корпусе деревьями зависимостей, в которых эллиптированным словам соответствуют особые узлы дерева, имеющие пустой текстовый элемент («фантом»). Специально для корпуса, для лучшего восприятия пользователями, был предложен ряд решений относительно формы представления эллиптических конструкций.

Упомянем два таких решения.

1) Восстановление некоторых типов синтаксического эллипсиса

В случаях, когда опущенные слова физически присутствуют в другой части предложения (например, в сочинительной конструкции с сочинительным сокращением), эти слова восстанавливаются. Например, в предложении

(5) *И приходится пересаживать детям почки, сердце от взрослых и только в подростковом возрасте, когда организм готов принять большой орган*

между словами *и* и *только* вставляется узел «ПЕРЕСАЖИВАТЬ» с соответствующим набором характеристик и пустым текстовым элементом. От таких «фантомных» слов проводятся все необходимые связи (см. рис. 5):

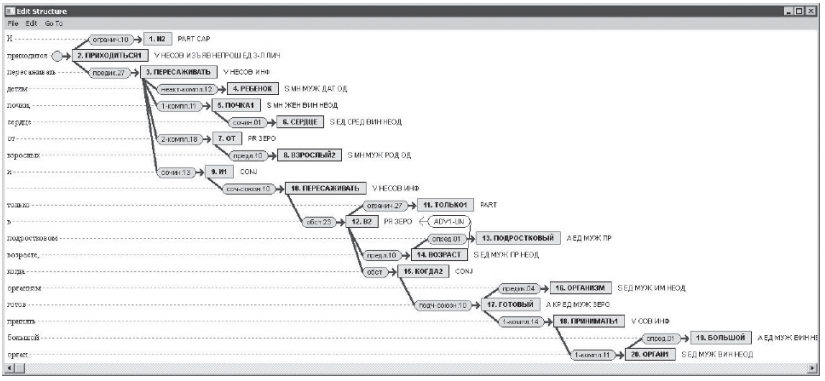


Рис. 5. Синтаксическая структура предложения (5) с восстановленным эллипсисом

При отсутствии восстановленного узла древесная структура предложения выглядела бы неестественно, а соответствующие синтаксические отношения невозможно было бы разумно интерпретировать.

Леммы восстановленных «фантомных» узлов совпадают с теми, которые уже встречались в предложении, а отдельные морфологические характеристики, например, род у глаголов прошедшего времени, могут меняться.

(6) *Отсюда он летит в Сочи, и сразу обратно в Уфу, на чемпионат Башкирии, а 5 апреля стартует чемпионат России в Ханты-Мансийске*

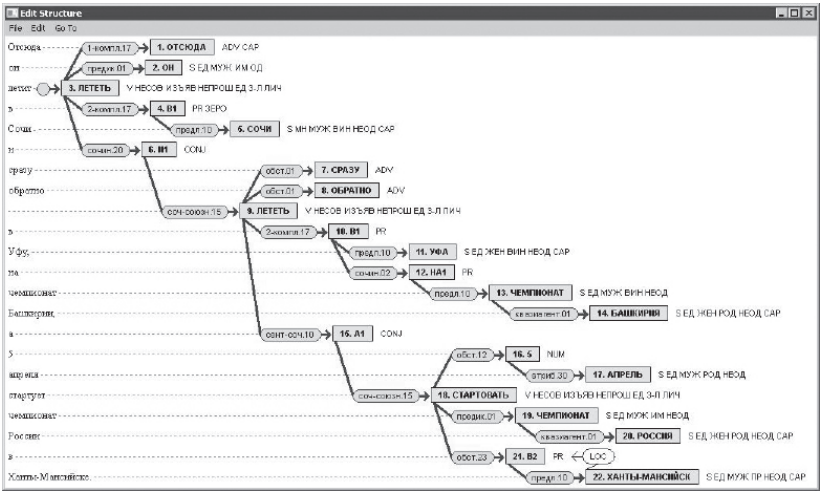


Рис. 6. Синтаксическая структура предложения (6) с восстановленным эллипсисом

В СинТагРус’е встречаются и достаточно курьезные примеры, понятные человеку из контекста, но не восстанавливаемые автоматически. Для того, чтобы получить приличную структуру второй фразы в тексте

(7) *Например, пшеничная мука первого сорта с 12 июня прошлого года до 4 марта нынешнего выросла в цене на 34 процента. Высшего — на 44%, ржанообдирочная мука — на 73%.*

пришлось, опираясь на данные предыдущей фразы из этого текста, дописать 8 «фантомных» узлов, среди них не только глаголы, но и имена существительные и предлоги:

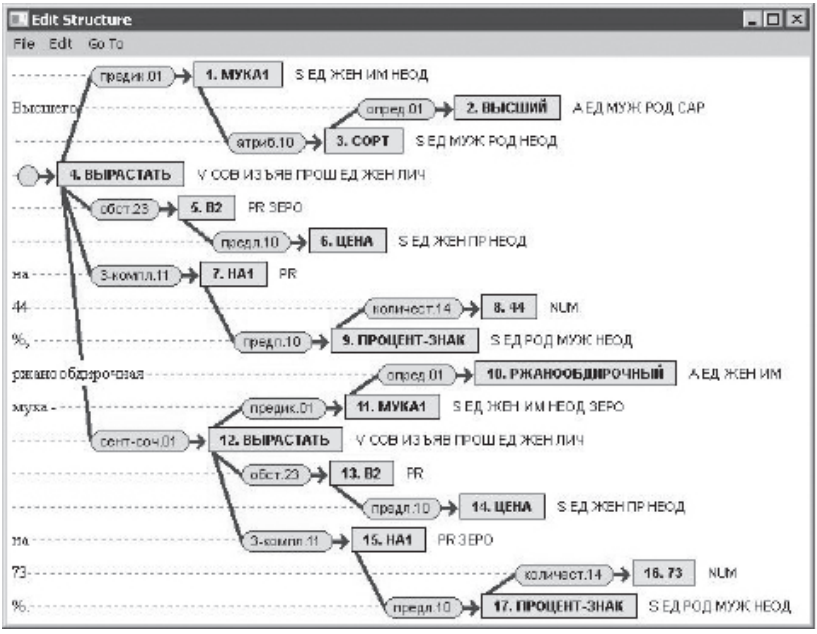


Рис. 7. Множественные восстановленные узлы в высокоэллиптичном предложении

Приведём для ясности восстановленную фразу целиком; здесь фантомные узлы написаны светлым шрифтом, а реально фигурирующие слова выделены жирным шрифтом:

(7′) *Мука* ***Высшего*** *сорта выросла в цене —* ***на 44%****,* ***ржанообдирочная мука*** *выросла в цене —* ***на 73%****.*

2) Использование специальных фиктивных слов для частичной нормализации высокоэллиптичных предложений

Этот прием используется в тех случаях, когда в предложении «опущен» глагол некоторой размытой семантики, так называемый неопределенный глагол, как, например, в предложении

(8) *Ждешь беду отсюда, ан из-за угла тебя мешком*

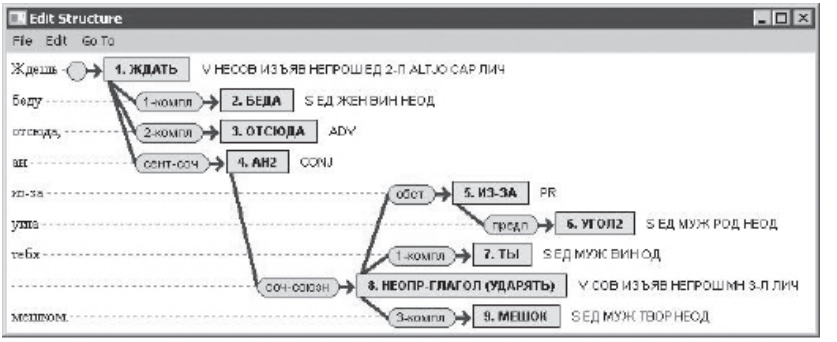


Рис. 8. Синтаксическая структура (8) с «фантомным» глаголом размытой семантики «ударять»

Приведем еще два примера фраз с эллипсисом, фигурирующих в корпусе:

(9) — *Как сами, так делают что хотят, а порядочному человеку — выговор*

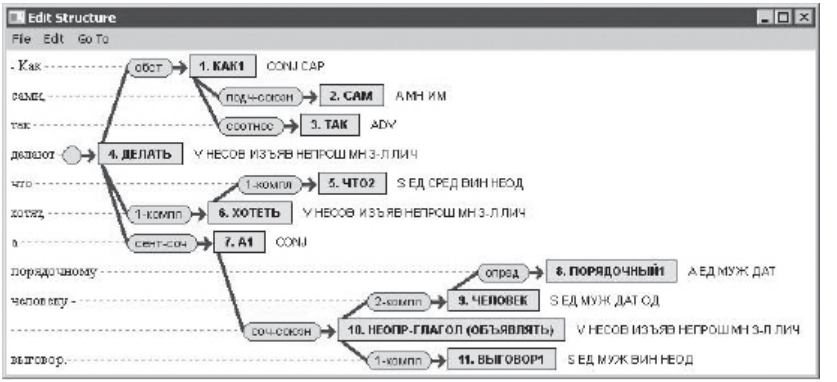


Рис. 9. Синтаксическая структура (9) с «фантомным» глаголом размытой семантики «объявлять»

(10) *Но это, когда разведчик уже лежал на снегу*

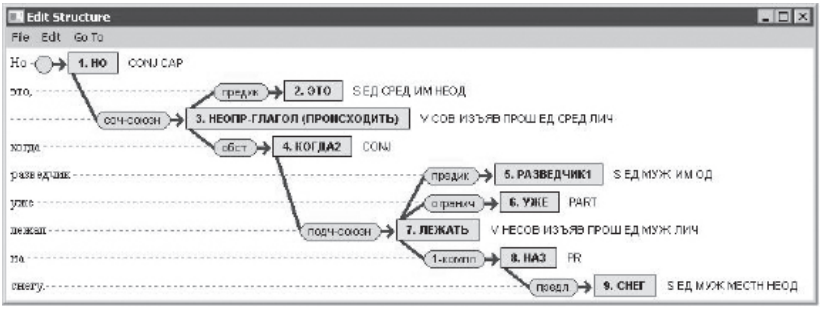


Рис. 10. Синтаксическая структура с «фантомным» глаголом размытой семантики «происходить»

Во всех таких случаях в предложение добавляется глагольная вершина, ей приписываются наиболее подходящие и естественные грамматические характеристики, а в качестве леммы пишется НЕОПР-ГЛАГОЛ (неопределенный глагол) и затем в скобках глагол, который является уместной и естественной гипотезой. Так, в предложении (8) после слова *тебя* добавляется узел с леммой НЕОПР-ГЛАГОЛ (УДАРЯТЬ), в предложении (9) после слова *человеку* — НЕОПР-ГЛАГОЛ (ОБЪЯВЛЯТЬ), а в последнем предложении после *это* — узел с леммой НЕОПР-ГЛАГОЛ (ПРОИСХОДИТЬ).

3. Использование корпуса СинТагРус в лингвистическом процессоре ЭТАП-3.

3.1. Регрессионное тестирование парсера ЭТАП-3

Как было показано в разделе 1, важнейшим этапом построения корпуса является создание полных древесных синтаксических структур каждого предложения с помощью автоматического синтаксического анализатора (парсера) системы ЭТАП-3, которые затем проверяются и при необходимости корректируются лингвистами-экспертами. При этом оказывается, что часть синтаксических структур не требует коррекции и вносится в корпус по результатам проверки безо всяких изменений. По мере развития корпуса и модернизации самого парсера доля таких предложений в корпусе постоянно увеличивается и сейчас составляет около 35% (приблизительно 20 000 предложений). Совокупность таких предложений

удобно использовать в качестве своего рода золотого стандарта – эталона, относительно которого удобно тестировать текущее состояние парсера системы ЭТАП-3 и ресурсов, на которые он опирается (словари и различные совокупности правил).

Вкратце процедура регрессионного тестирования парсера с помощью корпуса СинТагРус (впервые предложенная в работе Богуславский и др., 2008) выглядит следующим образом. Время от времени весь корпус пропускается через парсер системы ЭТАП-3. Результаты работы парсера автоматически анализируются, и выделяется эталонная совокупность предложений, структуры которых полностью совпадают со структурами отредактированного экспертами корпуса. Эту совокупность предложений мы называем текущим регрессионным тестом. Далее, после каких-либо изменений, внесенных в словари или правила, которые разработчики считают чувствительными (это может быть добавление новых слов или новых значений слов, редактирование существующих или создание новых правил), текущий регрессионный тест пропускается через парсер системы ЭТАП-3 в его актуальном состоянии. Автоматическое сравнение результатов работы предшествующего состояния парсера и его актуального состояния на текущем регрессионнном тесте позволяет увидеть все последствия сделанных изменений. Некоторые из этих изменений являются нежелательными и оказываются следствием ошибочных манипуляций с ресурсами парсера, которые приходится пересматривать; заметим, что иногда такие изменения носят просто-таки разрушительный характер, и своевременное их устранение жизненно необходимо. В то же время другая часть изменений, напротив, является желательной: в результате учета этих изменений совершенствуется как корпус, так и парсер системы, которые оказываются благотворно влияющими друг на друга. Подчеркнем, что объем текущего регрессионного теста вполне достаточен, чтобы он уверенно “откликался” даже на небольшие изменения ресурсов корпуса, т.е. словарей и правил, а также на изменения в алгоритмическом и программном обеспечении системы. Наконец, регрессионное тестирование позволяет оптимизировать выбор опций и настроек парсера, которые оказываются наиболее удобными для создания размеченного корпуса.

3.2. Создание гибридного парсера на основе парсера ЭТАП-3 и СинТагРус'а.

В последние несколько лет активно развивается подход, связанный с использованием статистических данных корпуса СинТагРус непосредственно в парсере лингвистического процессора ЭТАП-3. На основании данных корпуса вырабатываются оценки вероятностей альтернативных вариантов для элементов морфосинтаксической структуры, а именно оценки вероятностей различных лексико-морфологических интерпретаций одного и того же слова предложения и оценки вероятностей различных синтаксических связей, входящих в один и тот же узел дерева зависимостей. На основе этих данных парсер ищет дерево зависимостей, удовлетворяющее всем требованиям грамматических правил системы и имеющее максимальную результирующую оценку вероятности. Парсер ЭТАП-3, поддерживаемый статистикой корпуса, разумно считать своего рода гибридной системой, использующей комбинацию правилового подхода, на котором базируется исходный парсер ЭТАП-3, и статистическими данными.

Как показали масштабные эксперименты, гибридный подход ощутимо повышает качество синтаксических структур, получаемых парсером (см. Boguslavsky et al. 2011, 2014). Однако для дальнейшего совершенствования гибридного парсера требуется дополнительная исследовательская и экспериментальная работа.

4. Исследовательские задачи, решаемые с участием корпуса СинТагРус

По специальному лицензионному соглашению около 50 исследовательских групп и институтов получили в свободное пользование материалы корпуса СинТагРус.

Следует прежде всего отметить, что СинТагРус был использован для создания успешного работающего статистического Malt-парсера для русского языка (Nivre et al. 2008, Bohnet et al. 2013).

Сотрудники Исследовательского центра Samsung использовали СинТагРус для обучения и тестирования алгоритма по разрешению омонимии при помощи статистических методов. Статистические методы с успехом применялись, в частности, к извлечению темпоральных выражений и анализу именных групп, при этом подкор-

пус для машинного обучения был составлен на базе СинТагРус (Kudinov et al. 2014, Muzychka et al. 2014).

В компании Google СинТагРус и другие аналогичные корпусы текстов, размеченные синтаксическими зависимостями (для разных языков) использовали в целях обучения и тестирования многоязычного статистического парсера. Сообщается о том, что благодаря размеченным корпусам относительную ошибку анализа удавалось в ряде задач уменьшить на 13%, а в случае задачи распознавания именованных сущностей – даже на 26% (Täckström et al. 2012).

Участники норвежской исследовательской группы CLEAR, в задачи которой входит изучение русского языка, Х. Экхофф и А. Бердичевский (Норвежский Арктический университет, г. Тромсё) в рамках проекта Birds&Beasts занимаются разметкой и исследованиями на материале синтаксического корпуса TOROT (старославянский, древнерусский, старорусский): https://nestor.uit.no/ и сравнением данных, полученных по этому корпусу, с современным состоянием языка, представленном в СинТагРус'е. В частности, составляются синтаксические профили русских глаголов.

Количественное исследование синтаксического поведения некоторых классов русских слов было проведено на материале корпуса СинТагРус китайскими исследователями Ван Юном, Лю Хайтао (2013).

СинТагРус используется для целей классификации различных лингвистических объектов. Например, в работе Баранова 2013 корпус был использован для машинного обучения системы, определяющей уровень сложности русских предложений для целей обучения русскому языку как иностранному.

Наконец, синтаксически размеченный корпус русского языка может быть использован для составления лингвистических задач, направленных на поиск специальных конструкций (Шеманаева 2009).

5. Новейшее развитие корпуса СинТагРус и ближайшие перспективы

В последнее время разработчиками корпуса был проведен масштабный эксперимент по переводу фрагмента основной части Национального корпуса русского языка со снятой омонимией в формат СинТагРус'а. Цель эксперимента была двоякой: с одной стороны, планировалось проверить, насколько трудоёмкой и автоматизируе-

мой окажется работа по переводу морфологической разметки НКРЯ в морфологический формат СинТагРус'а, учитывая заметные расхождения морфологических трактовок обоих корпусов, а, с другой стороны, выяснить, насколько и как именно дополнительная информация о разрешенной неоднозначности лексических и морфологических элементов текста, присутствующая в НКРЯ, повлияет на качество синтаксического анализа, осуществляемого парсером ЭТАП-3. Содержание этой работы было подробно освещено в недавней статье Дяченко, Подлесская, Сизов, 2014; здесь излагаются ее основные результаты.

Следует отметить, что работа по переводу материалов размеченного корпуса из одного формата в другой предпринимались неоднократно, в том числе и с участием СинТагРус'а; ср., в частности, эксперименты по переводу СинТагРус'а в формализм HPSG (Avgustinova T., Zhang Y, 2010; Avgustinova 2011), а также эксперименты по переводу СинТагРус'а в формат Пражского корпуса зависимостей (PDT) для чешского языка (Mareček, Kljueva 2009), однако впервые материалы другого ресурса переводились в формат СинТагРус'а.

Мы исходили из предположения, что использование дополнительной информации такого рода в процессе автоматического построения синтаксической структуры предложения позволяет повысить качество получаемой структуры, тем более что у нас был успешный опыт аналогичного эксперимента с английским языком (Диконов, Дяченко 2010). В качестве источника информации использовался фрагмент НКРЯ, содержащий сведения о морфологическом разборе входящих в него слов со снятой вручную омонимией (будем для краткости называть эту часть НКРЯ Подкорпусом). Эксперимент проводился на фрагменте Подкорпуса, состоящем из текстов, написанных после 1950 г. (примерно 4,48 млн. словоупотреблений).

Между морфологическими разметками НКРЯ и СинТагРус'а существуют различия, которые не сводятся к простой замене имен меток, в то же время их можно преодолеть. В состав разметки подкорпуса входят морфологические характеристики и леммы. Мы рассчитывали, что учет этой информации поможет парсеру ЭТАП-3 разрешить морфологическую омонимию (в идеале — полностью), и тем самым существенно уменьшить число рассматриваемых аль-

тернативных вариантов разбора. Из оценки исходных данных следует, что около 83% неоднозначных разборов различаются морфологическими характеристиками или леммами и тем самым могут быть разрешены с использованием информации, взятой из Подкорпуса.

Оценить эффект данных из Подкорпуса на объем неоднозначностей, оставшихся после процедуры их разрешения, можно из следующей таблицы.

Словоупотребления, для которых неоднозначность снята полностью	1805555
Словоупотребления, для которых неоднозначность снята частично	21041
Словоупотребления, для которых ни один разбор парсера ЭТАП-3 не совпал с данными из Подкорпуса	171621
Всего словоупотреблений с неоднозначными результатами анализа	2507953

Табл. 1. Результаты разрешения неоднозначности.

Как видим, использование данных Подкорпуса позволяет уменьшить число неоднозначных морфологических разборов на 71%. Уменьшение числа неоднозначных разборов ожидаемо приводит к сокращению времени работы системы ЭТАП-3 с использованием данных Подкорпуса за счет сокращения числа рассматриваемых альтернативных гипотез, а именно: работа на фрагменте подкорпуса без учета его данных составила 37 ч. 15 м., а с учетом данных — 13 ч. 48 м.

Еще более показательным является сравнение качества синтаксического анализа некоторого фрагмента текста, использующего данные Подкорпуса, со стандартным СА.

Режим эксперимента	С данными Подкорпуса	Без данных Подкорпуса
Построено структур	255	
Сравнивалось слов	2740	2743
С верными частями речи и морфологическими характеристиками:	2632 (96%)	2528 (92%)
Верных лемм:	2570 (94%)	2519 (92%)

Режим эксперимента	С данными Подкорпуса	Без данных Подкорпуса
Верных структур (без учёта имён связей):	151 (60%)	148 (58%)
Верных структур (с учётом имён связей):	129 (50%)	123 (48%)

Табл. 2. Параметры качества синтаксического анализа

Приведенные в таблице 2 результаты показывают, что использование Подкорпуса при синтаксическом анализе приводит к существенному сокращению в разметке числа слов с неправильными морфологическими характеристиками, а также положительно влияет на качество построенных синтаксических структур. Правда, это изменение не столь значительно, как мы предполагали на ранних стадиях эксперимента. Возможно, на результат оказывают влияние еще не обнаруженные расхождения в разметке, которые, встречаясь достаточно редко, тем не менее ухудшают результат работы экспериментального синтаксического анализатора.

Достаточно высокое качество морфологического компонента и созданного на его основе корпуса позволит использовать результаты для обучения статистических программ, использующих морфологическую разметку — например, создать программу статистического морфологического анализа текстов в формате СинТагРус'а. В то же время недостаточно высокое качество построенных синтаксических структур, близкое к качеству результата работы ЭТАП-3 в обычном режиме, не позволяет применить полученный корпус в имеющемся виде для обучения статистического синтаксического анализатора, а также не дает существенного уменьшения трудозатрат при ручной разметке экспертами-лингвистами. Поскольку в ходе экспериментов были выявлены определенные резервы по усовершенствованию алгоритмов, то дальнейшие исследования, как мы надеемся, изменят ситуацию к лучшему.

Заметим, наконец, что в ближайшие планы разработчиков корпуса входит внесение разметки о различных микросинтаксических конструкциях русского языка и синтаксических фраземах. В результате СинТагРус окажется полезным и для исследований в области малого синтаксиса и фразеологии.

Литература

Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л., Иомдин Л. Л., Санников А. В., Санников В. З., Сизов В.Г., Цинман Л. Л. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы. // Национальный корпус русского языка: 2003-2005. М.: Индрик, 2005.

Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Санников В. З. Теоретические проблемы русского синтаксиса: Взаимодействие грамматики и словаря. Отв. ред. Ю. Д. Апресян. М.: Языки славянских культур, 2010.

Баранова Ю. Н. Автоматическая классификация типов предложений (на базе СинТагРус). Данные междисциплинарного научного семинара (МНС) «Методы интеллектуального анализа естественного языка», Нижний Новгород, 2013.

Богуславский И. М., Иомдин Л. Л., Валеев Д. Р., Сизов В. Г. Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов. // Международная научная конференция «Корпусная лингвистика», СПб., 2008а.

Богуславский И. М., Иомдин Л. Л., Митюшин Л. Г., Сизов В. Г. Длина синтаксических связей в русском аннотированном корпусе. // Международная научная конференция «Корпусная лингвистика», СПб., 2008б.

Диконов В. Г., Дяченко П. В. Эксперимент по построению синтаксической структуры английских предложений с использованием заранее известных фрагментарных данных // Информационные технологии и системы (ИТиС'10). Сборник трудов 33-ой Конференции молодых ученых и специалистов ИППИ РАН. Геленджик, 18-26 сентября 2010 г. М.: ИППИ, 2010. С. 310-319. ISBN 978-5-901158-12-8.

Дяченко П. В., Подлесская О. Ю., Сизов В. Г. НКРЯ: основной корпус и СинТагРус, синтаксический анализ текстов со снятой морфологической омонимией. // Информационные технологии и системы (ИТиС'2014). Сборник трудов 38-ой Конференции молодых ученых и специалистов ИППИ РАН. Нижний Новгород, Россия, 2014.

Иомдин Л. Л., Иомдин Б. Л. Отрицание и валентности в русском языке (по корпусным данным). // Международная научная конференция «Корпусная лингвистика». СПб.: С.-Петербургский гос. университет, 2013. С. 281–291.

Казенников А. О. Сравнительный анализ статистических алгоритмов синтаксического анализа на основе деревьев зависимостей. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2010) Выпуск 9 (16). С. 157–162.

Казенников А. О., Куракин Д. В. Сравнительный анализ алгоритмов синтаксического разбора в рамках задачи анализа новостных потоков. // Труды XVIII Всероссийской научно-методической конференции Телематика '2011, СПб, 2011. С. 235–236.

Шеманаева О. Ю. Глубоко аннотированный корпус русских текстов как обучающий электронный ресурс // Информационные технологии и системы (ИТиС'09). Сборник трудов 32-ой Конференции молодых ученых и специалистов ИППИ РАН. Бекасово, 15–18 декабря 2009 г. М., 2009. С. 205–209.

Шеманаева О. Ю., Фролова Т. И. Лексико-функциональная разметка текстов в СинТагРус // Информационные технологии и системы (ИТиС'10). Сборник трудов 33-й конференции молодых ученых и специалистов ИППИ РАН. Геленджик, 18-26 сентября 2010 г. М.: ИППИ, 2010. С. 320–324.

Ван Юн, Лю Хайтао. Квантитативное исследование имени существительного в русском языке по его синтаксическим признакам. // Вестник Московского университета. Сер. 9. Филология, 2013 (5). С. 35–51.

Avgustinova T. Parallel Construction of Slavic Grammatical Resources. Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2011». М.: Изд-во РГГУ, 2011. Вып. 10(17). С. 41–50.

Avgustinova T., Zhang Y. Conversion of a Russian Dependency Treebank into HPSG Derivations. Proceedings of the 9 th International Workshop on Treebanks and Linguistic Theories (TLT'9), 2010.

Apresjan Ju., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // MTT 2003, First International Conference on Meaning-Text Theory. Paris, Ecole Normale Superieure, Paris, June 16-18 2003, pp. 279–288.

Boguslavsky I. SynTagRus – a Deeply Annotated Corpus of Russian // Les émotions dans le discours. Emotions in Discourse. Eds. Peter Blumenthal, Iva Novakova, Dirk Siepmann. Peter Lang Edition, 2014. P. 367–381.

Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N. Dependency Treebank for Russian: Concept, Tools, Types of Information. // Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000). San Francisco: Kaufmann, pp. 987–991.

Boguslavsky I., Iomdin L., Timoshenko S., Frolova T. Development of the Russian Tagged Corpus with Lexical Functional Annotation // Metalanguage and Encoding Scheme Design for Digital Lexicography. MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia. April 15-16, 2009. Bratislava, 2009, p. 83–90.

Boguslavsky I., Iomdin L., Tsinman L., Petrochenkov V. Rule-Based Dependency Parser Refined by Empirical and Corpus Statistics // Proceedings of the International Conference on Dependency Linguistics (Depling'2011). Barcelona, September 5-7, 2011, p. 318–327.

Boguslavsky I., Iomdin L., Petrochenkov V., Sizov V., Tsinman L. A Case of Hybrid Parsing: Rules Refined by Empirical and Corpus Statistics // In: Computational Dependency Theory. Vol. 258 of Frontiers in Artificial Intelligence and Applications. Eds. Kim Gerdes, Eva Hajicova, Leo Wanner, 2014. pp. 226-240. IOS Press, ISBN 978-1-61499-351-3 DOI 10.3233/978-1-61499-352-0-226.

Bohnet B., Nivre J., Farkas R., Ginter F., Hajic J. Joint Morphological and Syntactic Analysis for Richly Inflected Languages // Transactions of the Association for Computational Linguistics, 2013. Vol. 1. pp. 415–428.

Iomdin L. Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact. // Jazykovedné štúdie XXXI. Bratislava: Jazykovedný ústav Ľ.Štúra Slovenskej akadémie vied, 2014. pp. 136–146.

Kudinov M., Romanenko A., Piontkovskaya I. Conditional Random Field in Segmentation and Noun Phrase Inclination on Tasks for Russian // Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2014». М.: 2014. С. 315–324.

Mareček D., Kljueva N. Converting Russian Treebank SynTagRus into Praguian PDT Style // Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages 2009, Bulgaria, pp. 26–31.

Muzychka S., Romanenko A., Piontkovskaya I. Conditional Random Field for Morphological Disambiguation in Russian // Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2014». М.: РГГУ, 2014. С. 474–483.

Nivre J., Boguslavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian. // In Proceedings of the 22nd International Conference on Computational Linguistics (COLING08), Scott D., Uszkoreit H. (eds.), 2008, pp. 641–648.

Podlesskaja O., Frolova T. Tagging lexical functions in Russian texts of SynTagRus // Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог'2011». М.: Изд-во РГГУ, 2011. Вып. 10(17). С. 207–218.

Täckström O., McDonald R. and Uszkoreit H. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure // North American Association for Computational Linguistics (NAACL), 2012.

[1]P.V. Dyachenko, [2]L.L. Iomdin, [3]A.V. Lazursky, [4]L.G. Mityushin, [5]O.Yu. Podlesskaya, [6]V.G. Sizov, [7]T.I. Frolova, [8]L.L. Tsinman

[1,2,3,4,5,6,7,8]A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences

[1,2,3,4,5,6,7,8](Russia, Moscow)

[1]pavelvd@iitp.ru, [2]iomdin@iitp.ru, [3]lazur@iitp.ru, [4]mit@iitp.ru, [5]olga@iitp.ru, [6]sizov@iitp.ru, [7]tfrolova@iitp.ru, [8]cinman@iitp.ru

A DEEPLY ANNOTATED CORPUS OF RUSSIAN TEXTS (SYNTAGRUS): CONTEMPORARY STATE OF AFFAIRS

The paper discusses the main features, principles of creation and parameters of the syntactically tagged corpus of Russian texts, SynTagRus. In addition to syntactic annotation of all sentences with dependency trees, the corpus contains information on arguments and values of lexical functions for words occurring in the sentence, as well as the data on lexical meanings of words. The article considers the subcorpus of sentences containing diverse types of ellipsis and discusses possible uses of the corpus for research tasks and applications.

Key words: SynTagRus, syntactically tagged corpus of Russian texts, dependency grammar, lexical functions, ambiguity resolution, ellipsis

References

Apresjan Ju. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. Z., Sizov V. G., Cinman L. L. [Syntactically and semantically tagged corpus of Russian: state of the art and prospects]. *Natsional'nyi korpus russkogo yazyka: 2003-2005* [Russian National Corpus: 2003–2005. Results and Prospects]. Moscow, Indrik Publ., 2005. (In Russ.)

Apresjan Ju., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT. *MTT 2003, First International Conference on Meaning-Text Theory*. Paris, Ecole Normale Superieure, Paris, June 16-18 2003, pp. 279–288.

Apresjan Ju. D., Boguslavskij I. M., Iomdin L. L., Sannikov V. Z. *Teoreticheskie problemy russkogo sintaksisa: Vzaimodeistvie grammatiki i slovarya* [Theoretical issues of Russian syntax: Interaction of the grammar and the dictionary]. Ju.D. Apresjan (Ed.). Moscow, Yazyki slavyanskikh kul'tur Publ., 2010.

Avgustinova T. Parallel Construction of Slavic Grammatical Resources. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2011)*. Issue 10 (17), Moscow: RSUH Publishers, pp. 41-50.

Avgustinova T., Zhang Y. Conversion of a Russian Dependency Treebank into HPSG Derivations. *Proceedings of the 9th International Workshop on Treebanks and Linguistic Theories (TLT'9)*, 2010.

Baranova Ju. N. [An automatic classification of sentence types (based on SynTagRus)]. *Dannye mezhdistsiplinarnogo nauchnogo seminara (MNS) "Metody intellektual'nogo analiza estestvennogo yazyka"*, Nizhnij Novgorod, 2013. (In Russ.)

Boguslavsky I. SynTagRus – a Deeply Annotated Corpus of Russian. *Les émotions dans le discours. Emotions in Discourse. Peter Blumenthal,* Iva Novakova, Dirk Siepmann (Eds.). Peter Lang Edition, 2014, pp. 367–381.

Boguslavsky I., Grigorieva S., Grigoriev N., Kreidlin L., Frid N. Dependency Treebank for Russian: Concept, Tools, Types of Information. *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*. San Francisco, Kaufmann, pp. 987–991.

Boguslavsky I. M., Iomdin L. L., Mitjushin L. G., Sizov V. G. [The

length of syntactic links in the Russian tagged corpus]. *Mezhdunarodnaya nauchnaya konferentsiya "Korpusnaya lingvistika"*, St.-Petersburg, 2008b. (In Russ.)

Boguslavsky I., Iomdin L., Timoshenko S., Frolova T. Development of the Russian Tagged Corpus with Lexical Functional Annotation. *Metalanguage and Encoding Scheme Design for Digital Lexicography.* MONDILEX Third Open Workshop. Proceedings. Bratislava, Slovakia. April 15-16, 2009. Bratislava, 2009, pp. 83–90.

Boguslavsky I., Iomdin L., Petrochenkov V., Sizov V., Tsinman L. A Case of Hybrid Parsing: Rules Refined by Empirical and Corpus Statistics. *Computational Dependency Theory*. Vol. 258 of Frontiers in Artificial Intelligence and Applications. Eds. Kim Gerdes, Eva Hajicova, Leo Wanner, 2014. pp. 226-240. IOS Press, ISBN 978-1-61499-351-3 DOI 10.3233/978-1-61499-352-0-226.

Boguslavsky I., Iomdin L., Tsinman L., Petrochenkov V. Rule-Based Dependency Parser Refined by Empirical and Corpus Statistics. *Proceedings of the International Conference on Dependency Linguistics* (Depling'2011). Barcelona, September 5–7, 2011, pp. 318–327.

Boguslavsky I. M., Iomdin L. L., Valeev D. R., Sizov V. G. [A syntactic analyzer of the ETAP system and its evaluation with the help of a deeply annotated corpus of Russian texts]. *Mezhdunarodnaya nauchnaya konferentsiya "Korpusnaya lingvistika"*, St.-Petersburg, 2008a. (In Russ.)

Bohnet B., Nivre J., Farkas R., Ginter F., Hajic J. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *Transactions of the Association for Computational Linguistics*, 2013. Vol. 1, pp. 415–428.

Dikonov V. G., Djachenko P. V. [An experiment of building syntactic structures of English sentences using previosly known fragmentary data]. *Informatsionnye tekhnologii i sistemy (ITiS'10). Sbornik trudov 33-oi Konferentsii molodykh uchenykh i spetsialistov IPPI RAN* [Information technologies and systems (ITAS'10). Proceedings of the 33-th conference of young scientists and experts of IITP RAS]. Gelendzhik, September 18-26, 2010. Moscow, IITP Publ., 2010, pp. 310–319. (In Russ.)

Djachenko P. V., Podlesskaja O. Ju., Sizov V. G. [National corpus of Russian: the basic corpus and SynTagRus, syntactic analysis of texts with resolved morphological ambiguity]. *Informatsionnye tehnologii i sistemy (ITiS'2014). Sbornik trudov 38-oi Konferentsii molodykh*

uchenykh i spetsialistov IPPI RAN [Information technologies and systems (ITAS'14). Proceedings of the 33-th conference of young scientists and experts of IITP RAS]. Nizhnii Novgorod, 2014. (In Russ.)

Iomdin L. Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact. *Jazykovedné štúdie XXXI*. Bratislava, Jazykovedný ústav Ľ.Štúra Slovenskej akadémie vied, 2014, pp. 136–146.

Iomdin L. L., Iomdin B. L. [The negation and valencies in Russian (according to corpus data)]. *Mezhdunarodnaya nauchnaya konferentsiya "Korpusnaya lingvistika"*, St. Petersburg, 2013, pp. 281–291. (In Russ.)

Kazennikov A. O. [A comparative analysis of machine learning dependency tree-based parsing algorithms]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Po materialam ezhegodnoi Mezhdunarodnoi konferentsii "Dialog"* [Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2010)]. Issue 9 (16), pp. 157–162. (In Russ.)

Kazennikov A. O., Kurakin D. V. [A comparative analysis of syntactic parsing algorithms within the framework of news flow analysis]. *Trudy XVIII Vserossiiskoi nauchno-metodicheskoi konferentsii "Telematika'2011"*, St. Petersburg 2011, pp. 235–236. (In Russ.)

Kudinov M., Romanenko A., Piontkovskaya I. Conditional Random Field in Segmentation and Noun Phrase Inclination on Tasks for Russian. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2014)*. Issue 13 (20), Moscow, 2014, pp. 315–324.

Mareček D., Kljueva N. Converting Russian Treebank SynTagRus into Praguian PDT Style. *Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages 2009*, Bulgaria, pp. 26–31.

Muzychka S., Romanenko A., Piontkovskaya I. Conditional Random Field for Morphological Disambiguation in Russian. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2014)*. Issue 13 (20), Moscow, 2014, pp. 474–483.

Nivre J., Boguslavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING08)*, Scott D., Uszkoreit H. (Eds.), 2008, pp. 641–648.

Podlesskaja O., Frolova T. Tagging lexical functions in Russian texts of SynTagRus. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2011)*. Issue 10 (17), pp. 207–218.

Shemanaeva O. Yu. [A deeply annotated corpus of Russian texts as a teaching electronic resource]. *Informatsionnye tekhnologii i sistemy (ITiS'09)*. Sbornik trudov 32-oi Konferentsii molodykh uchenykh i spetsialistov IPPI RAN [Information technologies and systems (ITAS'14). Proceedings of the 32-th conference of young scientists and experts of IITP RAS]. Bekasovo, December 15–18, 2009. Moscow, 2009, pp. 205–209. (In Russ.)

Shemanaeva O. Yu., Frolova T. I. [Tagging with lexical functions in SynTagRus]. *Informacionnye tehnologii i sistemy (ITiS'10)*. Sbornik trudov 33-oj Konferencii molodyh uchenyh i specialistov IPPI RAN. Gelendzhik, 18-26 sentjabrja 2010 g. Moscow, IPPI, 2010, pp. 320–324. (In Russ.)

Täckström O., McDonald R. and Uszkoreit H. Cross-lingual Word Clusters for Direct Transfer of Linguistic Structure. *North American Association for Computational Linguistics (NAACL)*, 2012.

Van Yun, Liu Haitao. [A quantitative study of the Russian noun according to their syntactic features]. *Vestnik Moskovskogo universiteta*. Ser. 9, Filologiya, 2013 (5), pp. 35–51. (In Russ.)

[1]А. И. Зобнин, [2]Г. В. Носырев
[1, 2]«Яндекс»
[1,2](Россия, Москва)
[1]alzobnin@yandex-team.ru, [2]nosyrev@yandex-team.ru

МОРФОЛОГИЧЕСКИЙ АНАЛИЗАТОР MYSTEM 3.0

В значительной части Национального корпуса русского языка морфологическая разметка накладывается на текст автоматически. Для этого используется разработанная в Яндексе программа Mystem с некоторой постобработкой результатов[1] (например, все несклоняемые существительные приобретают помету '0', двувидовые глаголы разделяются на отдельные парадигмы по видам и т. д.). Недавно вышла новая — третья — версия этой программы (см. https://tech.yandex.ru/mystem/). В этой статье мы представим обзор ее возможностей.

Ключевые слова: Mystem, морфология, Национальный корпус русского языка, Яндекс.

Задачи морфологического модуля в процессах автоматической обработки текстов — морфологический анализ и синтез. Анализом называют определение начальной формы (леммы) и грамматических признаков (части речи, рода, числа, падежа, и т. д.) слова. Синтез — генерация всех остальных форм в процессе словоизменения (но не словообразования).

В информационном поиске морфология имеет массу применений. Так, при поиске учитываются разные формы слова запроса; текст леммы может использоваться в качестве «ключа» в поисковых индексах; запросы при анализе могут обобщаться (нормализиро-

[1] Скрипты постобработки результатов морфологического анализа для разметки документов в НКРЯ доступны на https://github.com/ishalyminov/ruscorpora_tagging (автор — И. В. Шалыминов).

ваться) с точностью до лемм, и т. д. Известно, что именно морфология в поиске Яндекса явилась важным преимуществом над конкурентами в 2000-х годах.

В основе морфологического движка, как правило, лежит словарь. Эталоном для русского языка считается Грамматический словарь А. А. Зализняка [Зализняк, 1977]. Если задача анализа и синтеза ставится только для словарных слов, то она тривиальна: достаточно сложить слова в подходящую структуру данных (хеш-таблица, префиксное дерево) и приписать каждому из них его возможные заранее известные разборы. Однако любой словарь ограничен: в языке появляются новые слова, для которых тоже желательно получать гипотетические морфологические разборы. В флективных и агглютинативных языках это можно сделать, отыскав словарные слова с «похожими» аффиксами и окончаниями и проанализировав неизвестное слово по найденным образцам.

Программа Mystem, первая версия которой была написана 15 лет назад Ильей Сегаловичем и его коллегами, умеет строить такие гипотетические разборы для несловарных слов:

$ mystem -c -g

И хрюкотали зелюки как мумзики в мове

И{и} хрюкотали{хрюкотать?} зелюки{зелюк?|зелюка?} как{как} мумзики{мумзик?} в{в} мове{мов?|мово?}

Подробно технология изложена в статьях [Сегалович, Маслов, 1998, 7]. Опишем здесь кратко процедуру анализа. У всех словарных слов выделяются «основа» (неизменяемая общая часть) и «окончание» (изменяемая часть внутри парадигмы). Создается два трая[1]: для перевернутых окончаний и для перевернутых основ. Мы анализируем слово с конца, пытаясь найти его фрагмент сначала в трае окончаний, а оставшуюся часть — в трае основ. Каждую найденную пару окончания и основы проверяем на соответствие друг другу. Если основа и окончание друг другу соответствуют, то это подходящий разбор. Гипотезы для несловарных слов строятся так же, но перевернутая основа слова в этом случае находится в трае основ не полностью, а частично. Для всякой не полностью найденной

[1] Трай, префиксное дерево (англ. trie) — структура данных, хранящая текстовые ключи (и ассоциированные с ними значения) в виде дерева. Ключи группируются по общим префиксам.

основы требуется обойти все подветки в трае основ, чтобы найти все соответствующие ей словарные образцы.

Работа морфологического анализатора в этом случае соответствует недетерминированному конечному автомату. Недетерминированность заключается в том, что окончания и основы могут быть прочитаны в слове разными способами, и все их нужно перебрать. Разделение данных на основы и окончания позволяло хранить словарь очень компактно (несколько мегабайт), что в эпоху первой версии Mystem'а было актуальным.

Известно, что всякий недетерминированный конечный автомат может быть превращен в детерминированный (но с большим числом состояний). Объединяя два трая с окончаниями и основами вместе, мы увеличиваем объем данных в памяти, но ускоряем анализ. Теперь мы пытаемся найти в этом трае максимальное вхождение перевернутого слова и сразу получаем все его возможные разборы. Таким образом, несмотря на то, что несловарных слов у нас может существовать сколь угодно много, всех возможных вариантов разбора лишь конечное число, и все они указаны в этом едином трае. В этом заключается одно из отличий новой версии анализатора Mystem от предыдущих версий.

Морфологических разборов у одного слова может быть несколько. Это явление называется омонимией. Различают межлеммную омонимию с разными текстами лемм в разборах слова (*стали — сталь, становиться*), омонимию между различными частями речи с одинаковыми текстами лемм (*печь* — существительное и глагол), и омонимию между формами одной и той же леммы (*стол* — именительный и винительный падежи единственного числа). Строго говоря, имеется омонимия и между различными разборами с одинаковыми текстами леммы и частями речи, но с разными парадигмами: *кошка* — два существительных, одушевленное (с формой винительного падежа множественного числа *кошек*) и неодушевленное (с формой *кошки*); *лист* — два существительных с формами *листов* и *листьев*; *высыпаться* — два глагола с формами *высплюсь* (совершенного вида) и *высыплюсь* (несовершенного вида) и т. д.

В новой версии Mystem'а появилась возможность статистически снимать межлеммную омонимию. Задачу снятия омонимии можно решать как с учетом контекста слова, так и без него. В последнем

случае требуется просто упорядочить возможные разборы слова по вероятности их употребления.

Для ранжирования разборов без учета контекста в Mystem'е применяется обучение по подкорпусу со снятой омонимией Национального корпуса русского языка (объем — примерно 6 млн словоупотреблений). Просто упорядочить список разборов для каждого узла в трае в соответствии с их встречаемостью в корпусе — плохой вариант, так как корпус является достаточно разреженным, и не каждое слово в нем представлено. Поэтому мы вычисляем по корпусу следующие величины: частоту каждой морфологической схемы (парадигмы), частоту каждой основы слова внутри каждой схемы, и частоту каждого окончания внутри каждой схемы. Каждому разбору конкретного слова соответствует своя схема, точно определяющая границы основы и окончания. Мы предполагаем, что события «встретилась данная основа *stem*» и «встретилось данное окончание *flex*» слова *word* при фиксированной схеме разбора *scheme* независимы. Поэтому условную вероятность разбора слова *word* по схеме *scheme* можно определить по формуле Байеса:

$$P(scheme \mid word) = P(word \mid scheme)P(scheme) / P(word) = \\ = P(stem \mid scheme)P(end \mid scheme)P(scheme)) / P(word).$$

Этот подход представляет собой наивный байесовский классификатор. Его преимущество в том, что, во-первых, данные о частотах факторизованы по основам и окончаниям (а потому их можно эффективно упаковать), а во-вторых, в том, что разреженность корпуса и низкие частоты отдельных форм слов уже не представляют проблемы: эти данные будут сглажены за счет других форм с такой же основой и других слов с таким же окончанием внутри данной схемы. Кроме того, полученные частоты дополнительно сглаживаются (например, простейшим методом Лапласа [Jurafsky, Martin, 2009]).

Точность этой модели на русском языке достигает 95,9% (по тексту леммы), в то время как точность простого выбора леммы с самой частотной схемой (baseline) равна 90%. Еще раз обратим внимание, что здесь не используется контекст слова.

Для языков, для которых не доступны морфологически размеченные корпуса, можно пытаться обучать модель на частотном словаре. Здесь можно добиться ранжирования регулярных омонимов за

счет распределения частот основ и окончаний неомонимичных слов из тех же парадигм.

Вероятности бесконтекстного ранжирования можно получить с помощью опции --weight:

$ mystem -c -g --weight
еду
еду{ехать:0.76|еда:0.24}
дне
дне{день:0.75|дно:0.25}
айпад
айпад{айпад?:0.88|айпада?:0.12}

Кратко остановимся на задаче снятия омонимии с учетом контекста. Частный случай этой задачи — расстановку частей речи — обычно решают с помощью скрытых марковских моделей (HMM) или марковских моделей максимальной энтропии (MEMM), обученных на размеченном корпусе (см., например, [Jurafsky, Martin, 2009]). Среди более сложных моделей стоит отметить метод условных случайных полей (CRF) (попытка применить его для русского языка описана в [Антонова, Соловьев,2013]).

В работе [Зеленков, Сегалович, Титов, 2005] Ю. Зеленков, И. Сегалович и В. Титов предложили компактную и очень быструю модель снятия межлеммной омонимии на основе так называемых нормализующих подстановок. Эта модель на самом деле может быть вписана в схему MEMM. В Яндексе она использовалась при индексации документов. В настоящее время разработана также модель снятия омонимии, основанная на нормализующих подстановках, но учитывающая ряд дополнительных факторов и применяющая для машинного обучения технологию MatrixNet [Технология MatrixNet], которая и используется в Mystem 3.0. Эта модель также обучалась на Национальном корпусе русского языка.

Нормализующей подстановкой для пары словоформа — лемма является пятибуквенное (в [Зеленков, Сегалович, Титов, 2005] — трёхбуквенное) окончание словоформы, для которого указывается, как надо заменить последние буквы словоформы (сколько букв стереть, что дописать), чтобы получить лемму. Например, для пары *люди — человек* нормализующей подстановкой будет *люди(4человек)*, а для пары *научно — научный — аучно(1ый)*. Фак-

тором для модели является частота встречаемости нормализующей подстановки для текущего слова совместно с его контекстом. Как и в [Зеленков, Сегалович, Титов, 2005], под контекстом понимаются трёхбуквенные окончания слов-соседей (двух ближайших слов слева и двух справа). Другим важным фактором является бесконтекстный вес разбора (см. выше).

В качестве дополнительных параметров используются: логарифм частоты пары словоформа — лемма, частоты пар контекст — часть речи, а также ряд бинарных факторов, связанных с наличием определённых тегов у текущего и предыдущих слов. Частотные факторы используются не в чистом виде, а дополнительно нормализуются относительно всех возможных кандидатов для данного разбора.

Режим снятия омонимии включается опцией -d:

$ mystem -c -g -d

Второй том «Мёртвых душ» стоит на полке.

Второй{второй} том{том} «Мёртвых{мертвый} душ{душа}» стоит{стоять} на{на} полке{полка}.

Среди прочих возможностей Mystem, появившихся в третьей версии, отметим возможность вывода результатов в форматах, удобных для автоматической обработки (json и xml), режим генерации всех возможных несловарных разборов (опция --generate-all), а также возможность подключить пользовательский словарь с особыми разборами конкретных слов (опция --fixlist). Примеры можно посмотреть на странице программы https://tech.yandex.ru/mystem/.

Литература

Антонова А. Ю., Соловьев А. Н. Использование метода условных случайных полей для обработки текстов на русском языке. // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2013». М.: РГГУ, 2013. С. 39-52. http://www.dialog-21.ru/digests/dialog2013/materials/pdf/AntonovaAJ.pdf (дата обращения 23.05.2015)

Зализняк А. А. Грамматический словарь русского языка. Словоизменение. М., 1977.

Зеленков Ю., Сегалович И., Титов В. Вероятностная модель снятия морфологической омонимии на основе нормализующих подста-

новок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-2005». URL: http://tech.yandex.ru/research/2005/34/. (дата обращения 23.05.2015)

Сегалович И., Маслов М. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов. // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог-1998». URL: http://download.yandex.ru/company/DLG98-MM2.pdf. (дата обращения 23.05.2015)

Технология MatrixNet. https://yandex.ru/company/technologies/matrixnet/ (дата обращения 23.05.2015)

Jurafsky D., Martin J. Speech and Language Processing. 2009.

Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. MLMTA-2003. http://download.yandex.ru/company/iseg-las-vegas.pdf (дата обращения 23.05.2015)

[1]A. I. Zobnin, [2]G. V. Nosyrev

[1, 2]Yandex

[1, 2](Russia, Moscow)

[1]alzobnin@yandex-team.ru, [2]nosyrev@yandex-team.ru

MORPHOLOGICAL ANALYZER MYSTEM 3.0

The large part of the Russian National Corpus has automatic morphological markup. It is based on the morphological analyzer Mystem developed in Yandex with some postprocessing of the results (for example, all indeclinable nouns acquire the tag '0', verbs are divided into separate paradigms by aspect, etc.). Recently a new (third) version of Mystem has been released (see https://tech.yandex.ru/mystem/). In this article we give an overview of its capabilities.

Key words: Mystem, morphology, Russian National Corpus, Yandex.

References

Antonova A. Yu., Solov'ev A. N. [Using Conditional Random Fields for processing Russian texts]. *Trudy mezhdunarodnoi konferentsii "Dialog–2013"* [Proceedings of International Conference "Dialog–2013"]. Moscow, PSUH Publ., 2013, pp. 39-52, available at http://www.dialog-21.ru/digests/dialog2013/materials/pdf/AntonovaAJ.pdf (accessed 23.05.2015) (In Russ.)

Jurafsky D., Martin J. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. 2nd edition. Prentice-Hall, 2009.

Segalovich I., Maslov M. [Russian morphological analysis and synthesis]. *Trudy mezhdunarodnoi konferentsii "Dialog–1998"* [Proceedings of International Conference "Dialog–1998"]. Moscow, 1998, available at http://download.yandex.ru/company/DLG98-MM2.pdf. (accessed 23.05.2015) (In Russ.)

Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. *Proceedings of MLMTA-2003*, available at http://download.yandex.ru/company/iseg-las-vegas.pdf. (accessed 23.05.2015)

Tekhnologiya MatrixNet [MatrixNet Technology], available at https://yandex.ru/company/technologies/matrixnet/ (accessed 23.05.2015)

Zaliznjak A. A. *Grammaticheskij slovar' russkogo jazyka. Slovoizmenenie* [Grammatical dictionary of the Russian language. Inflection]. Moscow, Russkij Jazyk Publ., 1977.

Zelenkov Yu., Segalovich I., Titov V. [Probabilistic morphology disambiguation model based on normalizing substitutions and neighbour words' order]. *Trudy mezhdunarodnoi konferentsii "Dialog–2005"* [Proceedings of International Conference "Dialog–2005"]. Moscow, 2005, available at http://tech.yandex.ru/research/2005/34/. (accessed 23.05.2015) (In Russ.)

II. Национальный корпус русского языка в научных исследованиях и образовании

II. The Russian National Corpus in Scientific Research and Education

[1,2]Е. В. Рахилина
[1]Национальный исследовательский университет
«Высшая школа экономики»
[2]Институт русского языка им. В. В. Виноградова РАН
(Россия, Москва)
rakhilina@gmail.com

СТЕПЕНИ СРАВНЕНИЯ В СВЕТЕ РУССКОЙ ГРАММАТИКИ ОШИБОК[1]

На корпусном материале в статье рассматриваются распространенные ошибки в русской конструкции сравнения — с теоретической и типологической точек зрения. Используются данные корпуса Академического письма и Учебного корпуса, содержащего тексты изучающих русский язык как иностранный. Выявлены «слабые места» и «точки роста» этой конструкции. Показано, что наблюдаемая вариативность и изменения в узусе в целом ожидаемы: они упрощают имеющуюся в русском модель сравнения в соответствии с другими типологически релевантными конструкциями.

Ключевые слова: сравнительные конструкции, грамматическая норма, грамматика ошибок, русский язык как иностранный, корпус эритажных текстов

Введение

Привычно представление о грамматике как о каноническом своде правил, который описывает те значения в нормативном варианте данного языка, которые выражаются грамматическими средствами. Традицию такого описания для русского заложил еще М. В. Ломо-

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2015 году.

носов, и она исправно воспроизводилась до последнего издания Академической грамматики русского языка (1980). В сегодняшних условиях, используя новые возможности, которые получила лингвистика, естественно было бы обновить контекст описываемых в канонической грамматике явлений за счет диахронического, типологического, корпусного и статистического «расширений» каждого из них — иными словами, учесть происхождение и прежний круг употреблений всех грамматических маркеров, их типологические особенности, статистику употреблений по жанрам и в соотнесении с другими показателями. Полностью эту программу выполнить нелегко, хотя до некоторой степени она уже отражается в новом проекте русской корпусной грамматики rusgram.ru, которая создается на базе НКРЯ и других ресурсов.

Есть, однако, и более радикальные проекты, принципиально меняющие точку зрения на грамматику как таковую, лишающие ее, так сказать, привычного облика. Среди них — идея наряду с грамматической нормой для каждой грамматической зоны описывать еще и не-норму, то есть наиболее характерные, частотные, стандартные для нее отклонения-«ошибки».

Сама по себе идея грамматики ошибок не так уж нова: полноправным объектом исследования ошибки предложил считать в своей знаменитой книге еще Анри Фрей в 1929 году (см. недавно вышедший перевод Фрей 2006). В отношении русского языка большая работа в этом направлении ведется Отделом современного русского языка ИРЯ им. В. В. Виноградова РАН, сотрудники которого уже давно изучают особенности языка русского зарубежья и разговорной нормы в этом отношении — т. е., собственно, отклонения от стандарта — или «ошибки» в этих вариантах русского (см., например, Гловинская и др. 2008). Серьезный всплеск интереса к разговорной речи произошел сравнительно недавно — нельзя не упомянуть работы Степанова 2008, Русакова 2013 и др. Параллельно ведутся работы над изучением детской речи (см. прежде всего работы группы С. Н. Цейтлин, а также Воейкова 2011 и др.) и языковым поведением иностранцев, изучающих русский.

Несмотря на очевидный количественный рост и бесспорную значимость работ такого рода, они все еще остаются фрагментарными. Между тем для русского (как и для других языков) нужна такая «грамматика ошибок», которая бы последовательно отражала

системные отклонения в узусе самых разных типов говорящих. Эта задача (в другом контексте) обсуждалась еще в работе Mustajoki et al. 2010, который предлагал свой список лингвистически интересных нестандартных носителей. В него входили в том числе:

- Студенты, изучающие русский язык как иностранный (РКИ),
- Носители регионального русского,
- Носители так называемого «эритажного», или унаследованного русского (heritage Russian), на котором говорят дети эмигрантов, выросшие в иноязычной среде и унаследовавшие навыки родного языка от родителей,
- Дети и подростки,
- Взрослые носители стандартного русского в определенных обстоятельствах (например, студенты, овладевающие академическим письмом) и др.

С педагогической точки зрения изучение ошибок той или иной группы говорящих помогает усовершенствовать методику обучения языку: они дают информацию о том, в каких случаях и с какой частотой можно ожидать данные отклонения от нормы, позволяют их предвидеть, ранжировать и, в конечном счете, эффективнее преодолевать.

С теоретической точки зрения изучение ошибок тоже полезно: ошибки дают представление о действующих языковых правилах (см. нашу статью Рахилина 2014). Еще со времен структурализма лингвисты стали использовать так называемые отрицательные примеры (или «примеры со звездочкой»), для иллюстрации того, как нельзя сказать, обозначая границы, в которых действуют обсуждаемые закономерности. Обычно «неправильные» примеры искусственны: они специально сочиняются исследователем, который с их помощью защищает придуманное им правило. Между тем ошибки носителей предоставляют не сконструированный, а естественный отрицательный материал — можно сказать, «экологически чистые» примеры спонтанного поведения говорящего, нарушающего принятую языковую норму. Другое дело, что правила, которые они иллюстрируют, далеко не всегда известны и очевидны, они реконструируются именно благодаря сопоставлению с очевидными отклонениями от них, которые как раз представляют ошибочные примеры. Таким образом, в лингвистических исследованиях ошибки можно рассматривать как инструмент для поиска языковых правил — своего рода «правило-искатель».

Современная постановка проблемы требует, чтобы изучение языковых явлений опиралось на структурированные массивы данных, которые в частности, могут стать базой необходимой статистики — в нашем случае, это корпус текстов соответствующих носителей, представляющий свойственные им отклонения от стандартной нормы. Такие корпуса принято называть Learner corpora. Такой ресурс http://web-corpora.net/RussianLearnerCorpus создается в рамках НКРЯ. Пока он не очень большой и включает в себя корпус текстов эритажных носителей с доминантным английским (американский вариант)[1], несколько десятков текстов от студентов РКИ, а также корпус Академического письма (студенческие письменные работы носителей русского языка — эссе, курсовые и под. различной профессиональной тематики).

В зависимости от вида текстов, меняется тип правил, которые можно обнаружить этим инструментом. Например, изучение ошибок-калек на подкорпусе эритажных носителей и РКИ позволяет сравнивать организацию лексических систем разных языков и выявлять принципы перестройки конструкций в условиях интерференции языков, и такие исследования нами уже проводились, см. Выренкова и др. 2014.

Другая задача — описать постепенную перестройку языковых конструкций как естественный процесс без непосредственного внешнего влияния. Решить ее можно только выстроив целую линейку корпусных ресурсов.

Прежде всего, для этого нужны корпуса детской речи (как правило, детей до 6 лет), они помогают реконструировать стадии освоения языка ребенком. Для русского языка корпуса детской речи созданы (в виде базы данных), хотя и в недостаточном объеме, в рамках проекта CHILDES. Отсутствующий для русского ресурс — это корпус речи подростков, от 7–8 лет, причем прежде всего устной

[1] http://web-corpora.net/RussianLearnerCorpus/search. Корпус создан как часть Национального корпуса русского языка в рамках сотрудничества с National Heritage Language Resource Center (UCLA) и Language Science Lab Гарвардского университета (под руководством М.С. Полинской) и исследовательской группы Направления фундаментальной и прикладной лингвистики филологического факультета НИУ ВШЭ под руководством Е.В. Рахилиной. Главный конструктор корпуса - Тимофей Архангельский. В состав корпуса входят тексты Учебного корпуса академической речи (Russian Learner Corpus of Academic Corpus), разработанного О.В. Киселевой и А.А. Алсуфиевой, а также тексты, собранные И.Е. Дубининой и Е.С. Деньгубом.

речи — как кажется, он позволил бы наблюдать сразу и процессы конвенционализации синтаксических структур, которые берут начало в речи детей, и процессы языковых инноваций, свойственные взрослым носителям: не исключено, что главными инициаторами языковых изменений является именно это поколение носителей.

Вообще, корпус разговорной речи (к настоящему моменту больше 10 млн. словоупотребений в НКРЯ — это очень много для ресурса такого качества, но и мало — для масштабных исследований) должен быть в перспективе стратифицирован по возрастам, точно так же, как сейчас исторический корпус разбит на подкорпуса по историческим периодам: тогда будет видна динамика изменений.

Пока мы можем только реконструировать этот процесс, т. е. оглядываться назад (а не смотреть вперед) и описывать то, что фактически в нашем языке уже произошло — опираясь прежде всего на частотные ошибки взрослых в родном языке, в частности, те, которые представлены в Корпусе академического письма. Они составят материал настоящего исследования.

Мы покажем, что, во-первых, в соответствии с идеями А. Фрея, такие «ошибки» действительно следует рассматривать как маркеры нестабильных языковых зон, или очагов потенциальных языковых изменений в грамматике данного языка. И, во-вторых, что здесь корпус может дать стимул не только для новых лингвистических задач, но и для новых подходов к исследованию уже известной проблематики — например, сравнительных конструкций, о которых пойдет речь в этой статье.

2. Ошибки в сравнительных конструкциях

Степени сравнения заведомо входят в группу лидеров по количеству и разнообразию ошибок, особенно в устной речи, когда носители языка недостаточно контролируют свое языковое поведение. Мы остановимся на трех важных, на наш взгляд, и наиболее характерных отклонениях от грамматической нормы в этой области, см. типичные примеры вида (1–3).

(1) *Эти ноутбуки потребляли меньше энергии, но были* ***менее компактнее*** *по объему*.

(2) *И прибыль от разрушения гораздо более заметна и быстра,* ***нежели чем*** *от строительства*.

(3) *В русском языке семантический диапазон данного слова* ***чрезвычайно широк, нежели*** *в английском*.

Во всех трех примерах норма, очевидно, нарушена: конструкция не соответствует правилам русского языка. Правильную конструкцию можно было, в духе теории Construction Grammar, описать, как это сделано в близкой по тематике работе [Haspelmath M., Buchholz 1998], в виде таблицы. Ее вертикальные входы обозначают переменные или постоянные (лексически фиксированные) места в конструкции, которые в следующей строке иллюстрируются конкретным языковым примером, см. Таблицу 1.

X	(COMP1A)	Q	(COMP1Ba)	(COMP2)	Y	(COMP1Bb)
ПРИБЫЛЬ ОТ РАЗРУШЕНИЯ	БОЛЕЕ	ЗАМЕТНА	–ЕЕ /-Е	ЧЕМ / НЕЖЕЛИ	ПРИБЫЛЬ ОТ СТРОИТЕЛЬСТВА	РОД.П

Таблица 1. Структура сравнительной конструкции на примере (2’)

Нашим примером будет предложение (2) в исправленном и упрощенном варианте (2’):

(2’) *Прибыль от разрушения более заметна,* ***чем*** *прибыль от строительства*.

В нем две именные группы, соответствующие сравниваемым объектам: сравниваемый субъект X (*прибыль от разрушения*) и эталон для сравнения Y (*прибыль от строительства*). Они маркируют границы конструкции, все остальные ее компоненты всегда находятся между ними, в противном случае предложение теряет грамматическую правильность, ср.:

(2’’) **Прибыль от разрушения* ***чем*** *прибыль от строительства более заметна* / **более, чем от строительства заметна* и под.

Кроме того, в конструкции участвует признак Q, по которому сравниваются X и Y, а также два показателя сравнения: COMP1 и COMP2. Есть два варианта оформления COMP1.

Вариант A: COMP1 — это аналитический показатель — наречия *более* или *менее*, которые выступают в препозиции к признаку Q: *более* / *менее заметный*.

Вариант B: показатель сравнения встраивается в морфологическую структуру Q как особый суффикс *-ее*/*-е*, ср.:

более заметный (A) VS. *заметнее* (B).

При выборе синтетической формы прилагательного остальной

состав конструкции может ничем не отличаться от аналитического варианта. Это случай (Ba):

заметнее, чем прибыль ~ более заметна, чем прибыль

Однако синтетическая форма может быть сопряжена с изменением падежа для эталона Y: с именительного на родительный. Это случай (Bb). Родительный падеж становится частью составного маркера сравнительной конструкции, построенной по модели COMP1B, показатель COMP2 с этим вариантом оформления не используется:

заметнее (**чем*) *прибыли*

Как известно, распределение аналитической и синтетической форм нежестко — они варьируются, ср. следующие примеры из НКРЯ:

(4) *Ваша власть, когда вы отказываете, гораздо заметнее, чем когда разрешаете*. [Владимир Войнович. Иванькиада, или рассказ о вселении писателя Войновича в новую квартиру (1976)]

(5) *Более заметен рост цен в нагорной части города.* [«Биржа плюс свой дом» (Н. Новгород), 2002.04.08]

Второй показатель сравнения COMP2 следует непосредственно за Q. Сам он всегда аналитический (основной вариант — союз *чем*) и может, как мы видели, использоваться как в паре с аналитическим COMP1 A, так и с синтетическим COMP1B, но только в варианте (a): падеж Y-а при нем никогда не меняется.

COMP2 тоже допускает варьирование: в современном русском языке, наряду с более частотным *чем*, используется и союз *нежели*.

Таким образом, с учетом описанной вариативности, есть по крайней мере пять возможностей, соответствующих семантике (2’), для оформления сравнительной степени:

(2a) *Прибыль от разрушения заметнее* ***прибыли*** *от строительства*;

(2b) *Прибыль от разрушения заметнее, чем* ***прибыль*** *от строительства*;

(2c) *Прибыль от разрушения заметнее, нежели* ***прибыль*** *от строительства*;

(2d) *Прибыль от разрушения более заметна, чем* ***прибыль*** *от строительства*;

(2e) *Прибыль от разрушения более заметна, нежели* ***прибыль*** *от строительства*.

Очевидно, что это конструкция, в которой должна происходить какая-то перестройка: она слишком громоздка, маркирование в ней избыточно (или два показателя, или один составной); вдобавок, оно высоко вариативно. Ясно, что говорящие будут пытаться уйти от такой сложной конфигурации, как-то упростив ее — этим и мотивированы ошибки, с которыми мы постоянно сталкиваемся в письменных текстах, интернете и устной речи: происходит поиск оптимального способа перестроить неудачную форму. Чтобы понять основные стратегии, по которым идет упрощение, вернемся к нарушениям нормы, представленным в предложениях (1–3) и последовательно рассмотрим их в сопоставлении с нормативным (2').

3. Стратегии ошибок

Стратегия 1 отражена в Таблице 2. Как видим, здесь реализуются обе типа COMP1 — сразу: и A, и B, т. е. и аналитический, и синтетический. Таким образом, первый показатель компаратива как бы удваивается, однако падежная маркировка Y не меняется:

X	COMP1A	Q	COMP1Ba	COMP2	Y	--
ПРИБЫЛЬ ОТ РАЗРУШЕНИЯ	БОЛЕЕ	ЗАМЕТНА	–ЕЕ /-Е	ЧЕМ	ПРИБЫЛЬ ОТ СТРОИТЕЛЬСТВА	--

Таблица 2. Ошибка первого типа: удвоение первого показателя

Такого рода примеров встречается необычайно много — собственно, знаменитое *более лучше*, широко обсуждавшееся в свое время в интернете и прессе, относится как раз к этому типу. Поражает лексическое разнообразие признаков, с которыми встречается удвоение первого показателя сравнительной степени — оно говорит об очень высокой продуктивности такой стратегии. Ср. следующие примеры из Яндекса:

(6) Р*ост объема продаж в секторе легковых авто в Москве намного* ***более заметнее, чем*** *в других регионах страны.*

(7) <…> *где вы видели, чтобы щенку или котёнку давали сладкий чай или соки — наверное природа* ***более умнее, чем*** *человек.*

(8) *А поскольку сладкая вода имеет более густую консистенцию, то и охлаждаться ей куда* ***более дольше, чем*** *обычной.*

(9) <…> *многие, наоборот, в зрелом возрасте выглядят куда* ***более красивее, чем*** *в юности.*

(10) *Черный цвет смотрелся как то* ***более естественнее, чем*** *на остальных, и картинка не блеклая.*

(11) *Компоненты табачного дыма делают кровь* ***более гуще, чем*** *обычно, и приводят к образованию тромбов.*

Стратегия 2, напротив, связана с удвоением второго показателя компаратива — первый при этом, в принципе, может быть как аналитическим, так и синтетическим, но без дополнительного падежного маркирования Y:

X	(COMP1A)	Q	(COMP1Ba)	COMP2	Y	--
ПРИБЫЛЬ ОТ РАЗРУШЕНИЯ	(БОЛЕЕ)	ЗАМЕТНА	(–ЕЕ /-Е)	НЕЖЕЛИ ЧЕМ	ПРИБЫЛЬ ОТ СТРОИТЕЛЬСТВА	--

Таблица 3. Ошибка второго типа: удвоение COMP2

Такие примеры встречаются в самых разных текстах, прежде всего, в новостных, но и в текстах академического характера, ср.:

(12)… *ВТБ имеет гораздо меньшую системную значимость* ***нежели чем****, например, Сбербанк…*

(13) *Ведь как прогнозируют медики, детей с ДЦП в Волгоградской области будет больше,* ***нежели чем*** *еще пять лет назад.*

(14) *Интересно отметить, что несколько чаще используются глаголы, описывающие увеличение численности преступлений,* ***нежели чем*** *снижение преступной деятельности.*

Любопытна история этого двойного союза (см. Розина 2014): данные НКРЯ показывают, что еще в XIX в. он функционировал как полноправный (см. ниже пример из Лескова) — и, еще более ранний, с вариативностью *нежели / нежели чем*, у Радищева:

(15) *— Что же, — говорю, — это гораздо разумнейше,* ***нежели чем*** *совсем плешкой остаться*. [Н. С. Лесков. Заячий ремиз (1894)]

(16) *Я лучше желаю, чтобы во младых летах ваших вы были распутны, расточительны, наглы,* ***нежели*** *сребролюбивы или же чрезмерно бережливы, щеголеваты, занимался более убранством, нежели чем другим*. [А. Н. Радищев. Путешествие из Петербурга в Москву (1779–1790)]

Был вариант и с обратным порядком слов: *чем нежели*, ср. следующий пример как стилизацию:

(17) *Поклоны «от головы до сырой земли», «жив и здоров, чего и вам желаю», а где-то невзначай, неуместно вплетется писакой: «Скучно,* ***чем нежели*** *допреже одному без семьи жить… »* [К. С. Петров-Водкин. Моя повесть. Часть 1. Хлыновск (1930)]

В XX в. двойной сравнительный союз был полностью вытеснен: остаются оба одиночных союза, с неясным распределением: *нежели* — скорее, в академических и бюрократических текстах, *чем* — более нейтральных и разговорных. Теперь, как видим, двойной союз возвращается, причем, судя по НКРЯ, в том числе и в разговорную речь, ср.:

(18) [№ 2, муж, 50] *Меньшими правами / конечно / обладает /* ***нежели чем*** *монарх / в свое время.* [Беседа с социологом на общественно-политические темы (Самара) // Фонд «Общественное мнение», 2003]

Стратегия 3 принципиально отличается от первых двух. Там упрощение достигалось за счет того, что устранялся выбор способа оформления для показателя компаратива: выбирать становилось не нужно, потому что использовались одновременно оба варианта оформления. При этом конструкция действительно становилась проще в употреблении — но длиннее.

Ясно, что должна быть и другая возможность упрощения: наоборот, через редукцию хотя бы одного из показателей — и она реализуется в стратегии 3. Редуцируется первый маркер компаратива, со сложной дистрибуцией, и остается только второй, причем по-прежнему допускается вариативность *чем / нежели* (см. Таблицу 4).

Интересно, что возможность такой редукции возникает только в контексте интенсификатора — в примере (2’) это *чрезвычайно*, но, как следует из следующих ниже примеров, лексические интенсификаторы могут быть и другими — в частности, *очень* или *слишком*.

X	INT	Q	--	COMP2	Y	--
ПРИБЫЛЬ ОТ РАЗРУШЕНИЯ	ЧРЕЗВЫЧАЙНО	ЗАМЕТНА	--	ЧЕМ / НЕЖЕЛИ	ПРИБЫЛЬ ОТ СТРОИТЕЛЬСТВА	--

Таблица 4. Стратегия 3: редукция COMP1

(19) *Разница в состоянии и настроении малыша стала* ***очень заметна, чем*** *было до операции.*

(20) *а вот Реал я думаю по уровню игры на данный момент* ***очень хорош, нежели*** *Ювентус…*

(21) *влага в центре держится* ***слишком долго, нежели*** *сверху и по краям.*

(22) *Реально нет, нет сил отвечать всем, улыбаться, смеяться, это продлилось* ***слишком долго, чем*** *я планировал.*

Таким образом, третья стратегия использует интенсификатор как дополнительное средство маркирования сравнительной конструкции, не свойственное ей изначально. Откуда берется эта идея у носителей языка?

Прежде чем отвечать на этот вопрос, обратим внимание на тот факт, что подобные конструкции, вообще говоря, широко распространены в языках мира. Наиболее типично для них использование минимальной морфологической техники: основание для сравнения выражается обычным предикатом без дополнительных показателей, а маркер сравнительной степени — глаголом типа ‘превосходить’, содержащим идею интенсивности (тип ‘дерево высокое превосходит дом’ = ‘дерево выше дома’). Эти конструкции обычно считаются разновидностью сериальных (ср., например, Aikhenvald & Dixon 2006) и засвидетельствованы почти во всех ареалах, где имеется глагольная сериализация (в частности, в языках Юго-Восточной Азии, Западной и Центральной Африки и др.). Выбранный способ выражения компаратива заставляет задуматься: ведь в принципе, в определенной степени тот же смысл содержит и русский показатель COMP1 *больше*, так что сам по себе такой способ его передачи не так уж экзотичен, а в каком-то смысле семантически очень естествен.

Другое соображение касается сравнения «дублирующей» и «редуцирующей» стратегий перестройки конструкций. Полу-

чается, что техника удвоения маркера, в общем, тоже сродни интенсификации, только в одном случае интенсификатором служит «готовый» аналитический показатель компаратива, дополнительный по отношению к синтетической форме, а в другом — в конструкцию встраивается не свойственный ей интенсив как таковой, с более прозрачной семантикой.

И, наконец, третье: удвоение маркеров (в нашем материале — COMP1 и COMP2 соответственно) — как известно, стандартное поведение интенсивов, у которых со временем ослабляется, стирается исходная семантика (вспомним известный цикл Есперсена). Такой путь уже прошла превосходная степень: еще недавно плеонастическая форма *наилучший*, как и ее аналитический аналог *самый лучший* считалась грамматически неправильной. Теперь эти запреты не действуют: в сегодняшнем узусе приемлемо даже *самый наилучший*.

(24) *Самый могучий,* ***самый наилучший****, ничего не скажешь — молодец!* [Комсомольская правда, 2010.12.01]

Ср., однако, и гораздо более старые примеры, в которых встречается эта форма — из Мельникова-Печерского, Золотницкого; см. из НКРЯ:

(25) *Почему нельзя не признаться, чтоб сей способ, то есть, добрый пример, для воспитания детей не был* ***самый наилучший***. [архиепископ Платон (Левшин). Слово о воспитании (1765)].

Более того, саму форму *более лучше* позволял себе в живом, разговорном стиле письма …И. А. Гончаров:

(26) *Письмо к Г. Божерянову — написано очень живо, бойко и куда* ***более лучше*** *письма к дежурному офицеру и Дежурной палатки!* [И. А. Гончаров. Письма великому князю Константину Константиновичу (1888)]

Все сказанное означает, что интенсификатор можно рассматривать как маркер русской сравнительной конструкции, и возникает он не случайно, а в согласовании с внутренней семантикой, которой мотивируется ее дальнейшее развитие.

4. Сравнение и интенсификация: другие ошибки

Связь русской конструкции сравнения с интенсификацией можно проследить и по так называемому эритажному подкорпусу, о котором мы уже говорили. Эритажные носители тоже чувствуют себя неуверенно при построении русской сравнительной степени: на 340 тыс. словоупотреблений в нем встретилось 104 «ошибки» на сравнительные конструкции, причем не кальки.

Они касаются в основном сравнительной степени наречий и распределяются по двум основным типам: (1) «ошибки» в интенсификаторах и (2) в выборе наречия[1].

(1) ***«Ошибки» в интенсификаторах***

Оказывается, дополнительные интенсификаторы сравнения довольно частотны в обыденной речи. В этом качестве используются:

- частица *всё* (*всё яснее*),
- эффект удвоения (*постепенно дела шли* (*всё*) *лучше и лучше*),
- наречия *гораздо, много, наиболее, еще*,

ср. следующий пример из корпуса с помеченной звездочкой ошибкой и правильной конструкцией в скобках:

(27) *Война на Дальнем Востоке — это не только различие политических взглядов России и Японии, не поделивших территорию на Востоке, это что-то *намного больше* (*гораздо худшее* / *много большее* / *много более сложное*).

Надо сказать, что интенсификаторы — это область, промежуточная между словарем и грамматикой, формат их описания не определен. В словаре фиксируются только значения самих этих единиц, но не правила поведения соответствующих конструкций. Грамматика же регистрирует общие правила выражения обязательных значений, в том числе сравнительной степени, но там не предусмотрена возможность регистрации всех лексических расширителей конструкций, да еще с учетом взаимодействия с контекстом. Поэтому где должны оказаться правила употребления дополнительных интенсификаторов — неясно, и их нет ни для сравнительной, ни для превосходной степени, хотя, как мы теперь

1 В этом и следующем разделе использован материал, собранный по эритажному корпусу студенткой второго курса НИУ ВШЭ Дарьей Кошелевой в рамках ее курсового проекта

видим из анализа корпусных данных, как раз они вызывают большие трудности у нестандартных говорящих. Ср., например:

(28) *На педагогический факультет подают *само меньше заявлений а без педагогов, школы и университеты не могут существовать* (вместо: *больше всего*).

(2) «Ошибки» в выборе наречия

Неправильный, с точки зрения канонической грамматики, лексический выбор наречия в корпусных примерах обусловлен тем, что с семантической точки зрения, все эти наречия, в сущности, оказываются квазисинонимическими лексическими функциями (в терминологии Московской Семантической Школы) с общим значением интенсива или аттенуатива, ср. следующие пары примеров с близким значением:

(a) *Человек там* ***больше зависит*** *от другого человека, чем от Бога, человек там закрывает собой Бога, и о Нем забывают*. [Владимир Шаров. Воскрешение Лазаря (1997–2002)]

(b) *Он беззащитнее других, он* ***сильнее зависит*** *от окружающих, его внутренний мир более раним.* (http://bigbilet.ru/place/lateatr/afisha/svobodnaya_lubov?ev=03B76C80E0740C2FE0504B5E01F54B73, интернет-форум, обсуждение спектакля «Свободная любовь» от 3 января 2014)

(a/b) *Но в данном случае этот прием* ***чаще*** */* ***больше*** *отвлекал от сюжета, чем помогал ему раскрыться*. [«Театральная жизнь», 2003.08.25]

(a/b) *Постепенно глазное пятно ещё* ***глубже*** */* ***больше*** *вдавливалось в эпителий*. [Знание — сила», 2003]

В такой ситуации неуверенными говорящими, к которым относятся эритажные носители, выбирается композициональная стратегия, см. подробнее [Выренкова и др. 2014]: как правило, они используют лексему с самой общей семантикой и самой свободной сочетаемостью — это наречие *больше* (*меньше*). Между тем, грамматически правильная замена контекстно связанного наречия на семантически генерализованные больше или меньше возможна далеко не всегда, и правила в этой зоне пока не прояснены. Ср. следующий пример из нашего корпуса:

(29) *В России человек *больше (! чаще) считается расистом из-за действий*

5. Адъективные и адвербиальные сравнительные конструкции

Как мы уже говорили, большинство ошибок на сравнительную степень делается при выборе наречий, а не прилагательных: выбор прилагательного почему-то не вызывает сомнений. Дело в том, что в адъективной конструкции действительно важна семантика признака, по которому сопоставляются объекты, а в адвербиальной конструкции, как правило, собственно признак не так важен: как мы только что видели, о многих случаях он имеет общее значение интенсива. Трудность в том, что при разных предикатах интенсивность выражается разными наречиями — именно в этом уже фразеологизованном распределении, хорошо известном носителям, и ошибаются эритажные говорящие.

Это обстоятельство имеет свои причины и свидетельствует о противопоставлении наречных (адвербиальных) и адъективных конструкций сравнения. Надо сказать, что в авторитетных грамматических описаниях (АГ 80, Князев 2007, Санников 2008 и др.) принято рассматривать адвербиальные сравнительные конструкции в одном ряду с атрибутивными, как подчиняющиеся одним и тем же канонам. Кажущийся структурный параллелизм между адъективными и адвербиальными иллюстрируют следующе примеры из НКРЯ:

(30) Атрибутивная:

X COMP ЧЕМ\НЕЖЕЛИ Y

Правительство в этом вопросе считает, что ***эволюция лучше, чем революция***. [А. Чудодеев.//«Итоги», 2003.02.04]

(31) Адвербиальная:

V1 COMP ЧЕМ\НЕЖЕЛИ V2

Мы ***лечили лучше, чем лечат*** *в Америке.* [В. Голяховский. Русский доктор в Америке (1984–2001)]

Между тем, данные того же корпуса показывают, что в подавляющем большинстве случаев адвербиальная «двухчастная» конструкция сравнивает не две разные ситуации (т. е. два разных предиката — по аналогии с адъективной, сопоставляющей два имени по одному признаку), а варианты одной и той же. Собственно, это видно уже из приведенных только что примеров: в адъективной конструкции имена разные (*эволюция* и *революция*), а в адвербиальной предикат один и тот же (*лечить*). В определенном контексте можно

представить и что-то вроде: мы *лечили лучше, чем чинили машины / пели / шили* и под., но ситуативный контекст для оправдания такой структуры должен быть достаточно причудлив, так что она встречается редко[1].

Часто встречаются адвербиальные конструкции, в которых сравниваемая ситуация имеет измененный по сравнению с эталоном набор участников или представлена в другом временном, пространственном или модальном плане. Новые участники требуют обязательного поверхностного выражения (ср.: *Петя поет лучше, чем Вася*) — новые обстоятельства, хотя семантико-синтаксический статус их столь же обязателен, как и у прочих участников, могут и не быть выражены поверхностно, если сравнение идет с настоящим моментом или уже упомянутым прошедшим. Ср. (a-b) как типичные контексты такого рода:

(a) **TEMP/LOC**: *Петя играет лучше* <чем Петя же — в другом месте или времени: *тогда, там …* >

(b) MOD: *осталось меньше, чем хотелось бы*
стало хуже, чем могло бы

С теоретической точки зрения примеры типа (a) не представляет трудности, в них действительно есть сходство с адъективными контекстами, в частности, допускается экспликация эталона непосредственно внутри клаузы (*чем вчера*) или анафорическая (*прошлый раз сыграл неудачно — сейчас играет лучше <чем прошлый раз>*).

Сложнее устроены модальные примеры, которые рассматривают сравниваемое событие целиком, приписывая ему некоторую оценку в сопоставлении с эталоном. Те, что приведены в (b), двухчастны, их части анафорически связаны, и эталоном оказывается гипотетическое событие, которое в обоих случаях в худшую сторону от-

[1] Вообще говоря, адвербиальная сравнительная конструкция, как и адъективная, не предусматривает никаких ограничений на сравниваемые объекты и события. Любопытно, что на самом деле, по-видимому, некоторые типы сравнения просто невозможны: сравнивается только подобное: трудно представить себе предложение типа **мы лечили лучше, чем ходили в поход*. Такого рода ограничения до некоторой степени описаны для метафор – сравнения на метафоры похожи, но не совсем, так что поиск ограничений такого рода на богатом корпусном материале должен дать любопытные результаты.

личается от сравниваемого: несвершившееся, как правило, лучше реального. Ср.:

осталось меньше, чем хотелось бы, <чтобы осталось>;

стало хуже, чем могло бы <быть>.

Впрочем, гипотетическое событие может само быть объектом сравнения с некоторым эталоном, тоже не выраженным поверхностно. Речь идет о дейктических употреблениях, в которых эталонное событие связано с настоящим моментом. Естественно, в этом случае эталон никак не выражен, он вводится косвенно, сравнительной степенью в контексте специальных лексико-грамматических показателей, прежде всего:

модальных глаголов, как в: *хочу* ***глубже*** *узнать <чем знаю сейчас>,*

сослагательного наклонения, как в: *который бы* ***лучше*** *соответствовал <чем тот, что сейчас>,*

целевого придаточного, как в: *чтобы* ***поменьше*** *зависеть <чем завишу сейчас>*.

Общая семантика здесь та же: гипотетическое событие лучше настоящего, которое, однако, поверхностно не выражено и служит точкой отсчета для общей оценки, т. е. эталоном.

6. Стертая семантика — новые ошибки

Все сказанное свидетельствует о том, что адвербиальная сравнительная конструкция устроена особенным образом. Семантически в ней, как мы видели, может и не быть (и часто не бывает) сравнения двух разных событий, а только усиление или ослабление выбранной ситуации, как правило, связанное с ее оптимизацией (или — деградацией).

Именно этим объясняются ошибки, связанные с выбором наречия, которые мы обсуждали в разделе (4), но и ошибки другого типа. Действительно, если нет сравнения как такового, это может способствовать грамматикализации и лексикализации наречий — они преобразуются в маркеры времени или дискурсивные слова. Понятно, что ход и направление этих процессов зависит от их исходной семантики. Лексикализованные маркеры времени возникают из дейктических наречий: *ранее, позднее, позже, раньше*; грамматикализованные дискурсивные слова — в первую очередь, из оценочных, ср.: *точнее, вернее, скорее, скорее всего*.

Правила их распределения не описаны (хотя среди них много квазисинонимов, ср. *позднее / позже, точнее / вернее* и др.). Конечно, для носителей языка они совершенно не нужны, потому что носители хорошо владеют сочетаемостными ограничениями частотных грамматикализованных или лексикализованных единиц, а вот эритажные говорящие (или студенты РКИ) испытывают понятные трудности при употреблении этих «стертых» наречий. Вот примеры характерных ошибок из нашего корпуса:

(32) *Во время нападения Наполеона на Москву (1812) пострадал и медицинский факультет МГУ <…> **Позже** <! **позднее**> он был приведен в полный порядок*.

(33) *Как я сказал **раньше** <! **ранее**>, я считаю себя гражданином мира.*

(34) *Пакт Рериха имеет образовательное назначение, и, <! **что**> важнее, художник пытался развить международные отношения через понятие культуры*.

Заключение

Традиционно считается, что для русского языка проблематика компаративной конструкции, которую мы выбрали здесь примером, не входит ни в какое сравнение по степени сложности ни с наклонением, ни с аспектом. Собственно, русская сравнительная конструкция давно и профессионально описана, см. АГ-1980, Князев 2007, 2011, см. также Воротников 1999 и др. Тем не менее, на многие из вопросов, которые мы задали в этой статье, проанализировав ошибки обычных и эритажных носителей, ответов нет. Не потому, что это слишком сложные вопросы, которые требуют значительных новых исследований, а потому, что далеко не все они сформулированы и осознаны как лингвистические задачи. Мы выбрали иной угол зрения — корпусное описание и другой инструмент поиска — корпуса ошибок — и эти задачи стали видны:

• Определение стратегий перестройки конструкции, из которых здесь мы остановились только на трех;

• Изменение в синхронии и диахронии (*чем нежели* и *нежели чем*);

• Распределение сравнительных союзов *чем* и *нежели*;

• Синтетическая и аналитическая формы сравнительной степени: есть ли дистрибуция?

• Генитив как часть сравнительного маркера (в сопоставлении с союзом *чем*);

• Значимость компонента интенсификации в семантике русского компаратива;

• Семантика и распределение дополнительных интенсификаторов (*всё, гораздо, еще* и др. под.);

• Роль удвоения в сравнительной конструкции (*хуже и хуже*);

• Адъективная и атрибутивная конструкции;

• Дейктичность адвербиального сравнения;

• Связь интенсификации и оценки в контекстах модального сравнения;

• Развитие дискурсивных слов из сравнительных наречий (*точнее, вернее*…);

• Семантика производных наречий в парах типа *позднее / позже* и *ранее / раньше*

• Семантические границы сравнения: почему не говорят **мы лечили лучше, чем ходили в поход*

Уже из этого предварительного списка видно, сколько еще предстоит сделать, чтобы получить полное представление о семантико-синтаксической специфике русских конструкций сравнения. Ошибки — это важный объект исследования, расширяющий наши лингвистические горизонты.

Литература

АГ 1980 — Русская грамматика. / Н. Ю. Шведова (ред.). М.: Наука, 1980.

Воейкова М. Д. Ранние этапы усвоения детьми именной морфологии русского языка. М., 2011.

Воротников Ю. Л. Степени качества в современном русском языке. М.: Азбуковник. 1999.

Выренкова А. С., Полинская М. С., Рахилина Е. В. Грамматика ошибок и грамматика конструкций: «эритажный» («унаследованный») русский язык // Вопросы языкознания, 2014, № 3, 3-19.

Гловинская М. Я., Галанова Е. И., Ермакова О. П., Занадворова А. В., Земская Е. А., Какорина Е. В., Касаткин Л. Л., Касаткина Р. Ф.,

Крысин Л. П., Кузьмина С. М., Розина Р. И. Современный русский язык: Активные процессы на рубеже XX-XXI веков. М.: Языки славянских культур, 2008.

Князев Ю. П. Грамматическая семантика. Русский язык в типологической перспективе. М.: Языки славянских культур, 2007.

Князев Ю. П. Сравнительная степень. Для проекта «РусГрам». 2011

Рахилина Е. В. Грамматика ошибок. В поисках констант // Язык, константы, переменные. Памяти А.Е. Кибрика. СПб, Алетейа, 2014. С. 87-95.

Розина Р. И. Активные процессы в современной русской речи // Труды Института русского языка им. В.В. Виноградова. Вып. 2. М., 2014. С. 221-234.

Русакова М. В. Элементы антропоцентрической грамматики русского языка. М.: Языки славянских культур, 2013.

Степанова С. Б., Асиновский А. С., Богданова Н.В. Русакова М. В., Шерстинова Т. Ю. Звуковой корпус повседневного общения «Один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии. Сборник по материалам ежегодной международной конференции «Диалог», вып. 7 (14), М.: РГГУ, 2008. С. 488-495.

Фрей А. Грамматика ошибок. Комкнига, 2006.

Aikhenvald, A. and Dixon, R. (eds.). Serial verb constructions: a cross-linguistic typology. Oxford: OUP, 2006.

Haspelmath, Buchholz. Haspelmath M., Buchholz. O. Equative and similative constructions in the languages of Europe. // van der Auwera (ed.) Adverbial constructions in the languages of Europe. (Empirical Approaches to Language Typology / EUROTYP, 20-3). Berlin: Mouton de Gruyter, 1998, pp. 277-334.

Mustajoki A., Protassova E., Vakhtin N. Instrumentarium of linguistics: sociolinguistic approach to non-standard Russian. Helsinki, 2010.

Polinsky M., Kagan O. Heritage languages: In the 'wild' and in the classroom. Language and Linguistics Compass 1(5). 2007.

Protassova E., Korneev A. Санкт-Петербургская школа онтолингвистики. К юбилею доктора филологических наук профессора С.Н. Цейтлин. / Т.А. Круглякова (отв. ред). СПб: Златоуст, 2013. С. 201-211

[1,2]***Ekaterina V. Rakhilina***
[1]National Research University "Higher School of Economics"
[2]Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
rakhilina@gmail.com

RUSSIAN COMPARATIVE CONSTRUCTIONS IN THE ASPECT OF ERROR GRAMMAR

The paper considers the errors characteristic to the usage of the Russian comparative construction in the theoretical and typological perspective. The research uses the data of the Academic writing corpus and the Russian Learner Corpus containing texts of Russian as foreign language speakers. The “weak points” and “growth points” of the construction are detected. It is demonstrated that the variations and the deviations from the norm registered in the corpora and other resources are quite predictable. They simplify the Russian comparative construction into the other typologically relevant patterns.

Key words: comparative constructions, grammatical norms, error grammar, the Russian as a foreign language, the Russian Learner Corpus, heritage Russian

References

Aikhenvald, A. and Dixon, R. (Eds.). *Serial verb constructions: a cross-linguistic typology*. Oxford, OUP, 2006.

Frei A. *Grammatika oshibok* [Error Grammar]. Komkniga Publ., 2006.

Glovinskaya M. Ya., Galanova E. I., Ermakova O. P., Zanadvorova A. V., Zemskaya E. A., Kakorina E. V., Kasatkin L. L., Kasatkina R. F., Krysin L. P., Kuz'mina S. M., Rozina R. I. *Sovremennyi russkii yazyk: Aktivnye protsessy na rubezhe XX-XXI vekov* [Modern Russian language: Active processes at the turn of XX-XXI centuries]. Moscow, Yazyki slavyanskikh kul'tur Publ., 2008.

Haspelmath Buchholz. Haspelmath M., Buchholz O. Equative and similative constructions in the languages of Europe. van der Auwera (ed.) Adverbial constructions in the languages of Europe. (Empirical

Approaches to Language Typology / EUROTYP, 20-3). Berlin, Mouton de Gruyter, 1998, pp. 277-334.

Knyazev Yu. P. *Grammaticheskaya semantika. Russkii yazyk v tipologicheskoi perspektive.* [Grammatical semantics. Russian language in a typological perspective]. Moscow, Yazyki slavyanskikh kul'tur Publ., 2007.

Knyazev Yu. P. *Sravnitel'naya stepen'. Dlya proekta «RusGram»* [The Comparative degree. For "RusGram " project]. 2011.

Mustajoki A., Protassova E., Vakhtin N. Instrumentarium of linguistics: sociolinguistic approach to non-standard Russian. Helsinki, 2010.

Polinsky M., Kagan O. Heritage languages: In the 'wild' and in the classroom. *Language and Linguistics Compass* 1(5). 2007.

Protassova E., Korneev A. *Sankt-Peterburgskaya shkola ontolingvistiki. K yubileyu doktora filologicheskikh nauk professora S.N. Tseitlin* [The St. Petersburg school of ontolinguistic. To the anniversary of Doctor of philological Sciences, Professor S. N. Zeitlin]. T.A. Kruglyakova (ed.). St. Petersburg, Zlatoust Publ., 2013, pp. 201-211.

Rakhilina E. V. [Error Grammar. Searching the constants]. *Yazyk, konstanty, peremennye. Pamyati A.E. Kibrika* [Language, constants, variables. In Memory of A. E. Kibrik.]. St. Petersburg, Aleteia Publ., 2014, pp. 87-95. (In Russ.)

Rozina R. I. [Active processes in modern Russian speech]. *Trudy Instituta russkogo yazyka im. V.V. Vinogradova* [Proceedings of the V. V. Vinogradov Russian Language Institute]. Issue 2. Moscow, 2014, pp. 221-234. (In Russ.)

Rusakova M. V. *Elementy antropotsentricheskoi grammatiki russkogo yazyka* [Anthropocentric Elements of Russian grammar]. Moscow, Yazyki slavyanskikh kul'tur Publ., 2013.

Russkaya grammatika [Russian Grammar]. N. Yu. Shvedova (ed.). Moscow, Nauka, 1980.

Stepanova S. B., Asinovskii A. S., Bogdanova N. V. Rusakova M. V., Sherstinova T. Yu. [Sound corpus of everyday communication "One speech day": concept and state of formation]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Sbornik po materialam ezhegodnoi mezhdunarodnoi konferentsii «Dialog»* [Computational linguistics and intellectual technologies. Collection of proceedings of the annual international conference "Dialogue"]. Issue. 7 (14), Moscow, RSUH

Publ., 2008, pp. 488-495. (In Russ.)

Voeikova M. D. *Rannie etapy usvoeniya det'mi imennoi morfologii russkogo yazyka* [Early stages of children's mastering the nominal morphology of the Russian language]. Moscow, 2011.

Vorotnikov Yu. L. *Stepeni kachestva v sovremennom russkom yazyke* [The degree of quality in the modern Russian language]. Moscow, Azbukovnik Publ., 1999.

Vyrenkova A. S., Polinskaya M. S., Rakhilina E. V. [Error Grammar and Construction Grammar: "Heritage" ("inherited") Russian language]. *Voprosy yazykoznaniya*, 2014, no. 3, pp. 3-19. (In Russ.)

С. О. Савчук
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
savsvetlana@mail.ru

КОРПУС КАК ИНСТРУМЕНТ ДЛЯ ИССЛЕДОВАНИЯ ОСОБЕННОСТЕЙ ФУНКЦИОНИРОВАНИЯ РУССКОГО ЯЗЫКА В РЕГИОНАЛЬНОЙ ПРЕССЕ[1]

В статье рассматривается опыт использования инструментов корпуса региональных и зарубежных СМИ в выявлении лексико-семантических различий функционирования новых слов в русскоязычных газетах Гродненского региона в сравнении с российскими газетами. Излагается методика исследования и результаты ее применения на примере анализа несловарных прилагательных.

Ключевые слова: региональные СМИ, лексико-семантические особенности, регионально специфическая лексика, корпусное исследование.

Введение

Принято считать, что основным источником языковых инноваций, и в частности пополнения словарного состава, является повседневная разговорная речь как в устной, так и в письменной форме. Этим объясняется внимание лингвистов к материалу устной речи и к функционированию русского языка в интернете — к общению в социальных сетях, блогосфере и других видах электронной коммуникации. Не менее важным каналом, через который идет по-

[1] Работа выполнена при поддержке РГНФ (гранты № 13-24-01004 и № 15-04-12018) и РФФИ (грант № 15-06-04334).

полнение литературного языка прежде всего книжной лексикой, являются газеты и журналы. В газете отражаются явления в разных сферах жизни, интересные массовому читателю. Книжная лексика, связанная с освоением новых реалий, попадая на страницы газет, может рассматриваться как свидетельство того, что новое слово (даже будучи термином), уже вышло за пределы специального языка и стало частью общелитературного языка.

С другой стороны, разговорная стихия также проникает в газетный язык и через него получает статус кодифицированного слова. Через газетный текст осуществляется отбор разговорных, просторечных и жаргонных слов в общий фонд литературного языка — первоначально в качестве стилистически маркированных средств, которые в дальнейшем проживают в языке разную судьбу. Часть из них нейтрализуется, полностью усваивается литературным языком, другие сохраняют ту или иную разговорную или жаргонную окраску и используются в качестве средств выразительности, но значительная часть так и не получает в общем языке широкого распространения и остается в качестве слова-маркера, приметы эпохи.

Цель настоящего исследования состоит в поиске региональной специфики в употреблении новых слов. Новыми мы будем считать слова, которые вошли в литературный язык и закрепились в нем в последние 30 лет, с 1990-х годов, то есть с начала функционирования русского языка на территории отдельных государств в постсоветский период. Часть из них зафиксирована в толковых словарях, однако большинство из них в словарях еще не представлены. Поэтому основные свойства слова — структура значения, сочетаемостные свойства — выявляются прямо из контекстов употребления, для чего корпусные инструменты очень удобны: они обеспечивают конкорданс, распределение слова по датировке текстов, в которых оно встречается, метатекстовую информацию. Новые слова, в особенности слова с региональной спецификой, с большой вероятностью могут быть обнаружены среди так называемых несловарных словоформ, не имеющих соответствующих лемм в словаре морфологического анализатора. Формальным инструментом отбора новых слов служит имеющаяся у них помета bastard, которая присутствует в каждом гипотетическом разборе несловарной словоформы. Эта помета является чрезвычайно полезной: она используется как для коррекции корпуса (поскольку в число неопознанных слов попада-

ют слова с опечатками), так и для пополнения словника морфологического анализатора и создания словарных ресурсов. В частности, на основе анализа списка несловарных слов и приписывания им морфологических характеристик был составлен небольшой электронный Грамматический словарь новых слов [Гришина, Ляшевская 2008].

Имена прилагательные в региональном корпусе

Предметом исследования в данной статье стали словоформы-прилагательные, отсутствующие в словаре анализатора. Исследование проводилось на материале Корпуса региональной и зарубежной прессы, устройство которого описано в статье [Савчук 2015] в настоящем сборнике. Для наших целей было выделено четыре подкорпуса.

1) Гродненский корпус (ГК), общий объем 1,67 млн словоупотреблений;

2) Эталонный корпус российских региональных газет 2010-х годов (РР-2010-е), общий объем 1,9 млн словоупотреблений;

3) Региональные выпуски газеты «Комсомольская правда» (КП-регионы-2010-е), объем 5,4 млн словоупотреблений;

4) Российские региональные газеты 1990–2000-х годов (РР-1990–2000-е), общий объем 2,2 млн словоупотреблений.

В каждом из подкорпусов были получены списки словоформ прилагательных, отсутствующих в словаре анализатора, в составе сочетаний «прилагательное + существительное». Общее количество словоформ по разным подкорпусам приведено в Таблице 1. В дальнейшем эти списки были подвергнуты сплошному анализу, в результате которого в каждом из них было выделено несколько групп.

1) Прилагательные, связанные с именами собственными. Эта группа включает в себя прилагательные, которые входят в состав собственных имен — названий или являются оттопонимическими или отономастическими образованиями. Такая лексика обычно находится за пределами лингвистических словарей, поэтому она отсутствует в корпусном словаре анализатора.

2) Прилагательные, входящие в состав терминологических сочетаний или образованные от терминов-существительных. Вклю-

ченные в эту группу лексемы отсутствуют в словаре анализатора по причине «узкой специализации», однако представлены в текстах корпуса и в специальных словарях.

3) Новые слова. В эту группу включены слова, которые встречаются в других текстах НКРЯ не ранее 1990-х годов. Часть из них имеет бóльшую частоту в газетном корпусе, следовательно, можно относить их к газетной лексике, другая часть показывает большую частоту в основном корпусе и, следовательно, тяготеет к другим функциональным сферам — научной, официально-деловой, разговорной др.

4) Региональные слова — это новые слова, которые зафиксированы только в региональном корпусе и не встретились в газетном и основном корпусах. В данной статье они будут рассмотрены наиболее тщательно, чтобы выяснить, случайно или неслучайно выделение этой группы и какими факторами оно обусловлено.

5) Прочие. Сюда отнесены все слова, которые не попали в основные 4 группы: опечатки и ошибки, существительные, которые получили неправильный морфологический разбор (например, фамилии), регулярные образования-адъективы, сложные образования, которые обычно не включаются в словари (например, *четырехэтажный, четырехподъездный*, и пр.).

Соотношение несловарных слов всех перечисленных категорий в разных корпусах приведено в Таблице 1. Показано соотношение вхождений слова, то есть общего количества словоформ каждой группы, но не разных лексем.

Таблица 1

Корпус	**Собственные имена**	**Термины**	**Новые слова**	**Региональные слова**	**Прочие**	**Общее кол-во с/ф**
ГК	74%	7,5%	1,6%	1,9%	15%	1610
РР-2010-е	24%	10%	18%	4%	44%	442
КП-регионы 2010-е	37,8%	13%	13%	0,2%	36%	1309
РР-1990-2000-е	29%	11,2%	4,8%	6%	49%	1860

I. Прилагательные, связанные с именами собственными

Сразу бросается в глаза отличие гродненского подкорпуса от всех остальных по структуре и составу имен разных категорий «зоны расширения»: ¾ несловарных прилагательных составляют прилагательные, связанные с именами собственными, тогда как в остальных корпусах величина этой группы не превышает 40%. Одно из объяснений этого имеет, скорее всего, технический характер: в словаре морфологического анализатора отсутствуют собственные имена белорусских реалий, поэтому не опознаются как сами эти имена, так и все образованные от них прилагательные. В то же время значительная часть собственных имен российских реалий в словаре анализатора есть, и потому соответствующие прилагательные опознаются как словарные и в список не попадают.

Другое объяснение большой доли прилагательных, связанных с именами собственными, — высокая частотность этой группы лексики в гродненских газетах. 1193 словоформы (74%) представляют всего 137 лексем. Средняя абсолютная частота слов в этой группе — 8,6 вхождений на одну лексему, при этом отдельные оттопонимические лексемы очень частотны: *берестовицкий* — 286 вхождений, *свислочский* — 193, *ивьевский* — 124 вхождения. Слова этой группы встречаются в составе следующих наименований:

• названия районов и органов власти: *Берестовицкий район, райисполком, Дятловский район*;

• названия учебных заведений: *Берестовицкая ДЮСШ, училище, Массолянский ясли-сад-средняя школа*;

• названия предприятий и учреждений: *Берестовицкая птицефабрика, филиал, РОВД, быткомбинат, элитхоз, лесничество*;

• жителей: *берестовицкие спасатели, филателисты, связисты*;

• другое: *«Берестовицкая лыжня», «Зельвенский журавлик», «Ивьевская масленица», Занеманское предместье, Волковысская епархия, сцена*.

Для сравнения рассмотрим лексемный состав прилагательных, связанных с именами собственными, в подкорпусе КП-регионы-2010-е, в котором их доля среди российских подкорпусов наиболее высока (38%). 496 словоформ представляют 425 лексем, то есть средняя частота близка к 1. Прилагательные встречаются

в составе различных наименований, самые частотные из них *Яблоновый Овраг, Заельцовский район* (по 10 вхождений), *Которосльная набережная, мост* (7 вхождений), *Филейское кладбище* (6), *Щегловская засека* (5), *Нарымский квартал, остановка* (5 вхождений) и под.

Объяснение таким диспропорциям надо искать в свойствах корпусов, в различиях между ними. «Локальность» Гродненского корпуса создается а) небольшим количеством изданий (в нем всего 6 газет), б) принадлежностью всех изданий к одному региону и в) преобладанием в корпусе районных газет, ставящих в центр внимания местные новости. Все это обусловливает высокую частотность ограниченного набора собственных имен и их производных. Корпуса российских газет по всем этим параметрам «локальности» противопоставлены гродненскому корпусу, что и дает иные показатели для слов рассматриваемой категории[1].

Другая причина может иметь отношение к тематическим и жанровым особенностям текстов, входящих в состав разных корпусов. На наш взгляд, сравнительно высокий процент официальных документов, публикуемых на страницах районных газет гродненского корпуса, объясняет присутствие большого количества официальных наименований, включающих в свой состав прилагательные, что и обеспечивает их высокий частотный показатель.

II. Специальная лексика

Обратимся к другой категории несловарных слов, которые мы для краткости назвали терминами, а по сути эта группа относится к специальной лексике. Здесь мы также видим отличие в количественных показателях: в гродненском корпусе к разряду терминов отнесено около 7,5% словоформ, в российских корпусах — не менее 10%. Рассмотрим более подробно лексемный состав этой группы.

[1] Напомним также о «техническом» факторе уменьшения доли неопознанных слов, которое обеспечивается присутствием в словаре анализатора значительного количества собственных имен для российских реалий.

Гродненский корпус

Тематическая область	Количество с/у	Примеры
Сельское хозяйство	49	*зимнестойловый период, молочнотоварная ферма, комплекс, сахароперерабатывающий, сукозная козоматка*
Здоровье и медицина	39	*абдукционный, бескапсульные трахинеллы, ротавирусный, антиретровирусный, маммологический, предэпидемический*
Техника, производство	18	*когенерационная установка, плоскошовная машина, обрезная доска, предизолированные трубы*
Право, администрация и управление	4	*заявительный принцип, общенадзорная проверка, подучетные лица*
Прочее	10	*мотоманевренный, практикоориентированный, легкоконный*

Корпус-эталон российских СМИ 2010-х годов (РР-2010) и КП-регионы

Тематическая область	Количество с/у	Примеры
Здоровье и медицина, медтехника	105	*атопический дерматит, гомодиализные процедуры, гипоаллергенный режим, кардиотоксическое действие, срезовый томограф, транскраниальная магнитотерапия*
Право, криминал	30	*доследственная проверка, заявительное право, правоустанавливающие документы, общедолевая собственность*
Техника, производство, строительство	23	*пароконвекционный, газопоршневой, хромато-масс-спектрометры, безригельный каркас, гиперпрессованный кирпич, олеумный завод, непросадный грунт*
Сельское хозяйство	12	*органоминеральное удобрение, сидеральные травы, позднеспелый сорт, сукрольная крольчиха*

Тематическая область	Количество с/у	Примеры
Транспорт и коммунальное хозяйство	8	*асфальтобетонное покрытие, безбордюрная среда, аэротепловизионное обследование сетей*
Бизнес, финансы	7	*коллекторское агентство, скоринговая модель, клиентоориентированная линейка*
Прочее	42	*авианесущий крейсер, абровые ткани, аддитивное поведение, полнокупольное шоу*

Как следует из сравнения тематического состава специальной лексики, в гродненском корпусе лидирует тема сельского хозяйства и заметное место занимает лексика таких тематических областей, как медицина, техника и производство, право. В российских корпусах почти половина специальной лексики сконцентрирована в тематической области медицины (включая медицинскую технику и технологии), другие значимые темы — право и криминал, техника и производство, в том числе строительство, сельское хозяйство, транспорт и коммунальное хозяйство, бизнес и финансы.

Как представляется, одно из объяснений этих различий связано с количественным распределением текстов по тематике в гродненском и российском корпусах, которое показано на Рис. 1. Корреляция между тематическим составом текстов (см. статью Савчук в настоящем сборнике) и составом специальной лексики прослеживается отчетливо[1]. По количеству текстов сельскохозяйственной тематики гродненский корпус значительно превосходит российский, по остальным из перечисленных тематических областей (исключая медицину) преобладание у российского корпуса.

[1] Однако интересно различие по этому показателю между российскими корпусами. Корпус-эталон региональных российских СМИ содержит 44 термина, а СМИ 1990-2000-х – 214 терминов, то есть в 4 раза больше. Сопоставимо с последним и количество специальной лексики в КП-регионы – 173 словоформы.

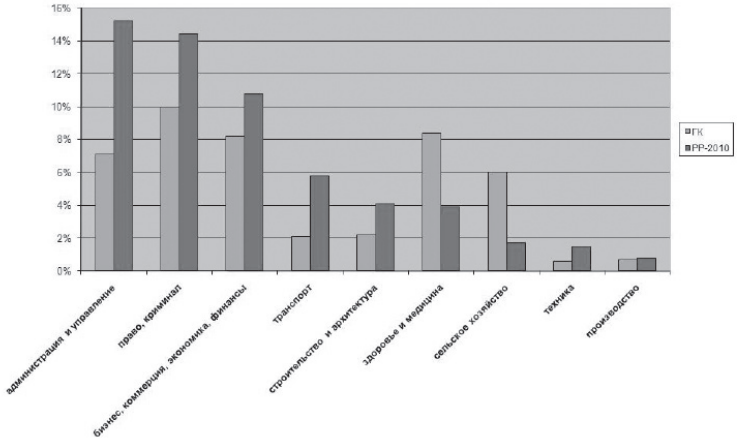


Рис. 1. Тематическое распределение текстов в Гродненском и российском корпусах

Что касается медицины, то здесь, по-видимому, такая прямая зависимость (чем больше текстов — тем больше терминов) не действует. В гродненском корпусе текстов медицинской тематики больше, но посвящены они в основном профилактике здорового образа жизни, здоровому питанию, охране материнства и детства, оценке экологической обстановки и под., ставят просветительские цели и рассчитаны на рядового читателя. Такие установки требуют от авторов публикаций достижения простоты и доступности изложения, сводя к минимуму употребление специальной терминологии. Рост количества медицинских терминов в газетном тексте происходит в основном за счет названий медицинских препаратов, процедур и техники, что мы и наблюдаем в российских газетах.

III. Новая лексика

К этой группе мы отнесли слова, которые не входят в Грамматический словарь (но могут быть зафиксированы в других словарях), и встречаются в текстах НКРЯ, датированных не ранее 1990-х годов. Интересно рассмотреть эту группу с разных точек зрения:

1) пересечение словников новых слов в гродненском и российском корпусах;

2) тематическое и стилистическое распределение новых слов;

3) абсолютная частота в НКРЯ — в основном и в газетном корпусах;

4) словарная фиксация.

III.1. Новая лексика, общая для всех подкорпусов.

При сравнении списков новых слов из трех корпусов — гродненского корпуса, корпуса региональных СМИ 2010-х годов и региональных выпусков КП — были выделены общая и непересекающаяся части. Пересекающаяся часть списков приведена в таблице 2.

Таблица 2.

Тематический класс	**Слово**	**Частота в НКРЯ (основной корпус)**	**Частота в НКРЯ (газетный корпус)**	**Фиксация в словарях[1]**
Экономика, финансы, бизнес	*зарплатный*	72	558	Орф
	среднеобластной	13	92	Орф
Техника и технологии	*интернетный*	19	18	Ефремова
Реклама	*имиджевый*	114	745	Орф, Ефремова
Правовая сфера	*контрафактный*	128	907	Орф, Ефремова
	патронатный	16	208	Орф
Городское хозяйство	*противогололедный*	8	189	Орф, БТС

[1] **Орф** – Русский орфографический словарь Российской академии наук. / Отв. ред. В. В. Лопатин. Электронная версия, «ГРАМОТА.РУ», 2001–2007; **БТС** – Большой толковый словарь русского языка/ Под ред. С. А. Кузнецова. СПб.: Норинт, 1998; **Ефремова** – Ефремова Т. Ф. Новый словарь русского языка. Толково-словообразовательный. – М.: Русский язык, 2000.

Тематический класс	**Слово**	**Частота в НКРЯ (основной корпус)**	**Частота в НКРЯ (газетный корпус)**	**Фиксация в словарях[1]**
Досуг	*досуговый*	77	279	Орф
	мультяшный	23	127	Орф
Спорт	*паралимпийский*[1]	6	564	Орф
	лыжероллерный	3	13	—
Дом и домашнее хозяйство	*бутилированный*	42	252	Орф

Новые слова, встретившиеся во всех трех подкорпусах, как видно из таблицы, принадлежат к разным тематическим областям и могут считаться уже освоенными. Они имеют достаточно высокую частоту и зафиксированы в словарях (за исключением прилагательного *лыжероллерный*, которое, вероятно, попало в зону всеобщего внимания журналистов в связи с возникновением и распространением нового вида спорта[2]).

Однако сочетаемость этих слов в гродненском и российских корпусах в отдельных случаях несколько отличается. Так, например, круг существительных, с которыми сочетается слово *зарплатный* в гродненском подкорпусе, ограничен словами *карта, счет, проект,* то есть употребление связано с описанием отношений в банковской сфере. В российских газетах этот круг шире: помимо перечисленных существительных[3] в круг сочетаемости входят *долги, налоги,*

[1] В Орф дается другой вариант этого слова – *параолимпийский*, не совпадающий с вариантом в официальном названии игр в русском языке, который представляет собой кальку с английского *Paralimpic Games.* Официальное написание оказалось предпочтительным в речевой практике: соотношение вариантов *паралимпийский / параолимпийский*, по данным газетного корпуса, – 564/85.

[2] Лыжероллерный спорт представляет собой скоростной бег на лыжероллерах. Зародился лыжероллерный спорт в Италии в 90-х годах.

[3] К которым добавляется еще *клиент* – 'участник зарплатного проекта', 'владелец зарплатной карты': *Итого максимальная скидка для* ***зарплатного клиента*** *с «хорошей» кредитной историей составит 4,75 процентных пункта от базовой ставки банка.* [Ирина Фролова. Хотите купить квартиру, дачу и автомобиль в придачу? Сбербанк поможет! (КП, 2011.03.02)].

часть, цифра, которые вводят прилагательное в сферу обозначения экономических отношений «работодатель — работник» и «работодатель — государство».

Что касается ***зарплатных долгов****, то здесь мы в округе на второй лидирующей позиции по долгам в промышленности и обрабатывающих производствах.* [Татьяна Румянцева. В Тверской области много увольняют да и платят неохотно (КП 2011.04.21)]. *На сайтах и бухгалтерских форумах, которые просматриваю по долгу службы изобилие информации, как уходить от* ***зарплатных налогов****, — поделился Полежаев.* [Татьяна Румянцева. В Тверской области бизнес снова ищет серые налоговые схемы (КП 2011.02.15)]. *Впрочем, не стоит обольщаться — привлекательную* ***зарплатную цифру*** *республике делает наша газовая и нефтяная отрасли, точнее, их высокооплачиваемые специалисты.* [Антон Забоев. Рассчитывая на высокие зарплаты, молодежь Коми идет в юристы и экономисты (КП 2011.04.29)]. *Сейчас поговаривают о сокращении персонала и уходе в тень* ***зарплатной части*** *многих компаний.* [Татьяна Белоусова, Семен Руденко. В новом году амурчанам придется вновь затянуть пояса (КП, 2011.01.10)].

В гродненском корпусе слово *контрафактный* сочетается с обозначением продукции, выпущенной незаконно: ***контрафактный*** *товар, продукция, сигареты, одежда.* В российском корпусе наряду с существительными того же ряда — *товар, продукция, алкоголь, DVD-диск, экземпляр* — встречается слово *фирма*, что свидетельствует о расширительном толковании значения прилагательного *контрафактный*: оно выступает уже не как синоним к слову *поддельный, пиратский*[1], а как синоним к слову *серый* (использующий "серые" схемы с целью уклонения от уплаты налогов). *Не связывайтесь с серыми страховщиками. — Иногда покупатели отказываются страховать авто в салоне, сами ищут какие-то* ***контрафактные фирмы*** *в надежде сэкономить, — рассказывает Олег Ширин. — Но экономии не получается — случись какая-нибудь царапина, и выясняется, что страховые полисы недействительны, фирма уже не существует, воз-*

[1] *Сотрудники кировской полиции по жалобам покупателей “накрыли” в городе магазин, торгующий под маркой известного бренда контрафактным, а проще говоря – пиратским, товаром*. [Полицейская облава на пиратский ширпортреб (Вятский край, 2012.12.12)].

никают проблемы. [Елена Калинина. Автомобиль — не роскошь! (КП, 2011.04.21)]

Таким образом, значительная часть новых прилагательных в сравниваемых корпусах совпадает: половина слов из гродненского списка и около 20% из корпусов российских газет.

III.2. Новая лексика в корпусе гродненских газет.

Теперь обратимся к несовпадающим частям списков. Прилагательные из гродненского корпуса относятся в основном к следующим тематическим группам.

Экономика, финансы, бизнес: *грантовский* (1/0)[1],

Досуг, зрелища и развлечения: *попкорновый* (0/10),

Мода: *макияжный* (4/1)

Политика: *контрдиверсионный* (1/1), *постъядерный* (1/4).

Общество: *белорусскоязычный* (2/9).

Ряд несловарных слов в гродненском корпусе относится к оценочной разговорной и жаргонной лексике: в корреспондентских публикациях встретилось *нажористый* (7/1), *8-гиговый* (0/3), в комментариях читателей *быдляцкий* (5/0).

Как видим, все эти слова, отсутствующие в корпусе российских региональных газет, имеют низкую частоту и в НКРЯ, однако говорить об их региональной гродненской специфике, вероятно, не приходится. Слово *нажористый* зафиксировано в Словаре московского арго В. С. Елистратова в значении 'сочный, яркий, необычный, колоритный'.

Можно отметить метонимический сдвиг в значении слова *попкорновый*, однако он идет в том же направлении, что и в примерах из НКРЯ. В текстах НКРЯ наряду с сочетанием, реализующим прямое значение (*попкорновый хруст*), гораздо шире представлено значение переносное: *попкорновый фильм, попкорновая публика, попкорновая роль*, так что сочетание *попкорновый сценарист* из гродненской газеты вписывается в этот ряд: *Забавно и вместе с тем тоскливо наблюдать нынешние трепыхания голливудских* ***попкорновых сценаристов****: «Фильмы про крутых дружных парней были?»* [Link. Значит, война (2012.02.10)].

А вот употребление слова *макияжный* вызывает некоторое со-

[1] В скобках для сравнения указана частота в основном и газетном корпусах НКРЯ.

мнение. Привычная сочетаемость этого прилагательного — *макияжный набор, карандаш, зеркальце*, в которой реализуется значение «предназначенный для нанесения макияжа или используемый при нанесении макияжа», синонимом может служить *гримерный*. В словаре Ефремовой оно определяется как ‘соотносящийся с сущ. макияж, связанный с ним’. В «Бераставіцкой газете» встречаем: *Заретушируют, где надо. И с обложки / Уже манит портрет прелестным лоском!* ***Макияжный глаз*** *влечет к киоскам, Чтобы дивам дивным подражая, Женщины косметику скупая, Все сражались со своей природой И гонялись за химерной модой* [Нина Судник. Дивы (Бераставіцкая газета, 2012.04.09)]. Здесь прилагательное обозначает «имеющий на себе макияж» и синонимично прилагательному *загримированный*.

III.3. Новая лексика в корпусе российских газет.

Список новых несловарных слов в российских газетах несколько больше, наиболее частотные из этих слов будут перечислены ниже. По тематическим областям они распределяются следующим образом[1].

Экономика, финансы, бизнес: ***аутсорсинговый*** (17/19), ***банкротный*** (2/12), ***бизнесовый*** (8/20), ***подушевой*** (0/20), *хозбытовой* (0/4), ***энергосбытовый*** (8/188), ***клининговый*** (0/58), *счетовой* 'связанный со счетами' (0/1), *единоразовый* (0/65).

Техника и технологии: *софтверный* (13/149), ***роуминговый*** (5/61), *контраварийный* (5/9), ***антивандальный*** (7/102). Последнее имеет отношение не только к технике, но и к общественной жизни (защите техники и изделий от людей), см.: *антивандальное покрытие, защита, материалы* и под.

Городское хозяйство: *внутридворовый* (8/58), *общедомовый* (2/260), а также менее частотные *внутриподъездный, прибордюрный, мусоровывозящий*. Тема коммунального городского хозяйства более актуальна для городских жителей, в то время как «дом и домашнее хозяйство» — для сельских. Возможно, активность этой группы связана с тем, что в российском подкорпусе преобладают го-

[1] Жирным шрифтом выделены не зафиксированные в гродненском корпусе слова, которые встретились как в эталонном корпусе, так и в региональных выпусках КП, светлым шрифтом – только в одном из двух корпусов. В скобках указана частота слова в основном и газетном корпусах НКРЯ.

родские и областные газеты, в то время как в гродненском — сельские районные газеты.

Политика: ***мэрский*** (65/430), *подковерный* (90/308), ***коррупциогенный*** (1/33), ***макрорегиональный*** (0/4).

Из других групп нужно отметить слова ***турагентский*** (10/15), *мангальный* (0/19), ***фанатский*** (25/1044), ***винтажный*** (15/95), *евроотремонтированный* (2/3).

Присутствие ряда прилагательных в российском списке и отсутствие их в гродненском может быть связано с тем, что они отражают новые предметы и явления, которые получили неодинаковое распространение в разных регионах. В других случаях отсутствие прилагательного связано не с новизной мотивирующего существительного, а с тем, что несловарному прилагательному предпочитаются более традиционные нормативные средства. В этих случаях можно говорить о значимом отсутствии. Приведем два примера.

Клининговый (от *клининг* — специализированная услуга по профессиональной уборке помещений и поддержанию чистоты) вошло в обиход потому, что услуги клининговых компаний в отдельных российских регионах получили массовое распространение, в том числе и в жилищно-коммунальном хозяйстве. В гродненских газетах вместо этого слова используются привычные *уборка территорий, уборщик* и под. Хорошей иллюстрацией к сказанному может быть сопоставление двух цитат из российской и гродненской газеты.

*9 февраля в пресс-центре «КП» собрались руководители жилищных управляющих компаний (их по старинке величают ЖЭКами или ЖЭУ). За «круглым столом» городские коммунальщики обменивались положительными примерами работы: разумеется, есть в их примере и такие. Кто технику для расчистки дворов закупает, кто — **меняет запойных уборщиц** на **клининговые компании**.* [Илья Архипов. «Владимирские крыши никогда так нет текли» (КП Владимир 2011.02.11)].

*Поступили вопросы в адрес РУП ЖКХ относительно наличия **службы по ландшафтному дизайну**, **уборки подъездов**, соблюдения параметров при подаче жильцам горячей воды. Директор РУП ЖКХ К.М. Борель пояснил, что в связи с отсутствием финансирования предприятие не имеет возможности дополнительно содержать **единицу по благоустройству придомовой территории**. Содержание объектов зеленого хозяйства производится самими до-*

мовладельцами. ***Уборка подъездов*** *многоквартирных жилых домов производится по графику: подметание — 2 раза в неделю, влажная уборка — 1 раз в неделю.* [Ядвига Курило. Ответ от первого лица (Свіслацкая газета 2012.12.03)].

Мэрский (*выборы, праймериз, кресло, руки, высказывание, дезинформация*). Слово получило широкое распространение в российских СМИ. *Персианов ответил за вольности возбуждением против него уголовного дела и в конечном счете лишением* ***мэрской*** *должности.* [Виктор Дятликович, Дмитрий Карцев. Охота на мэров // «Русский репортер», 2011]. *Отказаться от* ***мэрского*** *кресла ради научного бизнеса* [Дарья Золотухина, Виктор Сиднев. «Лучший капитан клуба» // «Русский репортер», 2011]. *В некоторых округах можно заметить даже откровенную конкуренцию между «****мэрскими*** *людьми».* [Александр Крутов. Ученики Макиавелли (2003) // «Богатей» (Саратов), 2003.05.29]. В белорусских газетах прилагательное, имеющее разговорный и потому слегка фамильярный оттенок, вообще не используется, а для передачи необходимого смысла используются генитивные конструкции. Ср. *мэрские люди* — и *…представил заместителей мэра*; *согласно мэрскому постановлению* — и *по заверениям мэра Минска, саму станцию «Октябрьскую» откроют…*

Что касается региональных выпусков КП, то список новых слов-прилагательных в них еще шире. Возможно, это объясняется тем, что и объем этого подкорпуса значительно больше (5,4 млн против 1,9 млн в российском и 1,6 млн в гродненском подкорпусах), хотя по количеству анализируемых несловарных словоформ он сопоставим с гродненским (1309 и 1610). Среди самых частотных отметим *коллекторский* (30/344), *инфокоммуникационный* (3/96), *трендовый* (17/24), *манипулятивный* (35/47), *новодельный* (5/11). Обращает на себя внимание то, что расширение словника прилагательных в региональных выпусках «Комсомольской правды» идет по двум направлениям: с одной стороны, отмечается прирост лексем (даже по сравнению с региональными российскими изданиями), имеющих книжную окраску и относящихся к социально-политической, экономической и производственно-технической сферам (*видеопроекционный, социоэкономический, среднегосударственный, субъектовый, эксклавный, антимафиозный*), с другой стороны, добавляется лексика разговорного характера, часто с выраженным

оценочным компонентом (*виповский, тусовый, бутиковый, маньячный, драйвовый, второлиговый*). Как кажется, здесь на уровне лексики проявляется специфика региональных выпусков федеральных СМИ как самостоятельного типа издания[1], которая состоит в том, что интерес читательской аудитории обеспечивается сочетанием новостей разного уровня — местных, региональных, «у соседей», федеральных, при этом местные проблемы часто рассматриваются в контексте общегосударственных. В стилистическом плане региональные выпуски ориентируются на общие принципы, принятые в центральном издании: для КП это использование приемов эмоционального воздействия на аудиторию, прежде всего молодежную, — выбор оценочных, разговорных, сниженных средств, модных слов, языковой игры и под.

IV. Региональная лексика.

К этой группе мы отнесли прилагательные, которые встретились только в текстах определенных региональных газет, но ни разу не зафиксированы ни в основном, ни в газетном корпусах НКРЯ и ни в одном из перечисленных выше словарей. Эта группа неоднородна по своему составу. Часть слов действительно имеет региональную специфику, часть слов, скорее, можно отнести к области авторского журналистского словотворчества, а часть уникальных образований придется квалифицировать как ошибочные употребления. Поскольку частотность этих слов в корпусе мала (в комментариях ниже количество вхождений приводится в скобках после лексемы), для подкрепления своих суждений мы будем использовать данные «большого Интернета». Начнем с гродненского корпуса.

IV.1. Гродненский подкорпус.

IV.1.1. На роль регионализмов могут претендовать следующие прилагательные.

Горпоселковый (2) ‘имеющий отношение к горпоселку’: *Второе место досталось Валерии Василинчик, медсестре поликлиники, и третье — Татьяне Лобач, медсестре Юратишковской* ***горпосел-***

[1] В региональном корпусе российских и зарубежных СМИ учтена возможность выбора региональных выпусков федеральных изданий в отдельный подкорпус для изолированного изучения.

ковой *больницы*. [Е. Дешко. Лучшая медсестра работает в хирургии («Іўеўскі край» 2012.05.25)]. Слово имеет ярко выраженную региональную специфику и используется на территории Беларуси. В Беларуси посёлки городского типа делятся на три категории — городские поселки, курортные поселки и рабочие поселки. От сочетания *городской поселок* способом аббревиации образовано существительное *горпоселок*, и производное отсубстантивное прилагательное используется даже в составе официальных названий: *Старобинская* ***горпоселковая*** *детская библиотека-филиал № 4*; *Уречский* ***горпоселковый*** *Дом Культуры*; *Городищенский* ***Горпоселковый*** *Исполнительный Комитет*; *Константиново — деревня в Свирском* ***горпоселковом*** *совете*.

Прикостельный (1) ‘находящийся при костеле’. *Сегодня ксендз-декан думает о том, как облагородить* ***прикостельную*** *площадку, увести дождевые воды, улучшить брусчатое покрытие и т. д.* [В. Поболь. Путь с Богом «Іўеўскі край», 2012.03.27]. Судя по контекстам, слово имеет широкое распространение на территории Беларуси, а также Западной Украины и Польши: ***Прикостельное*** *кладбище XVI–XIX век.* (Гродно); *Действительно парк в* ***прикостельной*** *зоне поражает своей ухоженностью.* (Брест); ***Прикостельное*** *кладбище продвигается в северном направлении к месту расположения бывшего дворца*. (Лида); *Во дворе у входа — скульптура Ефросиньи Полоцкой. От* ***прикостельной*** *площади радиально расходятся улицы с аккуратными домиками и ухоженными подворьями*. (Слободка); *В реестр памятников старины внесено 198 объектов, в том числе 35 костёлов, 5 часовен, 5 кладбищ (евангелистское, меннонитское и еврейское) и 13* ***прикостельных*** *кладбищ* (powiat.elblag.pl); *Епископ Львовский Марьян Бучек освятил новое распятие над входом в храм и новый памятник, на котором в полный рост над* ***прикостельной*** *площадью возвышается фигура Папы Римского.* (http://focus.ua/society/8487). Слово может использоваться и российскими авторами для обозначения местных реалий в текстах о Польше, Западной Украине, Беларуси: *В ноябре 1986-го в Варшаве костелы и* ***прикостельные*** *здания были оазисами, где можно было услышать поэтов, прозаиков, литературоведов.* (В. Британишский. Речь Посполитая поэтов, «Вопросы литературы» 2000, № 1).

Проводный в сочетании: *Проводный день* (3) ‘название религиозного праздника Радуницы’. *Вечером многие собирались за празд-*

*ничным столом и продолжали праздновать Пасху, точнее, устраивали ее проводы, так как одно из названий этого дня — **Проводный** день, или **Проводная** Радуница.* [БелТА. Радуница (2012.04.24)]. Слово довольно редкое, зафиксировано на территории Беларуси. *Кривая середа. Проводный день. Воловские проводы. Проводная радовница* [С. М. Толстая, Полесский народный календарь, 2005].

Наднеманский (1) ‘в верховьях реки Неман’. *Подол плавно спускается от вершины горы с готической башней бернардинского костела к самому краю **наднеманских** песков*. [Андрей Чернякевич, Андрей Вашкевич. «Красная книга Гродно»: Улица Подольная («Вечерний Гродно», 2012.04.11)]. Оттопонимическое прилагательное, широко распространено на территории Гродненщины. ***Наднеманский** край (Лида — Березовка — Новогрудок) Усадебный комплекс «Каралінскі фальварак Тызенгаўза» всегда рад новым гостям и посетителям! Восходящая звезда **наднеманского** туризма* (реклама). *Наднеманская Мадонна. Экскурсионная программа знакомит с жизненным и творческим путем Элизы Ожешко*. [http://mercana.by]. *Известно, что еще в 1530 году король Польши и Великий князь Литовский Сигизмунд подтвердил витебскому воеводе Ивану Сапеге право владения на приобретенные им дворцы над Неманом в Гродно. <...> На этом плаце вскоре был возведен новый каменный **наднеманский** дворец, который в перестроенном виде сохранился до нашего времени*. [Николай Шлык. Бывший Гродненский пивзавод станет культурно-развлекательным центром. «Гродзенская праўда», 03.04.14].

Почему мы не отнесли это слово к группе I (прилагательных, связанных с именами собственными)? Потому что в данном случае региональный характер имеет сама словообразовательная модель: название территории, расположенной в верхнем течении реки, и соответствующее оттопонимическое прилагательное образуется от основы гидронима с помощью приставки *над-* и суффикса *–j-* для существительного и *-ск-* для прилагательного. Эта модель используется для исторических названий регионов на западнославянских территориях (современных Беларуси, Украины, Молдовы, Польши) и не используется на территории России: *Наддвинский* край, *Наддднепровское* шоссе, «*Нагднестрянская* правда», *Надвислянский* этнографический парк (польск. Nadwiślański Park Etnograficzny), *Надсожье*, *Надприпятье* с сопредельными территориями (Средним

Надбужьем и др.). В то же время модели с приставками *по-* и *при-* являются универсальными: ср. *Побужье*, *Понеманье*, *Поднестровье*, *Поднепровье*, *Посожье* и *Поволжье*, *Подонье*, *Покамье*, *Поангарье*, *Поуралье* и др.

Памфлетовский (4) ‘относящийся к театру «Памфлет»’. *Мы можем только приоткрыть занавес, скрывающий от глаз этот* ***памфлетовский*** *путь длиною в два десятилетия*. [Наталья Турко. Театр «Памфлет»: приоткрывая занавес («Свіслацкая газета» 2012.04.03)]. Все 4 вхождения встретились в двух статьях одного автора. Это прилагательное можно отнести к категории образований от имен собственных, но оно имеет региональную привязку, поскольку театр «Памфлет» находится на Гродненщине, а других объектов с таким названием пока нет.

Полонистический (1) ‘относящийся к полонистике’. *Специальность «Современные иностранные языки (преподавание польского и английского языков)» продолжает традиции* ***полонистического*** *образования на Гродненщине.* [Служба новостей «ВГ». Вступительная кампания в ГрГУ имени Янки Купалы («Вечерний Гродно» 2012.05.17)].

Это прилагательное относим к регионализмам на основании частотного критерия: оно более употребительно на территории Беларуси и Литвы, чем на российских территориях. *«****Полонистический*** *центр учителей» в г. Люблине. Велись кружки по интересам:* ***полонистический****, исторический, драматический, краеведческий, спортивный с секцией трезвенников и некурящих* [http://sch1.kobrin.edu.by/main.aspx?guid=2121]. *Польская литература изучается в Литве также в Вильнюсском университете, где существует* ***Полонистический*** *центр. [А. И. Баранов (Вильнюс). Польская литература в Литве.* // Славянские языки и литературы в синхронии и диахронии Материалы международной научной конференции. М., МГУ 26–28 ноября 2013 г.].

Остальные прилагательные, первоначально отнесенные нами к группе регионализмов, при дальнейшем анализе были из нее исключены, поскольку обнаружились другие причины их отсутствия в НКРЯ. Они имеют невысокую частоту, но все-таки встречаются в публикациях на российских сайтах, следовательно, малоупотребительность этих слов объясняется другими факторами.

IV.1.2. Невысокая частота объясняется *новизной* слова.

Краснокрестный ‘относящийся к организации «Красный крест»’: *краснокрестный* актив, организация, идеи, мероприятия. *Светлана Николаевна рассказала, что на протяжении майских дней, когда по всей республике проходил месячник Красного Креста, волонтеры и активисты* ***краснокрестной*** *организации работали под девизом «Первая помощь спасает жизни»*. [Лариса Самарина. Под знаком добра и милосердия («Свіслацкая газета» 2012.06.18)]. *Институционализация красно-крестного движения на Урале. Оформление красно-крестной организации происходило под влиянием западных идей и собственных национальных традиций.* [http://www.moluch.ru/conf/hist/archive/117/5619]. *За два месяца через краснокрестный пункт помощи прошло более 3-х тысяч беженцев.* [http://www.rostov-redcross.ru/about/rostov/] Кемерово: Краснокрестный Кузбасс празднует 145-летие РКК.

IV.1.3. Невысокая частота может объясняться *окказиональностью* слова.

Окказиональные слова в основном используются в неформальной коммуникации на форумах, но могут употребляться наряду с официальными названиями в рекламно-информационных текстах.

Стеклоомывательный (1) ‘предназначенный для омывания лобового стекла автомобиля’. *Правильно! Продавать* ***стеклоомывательные*** *жидкости строго из этила! Нечего травить всякой гадостью любителей и профессионалов (принимать внутрь)*. [Вредные стеклоочистители снимут с продаж / Комментарии (Вечерний Гродно, 2012.02.29)]. Официальное название этой детали в системе очистки стекол автомобиля жидкостями — *бачок стеклоомывателя / омывателя ветрового стекла*. На российских сайтах встречаем разные варианты наименований: *Мужики, как снимается* ***стеклоомывательный*** *бачок? Пятна под машиной могут возникать из-за неисправностей в системе* ***омывателей*** *стекол. Тут всему виной негерметичный* ***стеклоомывательный*** *бачок.* ***Омывательный бак*** *выполнен из пластика, потому отремонтировать его можно, используя обычный паяльник.* ***Бачок стеклоомывателя*** *ветрового стекла для автомобилей Chevrolet Orlando (Шевроле Орландо), выпускаемых с 2009 года* и др. Со временем соотношение между но-

менклатурными, литературными и разговорными наименованиями предметов урегулируется, и слово может перейти в разряд специальной лексики.

IV.1.4. Невысокая частота объясняется *несоответствием слов нормам литературного языка.*

Некоторые новые слова до кодификации в словарях могут существовать в двух или нескольких вариантах, один из которых более употребителен и при включении в словари будет признан нормативным, а другой, менее употребительный, будет считаться лексической ошибкой.

Противокражевый (1) ‘предназначенный для защиты от краж’. (О новом магазине «Хозкультторары»): *Здесь современная система видеонаблюдения, вот-вот будет установлена и* ***противокражевая*** *система.* [Мария Драпеза. В Берестовице открылся новый магазин («Бераставіцкая газета» 2012.05.07)].

Слово употребляется в основном в рекламно-информационных материалах (часто очень низкого качества), при этом оба варианта — *противокражный* и *противокражевый* — могут встречаться в одном тексте: ***Противокражевые*** *системы. Выгод на самом деле множество. Самое основное это естественно защита от краж товара. Также магазины экономят на сотрудниках, ведь* ***противокражное*** *оборудование можно смело назвать некой своеобразной постоянно бодрствующей и находящейся на чеку охраной, которой не требуется платить зарплату и не требуется отдых и обеденный перерыв.* [http://slabo-tochka.ru/]. Нормативным следует считать прилагательное с суффиксом *-н-*, что подтверждается и соотношением вариантов *противокражевый / противокражный*: в российском интернете оно составляет примерно 1 к 670. В газетном корпусе отмечено 2 вхождения прилагательного *противокражный*.

Посиндромальный (1) ‘связанный с синдромом’. *В таких случаях достаточно оказать лишь* ***посиндромальную*** *терапию — сбить температуру обычными жаропонижающими препаратами, которые есть в каждой домашней аптечке*. [Ника Кожух. Прививка: за и против («Перспектива» 2012.04.27)]. По-видимому, это профессионализм, которому в нормативном употреблении соответствует термин *синдромальный* (*синдромальный подход, диагноз*).

Встречается также на российских медицинских сайтах и форумах: Нормативных актов, в которых бы оговаривался посиндромальный тонкий подход к такого рода ограничениям, в нашей стране нет [http://newpsyhelp.ru/schizophrenia.html].

Иммунальный (1) ‘относящийся к иммунитету’. *После ее окончания девушка осталась работать в институте младшим научным сотрудником отдела эпизоотологического и* ***иммунального*** *мониторинга*. [Наталья Черкес. В творческом поиске («Перспектива» 2012.01.23)]. Прилагательное употребляется наряду с нормативным *иммунный* как его синоним и в значении ‘укрепляющий иммунную систему’. Встречается на российских и зарубежных русскоязычных сайтах медицинской тематики. *Это* ***иммунальный*** *препарат, содержащий интерферон. Поддерживает иммунитет, помагает не заболеть или поскорее выздороветь.* [http://www.babyblog.ru]. *Вот такой я уникальный, Сладкий, вкусный,* ***иммунальный****, И полезный, и красивый Куст смородинки счастливый! Избавление от алкогольной зависимости. Избавление от аллергии и оздоровление* ***иммунальной*** *системы* [http://народныйлекарь.рф]. Зафиксировано даже в научном тексте: *Клинические признаки* ***иммунальных*** *нарушений и мембранные свойства мононуклеарных фагоцитов у больных рассеяным склерозом*. [АКД к.м.н. Казань: 2005][1]. Приведенные выше употребления не могут считаться нормативными.

Командно-штабный (1) *Игорь Гедич закончил* ***командно-штабный*** *факультет Военной академии РБ с отличием и золотой медалью.* [Служба новостей «ВГ». У мэра новый заместитель, у «Фармации» — начальник, а у театра — директор («Вечерний Гродно» 2012.03.28)]. Вариант, соответствующий современной норме, — *штабной*, вариант *штабный* считается устаревшим (зафиксирован в словаре Даля), однако встречается в специальных текстах: *Командно-****штабный*** *модуль. Предназначен для обеспечения в полевом районе необходимых условий служебной деятельности* [skc.ru]. *Командно-****штабный*** *колледж армии США* [http://enews.md/]. *Оборонно-****штабный*** *колледж им. Раковского София, Болгария* ***Штабный*** *состав лаборатории* [spmi.ru].

[1] Возможно, прилагательное образовано от существительного *иммунал* — названия известного препарата для укрепления иммунитета.

IV.2. Российский подкорпус.

Анализ группы условно региональной лексики в подкорпусе российских газет 2010-х годов не обнаружил в ней прилагательных с региональной спецификой. Слова из этого списка, отсутствующие в НКРЯ, встречаются в публикациях на российских и зарубежных русскоязычных сайтах, при этом зависимость от регионального фактора не просматривается. Группируются они следующим образом.

IV.2.1. Специальная лексика.

Пляжеобразующий (1) *Найден нефтешламм, они берут гальку, которая является* ***пляжеобразующим*** *материалом для береговой полосы Туапсинского района, складируют на свалку, сейчас как бы аккуратно это все пересыпают и говорят: «Ребят, у нас все чисто».* [Татьяна Гусельникова. Свалка законов-2 («НГ Кубани» 2012.11.26)].

Высокопотерьный (1) *Неслучайно в первую очередь взялись за так называемые* ***высокопотерьные*** *линии.* [Василий Дубовский. А не воруй! («Новгородские ведомости» 2012.11.20)]. Не признается литературной нормой, однако употребляется в специальных производственно-технических текстах и в деловой документации.

IV.2.2. Новые слова, которые в настоящее время не имеют широкого распространения.

Антивыборный (1) ‘направленный против выборов’. *Таким образом,* ***антивыборная*** *агитация с призывом портить бюллетени и голосовать против всех (несмотря на отсутствие такой строки в бюллетене) сыграла на пользу партии власти.* [Арнольд Ларионов. Займемся политикой? («Байкальские Вести» 2012.02.02)].

Предотставочный (1) ‘перед отставкой’. *На начальном этапе финальной атаки на Дмитрия Мезенцева этот вариант действительно обсуждался, но в итоге превратился в пропагандистскую торпеду для публичного расшатывания позиций иркутского губернатора и формирования «критической массы» для отставки Мезенцева и, что немаловажно,* ***предотставочной*** *атмосферы в коридорах областной власти*. [ИА Альтаир. Иркутские альтернативы (2012.12.02)].

Соцэкономический (1) ‘социально-экономический’. *Он пояснил, что можно снизить финансовые аппетиты области, но при этом возможен спад некоторых* ***соцэкономических*** *показателей.* [Смоленский бюджет страдает от долгов и дефицита («Рабочий путь» 2012.06.21)].

Взаимоинтересный (1) ‘представляющий взаимный интерес’. *В процессе* ***взаимоинтересного*** *разговора журналистами поднимались различные проблемы, и начальник УФСБ достаточно подробно и обстоятельно (насколько это было возможно) отвечал на все вопросы.* [Юрий Удоденко. Игорь Ахримеев: На то мы и спецслужба! («Байкальские Вести» 2012.01.12)]. Слово получило широкое распространение в бизнес-среде и сейчас стало чем-то вроде штампа в деловой публицистике: *Это взаимовыгодный,* ***взаимоинтересный*** *проект с индивидуальным подходом будет разработан нашим менеджером в тесном контакте с вашим представителем*. [salon-studio7.ru]. *Мы удовлетворены тем, что в текущей рыночной ситуации мы смогли найти с инвесторами* ***взаимоинтересный*** *уровень доходности ценных бумаг банка "ГЛОБЭКС", — прокомментировал Алексей Струнилин* [http://www.ibk.ru/]. *Разговор у нас состоялся очень живой, профессиональный,* ***взаимоинтересный***. *С заводом не расстаемся!* [Волга и Дон 20.07.2012].

Латочный (1) ‘произведенный с помощью латок, заплаток’. *Сейчас, спустя полгода, снова начались работы по асфальтированию данного участка автодороги, для восстановления дорожного покрытия была выбрана тактика «****латочного*** *ремонта»* [Дорога на улице Бородинской снова в аварийном состоянии (VL.ru (Владивосток) 2012.06.06)].

В газетном корпусе зафиксировано единственное употребление слова: *И обрушение штукатурки — это результаты неудачного* ***латочного*** *ремонта фасада*. [Виктория Анисимова. В Волгограде из-за строительства кафе в доме обрушилась стена // Комсомольская правда, 2009.09.23]. Более широкое распространение в газетных публикациях, в том числе и в гродненских газетах, нашло другое выражение — *ямочный ремонт* (частота в основном и в газетном корпусах — 5/142).

Помолвочный (1) ‘предназначенный для помолвки’. *Не забудьте о* ***помолвочном*** *кольце. Можно заранее купить его в ювелирном магазине, а можно и сделать своими руками из железной проволоки*

от шампанского, которое вам подадут в честь такого знаменательного события. Будет смешно и не так серьезно, но какая разница, ведь главное — помолвка состоялась! Где: «Роскошь», Русская, 19 а, тел. [Как сделать 8 Марта незабываемым//«Комсомольская правда», 2011.03.03]. Слово получило широкое распространение в рекламно-информационных текстах, предлагающих ювелирные украшения. *Заказать* ***помолвочное*** *кольцо с бриллиантом. Золотые кольца для помолвки, недорого. Второе значение* ***помолвочных*** *перстней менее прагматично. Украшение является подтверждением серьезных намерений мужчины по отношению к избраннице. Поэтому, только женщины носят* ***помолвочные*** *кольца. Каталог обручальных перстней включает модели для представителей обоих полов, а в подборке* ***помолвочных*** *изделий только девичий дизайн.* [http://tvoi-uvelirr.ru/pomolvochnye-kolca/].

IV.2.3. Новые значения.

Ведомостной (1) *Чуть меньше* ***ведомостной*** *стаж у главного редактора*. [Дмитрий Вертков. Всё впереди! («Новгородские ведомости» 2012.11.16)].

Шестикадровый (1) *Тем не менее главная героиня «Женщин» воплощает все ваши «****шестикадровые****» роли*. [Варвара Сиянова. Галина Данилова: «Все свои детские мечты я воплотила в «6 кадрах» («Амурская правда» 2012.10.02)]

IV.2.4. Авторские неологизмы.

Полуопилочный (1) ‘пополам с опилками’ *В шесть подъём и перекличка, мой номер — 7014, потом десятичасовой рабочий день на конвейере, питание — баланда на капусте и 170 граммов* ***полуопилочного*** *хлеба — известная история,— пожимает плечами Валентина Степановна.* [Валентина Васильева, Мария Клапатнюк. Огонь, вода и медные трубы («Новгородские ведомости» 2012.12.01)].

Пленумный (1) ‘относящийся к пленуму’ *Работа пленума была грамотно поставлена и тщательно срежиссирована Сталиным. <…> В эпилоге «****пленумной*** *драмы» главную роль сыграл режиссер-постановщик.* [Владимир Григорьев. И закрутился вал репрессий… («Байкальские Вести», 2012.03.12)].

Перезагорелый (1) *Еще я не люблю **перезагорелых** девушек. Мне самой противопоказан загар*. [Ольга Крыльцова. «Реальные пацаны»: Настоящая королева должна быть и умницей, и красавицей (КП 2011.04.29)].

Блоговый (2) *На прошлой неделе самыми популярными материалами на сайте novved.ru стали статьи «Банда угонщиков» Светланы Дубовицкой о преступной группе, угонявшей автомобили и «Блогу **блогово**» Марии Клапатнюк о встрече новгородских интернет-пользователей.* [Анна Огороднийчук. Калейдоскоп реплик («Новгородские ведомости» 2012.11.24)].

Если первые два примера являются образованиями ad hoc, типичными для разговорной речи и не отличающимися особой выразительностью, то последние два слова демонстрируют языковую игру.

IV.2.5. Варианты, не соответствующие нормам литературного языка.

Приютовский (1) вместо нормативного *приютский*: *«Собачий папа», Юрий Петрович Игнатенко, строил приют для бездомных четвероногих на собственные деньги, но погряз в долгах и заявил — мол, помогать ему никто не хочет, и от безысходности собирается усыпить 130 **приютовских** псов…* [Дарья Муратшина. В Ухте два спектакля собрали 125 тысяч рублей в пользу бездомных собак (КП 2011.04.29)].

Поствакционный (1) *В ходе прокурорской проверки выяснилось, что в некоторых регионах Российской Федерации в период прививочной компании в прошлом году уже регистрировались массовые случаи **поствакционных** реакций среди детей на вакцину против клещевого вирусного энцефалита «ЭнцеВир»*. [Елена Кондратьева. Врачи поставили эксперимент на детях: школьникам вкололи вакцину, не прошедшую клинические испытания (КП 2011.04.18)]. Существует несколько вариантов этого термина: *поствакцинальный, поствакцинный* (восходят к англ. postvaccinal), *поствакцинационный* (соответствуют англ. postvactination), *поствакциональный, послевакцинальный*. Все они имеют сходную сочетаемость: иммунитет, синдром, энцефалит, абсцесс, реакция, осложнения и пр.). Наиболее употребительным является *поствакцинальный*.

Разноотрубленный (1) ‘о спиленных ветвях деревьев’: … *я увидела вместо живых Платанов… белые скелеты с голыми черепами и воздетыми к небу* ***разноотрубленными*** *руками, совершенно голые обрубки — вот всё, что осталось*… [Обрезание по-туапсински (НГ Кубани 2012.12.17)]. Этот авторский неологизм, который используется с целью создания эмоционального образа, следует рассматривать как случай неудачного употребления.

Выводы

Метод сплошного анализа несловарных словоформ с помощью корпусных инструментов на материале прилагательных позволил выявить следующее. Метод успешно работает и помогает обнаружить лексико-семантические и стилистические особенности газетных текстов. С помощью корпусного инструментария удалось обнаружить как сходство в функционировании новых слов в газетах разных регионов, так и отличия между ними. Региональные отличия проявляются на нескольких уровнях.

• Регионально специфические единицы — прилагательные, образованные от существительных, называющих местные реалии. В гродненских газетах их оказалось больше, в российских газетах 2010-х годов не оказалось вообще. Однако это не значит, что их нет, они могут обнаружиться в других коллекциях[1].

[1] В подкорпусе российских газет 1990-2000-х годов прилагательное *байкалистский* может служить примером иркутского (или восточно-сибирского) регионализма: *На* ***Байкалистском*** *маршруте их обошла связка одной из сильнейших альпинисток России Светланы Шульгиной и Игоря Шапошникова.* [Валентин Брянский. В связке – надежные друзья // «Восточно-Сибирская правда» (Иркутск), 2003.06.07]. Другой пример – прилагательное *наукоградский* ‘имеющий отношение к наукограду’ (ср. рассмотренное выше *горпоселковый*). Поскольку подавляющее большинство муниципальных образований, имеющих статус наукоградов, исторически было сосредоточено в Московской области, прилагательное активно используется в местной подмосковной (а также калужской) прессе. *…План* ***наукоградских*** *мероприятий и бюджет утверждены задним числом, после того как уже израсходованы* ***наукоградские*** *субвенции*. [Людмила Пирогова. Столкновение мнений // «Встреча» (Дубна), 2003.06.11]. Ср. также из публикаций в интернете: ***наукоградский*** *проект, процесс, статус*; ***«наукоградские»*** *деньги*, ***наукоградские*** *старты* (Фрязино), ***наукоградское*** *движение* (Дзержинский), ***наукоградский*** *эксперимент*, *Первая* ***Наукоградская*** *Типография* (Обнинск), ***наукоградская*** *битва* (Жуковский), ***наукоградские*** *телодвижения Королёва* (Королёв). Центральными СМИ это прилагательное, судя по всему, пока не освоено: во всем газетном корпусе встретилось единственное прилагательное – *наукоградовский* [Известия, 2006.09.11].

• Региональная специфика проявляется и на уровне пополнения состава новыми словами. Пополнение идет двумя основными потоками — освоение специальной лексики и включение разговорных элементов. Белорусские газеты используют эти потоки менее активно, чем российские. Объясняется это несколькими причинами.

Во-первых, различия в интенсивности использования соответствующих наименований может быть обусловлено разным распространением соответствующего явления в России и Беларуси. Во-вторых, причина может заключаться в стилистических предпочтениях. Как кажется, для гродненских журналистов характерны более консервативные представления о нормах публицистического стиля — отсюда ограничения в использовании разговорных и оценочных средств. В-третьих, играет роль ориентация на читательскую аудиторию. У гродненских газет она по преимуществу сельская, и необходимость сохранять доступность и полезность публикаций накладывает ограничения на выбор тем и использование терминов.

• Региональная специфика отмечается в использовании имен собственных и прилагательных, образованных от них. Здесь между гродненскими и российскими газетами обнаружились резкие количественные различия, которые еще ждут своего изучения. Одно из объяснений этому — обилие официальных наименований в документах официально-делового стиля, процент которых в гродненских районных газетах значительно выше.

Помимо регионально-специфической лексики среди уникальных единиц, отмеченных в региональных корпусах и отсутствующих в НКРЯ, с помощью использованного метода удалось обнаружить единицы трех категорий: (1) новые слова, которые появились недавно и не зафиксированы в словарях; (2) примеры индивидуально-авторского творчества, случаи языковой игры; (3) речевые ошибки, связанные с неточностью словоупотребления и с нарушением лексической и фразеологической сочетаемости.

Новые слова не частотны, не получили массового распространения в СМИ, но используются в интернет-публикациях, в частности, в рекламных текстах и в электронной коммуникации. Станут ли эти новообразования массовыми, покажет время. Однако наряду с наблюдением, требуется внимательное их рассмотрение специалистами, квалифицированная оценка и рекомендации по использованию.

Словотворчество и языковая игра как средства повышения выразительности текста по определению не могут претендовать на массовое распространение, в противном случае они приобретают характер журналистского штампа или переходят в разряд стандартных языковых средств. Стилистический анализ средств выразительности с точки зрения точности и уместности позволит сформулировать рекомендации, способствующие повышению качества газетного текста.

В целом можно сказать, что на нашем материале удалось обнаружить различия между корпусами, которые касаются использования конкретных лексем, но не словообразовательных и семантических моделей при образовании новых слов и семантической деривации. Не удалось обнаружить какой-либо региональной специфики и в тех нарушениях нормативной сочетаемости, которые приводят к возникновению ошибок. Различия носят в основном количественный характер, что было показано в статье Г. И. Кустовой в настоящем сборнике, и, по-видимому, связаны не с особенностями языковой ситуации в разных регионах, а с уровнем языковой компетенции авторов.

В связи с этим можно говорить и о практической ценности исследований на основе регионального газетного корпуса. Полученные результаты опираются на изучение реальных текстов, выявляют их сильные и слабые стороны. На их основе должны быть разработаны рекомендации для журналистов и редакторов, тренировочные упражнения и задания для студентов, обучающихся этим профессиям, которые будут способствовать предотвращению ошибок и улучшению качества текстов.

Литература

Орф – *Русский орфографический словарь* Российской академии наук / Отв. ред. В. В. Лопатин. Электронная версия, «ГРАМОТА. РУ», 2001–2007

БТС – *Большой толковый словарь русского языка* / Под ред. С. А. Кузнецова. СПб.: Норинт, 1998

Ефремова – *Ефремова Т. Ф.* Новый словарь русского языка. Толково-словообразовательный. – М.: Русский язык, 2000

Гришина Е. А., Ляшевская О. Н. Грамматический словарь новых

слов русского языка. Электронный ресурс. http://dict.ruslang.ru/gram.php?

Кустова Г. И. Лексико-семантическая и социокультурная специфика русской речи Гродненщины (на материале текстов СМИ) // Труды Института русского языка им. В. В. Виноградова. Вып. 6. М., 2015

Савчук С. О. Корпус региональных газет России и зарубежья // Труды Института русского языка им. В. В. Виноградова. Вып. 6. М., 2015

S. O. Savchuk
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
savsvetlana@mail.ru

CORPUS AS A TOOL FOR STUDYING THE RUSSIAN LANGUAGE USAGE IN REGIONAL PRESS

The paper describes the results of corpus-based analyses of lexical-semantic and stylistic differences of functioning new words in Russian-language Belarusian newspapers in comparison with the Russian regional press. A sampling of adjectives not included in grammar dictionary of morphological analyzer and fixed in texts not earlier than the 1990-ies was retrieved with the help of corpus tools. Semantic, stylistic and frequency characteristics of these adjectives were compared in different sub-corpora of the RNC: Grodno region newspapers, Russian regional newspapers, main corpus and large newspaper corpus. A significant amount of new words are encountered both in Belarusian newspapers and in Russian regional press, but in Russian newspapers some of these words have a wider range of collocations indicating their more complex semantic structure. Lexemes that met only in Grodno subcorpus or only in the Russian regional subcorpus can be subdivided into several groups: (1) region-specific vocabulary relating to local realities, these words are represented in Grodno subcorpus more widely, (2) special terminology, and (3) conversational vocabulary, the latter two groups are represented to a greater extent in Russian media. In addition, there were distinguished

the following groups: (4) low-frequency new words that are not fixed in the RNC and in dictionaries, (5) authorial neologisms, (6) speech errors associated with impaired accuracy of usage and lexical compatibility.

The corpus-based analyses revealed some regional differences in functioning of specific lexical items or specific groups, and similarity in using models for word formation and semantic derivation.

Key words: regional media, lexico-semantic features, region-specific vocabulary, corpus-based study

References

Russkii orfograficheskii slovar' Rossyskoi akademii nauk [Russian Orthographic Dictionary]. V. V. Lopatin (ed.). Electronic resource, available at http://gramota.ru/slovari

Bol'shoi tolkovyi slovar' russkogo yazyka [The big explanatory dictionary of the Russian language]. S. A. Kuznetsov (ed.). St. Petersburg, Norint Publ., 1998.

Grishina E. A., Lyashevskaya O. N. *Grammaticheskii slovar' novykh slov russkogo yazyka* [Grammatical dictionary of new words of the Russian language]. Electronic resource. Available at http://dict.ruslang.ru/gram.php?

Efremova T. F. *Novyi slovar' russkogo yazyka. Tolkovo-slovoobrazovatel'nyi* [New dictionary of the Russian language. Explanatory-derivational]. Moscow, Russkii yazyk Publ., 2000

Kustova G. I. [Lexical-semantic and socio-cultural specificity of Russian speech of Grodno (on the material of media texts)]. *Trudy Instituta russkogo yazyka im. V. V. Vinogradova* [Proceedings of V. V. Vinogradov Russian language Institute], no. 6. Moscow, 2015. (In Russ.)

Savchuk S. O. [The corpus of regional russian-language newspapers in Russia and abroad]. *Trudy Instituta russkogo yazyka im. V. V. Vinogradova* [Proceedings of V. V. Vinogradov Russian language Institute], no. 6. Moscow, 2015. (In Russ.)

[1,2]Г.И. Кустова
[1]Институт русского языка им. В.В. Виноградова РАН
[2]Национальный исследовательский
Томский государственный университет
(Россия, Москва)
galinak03@gmail.com

ЛЕКСИКО-СЕМАНТИЧЕСКАЯ И СОЦИОКУЛЬТУРНАЯ СПЕЦИФИКА РУССКОЙ РЕЧИ ГРОДНЕНЩИНЫ (НА МАТЕРИАЛЕ ТЕКСТОВ СМИ)[1]

Описаны языковые явления, отражающие социокультурную ситуацию Гродненщины: экспансия новой лексики, происходившая в постсоветскую эпоху, с одной стороны, и сохранение советизмов и оборотов официально-делового стиля советского периода, с другой стороны.

Проанализировано использование в газетных текстах сленговой лексики, отражающей установку журналистов на «приближение» к читателю. Исследованы неточные и нестандартные словоупотребления в лексике и нарушения сочетаемости в рамках коллокаций.

Ключевые слова: социокультурная ситуация, язык СМИ, сленговая лексика, нестандартные употребления лексики

В настоящей работе представлены результаты анализа языковых и социокультурных особенностей белорусских СМИ, главным образом — СМИ Гродненщины, которые рассматривались нами как важный источник сведений о функционировании русского языка

[1] Работа выполнена при поддержке РГНФ, проект № 13-24-01004 «Лексико-семантическая и социокультурная специфика русской речи Гродненщины (на материале текстов СМИ)». Примеры извлечены из НКРЯ и корпуса региональной и зарубежной прессы НКРЯ.

в республике Беларусь. В качестве материала исследования использовался специально созданный в рамках проекта совместно с гродненскими лингвистами электронный корпус региональной прессы, далее — региональный корпус. В него вошли белорусские газеты — далее мы будем называть эту часть корпуса белорусским подкорпусом, — и российские газеты, далее мы будем эту часть корпуса называть российским подкорпусом. Кроме того, в некоторых случаях мы будем обращаться к материалам основного и газетного корпусов НКРЯ, далее — основной корпус НКРЯ и газетный корпус НКРЯ соответственно.

Подробные сведения о конкретном составе регионального корпуса и количественные характеристики его составляющих имеются в публикациях [Савчук 2015 а] и [Савчук 2015 б] в наст. сборнике. Для нашей темы существенно отметить, что корпус белорусских газет неоднороден по составу. В него вошли (1) районные газеты, издающиеся в Гродненской области: «Островецкая правда», «Берестовицкая газета», «Ивьевский край», «Праца» и «Свислочская газета», (2) городские газеты г. Гродно «Вечерний Гродно» и «Перспектива» и (3) региональное издание «Комсомольской правды» — «Комсомольская правда» в Беларуси (издается в Минске). В такой разнородности есть свои преимущества: можно проводить не только сопоставительный анализ белорусского и русского подкорпуса в целом, но и исследовать различия внутри белорусского подкорпуса. Далее мы будем различать внутри белорусского подкорпуса Гродненский корпус (ГК) и «Комсомольскую правду» в Беларуси (КПБ), поскольку их языковые характеристики по некоторым параметрам не совпадают.

Региональная пресса представляет интерес по крайней мере в двух отношениях. Во-первых, она лучше отражает языковую и социокультурную ситуацию конкретного региона, во-вторых, она в меньшей степени находится под воздействием процессов, связанных с унификацией и стандартизацией языка (по сравнению с центральными СМИ), равно как и под воздействием самих центральных СМИ как источника литературной нормы, а с другой стороны, испытывает влияние местных языковых особенностей.

Русскоязычные СМИ республики Беларусь представляют с этой точки зрения специальный интерес и являются ценным источником для изучения языковых и социокультурных процессов в силу осо-

бой языковой ситуации, сложившейся в Беларуси. В Беларуси существуют русскоязычные, белорусскоязычные и «смешанные» газеты (а также радио- и телеканалы), в школах и вузах преподавание ведется как на русском, так и на белорусском языке. Жители Беларуси, в том числе авторы публикаций, находятся, таким образом, в условиях взаимодействия и взаимовлияния двух языков.

Кроме того, в Беларуси постсоветского периода не было столь радикального разрыва с советским прошлым и в какой-то мере сохраняется ориентация на советские ценности, что отражает, в частности, язык СМИ, особенно региональных.

Дальнейшее изложение будет состоять из двух частей.

Сначала рассмотрим,— в основном, на материале ГК, но с привлечением материалов КПБ и российского подкорпуса,— как отражается в языке СМИ социокультурная ситуация, как процессы, происходящие в обществе, изменения в политической жизни, экономике и культуре воздействуют на производимые этим обществом тексты — в нашем случае на тексты СМИ. Затем — на материале ГК — рассмотрим более конкретные языковые явления в сфере лексики и фразеологии.

Отражение социокультурной ситуации в языке СМИ Гродненщины

Одной из ярких особенностей белорусской социокультурной ситуации (а в какой-то мере и российской, особенно если судить по региональной прессе) является то, что в ней соседствуют две разнонаправленные тенденции. Одна тенденция — «инновационная»: это социальные и культурные новшества, связанные с распадом СССР: изменение политического устройства, развитие рыночных отношений, появление новых профессий, влияние западной культуры, распространение новых технологий, прежде всего — интернета; на уровне языка эти процессы отражаются в экспансии иноязычной (прежде всего — англоязычной, но не только) лексики и фразеологии. Другая тенденция — «инерционная»: попытка законсервировать (или реанимировать) ценности и идеалы советского прошлого, а также (что особенно важно для нашей темы) сохранить сам язык советской эпохи (о соответствующих процессах в российской прессе см., в частности, [Ермоленкина 2011]; [Вепрева 2005]; [Кузьмина

2007]; [Кузьмина, Абросимова 2013]; [Мокиенко, Никитина 1998]; [Ностальгия 2011]; [Резанова 2009]; [Советское прошлое 2009]; [Стернин 2004]). Обе тенденции хорошо прослеживаются именно в региональных СМИ — как российских, так и белорусских (в центральных российских СМИ в первые постсоветские годы господствовала установка на полный разрыв с прежней эпохой, сознательный отказ от старых идеологем и того специфического языка («советского новояза»), на котором эти идеологемы транслировались; впрочем, в последнее время консервативные тенденции коснулись и центральных СМИ).

Что касается «инновационной» тенденции, гродненские СМИ вполне освоили не только ту лексику, которая активно входила в русский язык в первые постсоветские (т. е. в 1990-е) годы: *байкер, бизнес-план, блокбастер, бренд, видеоролик, дилер, имидж, инвестор, инвест-проект, коррупция, креатив, менеджер, ньюсмейкер, онлайн-* (*-сервис, -конференция, -консультирование*), *от кутюр, офис, пиар, предоплата, презентация, приватизация, провайдер, рейтинг, слоган, тренд, уик-энд, фитнес, флаер, холдинг, эскапизм* и т. п., — но и лексику, которая стала осваиваться русским языком сравнительно недавно: *гаджет, фаер-шоу, флешмоб, франшиза* (в НКРЯ эти слова зафиксированы уже после 2000 г.).

Некоторые слова, зафиксированные в российском подкорпусе, в ГК пока еще не встретились: *аутсорсинг, блендер, бэдж (бэйдж), демпинг, кейтеринг, легитимность, маржа, мерчендайзер, оффшор, постер, промоутер, промоушен, ребрендинг, рестайлинг, саспенс, секьюрити, селебрити, спойлер, тостер, фрилансер, хеппиэнд, шейкер* (при этом *аутсорсинг, блендер, бэдж (бэйдж), постер, рестайлинг, саспенс, секьюрити, селебрити, фрилансер* есть в КПБ). Однако надо иметь в виду, что ГК включает тексты только до 2012 г.

Слово *ваучер*, столь значимое в российской жизни 1990-х годов, в белорусских СМИ употребляется в связи с туризмом, а главное — не имеет тех идеологических коннотаций, которые у российских граждан старших поколений связаны с процессом приватизации.

Интересно, что в белорусском подкорпусе встречаются новообразования (*инвест-уик-энд, бизнес-идея, агроэкотуризм*), которые в российском региональном подкорпусе и в основном корпусе НКРЯ не встречаются или встречаются крайне редко: *инвест-уик-*

энд не встречается ни в НКРЯ, ни в российском региональном подкорпусе; *бизнес-идея* во всем 400-миллионном (если включать газетный корпус) НКРЯ встречается 6 раз.

В белорусском подкорпусе широко представлено слово *агроэкотуризм*, которое отражает явление, специфичное именно для Беларуси. Ни в основном корпусе НКРЯ, ни в российском подкорпусе регионального корпуса это слово не встречается. Есть лишь одно употребление в газетном корпусе НКРЯ, но не в газетной публикации, а в новостном сообщении информагентства «Новый регион — Екатеринбург».

Наряду с появлением новых слов, многие «старые» слова (и это важно именно в социокультурном аспекте) получают, так сказать, второе рождение и нередко «меняют знак»: слова типа *бизнес, биржевые спекулянты, коммерсант, карьера, амбициозный* утратили свои отрицательные коннотации и стали вполне нейтральными.

Интересно проследить по материалам НКРЯ судьбу слова *шоу*. В первые десятилетия советского периода оно встречается только как фамилия *Шоу*. С 60-х годов оно употребляется как обозначение явления западной жизни (*бродвейское шоу*), иногда с отрицательной оценкой (*пошловатое шоу*). Сейчас оно употребляется как нейтральное название соответствующего жанра (и в российских, и в белорусских СМИ), ср. *Цыганское шоу «Аллюр» своим ярким представлением поднимет настроение трудовому коллективу СПК «Свислочь»* [«Перспектива», Гродненская область, 2012.11.08] (разумеется, *шоу*, как и ряд других «зрелищных» существительных — *балаган, оперетка, цирк, кино*,— может иметь и не-нейтральное, оценочное значение, но в белорусском подкорпусе оно не встретилось). Интересно, что это слово может восприниматься как новое, появившееся в 90-е годы, хотя, как мы видели, встречалось в текстах более раннего периода, ср.: *Забугорное словечко «шоу» вошло в наш быт едва ли не одномоментно с такими ругательными, как ваучер, бизнес, дефолт и пресловутый Чубайс*. [Валерий Туркин. Много шоу из ничего («Красное знамя» (республика Коми), 2013.05.13)],— т. е. до 1990-х гг. в качестве обозначения явления западной жизни оно не воспринималось автором публикации как слово русского языка (пусть и заимствованное).

Другой важной чертой многих публикаций ГК,— отражающей «инерционную» (консервативную) тенденцию,— является исполь-

зование слов и оборотов, характерных для языка советского периода,— для статьей и книг общественно-политической тематики, официальных речей советских руководителей, партийных документов и т. д. (исследователи языковых изменений постсоветского периода называют их советизмами (лексическими, семантическими, коннотативными), ср. [Кузьмина, Абросимова 2013: 63–67], см. также [Вепрева 2002], [Ермоленкина 2011]; [Кузьмина 2007], [Мокиенко, Никитина 1998], [Ностальгия 2011]; [Резанова 2009]; [Советское прошлое 2009]): *работа по патриотическому и духовно-нравственному воспитанию молодых граждан; идет активная культурно-просветительская и спортивно-массовая работа с молодежью; патриотическое воспитание молодежи; пропагандировать здоровый образ жизни; стремление большинства граждан Беларуси принимать деятельное участие в общественно-политической жизни страны, в реализации курса ее развития; трудовые успехи; почетные гости дали высокую оценку проведенной работе по созданию условий для качественного медицинского обслуживания населения; наши цели и задачи остаются неизменными — быть опорой государства, консолидирующей силой общества в интересах народа и процветания страны; сегодня наша основная цель — совершенствовать формы и методы работы.*

Причиной использования «советской фразеологии» не обязательно является сознательная идеологическая установка автора. Чаще всего такие языковые предпочтения объясняются, по-видимому, именно инерционным сохранением старого газетного языка.

Встречаются также обороты из канцелярского и — шире — официально-делового языка: *документ принят в целях совершенствования системы государственной адресной социальной помощи и усиления государственной поддержки населения; вовлечь в спортивно-массовые и оздоровительные мероприятия как можно больше учащихся; банное обслуживание населения; инвалиды имеют право на бесплатное оздоровление раз в два года; стандарт по количеству помывочных мест на 1 тыс. населения; увеличился охват учащихся занятиями по интересам; организован подвоз в школы 702 детей; наращивать производство продукции.*

Наряду с «советской» и официально-деловой лексикой выделяется такая группа оборотов, которую можно назвать газетными штампами.

Под газетными штампами мы понимаем выражения, которые призваны заменить нейтральную номинацию на образную или просто более выразительную: *люди в белых халатах* (врачи); *работники прилавка, черное золото* (нефть); *засыпать* [столько-то тонн зерна] *в закрома Родины* (собрать урожай); *вести с полей; трудовой подвиг, трудовая вахта, ударный труд, труженики села, слуги народа* и т. п. Первоначально такие выражения использовались для «оживления» текста и усиления воздействия на аудиторию, однако в результате многочисленных повторений превратились именно в навязчивые штампы (в некоторых случаях чуть ли не обязательные)[1].

В первые годы после распада СССР такие выражения широко использовались в прессе в ироническом ключе как знак ушедшей эпохи. Что касается их неиронического употребления, то в центральной прессе и электронных СМИ они избегались, однако в региональной прессе продолжали бытовать, поскольку контингент журналистов там обновился в меньшей степени, и авторы пользовались газетными штампами как привычной частью своего профессионального языка.

Что касается районных газет, которые составляют бóльшую часть белорусского подкорпуса), то есть еще одна причина использования подобных штампов. Авторами статей в районных газетах нередко являются не профессиональные журналисты, а сами участники описываемых событий: заметки о текущих хозяйственных делах или мероприятиях с участием местных руководителей пишут представители администрации, заметки о школьных конкурсах и других событиях школьной жизни — учителя и т. д. В отличие от профессиональных журналистов, у таких авторов, по-видимому, нет установки избегать употребления газетных штампов,— скорее наоборот, они считают, что это необходимый атрибут языка газеты, и хотят подчеркнуть свое владение «газетным языком», принадлежность к сфере СМИ (надо заметить, что многие подобные обороты, не связанные напрямую с реалиями советской эпохи, продолжают употребляться и в российской прессе, в том числе в центральных российских газетах). В Гродненском корпусе встретились такие газетные штампы, как *подхватить начинание, ждать с особым вол-*

[1] Некоторые из этих выражений отражены в словаре языка газеты, радио и телевидения, см. [Солганик 2002].

нением, стоять на пороге <какого-л. события>, *творческая задумка*, *найти живой отклик*:

Начинания *гродненцев* ***подхватили*** *и другие регионы нашей страны;* ***С особым волнением*** *ждали визита ученики, изучающие этот язык на школьных факультативах; Сегодня РОО «Белая Русь»* ***стоит на пороге значительного события****: грядут перемены в ее статусе; Многие* ***творческие задумки*** *позаимствовали у учреждений образования и другие участники конкурса, что только положительно отразилось на общем облике города; В течение двух дней в библиотеке проходили разнообразные культурные мероприятия, которые нашли* ***живой отклик*** *у всех принявших в них участие.*

Рассмотрим один из типичных газетных штампов — выражение *найти живой отклик*. По материалам НКРЯ можно проследить его историю. В XIX в. выражение встречается в художественной литературе в более узком значении эмоциональной реакции — в вариантах *найти отклик в душе / в сердце*. В первой половине XX в. оно представлено, в основном, в мемуарных текстах, часто — непрофессиональных писателей. Наконец, во 2-й половине XX в. оборот практически полностью перекочевывает в газеты и у профессиональных литераторов почти не встречается, т. к. отчетливо ощущается его принадлежность к набору газетных штампов. В постсоветской прессе этот оборот также встречается, причем в газетах либерального направления — с ироническим подтекстом, как характеристика консервативного электората, приверженного старым ценностям.

Вообще, выражения *подхватить начинание, стоять на пороге* <какого-л. события>, *творческая задумка* и под. встречаются, конечно, не только в прессе, но и в художественных текстах. При этом в последних такая «газетная фразеология» обычно выполняет особую изобразительную функцию — служит знаком определенной эпохи с ее идеологическими установками, ценностями, общественной атмосферой.

Еще одна яркая черта языка современных СМИ — обилие разговорной и сленговой лексики, юмористических выражений, оборотов, содержащих игру слов (см., например, [Кузьмина, Абросимова 2013]). Эта черта связана не с отражением политико-экономических или социокультурных реалий, проявившихся в тех двух тенденциях, о которых сказано выше, а с изменением установок самих журнали-

стов, в частности — с установкой на сокращение дистанции между журналистом и читателем. Такие средства, как сленг, юмор, игра слов, призваны создать атмосферу непринужденного диалога с читателем, дружеского, неофициального общения, что само по себе отличает современные газеты от советской прессы, в которой был принят «строгий тон». В этом же ряду стоит использование общеизвестных прецедентных цитат из фильмов, песен, анекдотов, юмористических скетчей (особенно из выступлений А. Райкина, текстов М. Жванецкого), речей политических деятелей и т. п.

Знание прецедентных цитат объединяет жителей постсоветского пространства, особенно старших поколений, что активно используется в центральной прессе и в разговорных передачах электронных СМИ. Однако в региональной прессе, вопреки ожиданию, обилия прецедентных выражений не обнаруживается: например, «Таможня дает добро» (из фильма «Белое солнце пустыни») встретилось только в российском подкорпусе, «Вор должен сидеть в тюрьме» (из фильма «Место встречи изменить нельзя») — 2 вхождения в российском подкорпусе и 1 — в белорусском. Кроме того, и в русском (16 вхождений), и в белорусском (3 вхождения) подкорпусах встретилось известное выражение В. С. Черномырдина «Хотели — как лучше, а получилось — как всегда» (иногда — с вариациями, ср.: *Хотел как лучше, а получилось не очень…* [Вечерний Гродно, 2012.01.18]; *Только вот сам закон оказался не на высоте: хотели, как лучше, а получилось, как вчера* [Богатей (Саратов), 2003.11.06]; *И все это нужно бы как лучше, а не как по Черномырдину* [Красноярский рабочий, 2003]).

Что касается жаргонной и сленговой лексики и фразеологии, то она представлена в обоих подкорпусах. Сленговая лексика, считавшаяся в советский период недопустимой в языке печатных изданий, сейчас не просто получила широкое распространение в языке СМИ (а также в языке рекламы), но является одним из главных средств воздействия на читателя — в частности, благодаря установке журналистов (как печатных, так и электронных СМИ) говорить «народным» языком.

По существующим сейчас многочисленным словарям жаргонной и сленговой лексики (см., в частности, [Никитина 2009], [Курилова 2007]) мы сделали выборку слов, которые активно употребляются не только в СМИ, но и в современной кино- и телепродукции,

например, в сериалах (разумеется, из этой выборки была исключена грубая и ненормативная лексика). В итоговый список вошли 410 лексем. Под лексемой мы понимаем слово в определенном значении. Например, в выборку вошли 3 лексемы *слить* в разных значениях: *слить кого* (ср. *Сердюкова слили*); *слить* (*в прессу*); *слить* (*игру* — о договорных матчах). Были получены данные о вхождении этих лексем в белорусский подкорпус и в российский подкорпус. Число вхождений каждого слова было также установлено, однако как статистические данные эти цифры ценности не имеют: корпус региональной прессы пока еще слишком невелик по объему, чтобы по его материалам можно было делать количественные оценки. Для нас было принципиально установить, встречается ли слово в обоих подкорпусах, в одном из них, или ни в одном. Что касается количества вхождений, то оно интересно скорее не для сравнения подкорпусов, а для сравнительного анализа внутри каждого подкорпуса, т. к. показывает, какие из сленговых слов и выражений встречаются чаще, а какие — реже. Это в какой-то мере характеризует пристрастия журналистов, которые, возможно, отражают пристрастия всего языкового коллектива. Сленговая лексика в составе названий (например, *пацаны* в названии сериала «Реальные пацаны») в расчет не принималась.

Результаты можно разбить на 5 групп:

(1) 126 лексем не встретились ни в одном из подкорпусов;

(2) 7 лексем встретились только в белорусском подкорпусе;

(3) 159 лексем встретились только в российском подкорпусе;

118 лексем встретились в обоих подкорпусах, причем:

(4) 80 лексем встретились в российских газетах и в Гродненском корпусе (ГК). Причем часть из них встречается также в текстах «Комсомольской правды» в Беларуси, но нам было важно, что эти лексемы встречаются в нестоличных, в частности — районных, газетах;

(5) 38 лексем встретились в российских газетах и в текстах «Комсомольской правды» в Беларуси (КПБ), но не в ГК.

(1) Лексема не встретилась ни в одном из подкорпусов — 126:

башлять	заколебать	ментура	писюк
борзеть	запрессовать	мочалка	погоняло
баклан	засада	мулька	погремуха
бортануть	засандалить	мутить	поддавать
бухло	застрематься	надраться	полташ
бэха,беха	затихариться	наколбасить	полтос
винтить, свин-	канать	наколоть	полтосик
тить	катить (не катит)	нервяк	по-любому
вломить	кекс	ништяк	помахаться
вмазаться (нарк.)	колбасить	нычка	понтярщик
врубаться ('по-	копыта	обдолбанный	пруха
нимать')	крайняк	обнести	порожняк
втирать	крендель	обратка 1	прессануть
гаситься	кувыркаться	опупеть	приколоться
голяк	лавэ	отлуп	простава
дербанить	лажануться	отоварить	прохилять
дернуть 'выпить'	лажа	отпад	пруха
джинса	лажовый	отслюнить	путяга
днюха	лепила	оттоптаться	расколбасить
дуболом	линять	оттопыриться	расслабон
жмур	слинять	отыметь	салабон
жрачка	маза	опупеть	свинтить
забашлять	малява	палиться	свист
загаситься	махалово	педик	слинять
заиграть что	махач	перепихнуться	слиться
стебаться	терки	пестик	солидняк
сторчаться	тихариться	фуфел	ширево
стрематься	торчок	фуфло	шкериться
сухач	трындеть	хилять	шмалять
табло (по табло)	тухляк	чел	шмурдяк
тачила	угорать	чмошник	шняга
	усечь	чморить	шобла
		чувиха	

(2) Лексема встретилась только в белорусском подкорпусе — 7:

	ГК	КПБ
в лом	1	0
квасить	0	1
кранты	0	1
перекантоваться	1	0
прошвырнуться	1	0
салага	0	2
стремный	0	1

(3) Лексема встретилась только в российском подкорпусе — 159:

бомбила	10	кидать	8
ботаник	9	кидала	3
бумер	9	кидалово	2
вмазать, вмазанный	2	клава	1
впарить	6	клевый	7
впаять	6	клоун	3
втереть	1	комп	3
вставлять	1	кореш	4
выпендреж	1	косить под	5
вырубиться	2	косить от	8
гадюшник	4	косуха (куртка)	4
грохнуть 'убить'	7	крышевать	6
грузить	9	ксива	5
движуха	3	лапша (вешать лапшу)	2
депрессуха	2	ласты склеить	1
раздербанить	1	левак	1
жечь	6	ломануться	12
жлоб	5	ломовой	1
жмурик	1	лопатник	1
забойный	6	лопух	1
завалить	20	лохануться	2
зависнуть (находиться,		лям	2
проводить время)	3	ментовка	3
загогулина	3	мерс	14
задолбать	3	мобила	5
заказуха	6	мочилово	1
закидон	4	мочить	11
закосить	1	замочить	17
замести	1	замутить	5
заныкать	3	наварить	4
замутить	6	надыбать	3
запариться	1	накатить	2
запороть	3	нал	10
засланец	5	напряг	3
зашибать	1	нарисоваться	5
зелень	4	настучать (донести)	5
зомбоящик	1	непруха	1
игла (на игле)	4	несознанка	1
кантоваться	2	нехилый	11
капуста	3	нормалек	1
кент	1	облом	11
кидала	4	обломиться	3

обшмонать	2	репа	2
общак	8	ржач	1
огрести	3	сдуться	9
отвязанный	1	скин	9
отжать	2	слив	2
откат	44	слить (кого)	2
отмазать	8	слить (в прессу)	2
отметелить	1	слить (игру)	2
отстой	3	спалиться	1
оттянуться	5	стопудово	1
офигеть	3	стрелка	4
паленый	11	стукнуть (на кого)	1
палево	1	сфоткать	3
париться	2	сфоткаться	5
передоз	1	сходняк	1
поддатый	4	тащиться от	1
подкол	1	тормоз	1
подогнать	1	торчать	1
подмазать	1	трезвяк (= вытрезви-	
подрезать (кошелек)	1	тель)	1
подстава	29	троллить	1
подставиться	4	тупить	1
подсуетиться	10	туса	2
подкатить	1	тусить	3
подтянуться	15	тусовщик	11
попадос	1	туфта	2
поперло	1	уважуха	1
попилить (бюджет)	3	уделать	1
порвать	2	улетный	1
порешать	1	упырь	1
порнуха	8	хавка	2
предъява	1	шарашка	3
продинамить	1	шиза	1
пролет (в пролете)	2	ширяться	1
проставиться	1		
проставляться	1		
разводка	7		
раздербанить	1		
раздолбай	2		
распилить	3		
разогрев	11		
рассекать	16		
расслабляться	13		

Лексема встретилась в ОБОИХ подкорпусах — 118:

(4) Лексема встретилась в российском подкорпусе и Гродненском корпусе белорусского подкорпуса (некоторые лексемы — также в КПБ) — 80:

	Рос.	ГК	КПБ
бабло	9	1	
бабки	18	3	
врубиться	2	1	
дальнобойщик	57	9	10
дембель	17	1	
депрессняк	1	1	
драйв	11	51	2
жесть	9	1	
забить на	1	1	
зависать	7	1	
зависнуть (остановиться)	18	1	1
завязать	22	1	1
зажигать	26	2	3
заморачиваться	22	1	
засветиться	64	5	3
затариться	9	1	1
инфа	1	1	
кайф	30	2	5
качок	11	1	1
кинуть	39	1	3
крутой	113	9	15
лейбл	24	1	
липовый	88	3	3
мажор	5	1	
мандраж	7	3	
мент	52	1	5
мерин	3	1	
навороченный	33	2	4
наехать	12	1	
накрыться	5	1	
напрягать	36	2	3
наркота	21	1	1
настучать (побить)	3	1	
неслабо	4	1	
нехило	2	1	2

	Рос.	ГК	КПБ
обломаться	1	1	
обратка 2 (ответные действия)	1	1	
обуть	4	1	
общага	26	2	
окучивать	4	1	
откосить	8	1	
отмашка	16	2	
отмороженный	3	2	
оторваться	11	1	
отрываться	19	2	
отрубиться	2	1	
париться (где)	3	1	
пахать 'работать'	39	1	4
пацан	36	1	
пилить (бюджет)	2	1	
повестись	7	1	
подколоть	3	2	
подрезать (авто)	14	1	
подсадить на	5	1	
понты	11	1	
попса	14	1	
прессовать	11	1	
прикид	5	1	
прикол	38	4	
прикольный	21	4	
прокол	14	2	
пролететь	7	1	1
просечь	1	1	
профи	40	2	1
разборка	123	3	14
раскрутить	42	4	
раскрутка	41	1	2
релакс	5	1	2
респект	6	3	
ржачка	1	1	
срубить (бабки)	3	3	
стукач	2	2	
тусовка	57	5	5
урод	10	1	
фанат	266	8	
фарт	5	1	2

	Рос.	ГК	КПБ
фиг (в т.ч. на фиг, по фигу)	18	1	2
фишка	71	10	13
фраер	4	1	
халява	21	6	3

(5) Лексема встретилась в российском подкорпусе и в текстах «Комсомольской правды» в Беларуси (в ГК не встретилось) — 38:

	Рос	КПБ		Рос	КПБ
барабан (по бара-			пахать на	1	1
бану)	8	1	подставить	23	3
бухать	3	1	проколоться	2	1
гопник	9	1	просыхать (не		
горючка	18	3	просыхал)	1	1
дернуть 'сбежать'	2	1	развести	13	5
доставать	4	4	свалить	3	1
жратва	3	1	струхнуть	3	1
заморочка	6	1	тачка	37	2
косяк	10	2	трэш	4	1
липа	5	1	убойный	9	2
ломиться в / к /			фанатеть	1	1
внутрь	19	1	фанера	6	1
лох	20	3	фотка	8	1
наезд	18	1	френд	11	1
накосячить	1	1	чернуха	5	1
напряженка	20	1	чмо	2	1
немерено	9	1	чувак	7	1
непонятки	2	1	чумовой	1	1
облажаться	1	1	ящик	4	1
обчистить	11	2			

Итак, приведенные данные показывают, что в белорусском подкорпусе сам «фонд» сленговой лексики существенно меньше, чем в российском,— это чуть больше четверти (118) от общего исходного списка, причем часть этих лексем (38) встретилась только в КПБ.

Нестандартные употребления в лексике и фразеологии

В этом разделе мы обсудим лексические и стилистические неточности и нарушения в сочетаемости, которые касаются употребления отдельных слов и устойчивых выражений. Естественно, мы сможем обсудить не весь такой материал, а лишь некоторые примеры в качестве иллюстрации. Этот материал можно разбить на две группы.

НЕТОЧНОЕ или НЕУДАЧНОЕ СЛОВОУПОТРЕБЛЕНИЕ

Неточное или неудачное словоупотребление может иметь семантический или стилистический характер. Например, выражение, исходно имеющее отрицательную оценку, в следующих контекстах употреблено в положительном смысле:

Еще одно совместное увлечение членов семьи доставляет всем невероятное удовольствие. Это кролиководство. Больше всех оно нравится главе семьи, который сумел ***втянуть в это занятие и детей*** [Ивьевский край, 2012.11.27]; *Заниматься любимым видом спорта Михаил начал, когда учился в четвертом классе. Спустя пять лет получил звание мастера спорта по мотокроссу. До 21 года смог удержать звание чемпиона СССР в классе 500 куб. см. Именно Михаил* ***втянул в мотогонки*** *своего* ***младшего брата*** *Валерия* [Вечерний Гродно, 2012.09.20].

В этих примерах речь идет о чем-то нейтральном или даже положительном. Между тем выражение *втянуть кого во что* обычно имеет отрицательные коннотации (втягивают обычно во что-то плохое). Вот контексты из НКРЯ: *втянуть в конфликт, в закулисную свару, в историю, в скандал, в свои сети, в ловушку, в долги, в затраты, в водоворот, в кризис, в борьбу, в войну, в большую игру* [= интриги], *в криминал, в беду, в аферу, в авантюру*. Те немногие примеры, в которых существительное нейтральное (*втянуть в дискуссию, в спор, в историю, в разговор, в орбиту, в проект*), значат ‘против воли, при сопротивлении другого лица’, ср.: *Стенли попросил свою верную Роуз Мороуз соединить его с той самой «Путни продакшн», которую Штефан Чапский безуспешно пытался втянуть в их проект* [Василий Аксенов. Новый сладостный стиль (2005)].

Девятиклассница Тома мечтала быть археологом и к этому

готовилась. Но через год ***мечту изменила*** [Свислочская газета, 03.04.2012]

В НКРЯ сочетание *изменить мечту* не встречается. Глагол *изменить* сочетается с теми существительными из данного семантического класса (ментальные объекты и состояния, результаты ментальной деятельности), которые обладают признаком контролируемости, ср. *изменить решение, намерения, планы*.

[Ольга Викторовна, руководитель фермы] *никогда* ***не оставалась равнодушной и к состоянию молочного оборудования, санитарному порядку*** *в животноводческих помещениях, к своевременному кормлению рогуль* [Ивьевский край, 2012.05.07].

Выражение *не остаться равнодушным к чему-л.*, хотя и не относится к высокому стилю в строгом смысле, тем не менее стилистически не вписывается в контекст молочного оборудования и животноводческих помещений. Оно обычно описывает личную («душевную») реакцию человека на какое-л. событие, которое его эмоционально задевает, ср. контексты из НКРЯ: *не остаться равнодушным к чужому горю, к чужой беде, к этой истории, к участи Сербии*; ср. также примеры из белорусского подкорпуса: *Уезжая, мы от души благодарили хозяев за радушный прием, за доброту и открытость, за то, что* ***не остаются равнодушными к судьбе инвалидов*** [Свислочская газета, 2012.06.23]; *Не оставайтесь равнодушными. Помните, чужой беды не бывает* [Свислочская газета, 2012.03.19]. Если речь идет о выполнении служебных обязанностей, было бы более уместным нейтральное выражение (*ответственно / добросовестно относиться* и т. п.).

Молодой человек — профессиональный юрист, такая же профессия и у его жены, так что белорусское ***законодательство*** *они* ***знают не понаслышке*** (Берестовицкая газета, 12.06.2012].

Выражение *знать не понаслышке* обычно применяется не к знанию текстов, а к знанию жизни: оно употребляется в случаях, когда человек имеет личный опыт (а не опирается только на рассказы других людей), лично принимает участие в какой-то ситуации, деятельности: *Сиднев про Силиконовую долину знает не понаслышке: в конце 80-х он участвовал в проекте по учительскому обмену и год преподавал в Штатах математику в обычной американской школе* [«Русский репортер», № 37 (215), 22 сентября 2011, 2011]; *Валерий Чичканов, не понаслышке знакомый с вопросами управления на раз-*

ных уровнях — был вице-премьером правительства РФ, — рассказывает "Времени МН" о некоторых проблемах административной реформы [«Время МН», 2003.08.07]. Кроме того, в нашем примере есть стилистическая несогласованность разговорного наречия *понаслышке* с юридической лексикой.

Это человек, который отдает ***сполна*** *все свои знания и опыт ученикам, вкладывает в них всю свою душу и энергию* [Берестовицкая газета, 05.05.2012].

Сполна используется, когда речь идет о реальном или метафорическом возвращении долга или оплате чего-л. (*расплатиться / заплатить сполна* [Токарева Виктория. Своя правда // «Новый Мир», 2002]; *Лужков сполна рассчитался с депутатами, поддержав Бориса Ельцина и лишив Белый дом света, тепла, электричества и прочих благ цивилизации* [«Коммерсантъ-Власть», 1999]), либо когда речь идет о процессах, не контролируемых человеком: *И, поистине, получите сполна вы за деяния свои в Судный День* [Шамиль Аляутдинов. Мусульмане: кто они? (1997–1999)]; *Однако сполна вкусить все радости семейного досуга удаётся не всегда* [«Семейный доктор», 2002.12.15]. В контекстах, подобных приведенному, обычно используются интенсификаторы *целиком, полностью, без остатка*.

Сегодня Тамаре Ивановне трудно с ходу сказать, сколько спектаклей поставлено ею с того стартового 1992 года, сколькими дипломами и грамотами они отмечены, для скольких актеров из народа игра в театре ***становилась настоящим хобби****…* [Свислочская газета, 03.04.2012].

Интенсификатор *настоящий* предполагает некоторую шкалу, где есть иерархически более низкие и более высокие ступени, и присоединение этого интенсификатора либо свидетельствует о высоком статусе (*Он — настоящий профессионал*), либо призвано повысить статус характеризуемого предмета, например: *Этот актер-любитель играет как настоящий профессионал.* В данном случае нарушение состоит в том, что хобби занимает как раз низшую ступень в иерархии занятий: хобби — увлечение — профессия — призвание.

Могу сказать, что за прошлый год заработано 1 млрд. 291 млн. руб.

— Возникает ***логический вопрос****: как расходуются заработанные денежные средства?* [Ивьевский край, 06.03.2012], ср. *логично*

задать вопрос. В основном корпусе НКРЯ словосочетание *логический вопрос* не встретилось, однако в газетном корпусе НКРЯ есть 4 вхождения.

НАРУШЕНИЕ ЛЕКСИЧЕСКОЙ ИЛИ ФРАЗЕОЛОГИЧЕСКОЙ СОЧЕТАЕМОСТИ

Еще одна область, где наблюдаются языковые нарушения и неточности, — лексическая и фразеологическая сочетаемость.

Под лексической сочетаемостью мы, вслед за Ю. Д. Апресяном [Апресян 1974], будем понимать случаи, когда слово в качестве распространителей допускает одни лексемы и не допускает другие, хотя никакого семантического запрета при этом не нарушается, например, можно сказать *ошибся дверью / номером* (если человек вошел не в ту дверь или набрал не тот номер), но нельзя сказать **ошибся зонтиком / шляпой*, если человек взял не тот зонтик или надел не ту шляпу. С семантической точки зрения здесь ничего не нарушено, именно поэтому лексическая сочетаемость позже осваивается при изучении языка, раньше утрачивается при отсутствии языковой практики и т. д.

*Военную **балладу** о милосердии **поведала** учитель музыки Новодворской СШ Татьяна Захарко* [Свислочская газета, 14.02.2012].

Про балладу обычно говорят *исполнить*, а *поведать* можно *историю, легенду, истину, детские тайны, подробности* (контексты из НКРЯ).

*От профессиональной театральной режиссуры Тамару Ивановну (позже она окончила Белорусский государственный институт культуры и искусства) отделяли годы работы учителем музыки в одной из школ соседской России. Но и там ее творческая **натура грезила сценой*** [Свислочская газета, 03.04.2012] — *грезить* может человек, а не натура как параметр человека.

*С целью дальнейшего роста рядов объединения следует активизировать прием в члены «Белой Руси» достойных граждан, проявляющих **высокую гражданскую позицию** и имеющих значительные трудовые успехи* [Свислочская газета, 31.01.2012].

В основном корпусе НКРЯ встретился только 1 подобный пример (в региональном издании): *Фестиваль помог по крупицам воссоздать историю славных родов Строгановых, Лазаревых и Аба-*

мелек-Лазаревых; показать, как их благотворительность, высокая гражданская позиция и разумное управление уральскими имениями обрели широкую известность в России и за ее пределами [«Нефтяник» (Пермь), 2003.05.27], в газетном корпусе НКРЯ таких сочетаний нет. Абсолютное большинство сочетаний в основном корпусе — *активная гражданская позиция*; встречаются также *четкая, твердая, правильная гражданская позиция*.

25 лет не только народному театру, а театру на сцене Ивьевского районного Дома культуры вообще. Потому что все то, что было до того,— была художественная самодеятельность: инсценировки талантливые и бесталанные, отдельные ***пробы поставить*** *спектакли* [Ивьевский край, 06.04.2012]. С инфинитивом обычно употребляется *попытка*, ср.: *попытки поставить*. В основном корпусе встретились 5 примеров сочетания «*проба* + инфинитив»,— однако их нельзя считать показательными, т. к. они относятся к 1-й половине XX века и для современного языка являются устаревшими: *Восстанавливаю (была* ***проба «сократить»****) «Леф», уже «Новый».* [В. В. Маяковский. Я сам (1922–1928)]; *Потом еще остановка, опять воду берут, опять на следующем перегоне* ***проба отворить*** *дверь…* [В. А. Гиляровский. Мои скитания (1927)]; *Вместе с многократной* ***пробой воспитывать*** *по выработанному у нас методу сочетательный рефлекс у оглохших от контузии постепенно утраченный ранее слух восстанавливался* [В. М. Бехтерев. О сочетательно-рефлекторной терапии (1925)]; *Вот первый шаг, первая* ***проба исполнить*** *заповедь любви и ответить ласковым словом тому, кого ненавидишь* [П. Н. Краснов. От Двуглавого Орла к красному знамени (1922)]; *Вот как отозвалось на моем сердце первое личное дело, первая личная моя обязанность, первая* ***проба предъявить*** *размеры личной смелости жить на свете …* [Г. И. Успенский. Волей-неволей (1884)].

Всё это находится в нашем районе ***под постоянным вниманием*** [Берестовицкая газета, 03.12.2012] — ср.: *требует внимания* или *находится под контролем*.

В эти благодатные весенние дни, когда вся природа пробуждается для новой жизни, свой профессиональный праздник отмечают работники одной из самых сложных, во многом ***неблагодарной****, но такой необходимой* ***отрасли*** *нашей жизни — жилищно-коммунального хозяйства* [Берестовицкая газета, 23.03.2012]. Сочетание

неблагодарная отрасль ни в основном, ни в газетном корпусе НКРЯ не встретилось. Встречаются контексты: *неблагодарный труд, занятие, работа, дело, роль, задача*.

В ГК встречаются также нестандартные употребления в сфере фразеологической сочетаемости. Мы имеем в виду в первую очередь так называемые коллокации, сочетания, в которых один компонент выступает в свободном значении, а второй — во фразеологически связанном (см., в частности, [Виноградов 1977], [Апресян 1974], [Мельчук 1999], [Баранов, Добровольский 2014]): *принять решение, причинить ущерб, принести извинения, нанести обиду, оказать помощь*. Глаголы в таких сочетаниях полузнаменательные, семантически вырожденные, поэтому на их употребление нет собственно семантических запретов, тем не менее нельзя сказать **принести обиду* или **причинить помощь*. Наиболее распространенные случаи сочетаемостных нарушений в составе коллокаций, обнаруженные в ГК, связаны с глагольно-именными оборотами типа *принять решение* и со степенными определителями типа *жгучий брюнет* (разные слова, допускающие определение по степени, требуют разных показателей высокой степени, ср. *непреодолимое желание*, но *смертельная обида* (**непреодолимая обида*); *отъявленный двоечник*, но *круглый отличник* (**отъявленный отличник*) и т. д.).

Чуть-чуть уступили лидерам, но ***вызвали*** *тоже* ***много лестных слов*** *в свой адрес жители д. № 19 а по ул. Ленина — семья В. К. Яскулда и О. И. Ушкевич* [Свислочская газета, 16.01.2012].

С глаголом *вызвать* из класса речевых существительных обычно сочетаются нейтральные слова (*вызвать спор, дискуссию, восклицание*) или слова, содержащие отрицательную оценку: *вызвать возражения, жалобы, замечания, нарекания, насмешки, отповедь, пересуды, претензии, протест, разговоры, сплетни, толки, упреки, шквал критики*, и т. п.

Он ***высказал*** *свои искренние* ***поздравления*** *и пожелал всем ветеранам здоровья и долголетия* [Свислочская газета, 03.03.2012].

Стандартная сочетаемость глагола *высказать* с речевыми существительными — *благодарность, возражение, догадка, намек, осуждение, пожелание, предположение, претензия, соображение, упрек* (наибольшее количество вхождений — *высказать пожелание*). Существительное *поздравления*, как показывают данные НКРЯ, в современном языке практически не обслуживается никаки-

ми вспомогательными глаголами со значением речи. Оно выступает либо в составе перформатива *Прими* (*те*) *мои / наши поздравления*, либо с глаголами *послать, отправить, зачитать* [поздравление]. Сочетание *принести поздравления* встречается в текстах XIX и начала XX вв. Таким образом, существует объективная трудность поиска подходящего вспомогательного глагола для данного существительного.

В последние годы наряду со специальностями экономистов, юристов, менеджеров ***прочно котируется*** *профессия психолога* [Ивьевский край, 05.10.2012].

В основном и газетном корпусе НКРЯ такое сочетание не встретилось. Встречаются сочетания: *высоко, невысоко, очень, неплохо, плохо котируется*.

Итак, материал показывает, что расшатывание, размывание нормы происходит, так сказать, в ожидаемых участках языковой системы. Лексическая и фразеологическая сочетаемость — это та область, где не действуют семантические запреты, поэтому там нет надежных критериев вида «так сказать нельзя, потому что это абсурдно». Например, в отличие от сочетаний типа ?*неблагодарное происшествие, признание, неблагодарная путаница, причина, разведка* и т. п., которым вообще трудно придумать интерпретацию, сочетание *неблагодарная отрасль* вполне понятно на фоне выражений *неблагодарный труд, занятие, работа*, — и даже может рассматриваться как метонимическое стяжение (‘отрасль, где преобладает неблагодарный труд’).

Подобные сдвиги и нарушения именно в сфере лексикализованной и фразеологизованной сочетаемости наблюдаются в ситуациях, когда по каким-то причинам ослаблен «пресс» литературной нормы, — в языке эмиграции, в языке регионов, особенно двуязычных регионов. Кроме того, в региональной прессе нет такого «жесткого» редактирования, как в центральных СМИ. С другой стороны, такие нарушения могут объясняться не только недостаточным владением нормой, но и расширением сочетаемости в рамках постоянно происходящих языковых сдвигов, т. е. изменением самой нормы. Случаи нарушений в употреблении единиц, в том числе — нарушений в сочетаемости, дают ценный материал для лингвистических исследований, т. к. показывают точки роста, изменения и развития языка в исторической перспективе. В этом

смысле материал регионального корпуса представляет интерес не только с точки зрения описания региональных особенностей русского языка, но и с точки зрения теоретических и исторических исследований.

Литература

Апресян Ю. Д. Лексическая семантика. М., 1974.

Баранов А. Н., Добровольский Д. О. Основы фразеологии. М., 2014.

Вепрева И. Т. Языковая рефлексия в постсоветскую эпоху. М., 2005.

Виноградов В. В. Об основных типах фразеологических единиц в русском языке // В. В. Виноградов. Избранные труды. Лексикология и лексикография. М., 1977. С. 140–161.

Ермоленкина Л. И. Образ адресата региональной прессы: проекции советской информационно-коммуникативной практики СМИ // Ностальгия по советскому. Томск, 2011. С. 374–383.

Кузьмина Н. А. Деидеологизация или новая идеологизация? (идеологемы нового времени) // Язык и стиль современных средств массовой информации. М., 2007.

Кузьмина Н. А., Абросимова Е. А. Активные процессы в русском языке и коммуникации новейшего времени. М., 2013.

Курилова А. Д. Толковый словарь разговорного языка. М., 2007.

Мельчук И. А. Опыт теории лингвистических моделей «Смысл ↔ Текст». М., 1974/1999.

Мокиенко В. М., Никитина Т. Г. Толковый словарь Совдепии. СПб, 1998.

Никитина Т. Г. Молодежный сленг: Толковый словарь. М., 2009.

Ностальгия по советскому. Томск, 2011.

Резанова З. И. Советский интертекст в дискурсе региональных СМИ: трансформации и функции // Советское прошлое и культура настоящего. В 2-х тт. Екатеринбург, 2009. Т. 2. С. 76–90.

Савчук С. О. Корпус региональных газет России и зарубежья // Труды Института русского языка им. В. В. Виноградова. Вып. 6. М., 2015 а.

Савчук С. О. Корпус как инструмент для исследования особенностей функционирования русского языка в региональной прессе

// Труды Института русского языка им. В. В. Виноградова. Вып. 6. М., 2015 б.

Советское прошлое и культура настоящего. В 2-х тт. Екатеринбург, 2009.

Солганик Г. Я. Толковый словарь: Язык газеты, радио, телевидения. М., 2002.

Стернин И. А. Общественные процессы и развитие современного русского языка. Воронеж, 2004.

[1,2]Galina I. Kustova
[1]Vinogradov Russian Language Institute
of the Russian Academy of Sciences
[2]National research Tomsk State University
(Moscow, Russia)
galinak03@gmail.com

LEXICAL-SEMANTIC AND SOCIO-CULTURAL SPECIFICITY OF RUSSIAN SPEECH OF GRODNO (ON THE MATERIAL OF MEDIA TEXTS)

We describe the linguistic phenomena that reflect the socio-cultural situation of Grodno: expansion of the new vocabulary in the Post-Soviet period on the one hand, and preservation of sovietisms and official style collocations on the other hand.

We analyze the usage of slang vocabulary in newspaper texts, reflecting the journalists' to be closer to the reader. We investigate inaccurate and non-standard word usage and non-standard compatibilities in collocations.

Key words: socio-cultural situation, language of media, slang vocabulary, non-standard use of words

References

Apresjan Ju. D. *Leksicheskaya semantika* [Lexical Semantics]. Moscow, 1974.

Baranov A. N., Dobrovol'skij D. O. *Osnovy frazeologii* [Foundation of phraseology]. Moscow, 2014.

Vepreva I. T. *Yazykovaya refleksiya v postsovetskuyu epokhu* [Language reflection in the Post-Soviet period]. Moscow, 2005.

Vinogradov V. V. [The main types of phraseological units in Russian]. *Izbrannye trudy. Leksikologiya i leksikografiya* [Selected Works. Lexicology and Lexicography]. Moscow, 1977, pp. 140-161. (In Russ.)

Ermolenkina L. I. [The image of the addressee of the regional press: projection of the soviet information and communication media practices]. *Nostal'giya po sovetskomu* [Nostalgia for the Soviet]. Tomsk, 2011, pp. 374 – 383. (In Russ.)

Kuz'mina N. A. [Deideologization or new ideologization? (ideologemes of modern times)]. *Yazyk i stil' sovremennykh sredstv massovoi informacii* [Language and style of modern media]. Moscow, 2007. (In Russ.)

Kuz'mina N. A., Abrosimova E. A. *Aktivnye processy v russkom yazyke i kommunikatsii noveyshego vremeni* [Active processes in the Russian language and communication of recent times]. Moscow, 2013.

Kurilova A. D. *Tolkovyi slovar' razgovornogo yazyka* [The explanatory dictionary of spoken language]. Moscow, 2007.

Mel'chuk I. A. *Opyt teorii lingvisticheskikh modelei «Smysl ↔ Tekst»* [The experience of the theory of linguistic models «Meaning ↔ Text»]. Moscow, 1974 / 1999.

Mokienko V. M., Nikitina T. G. *Tolkovyi slovar' Sovdepii* [The explanatory dictionary of Sovdepiya]. St. Petersburg, 1998.

Nikitina T.G. *Molodezhnyi sleng: Tolkovyi slovar'* [Youth slang: The explanatory dictionary]. Moscow, 2009.

Nostal'giya po sovetskomu [Nostalgia for the Soviet]. Tomsk, 2011.

Rezanova Z. I. [Soviet intertext in the discourse of regional media: transformations and functions]. *Sovetskoye proshloye i kul'tura nastoyashchego* [Soviet Past and the culture of Present]. In 2 v. Ekaterinburg, 2009, v. 2, pp. 76–90. (In Russ.)

Savchuk S. O. [The corpus of regional Newspapers in Russia and abroad.] *Trudy Instituta russkogo yazyka* [Proceedings of the V.V. Vinogradov Russian Language Institute], no. 6. Moscow, 2015a (In Russ.)

Savchuk S. O. [Corpus as a tool for studying the Russian language usage in regional press]. *Trudy Instituta russkogo yazyka* [Proceedings of

the V.V. Vinogradov Russian Language Institute], no. 6. Moscow, 2015b. (In Russ.)

Sovetskoye proshloye i kul'tura nastoyashhego [Soviet Past and culture of Present]. In 2 v. Ekaterinburg, 2009.

Solganik G. Ya. *Tolkovyi slovar': Yazyk gazety, radio, televideniya* [The explanatory dictionary: Language of newspaper, radio and television]. Moscow, 2002.

Sternin I. A. *Obshchestvennye processy i razvitie sovremennogo russkogo yazyka* [Social processes and the development of modern Russian]. Voronezh, 2004.

Я. Э. Ахапкина
Национальный исследовательский университет
«Высшая школа экономики»
(Россия, Москва)
yana.akhapkina@gmail.com

РЕФЛЕКСИВНЫЕ ГЛАГОЛЫ «УБИРАТЬСЯ» И «ИГРАТЬСЯ»: КОДИФИКАЦИЯ И УЗУС[1]

В статье на материале Национального корпуса русского языка и интернет-узуса анализируется функционирование разговорных возвратных глаголов «убираться» и «играться»: выявляются типы значений и особенности контекстуальных употреблений единиц, традиционно трактуемых как ненормативные; сопоставляется речевая судьба рефлексивов — соотносительного и несоотносительного с исходным переходным глаголом. Как утверждает С. С. Сай, для функционирования русских возвратных глаголов существенны семантические отношения не только с производящими, но и с другими возвратными, внутри системы рефлексивов: имеются в виду ассоциативные связи в смысловых рядах, образуемых глаголами с постфиксом *-ся*. Мы пытаемся выяснить, насколько обоснована идея считать глагол «играться» ошибочным только потому, что производящий глагол «играть» в рассматриваемом значении не является переходным. В статье анализируются рефлексивы, образованные от непереходных глаголов; рассматриваются возвратные глаголы, окказионально сконструированные по аналогии с нормативными рефлексивами. Мы возвращается к идее С. П. Обнорского, считавшего *-ся* маркером непереходности, закрепляющим это значение

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ в 2015 году.

морфологически. Вероятно, глаголы «играться» и «убираться» в равной мере принадлежат разговорному регистру, реализуют естественные потенции языковой системы и более или менее отчетливо вступают в отношения семантического и прагматического распределения с производящими.

Ключевые слова: грамматика, норма, речевая практика, контекст, возвратность, рефлексивность, переходность, залог.

Постановка проблемы. В традиционных и ориентированных на традицию пособиях по культуре речи можно прочитать, что небезличное употребление рефлексивных глаголов «убраться» / «убираться» (*где-либо*, например, *в квартире*) и «играться» (*с кем-либо, во что-либо*, например, *с сестрой, в куклы*) в конструкциях активного залога относится к просторечию и противоречит литературной норме: контексты употребления этих возвратных образований признаются неправильными (**убираться в комнате*, **играться в игрушки*); каких-либо различий между указанными глаголами или градации некорректности высказываний с ними при этом не фиксируется [Орлова и др. 2014: 63–64; Русский язык и культура речи 2013: 13; Вариясова и др. 2007: 42].

Словари, однако, по-разному описывают эти глаголы. Так, для «играться» встречаются пометы *областное* [Ушаков 1935–1940, МАС 1985], *разговорно-сниженное* [Ефремова 2000], *разговорное* [Ефремова 2005], *народно-разговорное* [Кузнецов 1998]. В ряде источников глагол «играться» как синоним «играть» вовсе не обсуждается, например, в [Ожегов 1986], тогда как «убраться» присутствует, причем если в словаре Д. Н. Ушакова рассматриваемое употребление указано под номером 2 с пометой *разговорное*, то в словаре С. И. Ожегова [1986] с той же пометой, но уже под номером 1. МАС, наследуя словарю Д. Н. Ушакова, толкует «убираться-1» через коррелят совершенного вида, для которого рассматриваемое значение указывается вторым и помечается как *разговорное* (при этом не делается различия в пометах для «убраться в помещении» и «убраться восвояси»). Фиксирует «убираться» и игнорирует «играться» словарь [Бабенко 1999]. Таким образом, для традиционной лексикографии «убираться» оказывается более приемлемой единицей, чем «играться» (нет словарей, игнорирующих «убираться», но есть словари, игнорирующие «играться»; по-

меты при «играться» жестче: *областное*, *разговорно-сниженное*, *народно-разговорное*).

Для корпуса оба употребления маргинальны. В НКРЯ «играться» в рассматриваемом значении до 1900 года встречается всего 12 раз, следующие 12 раз — между 1900 и 1940, новые 12 употреблений фиксируются к 1980, затем 12 — уже к 1990, а с конца 20 века частота употребления растет: 12 в промежутке 1990–1995 (до первого употребления в 1995), 12 между 1995 и 1997, 12 с 1998 по 2000, с 2000 по 2012 — 59 раз (всего 143 употребления); в последние десятилетия лексема активнее используется вне передачи речи персонажа, без стилизации: *Играются уже не только правительства, политики, но и население.* [Ю. Богомолов. Победа будет за ними. Теленеделя (2001) // «Известия», 2001.09.28]; *Бумажными деньгами в ту пору игрались дети да старухи обклеивали ими крышки сундуков*. [Б. Екимов. Белая дорога (2001)]; **—** *Конечно, тогда она просто отказывается работать, начинает играться, все хватает*. [Ю. Зорина. Собачья работа (2001) // «Семья», 2001.11.14].

«Убираться» в рассматриваемом употреблении к 1900 году фиксируется только 7 раз, а между 1900 и 1910 не фиксируется ни разу, следующие 7 употреблений обнаруживаются только к 1940 году, новые 7 — к 1970, следующие 7 — к 1982, затем 7 к 1991, 7 с 1992 по 1996, 7 в 1996–1997, 7 к 1998 году, 7 с 1998 по 2000 (до первого упоминания в 2000), с 2000 по 2012 — 89 раз (всего 152 раза, значения ‘приводить себя в порядок’, ‘справляться с полевыми работами’ и ‘хлопотать по хозяйству’, разумеется, не учитываются, рассматривается только значение 'наводить порядок, заниматься очисткой помещения'). Итак, в НКРЯ «играться» в искомом употреблении встретилось 143 раза, «убираться» — 152, при этом частота обоих употреблений нарастает с конца XX века, в наше время «убираться» встречается, по данным корпуса, несколько чаще, чем «играться» (с 2000 по 2012 год 89 и 59 раз соответственно).

Современные источники противопоставляют эти глаголы как образованный от переходного (*убирать*) и от непереходного (*играть*) производящего соответственно, тем самым узаконивая первый (относя его к разговорной сфере литературного языка, периферии нормы) и отказывая в легитимизации второму. Глагол «убираться» трактуется как косвенно-возвратный, ситуация употребления интерпретируется как включение в сферу субъекта,

акцентируется сема «в интересах деятеля» [Другoвейко-Должанская], ср. аналогичное образование *прибраться* как пример косвенно-возвратного глагола без стилистических помет в справочнике [Улуханов 1998]. К этой группе относят возвратные глаголы, невозвратные корреляты которых требуют дополнения: *собираться / паковаться* ‘собирать / паковать (свои) вещи’, *парковаться* ‘парковать (свою) машину’, *заряжаться* ‘заряжать (свой) телефон’, *защититься* ‘защитить (свою) диссертацию’ (*убраться* — ‘убрать помещение’ — оказывается чуть шире, компонент принадлежности объекта деятелю здесь менее значим, ср. *переключиться* ‘переключить программу’). Е. Н. Никитина [2008] говорит в этой связи о корректности употреблений, подразумевающих объект, инкорпорированный в сферу человека (метонимически объяснима включенность в личную сферу *диссертации, машины, дома*). Однако *убираться* — ‘заниматься уборкой помещений профессионально’ — от такой инкорпорированности удаляется.

Наследуя традиции, заложенной А. А. Потебней и продолженной В. В. Виноградовым [Виноградов 1947], кодификационная практика видит в варианте «убираться» большую интенсивность действия по сравнению с производящим: «В совпадающих в значении парах *грозить — грозиться*, *стучать — стучаться*, *убирать — убираться* и т. п. возвратным глаголам присуще значение большей интенсивности действия, заинтересованности в его результате; ср.: *он постучал в дверь — он постучался в дверь, чтобы ему открыли*. Стилистически эти формы различаются тем, что формы на *-ся* свойственны сниженному стилю речи, например: *Одна баба с порогу своей избы погрозилась мне ухватом* (Тургенев)» [Розенталь и др. 1999], ср. [Шелякин 2000: 172].

«Играться» же трактуется как ненормативный дублет исходного «играть», существующий за гранью литературного языка. Так, рекомендация Д. Э. Розенталя категорична: «Не рекомендуется употреблять *играться* вместо *играть*» [Розенталь и др. 1999], ср. в позднейшем издании: отмечается «…разговорно-просторечный характер таких возвратных глаголов, как *играться* (вместо *играть*), *плеваться*, *стареться*, *тлеться*» [2001]. Достаточно ли словообразовательной аргументации для противопоставления разговорного, но допустимого «убираться» просторечному и нежелательному «играться»?

Несоотносительные рефлексивы. В работе [Сай 2008] отмечено, что строго деривационного подхода для описания рефлексивов (а значит, и для противопоставления корректных и некорректных образований с *-ся*) недостаточно, он должен дополняться семантическим. Очевидно, что в литературном языке существуют в принципе несоотносительные с невозвратными, тем более с переходными, глаголы с рефлексивным показателем (причем при относительной бедности словника таких глаголов их плотность в тексте весьма высока, до 50% по наблюдениям автора указанной работы):

— необратимые: *бояться, улыбаться, проснуться*;

— исторически соотносительные, но в современном языке утратившие невозвратный коррелят (постепенно вытеснившие его за последние 100–150 лет): *виднеться, прислушиваться, совещаться*;

— семантически несоотносительные (в разной степени и по разным моделям модифицирующие семантику производящего невозвратного, синхронно не имеющие с ним непосредственной связи): *плакаться*; ср. семантически несоотносительные с исходным переходным: *тащиться, плестись, жаловаться, водиться*;

— производные от возвратных (соответственно, непереходных) глаголов при помощи приставки: *засмеяться* от *смеяться*;

— производные от невозвратных глаголов при помощи циркумфикса: *разбегаться, накричаться, всмотреться, задуматься*.

В принципе, соотносительность требует соблюдения четырех параметров: формальной, лексической, семантической и синтаксической соотнесенности с исходным невозвратным глаголом [Генюшене 1983], отсутствие любого из них приводит к несоотносительности рефлексива с исходным невозвратным глаголом [Гото, Сай 2009]. Разнообразие нормативных моделей конструирования и частотность в текстах несоотносительных рефлексивов ослабляют деривационный аргумент, которым оперируют сторонники узко понимаемой кодификационной нормы: все указанные выше нормативные рефлексивные глаголы не соотносятся в современном языке с исходным переходным невозвратным.

Морфологическое уподобление: возвратные дублеты непереходных производящих. В речи носителей языка активен процесс морфологического уподобления: конструирования возвратных образований вместо нормативно невозвратных глаголов (по аналогии с семантически близкими возвратными): *эволюционироваться*

(ср. *изменяться, развиваться*). Более того, отмечается, что носители русского языка порождают и воспринимают возвратные образования без непосредственной опоры на невозвратные корреляты [Сай 2008].

Приведем некоторые примеры узуальных *-ся*-дублетов невозвратных глаголов: *Но я, хоть и комплексовался, но сильно не переживал*. [А. Режабек. Отчего умерла моя мама, Галина Щербакова. 2010]; *Всех* <кроликов> *съели, до одного. А шкурки на продажу хотели — да все попортили. Зачервивелись. Жалко*. [М. Степанова. Безбожный переулок. 2014]; *По ночам уже слыхать стало шаги, свет в лампе прыгал, дрожался*. [Б. Агеев. Душа населения. 2011].

Возвратные дублеты непереходных производящих образуют несколько групп. Во-первых, некогда равноправные, но вытесняемые из употребления невозвратными глаголами рефлексивы: *белеться, чернеться, алеться, темнеться* и под., ср. также изолированное употребление *миновать — миноваться* (*жизнь, пора миновалась*). Так, в НКРЯ (при 1379 употреблений *белеть*) зафиксировано только 146 употреблений *белеться*, в последние десятилетия они единичны.

Во-вторых, стремящиеся разойтись в сочетаемости и, возможно, в семантике с производящими варианты, в частности в позиции изолированного (абсолютного нераспространенного) употребления и, напротив, употребления с зависимыми: *кружиться* (*Кидаешь монетку, она кружится, кружится и проваливается вниз — Пассажирский самолет кружит над Нижним Новгородом и не может сесть*), *плеваться* (*Верблюд плюется — Больной плюет кровью*), *коситься* (*Злобно косится — Косит лиловым глазом*).

В-третьих, различающиеся интенсивностью, направленностью действия, контролируемостью ситуации (см. подробнее о роли агентивности в семантике высказываний с возвратным глаголом [Никитина 2008]): *стучаться, решиться, грозиться, светиться, мешаться, ругаться, целиться*. М. А. Шелякин так формулирует разницу между невозвратным и возвратным вариантами этого типа: в рефлексиве подчеркивается «сосредоточенность действия на его субъекте, часто с признаком интенсивности, усиления, с какой-то целью» [Шелякин 2000: 172].

В-четвертых, вытесняющие невозвратный коррелят: *хвастаться, тусоваться, колбаситься*. В случае с двумя последними, оче-

видно разговорными, глаголами возможно семантическое противопоставление акционального (гипотетически, семантически реципрокного или, как минимум, включающего идею совместной деятельности) невозвратного и эмотивного возвратного (ср. *злиться, радоваться, печалиться*) осмыслений.

Ненормативные узуальные *-ся*-образования поддаются шкалированию по степени корректности с точки зрения предписывающей нормы: от фиксированных в словарях как разговорные или областные до не зафиксированных словарями окказиональных — см. отмеченные в словарях как разговорные: *скислось* (в частности, в словаре Д. Н. Ушакова; ср. нормативные *створожилось, ссохлось*); узуальные несловарные: *протухлось, остылось, сгнилось, погаслось, треснулось*, см. [Цейтлин 1978; Сай 2008]. Подобные инновационные образования активны в детской речи [Цейтлин 1978]. Но и во взрослом узусе, причем в академической письменной речи, закрепляются синонимичные невозвратным дублеты: *варьироваться, коррелироваться, эволюционироваться* [Сай 2008].

Таким образом, возвратный дублет непереходного глагола может быть в истории языка вытеснен невозвратным, может постепенно вытеснить невозвратный из узуса или разойтись с ним в семантике и (или) сочетаемости. Временное сосуществование таких коррелятов показывает системную адекватность возвратного «играться».

Постфикс *-ся* как маркер непереходности. С. П. Обнорский [1960] полагал появление *-ся*-дублетов (*хвастаться, белеться, чернеться, темнеться, светиться, алеться, стучаться, звониться, плеваться, коситься*) следствием стремления глагола морфологически маркировать непереходность: возвратные глаголы в подавляющем большинстве непереходные (исключения типа *бояться, слушаться, дождаться, добудиться* немногочисленны, *-ся* в общем случае блокирует управление винительным падежом [Летучий 2010, 2013]). «Непереходные глаголы, как сказано, оканчиваются и на *-ть*, и на *-ться*. В раннюю пору соответственные глаголы на *-ть* вызвали появление дублетов на *-ться*. Подоплека этого явления была чисто аналогическая. Глаголы переходные все оканчиваются на *-ть*, глаголы на *-ся* — все непереходные. Отсюда неизбежное стремление языка — непереходные глаголы на *-ть* снабжать приметой, своего рода морфологической чертой, — частицей *-ся*. В словаре И. И. Срезневского это явление уже свидетельствуется

многими примерами. Ср. *нáсти* и *нáстися*, *въздремати* и *въздрематися*, *жалити* и *жалитися*, *молити* и *молитися* ("молити" глагол с ослабленной переходностью) и т. д., ср. еще укр. *постити* и наше *поститься*» [Обнорский 1960].

Ср. сходные рассуждения В. Н. Сидорова [Аванесов, Сидоров 1945]: «Подразделение глаголов на возвратные и невозвратные зависит от того, выражено ли у них грамматически или не выражено непереходное значение процесса. Глаголы возвратные — это глаголы с грамматически выраженной непереходностью, т. е. они указывают, что выражаемый ими процесс не обращен и не может быть обращен на прямой объект, выраженный существительным в вин. пад. без предлога, например: *мыться, одеваться, встречаться, сердиться, стучаться, чернеться* и т. д. В отличие от них, у невозвратных глаголов отсутствует указание на непереходность процесса, и потому они могут быть как переходными: *мыть* (руки), *одевать* (ребенка), *встречать* (делегацию), *сердить* (отца), так и непереходными: *стучать, чернеть* и др.». Соответственно, системного нарушения в образовании *-ся*-варианта непереходного глагола не наблюдается.

Семантизация возвратности. Значения рефлексивов описываются в традиционной русистике следующим образом [Улуханов 1998]: собственно возвратное (*умываться*); взаимно-возвратное (*целоваться*); безобъектно-возвратное — в другой терминологии активно-безобъектное (и примыкающее к нему пассивно-качественное), абсолютивно-потенциальное (*собака кусается, сахар плохо растворяется*); общевозвратное (указывает на психическое или физическое состояние субъекта: *сердиться*); косвенно-возвратное (направленное на личную сферу деятеля, совершаемое 'для себя': *запастись*); безличное (*хочется*). В современное описание [Летучий 2010] включены такие подтипы возвратных глаголов: рефлексивное (*папа бреется*); пассивное (*дом строится*); декаузативное (*мяч катится*); автокаузативное (*бросился на врага*); взаимное (*влюбленные целуются*); возвратно-каузативное (*стрижется у парикмахера*); рефлексивно-бенефактивное (*запасся дровами*); объектно-имперсональное (*собака кусается*); модально-пассивное (*дверь не открывается*); безлично-пассивное (*в работе на это указывается*); безличное модально-пассивное (*мне не спится*).

Глагол «играться». Опираясь на предложенные варианты описаний, можно выделить следующие употребления просторечного

глагола «играться» (не исключающие его синонимии и контекстуальной замены на «играть»: о семантическом компоненте широко понимаемой возвратности в структуре невозвратного глагола, т. н. немаркированной рефлексивности, см. [Ясаи 2005]).

1. Взаимное действие, ср. модель *телята бодаются, щенки дерутся*: ***брат с сестрой играются***:

А у нас-то ведь дело деется: Головами, батюшка, играемся. [Ф. Буслаев. Русский богатырский эпос (1887)] (в данном случае можно обсуждать корректность возможной замены невозвратным глаголом; «играться» как 'бороться' здесь дополняется оценочным осмыслением, ср. 'баловаться', 'забавляться'; «играть» приведет к потере смыслового оттенка); *— Ну, хорошо! будем играться в индейцев! — говорит Тема*. [Н. Гарин-Михайловский. Детство Темы (1892)]; *Мишка толечко раз меня побил, а то мы с ним хорошо игрались, — шепотом отвечала девочка…* [М. Шолохов. Тихий Дон. Книга IV (1928–1940)]; *Мы с медвежонком часто игрались, возились, боролись*. [Ф. Искандер. Морской скорпион (1977)]; *Барыня тут барину все от себя и рассказала: и как мы шли, и как встретили, и как игрались барышня с государем, и как сам отворил он нам двери…* [Н. Лухманова. Девочки (1894)]; *Деревня уже спала, только котенок со щенком все еще игрались, догоняя друг друга и громко смеясь*. [А. Курков. Школа котовоздухоплавания (1985–1990)]; *…жена мужа выгоняла раз 10, каждый раз приползал на пузе, лет пять они так игрались…* [Вопрос про себя (2012)].

На базе этого значения формируется употребление с идеей совместного действия: *Сяду на лавочке и смотрю, как она с детишками на песочке играется.* [В. Панова. Сентиментальный роман (1958)]; *Поп и дьякон, конечно, поют и читают, а около ограды хлопцы играются.* [М. Шолохов. Поднятая целина. Книга 1 (1958)]; *Детки в песочке игрались, детки машину нашли* [mirtesen.ru/url?e=pad_click&pad_page=1&blog_post_id…].

2. Действие / состояние, замкнутое в сфере субъекта, сосредоточенное в зоне его интересов, ср. модель *мальчик забавляется, малыш балуется*: ***младенец играется (с) погремушкой***:

Она склонила голову на плечо къ матери и играется ея ожерельемъ, а на него и не взглянетъ. [П. Кулиш. Черная рада (1846–1857)]; *Она немножко оживилась. — Он очень тих, все кулаками играется… Боюсь, как бы не умер…* [А. Осипович (Новодворский).

Карьера (1880)]; *Или шлепает мальчонка босыми ногами по грязи, или, заворотив подол, сидит в луже и играется камешками…* [М. Салтыков-Щедрин. За рубежом (1880–1881)]; *В это время я и познакомился с моим новым сынком, вот с этим, какой в песке играется*. [М. А. Шолохов. Судьба человека (1956)]; *Осчастливленный Толстый играл пробкой. Подбрасывал ее двупалыми клешнями и пробовал поймать*. [М. Петросян. Дом, в котором… (2009)]; *— Тут играйся, во дворе, — говорила она и глядела, как девчушка забавляется с куклою, напевая и баюкая ее*. [Б. Екимов. Сосед (2001)]; *Вот на ребенке в чистом виде — что есть страдание? Играл младенец на диване. Я сидел с краешку…* [Г. Гачев. Жизнемысли//Библиотека «Огонек», 1989]; *Сидит маленькая девочка в песочнице, играется. Подходит дяденька* [Коллекция анекдотов: чёрный юмор (1970–2000)].

3. Постоянное, характерное, типичное свойство, абсолютивно-безобъектное или пассивно-качественное употребление (на продуктивность модели образования рефлексивов этого типа в разговорной речи взрослых и в детской речи указывает С. Н. Цейтлин [1978]), ср. модель *крапива жжется, иголка колется, собака кусается*: ***ребенок играется***:

Толстая женщина — неудачница, если она, потолстев, остается женщиной, пытается играться в трепетные и страстные женские игры. [Е. Пищикова. Пятиэтажная Россия (2007)//«Русская Жизнь», 2008], *Не разрешают им, видите ли, в кукол на уроках играться в младших классах.* [Сегодня в топе блогов история учительницы (блог) (2008)]; *Десять опасных вещей, которыми игрались советские дети*. [hronika.info› Фоторепортажи]. В последнем случае замена невозвратным «играли» уничтожит иронию: возвратный глагол акцентирует идею нарушения запрета.

Этот тип значения актуализируется при сравнении (в качестве эталона сравнения выступает субъект, для которого игра является типичным занятием — ребенок, животное): *Ветер,* ***как новорождённый, играется*** *с мелкой искрящейся рябью.* [Т. Соломатина. Отойти в сторону и посмотреть (2011)], *Я вроде ниче не понимаю,* ***как маленькая, играюсь*** *с ей*. [В. Распутин. Последний срок (1970)]; *А Серега Баринцов бесился меньше всех… он был* ***вылитый кот и играл*** *во все, что хотел: наводил внезапные гонения на салабонов… дергал без устали шнурков…* [А. Терехов. Мемуары срочной

службы (1991)]; *Красные панды радовались снегу и* ***игрались как дети*** [gazeta.ua/ru/articles/animals/].

Во 2–3 употреблениях возвратный маркер приближает глагол к семантике состояния, противопоставляя ее акциональности непереходного варианта.

4. Взаимодействие сильного со слабым, направленное на пациенса действие (косвенная переходность), ср. модель *мать с детьми нянчится, возится, занимается*: ***кошка с мышкой играется***:

...у меня создавалось впечатление, что у нас на Арбате существуют генералы, которые «играются» армией, как дети кубиками… [В. Баранец. Генштаб без тайн. Книга 1 (1999)]; *Пока поют, забавно, не так забавно, как щекотно, будто с тобой, как с ребенком, кто-то играется*. [В. Распутин. Последний срок (1970)]; *Саня, как бы играясь руками девушки, снимал кольцо, а Эди заговаривал зубы владелице кольца.* [Э. Лимонов. Подросток Савенко (1982)]; *Не больно. Мы же с ним играемся. Ни слова больше не говоря, Валя схватил мальчишку за ноги и перевернул его вниз головой*. [Ю. Вяземский. Шут (1982)]; *Она недолго посмотрела, как кошка играется задушенной крысой у нее в ногах и подумала, что Грунька сытая, а то б сожрала уже крысу, не наигравшись вдосталь*. [О. Радзинский. В одном дворе (1985)].

У глагола «играться» легко формируется оценочный компонент употребления, более очевидный, чем у невозвратного коррелята: возвратным показателем маркируется действие, вызывающее неодобрение, пренебрежение, осуждение, как минимум,— снисходительность говорящего: *Я вам здесь слово божие объясняю, а вы в игрушечки играетесь.* [А. Куприн. На переломе (Кадеты) (1900)]; *Хватит шутки шутить и играться в молодежные игры!* [С. Гандлевский. НРЗБ // «Знамя», 2002].

Действие оценивается как легкомысленное, недостойное взрослого серьезного человека: *— Такие здоровые бугаи,— сказал Степка,— а в борьбу играются*. [Л. А. Кассиль. Кондуит и Швамбрания (1928–1931)]; *Оцепенелую Софью Андреевну заставил встрепенуться патетический стонущий гром — это гость, игравшийся с роялем, пошатнулся и умял в клавиатуру обе пятерни*. [О. Славникова. Стрекоза, увеличенная до размеров собаки (1995–1999)]; *Одесские «дочери офицеров» сжигали Путина и игрались в «донецких киборгов»*. [donbass.center/…/2547]. Так, «играться в карты» (за-

нятие, которое можно воспринимать как несерьезное развлечение) фиксируется поисковиком 21 900 раз, тогда как «играться в шахматы» (более серьезное занятие, интеллектуальный спорт) только 2 900 раз; при этом у невозвратного коррелята разница иная: «играть в карты» — 369 000 раз, «играть в шахматы» — 379 000 раз.

Возможный вариант экспрессивно-оценочной окраски — самоирония: *Я игрался в родственника из мафии: «Подонок!»* [М. Козаков. Актерская книга (1978–1995)], *Игрался в телекоммуникационный бизнес. Надоело в смерть — уволился*. [Что сейчас модно? (форум) (2005.06.07)], *Лежа в ванне, я долго игралась с гибким душем.* [М. Гиголашвили. Экобаба и дикарь (1998–2007) // «Зарубежные записки», 2009].

За счет той же негативной окраски возвратный вариант в узусе активно конкурирует с нормативным в идиомах (*играться в игры / игрушки, в бирюльки* — 'заниматься несерьезным делом'): *Вторые лениво игрались в бирюльки, создавая на пустом месте громкие репутации, затем теряя интерес к вчерашним героям…* [А. Архангельский. Господа Гексогены (2002) // «Известия», 2002.04.12]; *Не совершают они акт вандализма. И не играются в игрушки. Они постигают философию, хотят они того или нет.* [Т. Соломатина. Акушер-ХА! Байки (2009)].

«Играются» животные, дети, влюбленные и те, кто им уподобляется: *Коза Бубенчик весело бежала и, играясь, в шутку бодала ребят рожками.* [В. Постников. Путешествие Карандаша и Самоделкина (1995)]; *Наша кошка ужасно любит играться с елочным дождем*. [На заметку // «Даша», 2003]; *Тут бегала, игралась в придорожной пыли вольная, чудилось, бездомная ребятня*. [О. Павлов. Дело Матюшина (1996)]; *Детьми с утра не займешься, скажем так, на то они и дети, чтобы беситься, резвиться и играться с утра пораньше!* [Женщина + мужчина: Брак (форум) (2004)]; *А до осени валяй домой, с женой играйся, чай пей, — плохо тебе?* [Ю. Трифонов. Утоление жажды (1959–1962)]; *Ты просто играешься в игрушку. Я.: Да*. [Л. Якубович, А. Политковский, А. Мазепа. Очень мужской диалог о детях // «Общая газета», 1998]; *… новый курс нашей политики будет «твердый»: мол, давайте открывайте Второй фронт, хватит играться*. [Г. С. Эфрон. Дневники. Т. 2 (1941–1943)]; *— Вот что, липовый друг, я тебе не Литвинов, а ты мне не Чемберлен, и нам с тобой нечего в дипломатию играться.* [М. А. Шолохов.

Поднятая целина. Книга 2 (1960)]; … *правда и то, что Марамзин, однажды следуя рейсом Аэрофлота с востока на запад, играл фигой в кармане под облаками на всю страну*. [В. Губин. А вы куда, ребята? (2002) // «Звезда», 2003]. «Играются», притворяясь, создавая иллюзию: *Они его себе придумали и сами в него играются.* [Я люблю мир, которого вообще нет // «Дружба народов», 1999.09.15]. «Играют» на бирже, в политике при достойном поведении сторон, в серьезном спорте, в юриспруденции, в образовании, в искусстве.

В переносном употреблении значения 'забавляться', 'имитировать' трансформируются в (1) 'издеваться, проявлять власть': *Мы можем играться с ним: навещать или не навещать*. [В. Месяц. Лечение электричеством // «Урал», 2002]; … *у меня создавалось впечатление, что ... существуют генералы, которые «играются» армией, как дети кубиками: они не удосужились даже основательно проработать все детали создания МС на базе ВДВ с командующими видами Вооруженных сил и родами войск.* [В. Баранец. Генштаб без тайн. Книга 1 (1999)]; *Синьор Тодеро тащит у него на веревочке игрушечный макет декорации Алекси-Месхишвили. Сатрап играется.* [Прореха на человечестве // «Культура», 2002.03.25]; (2) 'экспериментировать, рисковать, действуя необдуманно и легкомысленно': *Поэтому иерархам РПЦ не стоило бы играться с авторитетом собственной институции и заживо хоронить патриарха*. [В. Портников. Кирилл и Мефодий // «ПОЛИТКОМ.РУ», 2003.05.20]; *Я бы не играл с подзаголовком «Опыт исторического психоанализа», так как из-за него теряется близкий сердцу автора идеал «монизма»*. [А. Асмолов. Назад — к методологии психологии // «Вопросы психологии», 2004.06.15]; *И пока Путин играется словами о реформах, в Вооруженных Силах происходит стремительный процесс «накопления» развала.* [В. Шурыгин. Армия эпохи «реставрации» // «Завтра», 2003.01.05]; *Как депутаты с баннером и скотчем в Верховной Раде «игрались»* [fakty.ictv.ua/ru/index/read-news/id/1552527].

Из переносного 'экспериментировать' в современной речевой практике возникает особое употребление: *играться* — 'тестировать устройство': *Пару месяцев назад играл с Панасоником — показался неплохой мыльницей.* [Выбор цифрового фотоаппарата (2007–2008)]; возможно, происходит это по аналогии с компьютерными играми: *Ты фар край-то установил? Играл в него уже? Нет еще.* [ICQ переписка Gladiator и Serge (2007)]. На базе этого употребле-

ния формируется вариант семантического перехода 'анализировать структуру объекта, изучать, разрабатывать —> применять объект, использовать на практике': *Чтобы правильно использовать их, нужно понять идеалы и страшилки данного конкретного народа и потом «играться» с ними*. [В. Ляпоров. Хождение в народ // «Бизнес-журнал», 2004.03.16]; <Япония> *Уран обогащает, плутоний выделяет, тритий собирает, литий-6 производит, с бридерами возится, с нейтронными источниками играется.* [коллективный. Иранская ядерная программа (2010–2011)].

Наконец, прослеживается слэнговое употребление — возвратность используется для снижения значимости названного процесса, формулируется этикетное самоуничижение говорящего: *3 часа играл с новым айфоном… Что я вам могу сказать? Слава богу, что Стив Джобс не дожил до этого бездарного безобразия*. [https://twitter.com/surkovrussia]; *весь день играл с новым лазерным нивелиром — понял, что перепад на участке около 50 см на 30 метрах. разметил размеры котлованов* [https://www.forumhouse.ru/threads/302053/]; *Играл с новым HSDPA модемом. Поставил в него симку из телефона, где она до этого благополучно работала.* [kiyan.dp.ua/misc/2011/02/17/o-strannov.html]; *С дизайном вывода буду играться сам* [forum.vamshop.ru › … › Коммерческие предложения].

Противопоставление возвратного и невозвратного коррелятов как нейтрального и оценочного отчетливо видно в рамках одного контекста: *Одни играли, другие — игрались*. [О. Лысенко. Футбол. 29 июня 2009 года, понедельник. 11:00. http://www.championat.com].

Таким образом, рефлексив «играться» занимает свою семантическую нишу, вписываясь как в систему значений возвратных глаголов (образуя, как минимум, реципрок, объектный имперсонал и рефлексивный бенифактив), так и в коммуникативную структуру оценочно маркированных высказываний.

Глагол «убираться». Для разговорного «убираться» обнаруживаются два основных варианта употребления.

1. Косвенно-возвратное (рефлексивно-бенефактивное): *Леонид был на службе, попа увезли куда-то с требой, а попадья убиралась в кухне.* [Д. Мамин-Сибиряк. Братья Гордеевы (1891)]; **—** *Наташка, будешь убираться в конторе, так пригляди, куда прячет Андрон Евстратыч ключ от железного сундука,* **—** *наказывала она перед отъ-*

ездом. [Д. Мамин-Сибиряк. Золото (1892)]; *Воспитанники встают рано и сами, все вместе, убираются, чистят камеры и, когда надо, моют полы*. [Ф. Достоевский. Дневник писателя. 1876 год (1876)]. Для этого значения важна сема «зоны субъекта», идея «действия, совершаемого в своих интересах»: … *прежде чем пойти умываться, в пижаме и лаковых ночных туфлях, стал убираться в комнате, убрал со стола, поставил на книжный шкаф недопитую бутылку красного вина, вазу с печеньем поставил на книжный шкаф, на нижнюю полку, перебрал на столе пепельницу, чернильницу, блокноты, книги.* [Б. Пильняк. Повесть непогашенной луны (1926)]. Ср. в переносном употреблении: *А вы не думаете, что люди не хотят убираться в стране потому, что понимают: после их уборки придет кто-нибудь другой и обязательно поломает их «снеговиков»?* [Е. Гусятинский, Н. Михалков. Разные уровни лжи // «Русский репортер», № 15 (143), 22–29 апреля 2010, 2010]

2. Типичное для субъекта действие, качественная характеристика субъекта по действию (безобъектно-возвратное, или объектно-имперсональное, употребление): *Не похоже ли это удивление на наивный разговор домашних хозяек, которые, изо дня в день убираясь в квартире, неизменно восклицают: «Откуда только проклятая пыль берется?* [А. Мариенгоф. Мой век, мои друзья и подруги (1956–1960)]; — *У нас одна женщина убирается в архиве,— вот и Анюша, тихо-тихо.* [М. Вишневецкая. Вышел месяц из тумана (1997)], *Катерина Петровна приходила из института, она у отца на кафедре двадцать лет убиралась, даже денег не хотела с меня брать*. [А. Берсенева. Полет над разлукой (2003–2005)]. Так, едва ли допустимо квалификативное **Она хорошо убирает, но плохо готовит*, при изолированном употреблении невозвратного «убирать», тогда как вариант с возвратным коррелятом вроде бы менее аномален, во всяком случае, употребителен, ср.: *Ну, вот в семье я лучше всех убираюсь*. [Эльвира Савкина. «Я — специалист по наведению порядка» // «Дело» (Самара), 2002.06.25].

Возвратный глагол активен в неопределенно-личном употреблении: *Или некому проследить, чтоб у начальника в кабинете убирались?* [Семен Данилюк. Бизнес-класс (2003)]; *Здесь давно не убирались, но со стены можно содрать плакат, а с крючка справа вверху — снять ремень генератора.* [коллективный. Отель «У погибшего альпиниста» (2008–2010)].

Заключение. В отношении глаголов «играться» и «убираться» прослеживаются три варианта оценки нормативности. Первый: оба глагола так же недопустимы, как «кушать» вместо «есть» применительно к действию взрослого здорового человека (вне императивных конструкций), «горячее кофе» (в среднем роде), «звОнит» (с ударением на первом слоге). Это ряд экземплификативов (типичных примеров) нарушения узко понимаемой нормы, причем в отношении каждого элемента этого ряда легко находится объяснение, проясняющее природу вариативности и широкое распространение в узусе употребления, признаваемого некорректным с точки зрения предписывающей нормы. Второй: «убираться» — естественный вариант периферии литературной нормы, единица разговорного регистра, глагол с косвенно-возвратной (ср. «стираться» ('стирать белье'), «закупаться», «стелиться» ('стелить постель'), «складываться») или объектно-имперсональной семантикой; тогда как «играться» лишен переходного коррелята и потому относится к просторечию, функционирует за гранью литературной нормы (этой интерпретации противоречит как активность несоотносимых с переходным коррелятом возвратных образований в рамках нормативного идиома, ср. *хвастаться, стучаться, коситься, грозиться, целиться*, так и семантическая и коммуникативная спецификация «играться» по отношению к исходному «играть»). Третий, наиболее вероятный: оба варианта принадлежат разговорному регистру, реализуют естественные потенции языковой системы и более или менее отчетливо вступают в отношения семантического и прагматического распределения с производящими.

Словари

Бабенко Л. Г. Толковый словарь русских глаголов: Идеографическое описание. Английские эквиваленты. Синонимы. Антонимы. М.: Аст-Пресс, 1999.

Кузнецов С. А. Большой толковый словарь русского языка. СПб.: Норинт, 1998.

Ефремова Т. Ф. Новый толково-словообразовательный словарь русского языка. М.: Дрофа, Русский язык, 2000.

Ефремова Т. Ф. Современный толковый словарь русского языка. В 3 томах. М.: АСТ, Астрель, Харвест, 2005.

МАС — Словарь русского языка. Под ред. А. П. Евгеньевой. В 4 томах. М.: Русский язык, 1985.

Ожегов С. И. Словарь русского языка. М.: Русский язык, 1986.

Ушаков Д. Н. Толковый словарь русского языка. М.: Гос. изд-во иностр. и нац. словарей, 1935–1940.

Литература

Аванесов Р. И., Сидоров В. Н. Очерк грамматики русского литературного языка. Ч. I. Фонетика и морфология. М.: Учпедгиз, 1945.

Вариясова Е. В., Комарова Л. Э., Рябкова Н. И., Сапрыкина Е. Н. Культура речи в сфере связей с общественностью. Учебное пособие для студентов вузов. Сургут: Изд-во СГУ, 2007.

Виноградов В. В. Русский язык: грамматическое учение о слове. М.: Учпедгиз, 1947.

Генюшене Э. Ш. Рефлексивные глаголы в балтийских языках и типология рефлексивов. Дисс. … докт. филол. наук. Вильнюс, 1983.

Гото К. В., Сай С. С. Частотные характеристики классов русских рефлексивных глаголов // Корпусные исследования по русской грамматике. / Ред.-сост. К. Л. Киселева, В. А. Плунгян, Е. В. Рахилина, С. Г. Татевосов. М.: Пробел, 2009. С. 184–223.

Друговейко-Должанская С. В. Как правильно убираться? // Культура письменной речи. Ликбез от gramma.ru-4. [Электронный ресурс]. Адрес доступа: http://gramma.ru/RUS/?id=13.55. Дата обращения: 8.07.15.

Летучий А. Б. Возвратность // Русская корпусная грамматика. 2010. [Электронный ресурс]. Адрес доступа: http://rusgram.ru/index. Дата обращения: 8.07.15.

Летучий А. Б. Не о многом, но много: особые синтаксические свойства лексем типа много // Глагольные и именные категории в системе функциональной грамматики. СПб.: Нестор-История, 2013. С. 149–159.

Никитина Е. Н. Акциональность / неакциональность возвратных глаголов и категория субъекта (к грамматической сущности категории залога). Дисс. … канд. филол. наук. М., 2008.

Обнорский С. П. Возвратные глаголы на -ся в русском языке // Известия Академии Наук СССР. Отделение литературы и языка. Т. XIX. Вып. 1. М., 1960. С. 19–21.

Орлова Н., Дмитриева О., Павлова Н. Русский язык и культура речи. Учебное пособие. М.: Флинта, 2014.

Розенталь Д. Э. Справочник по правописанию и литературной правке. Под ред. И. Б. Голуб. М.: Айрис-Пресс, Рольф, 2001.

Розенталь Д. Э., Джанджакова Е. В., Кабанова Н. П. Справочник по правописанию, произношению, литературному редактированию. М.: ЧеРо, 1999.

Русский язык и культура речи. Материалы для подготовки слушателей по специальности переподготовки 1–21 06 74 «Современный иностранный язык (экономическая деятельность)» учреждения образования «Белорусский торгово-экономический университет потребительской кооперации» (кафедра экономических и правовых дисциплин). Гомель, 2013.

Сай С. С. К вопросу о семантической деривации русских возвратных глаголов // Известия Российского государственного педагогического университета им. А. И. Герцена. 2008. № 65. С. 285–289.

Улуханов И. С. Возвратные глаголы // Караулов Ю. Н. (ред.) Русский язык. Энциклопедия. М.: БРЭ, Дрофа, 1998. С. 72–73.

Шелякин М. А. Справочник по русской грамматике. М.: Русский язык, 2000.

Цейтлин С. Н. Возвратные глаголы и детская речь // Проблемы теории грамматического залога. Л.: Наука, 1978. С. 128–130.

Ясаи Л. О морфологической немаркированности и проблеме наименования // Studia Slavica Hung. 2005. № 50. P. 371–378.

Yana E. Akhapkina
National Research University «Higher School of Economics»
(Russia, Moscow)
yana.akhapkina@gmail.com

REFLEXIVE VERBS «УБИРАТЬСЯ» (UBIRAT'SYA) AND «ИГРАТЬСЯ» (IGRAT'SYA): CODIFICATION AND DISCOURSE PRACTICE

The article based on the Russian National Corpus and Internet usage analyses functioning of colloquial verbs *ubirat'sya* and *igrat'sya*; types of meanings and peculiarities of contextual usage of these units tradi-

tionally interpreted as non-normative are under scrutiny; the usage of reflexives that refer and do not refer to the initial transitive verbs they are formed from is compared. According to S. Sai, the functioning of Russian reflexive verbs system is determined not only by semantic relations that link certain reflexive verbs with their non-reflexive derivational bases, but also by associative relations within different semantic groups of reflexive verbs, irrespective of their derivational characteristics. The article attempts to clarify if there are reasons to reject the verb *igrat'sya* as incorrect on the basis of its derivational base *igrat'* being intransitive verb. Other reflexive verbs motivated by intransitive verbs are analyzed including reflexives created occasionally by analogy with normative derivatives. In this analysis we apply S.P. Obnorsky's idea on *-s'a* as intransitivity marker making this meaning morphologically expressed. Most probably such verbs as *igrat'sya* and *ubirat'sya* belong to colloquial register, implement natural Russian linguistic system potentiality and are more or less explicitly differentiated from their derivational bases in terms of semantics and pragmatics.

Key words: grammar, norm, discourse practice, context, reflexivity, transitivity, voice

References

Babenko L. G. *Tolkovyi slovar' russkikh glagolov: Ideograficheskoe opisanie. Angliiskie ekvivalenty. Sinonimy. Antonimy* [Dictionary of Russian verbs]. Moscow, AST-Press Publ., 1999.

Kuznetsov S. A. *Bol'shoi tolkovyi slovar' russkogo yazyka*. [Big Dictionary of the Russian Language]. St. Petersburg, Norint Publ., 1998.

Efremova T. F. *Novyi tolkovo-slovoobrazovatel'nyi slovar' russkogo yazyka*. [New Explanatory and Derivational Russian Dictionary]. Moscow, Drofa Publ., Russkii yazyk Publ., 2000.

Efremova T. F. *Sovremennyi tolkovyi slovar' russkogo yazyka* [Modern Russian Dictionary]. In 3 v. Moscow, AST, Astrel', Kharvest Publ., 2005.

MAS – *Slovar' russkogo yazyka.* [Russian Dictionary]. A. P. Evgen'eva (ed.). In 4 v. Moscow, Russkii yazyk Publ., 1985.

Ozhegov S. I. *Slovar' russkogo yazyka* [Russian Dictionary]. Moscow, Russkii yazyk Publ., 1986.

Ushakov D. N. *Tolkovyi slovar' russkogo yazyka* [Russian Dictionary]. Moscow, Gos. izd-vo inostr. i nats. slovarei, 1935–1940.

Avanesov R. I., Sidorov V. N. *Ocherk grammatiki russkogo literaturnogo yazyka. Ch. I. Fonetika i morfologiya* [A sketch of Russian Grammar]. Moscow, Uchpedgiz, 1945.

Variyasova E. V., Komarova L. E., Ryabkova N. I., Saprykina E. N. *Kul'tura rechi v sfere svyazei s obshchestvennost'yu. Uchebnoe posobie dlya studentov vuzov* [The speech culture in public relations]. Surgut, Surgut State Univ. Publ., 2007.

Vinogradov V. V. *Russkii yazyk: grammaticheskoe uchenie o slove.* [The Russian language: grammatical doctrine of the word]. Moscow, Uchpedgiz, 1947.

Genyushene E. Sh. *Refleksivnye glagoly v baltiiskikh yazykakh i tipologiya refleksivov. Diss. ... dokt. filol. nauk* [Reflexive verbs in Baltics and the typology of reflexives]. Vil'nyus, 1983.

Goto K. V., Sai S. S. [Frequency characteristics of Russian reflexive verbs]. *Korpusnye issledovaniya po russkoi grammatike* [Corpus studies on Russian Grammar]. K.L. Kiseleva, V.A. Plungian, E.V. Rakhilina, S.G. Tatevosov (eds.). Moscow, Probel Publ., 2009, pp. 184–223. (In Russ.)

Drugoveiko-Dolzhanskaya S. V. [How should we "*ubirat'sya*"?]. *Kul'tura pis'mennoi rechi. Likbez ot gramma.ru-4*. Available at URL: http://gramma.ru/RUS/?id=13.55. (accessed on 8.07.15).(In Russ.)

Letuchii A. B. [Reflexivity]. *Russkaya korpusnaya grammatika.* 2010. Available at URL: http://rusgram.ru/index. (accessed on 8.07.15). (In Russ.)

Letuchii A. B. [Syntactic properties of words like "*mnogo*"]. *Glagol'nye i imennye kategorii v sisteme funktsional'noi grammatiki* [Verbal and nominal categories in the functional grammar]. St. Petersburg, Nestor-Istoriya Publ., 2013, pp. 149–159. (In Russ.)

Nikitina E. N. *Aktsional'nost' / neaktsional'nost' vozvratnykh glagolov i kategoriya sub"ekta* (k grammaticheskoi sushchnosti kategorii zaloga). [(Un)actionality of reflexive verbs and a subject category]. Phd. diss. Moscow, 2008.

Obnorskii S. P. [Reflexive verbs with -*sja* in Russian]. *Izvestiya Akademii Nauk SSSR. Otdelenie literatury i yazyka*, 1960, V. XIX. Issue 1, pp. 19–21. (In Russ.)

Orlova N., Dmitrieva O., Pavlova N. *Russkii yazyk i kul'tura rechi. Uchebnoe posobie* [The Russian language and the culture of speech]. Moscow, Flinta Publ., 2014.

Rozental' D.E. *Spravochnik po pravopisaniyu i literaturnoi pravke.* I. B. Golub (ed.) [Handbook of spelling and literary editing]. Moscow, Airis-Press, Rol'f Publ., 2001.

Rozental' D. E., Dzhandzhakova E. V., Kabanova N. P. *Spravochnik po pravopisaniyu, proiznosheniyu, literaturnomu redaktirovaniyu* [Handbook of spelling, pronunciation, literary editing]. Moscow, CheRo Publ., 1999.

Russkii yazyk i kul'tura rechi. Materialy dlya podgotovki slushatelei po spetsial'nosti perepodgotovki 1–21 06 74 "Sovremennyi inostrannyi yazyk (ekonomicheskaya deyatel'nost')" uchrezhdeniya obrazovaniya "Belorusskii torgovo-ekonomicheskii universitet potrebitel'skoi kooperatsii" (kafedra ekonomicheskikh i pravovykh distsiplin). [Russian language and the culture of speech]. Gomel', 2013.

Sai S. S. [The question of the semantic derivation of Russian reflexive verbs]. *Izvestiya Rossiiskogo gosudarstvennogo pedagogicheskogo universiteta im. A.I. Gertsena.* 2008, no. 65, pp. 285–289. (In Russ.)

Ulukhanov I.S. [Reflexive verbs]. Karaulov Yu. N. (red.) *Russkii yazyk.* Entsiklopediya. [Encyclopedia "The Russian language"]. Moscow, BRE, Drofa Publ., 1998, pp. 72–73. (In Russ.)

Shelyakin M.A. *Spravochnik po russkoi grammatike* [Guide to Russian grammar]. Moscow, Russkii yazyk Publ., 2000.

Tseitlin S.N. [Reflexive verbs and baby talk]. *Problemy teorii grammaticheskogo zaloga.* [Theoretical problems of grammar voice]. Leningrad, Nauka Publ., 1978, pp. 128–130. (In Russ.)

Yasai L. [About morphological unmarkedness and a nomination problem]. *Studia Slavica Hun*g. 2005, no. 50, pp. 371–378. (In Russ.)

Д. О. Добровольский
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
dm-dbrv@yandex.ru

КОРПУС ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ И СОПОСТАВИТЕЛЬНАЯ ЛЕКСИКОЛОГИЯ[1]

В статье рассматриваются языковые явления, представляющие серьезные трудности как для контрастивной лексикологии, так и для двуязычной лексикографии. Цель исследования — попытаться ответить на вопрос, насколько корпусы параллельных текстов могут оказаться полезным инструментом в описании особенностей функционирования форм русского языка в сопоставлении с другим, и наоборот. Работа основана на эмпирических данных русско-английского, русско-немецкого, а также англо-русского и немецко-русского параллельных корпусов НКРЯ. Анализ показал, что обращение к параллельным корпусам помогает выявить как нетривиальные семантические и дискурсивные особенности слов, не имеющих в других языках стандартных эквивалентов, так и специфику их кросс-лингвистических функциональных соответствий.

Ключевые слова: параллельный корпус, лексическая семантика, двуязычная лексикография, лингвоспецифичная лексика, русский язык, английский язык, немецкий язык

Цель данной работы — попытаться ответить на вопрос, насколько корпусы параллельных текстов могут оказаться полезным ин-

[1] Работа выполнена при поддержке РФФИ (грант 13-06-00403), РГНФ (гранты 15-04-00507 и 15-04-12018) и Фонда содействия развитию интернета «Internet Support Foundation». Особая благодарность Анне А. Зализняк за полезные замечания и советы.

струментом в решении задач контрастивной лингвистики, то есть в описании особенностей функционирования форм русского языка в сопоставлении с другим (например, с немецким и с английским, как в данной статье), и наоборот. В центре нашего внимания стоят лексические явления сопоставляемых языков, в частности проблемы межъязыковой эквивалентности, связанные с лингвоспецифичностью некоторых лексем, и способы их представления в двуязычных словарях. Материал параллельных корпусов провоцирует вопросы лексикографического представления[1] межъязыковых лексических коррелятов, особенно если подобные соответствия не являются эквивалентами в точном смысле.

1. Корпус параллельных текстов с английским и немецким языками в Национальном корпусе русского языка

Создание корпуса параллельных текстов в качестве одного из модулей Национального корпуса русского языка (НКРЯ) — актуальная лингвистическая задача, над решением которой вот уже больше десяти лет работает целый коллектив исследователей. Такой корпус необходим в первую очередь потому, что надежные, подкрепленные эмпирическими данными ответы на большинство вопросов сопоставительной лингвистики, в том числе лексикологии и лексикографии, могут быть получены только при наличии значительных по объему и общедоступных параллельных корпусов. Ср. подробнее [Добровольский, Кретов, Шаров 2005; Добровольский 2009].

Корпус параллельных текстов НКРЯ состоит в основном из множества художественных произведений с их переводами. В относительно небольшом объеме в корпус включены тексты СМИ. Корпус включает в себя, с одной стороны, оригинальные русские тексты с их переводом на другие языки, а с другой — иноязычные тексты с их переводом на русский. В дальнейшем мы будем говорить исключительно о русско-английских, англо-русских, русско-немецких и немецко-русских частях корпуса.

[1] Вопросы использования корпусных данных в — прежде всего, одноязычной — лексикографии обсуждаются в [Atkins, Fillmore, Johnson 2003]. Ср. также спецвыпуск журнала «Sprachtypologie und Universalienforschung», посвященный использованию параллельных корпусов в языковом сравнении и типологии [Cysouw, Wälchli 2007]. На эту публикацию мое внимание обратил Д.В. Сичинава.

1.1. Зачем нужен параллельный корпус?

Создание и использование параллельных корпусов представляется целесообразным как с практической точки зрения, так и с точки зрения развития корпусной лингвистики — одного из наиболее перспективных лингвистических направлений. Корпус параллельных текстов может быть эффективно использован как в различных лингвистических исследованиях сопоставительного характера, так и в исследованиях по теории перевода, сравнительному литературоведению, культурологии, автоматической обработке текста. В этой работе нас будут в первую очередь интересовать собственно лингвистические — и в том числе лексикографические — аспекты использования параллельных корпусов.

Обращаясь к корпусу, создатели двуязычных словарей получают весьма простой и эффективный инструмент сбора материала и эмпирической проверки своих гипотез, касающихся межъязыковой эквивалентности. Ценность этого инструмента определяется тем, что в лингвистике этап сбора материала является наиболее трудоемким и наименее творческим, а корпус параллельных текстов позволяет значительно сэкономить время и силы для творческого этапа работы.

Традиционная сравнительная лексикология и двуязычная лексикография характеризуется ориентацией на сопоставление более или менее изолированных языковых структур. Отрицательным последствием подобной ориентации является недостаточный учет узуса, то есть тех особенностей синтаксического и сочетаемостного поведения единиц языка, которые нельзя объяснить их системными признаками. Так, в принципе известно, что та или иная структура одного языка не может быть во всех контекстах переведена на другой с помощью своего стандартного эквивалента. В определенных контекстах язык L2 прибегает к другим способам описания соответствующей ситуации. Известно также, что не существует продуктивных правил, по которым можно было бы вывести подобные отклонения от «стандартной эквивалентности» из неких более общих принципов. Единственный способ описания подобных отклонений — это их тщательная фиксация на аутентичном материале. Только так можно построить исчерпывающие сопоставительные описания и создать словари, удовлетворяющие современным требованиям.

Корпус параллельных текстов представляет собой наиболее адекватный инструмент для выполнения этих задач. Та или иная языковая структура, интересующая исследователя, может быть найдена во всех представленных в корпусе контекстах с их переводами на соответствующий язык. Таким образом, исследователь получает в свое распоряжение набор аутентичных контекстов, представляющих интересующую его структуру в ее естественном окружении, а также самые разнообразные эквиваленты этой структуры в языке цели. Поскольку эти эквиваленты также оказываются встроенными в естественные контексты, на основе полученных с помощью параллельного корпуса материалов могут быть сделаны выводы о зависимости выбора эквивалента от типа контекста. Подобные результаты практически всегда расходятся с теми сведениями, которые мы можем почерпнуть из традиционных, созданных в «докорпусную эру» словарей, являясь тем самым нетривиальными.

Важным параметром, по которому языки могут различаться между собой, является степень употребительности определенных выражений. Так, некоторое выражение *A* языка L1 может стандартным образом переводиться на язык L2 с помощью выражения *B* — вполне корректного с точки зрения норм этого языка. Таким образом, выражения *A* и *B* оказываются эквивалентными в рамках системы соответствующих языков. Тем не менее, их функциональная эквивалентность часто представляется неполной. В частности, это имеет место в случае, когда одно из выражений оказывается в своем языке существенно более употребительным, чем его переводной эквивалент — в своем. Такие случаи хорошо прослеживаются при сопоставлении оригинальных текстов с их переводами на другие языки; ср., например, [Йокояма 2012: 152].

Исследование некоторого языкового явления на основе корпуса параллельных текстов (особенно если рассматриваются феномены языка L2)[1] может быть по ряду параметров противопоставлено исследованию этого явления на основе большого корпуса оригинальных текстов. Отличие оригинальных текстов от переводов заключается и в объеме (миллионы слов оригинальных текстов, производимых ежедневно, против относительно небольшого количества текстов, переводимых с иностранных языков), и в природе автор-

[1] Например, некоторое явление русского языка на материале англо-русского или немецко-русского корпусов.

ства (то есть в степени оригинальности и творческой свободы при порождении текста), а также в культурном контексте (переводные тексты обычно погружены в культуру исходного языка). Все эти факторы обеспечивают различия между исходными и переводными текстами по целому ряду параметров.

Между конкретными словами с близкими значениями в разных языках (которые принято считать эквивалентными) нет взаимно однозначных соответствий. В принципе это касается многих семантических классов слов. Каждое конкретное слово обладает своим уникальным набором сочетаемостных ограничений и предпочтений. Эта идея относится к числу традиционных теоретических положений лингвистики. То новое, что вносит обращение к корпусам параллельных текстов в решение подобных задач,— это возможность эмпирической проверки соответствующих гипотез на представительном материале.

1.2. Состав немецко-русского и русско-немецкого корпуса в НКРЯ

На данный момент в состав доступного online немецко-русского корпуса (наряду с небольшим массивом публицистических текстов, любезно предоставленных компанией ABBYY) входит около тридцати произведений художественной литературы. В стадии подготовки находится роман Теодора Фонтане «Эффи Брист» (Theodor Fontane — Effi Briest), а также статья Вальтера Беньямина «О понятии истории» (Walter Benjamin — Über den Begriff der Geschichte). Работа над русско-немецким корпусом была начата относительно недавно. Пока он существенно уступает по объему немецко-русскому. Сейчас (на июнь 2014 года) в открытом доступе находятся лишь пять публицистических и пятнадцать художественных текстов[1]. В ближайшее время корпус будет дополнен следующими произведениями русской классической литературы:

— Н. В. Гоголь:

Вечера на хуторе близ Диканьки — Abende auf dem Vorwerke bei Dikanjka und andere Erzählungen;

Портрет — Das Porträt;

[1] С составом корпуса можно ознакомиться на сайте http://www.ruscorpora.ru/mycorpora-para.html.

Записки сумасшедшего — Aufzeichnungen eines Irren;
Нос — Die Nase;
Невский проспект — Der Newskij-Prospekt;
Вий — Der Wij;
— Ф. М. Достоевский:
Братья Карамазовы — Die Brüder Karamasow;
Бедные люди — Arme Leute;
— А. С. Пушкин:
Капитанская дочка — Die Hauptmannstochter (в переводе Фреда Отто — Fred Ottow)[1];
— Л. Н. Толстой:
Анна Каренина — Anna Karenina;
Крейцерова соната — Die Kreutzersonate;
Семейное счастие — Glück der Ehe;
Хозяин и работник — Herr und Knecht;
— И. С. Тургенев:
Фауст — Faust;
— А. П. Чехов:
Дуэль — Ein Zweikampf;
— Н. С. Лесков:
Леди Макбет Мценского уезда — Die Lady Makbeth des Mzensker Landkreises;
Запечатленный ангел — Der versiegelte Engel
— И. А. Гончаров:
Обломов — Oblomow;
— Максим Горький:
Трое — Drei Menschen.

1.3. Состав англо-русского и русско-английского корпуса в НКРЯ

Работа над параллельными корпусами с английским языком, была начата существенно раньше, чем с немецким. Соответственно, объем англо-русского и русско-английского корпусов намного превосходит объем немецко-русского и, в особенности, русско-немец-

[1] Параллельный корпус по «Капитанской дочке» в другом немецком переводе, любезно предоставленный проекту НКРЯ А. Барентсеном, уже размещен на сайте НКРЯ.

кого. Сейчас в НКРЯ доступны online 146 англо-русских и 77 русско-английских текстов. Более подробно см. на сайте http://www.ruscorpora.ru/mycorpora-para.html. В стадии подготовки находятся следующие параллельные тексты (произведения русской литературы с их английскими переводами):

— Ф. М. Достоевский:
Братья Карамазовы — The Brothers Karamazov;
Белые ночи — White Nights;
Бесы — The Possessed or, The Devils;
Преступление и наказание — Crime and Punishment;
— В. В. Набоков:
Звонок — The Doorbell;
Письмо в Россию — A Letter that Never Reached Russia;
Подлец — An Affair of Honor;
Бритва — Razor;
Возвращение Чорба — The Return of Chorb;
Драка — The Fight;
Пассажир — The Passenger;
Путеводитель по Берлину — A Guide to Berlin;
Рождество — Christmas;
Ужас — Terror;
Сказка — A Nursery Tale.

В следующих разделах я попробую на конкретных примерах показать возможности использования корпуса для сопоставительного описания лексических единиц и, в частности, для исследования лингвоспецифичных слов и выражений того или иного языка.

2. Русское *нет* как лингвоспецифичное слово

Корпусное исследование лексических единиц разных языков, традиционно считавшихся эквивалентными, часто позволяет выявить ряд нетривиальных семантических различий. Так, в [Добровольский, Левонтина 2009б; Dobrovol'skij, Levontina 2009; 2012] было показано, что у русского слова *нет* есть специфические значения, которые отсутствуют у немецкого слова *nein* или у английского *no*. Прежде всего бросаются в глаза случаи, когда русскому *нет* соответствует немецкое *doch*, а не *nein*, то есть случаи ответа на вопрос, содержащий отрицание, или реакции на реплику с отри-

цанием. Ср.: (а) — *Ich habe gehört, ihr wart gar nicht weg?* — ***Doch***. / — *Я слышал, вы и не уезжали никуда?* — ***Нет***, *уезжали*. (б) — *Das stimmt nicht!* — ***Doch***. / — *Это неправильно!* — ***Нет***, *правильно*.

В русском языке для выражения положительного vs. отрицательного согласия-несогласия есть только два слова: *да* и *нет*. Причем в нормированной речи *да* употребляется только в случае «положительного согласия» (i), а во всех остальных случаях — *нет*. Ср.: (i) — *Будешь?* — ***Да***, *буду*; (ii) — *Будешь?* — ***Нет***, *не буду*; (iii) — *Не будешь?* — ***Нет***, *буду*; (iv) — *Не будешь?* — ***Нет***, *не буду*. Строго говоря, выбор *да* и *нет* в русском языке зависит не только от стимула (в частности, от того, содержит ли предшествующий вопрос отрицание или нет), но и от ситуации и контекстных условий. В реальной коммуникации в случаях типа (iii) и (iv) встречается и *да*. Е. А. Гришина [2011: 173] указывает на то, что «в живой устной речи не так уж много явлений, которые запрещены бескомпромиссно». Кроме того, из-за полифункциональности русского *нет* реакция типа (iii) и (iv) содержит повтор релевантного элемента стимула. Одиночное *нет* встречается в таких случаях крайне редко. Случаи типа (iii)) и (iv) обсуждаются также в [Добрушина 2003:182]. Ср. ее пример на тип (iv): — *Мне не звонили?* — ***Нет***. / ***Нет***, *не звонили*. Что касается реакций типа (iii), единственно возможным способом «введения положительной пропозиции в ответ на такие вопросы» Е. Р. Добрушина считает повтор релевантного элемента стимула, то есть форму — *Мне не звонили?* — ***Звонили***. Мне представляется, что реакция типа ***Нет***, *звонили* в реальной коммуникации вполне уместна. Таким образом, из четырех возможных случаев выражения положительного vs. отрицательного согласия-несогласия, как уже было сказано, в трех — (ii), (iii) и (iv) — более распространенным способом является *нет*, и только в одном — *да*[1].

В немецком языке (так же, как в скандинавских и во французском) на четыре принципиально возможные реакции приходятся три слова. Соответственно один из трех случаев, в которых в рус-

[1] В работе [Pope 1976: 118 ff.] подробно обсуждается система ДА-/НЕТ-реакций на стимулы с отрицанием и без отрицания в разных языках. В принципе, выбор мотивируется двумя параметрами: «положительное vs. отрицательное утверждение» и «согласие vs. несогласие». Возможны четыре комбинации этих признаков. Но языков, в которых существовало бы четыре разных слова для выражения этих смыслов, крайне мало. К ним, например, относится английский язык XVI века.

ском языке употребляется *нет*, требует не *nein*, а именно этого «третьего» слова — *doch*. Это тип реакции (iv) — см. примеры (а) и (б) выше. В английском языке, где нет специального слова, эквивалентного немецкому *doch* или французскому *si*, в подобных случаях употребляется не *no*, как можно было бы ожидать, исходя из русского, а *yes*. Напомним знаменитый предвыборный лозунг Обамы *Yes, we can!*, обсуждаемый в [Dobrovol'skij, Levontina 2012] в связи с особенностями употребления русского *нет* в контекстах такого типа. Этот лозунг должен рассматриваться как реакция на гипотетический скепсис оппонентов: *They say, we are not ready, we cannot.* ***Yes****, we can!* Ср. *Нам говорили, что мы не готовы, что мы не можем*. ****Да****, мы можем!* при норме ***Нет****, мы можем!*

Случаи несовпадения употреблений *нет*, *nein* и *no*, обусловленные принятыми в разных языках способами выражения положительного vs. отрицательного согласия-несогласия, весьма интересны с точки зрения семантической типологии, но достаточно тривиальны, коль скоро речь идет об исследовании феномена лингвоспецифичности корпусными методами. Менее самоочевидными, не поддающимися описанию с помощью достаточно регулярных правил представляются употребления слова *нет* в функции дискурсивной частицы. Имеются в виду употребления типа *Нет, ты только представь себе…* Ср. подробнее [Добровольский, Левонтина 2009а; 2009б; Dobrovol'skij, Levontina 2012].

Для ответа на вопрос, является ли русская дискурсивная частица *нет* лингвоспецифичной лексической единицей, мы пользуемся одним из методов, предложенных в [Бунтман и др. 2014]. Суть этого метода состоит в том, что для лексических единиц, потенциально относящихся к лингвоспецифичным, определяется число их переводных эквивалентов, после чего оценивается дисперсия. Этот метод применим только к тем лингвоспецифичным словам, которые относятся к категории безэквивалентной лексики, при том что существуют и лингвоспецифичные лексемы, не связанные с идеей безэквивалетности. Что касается *нет*, эта частица представляет собой достаточно сложный с теоретической точки зрения случай. С одной стороны, отнесение *нет* к категории безэквивалентной лексики может натолкнуться на интуитивный протест: ведь немецкое слово *nein* и английское *no* выглядят (по крайней мере на первый взгляд) как бесспорные эквиваленты русского *нет*. С другой стороны, как

раз в функции дискурсивной частицы или прагматического коннектора (ср. — *Ну ты и вырядился! — **Нет**, а одежда-то моя здесь причем?*) *нет* не имеет столь однозначных эквивалентов в других языках, а немецкое *nein* и английское *no* оказываются далеко не самыми частотными и прагматически адекватными соответствиями. Таким образом, описанный выше метод исследования представляется в этом случае принципиально применимым.

В дальнейшем (в разделе 2.2) мы сконцентрируемся на описании различных способов перевода дискурсивного слова *нет* на немецкий и английский языки. Что касается вычисления частоты отдельных переводных эквивалентов и определения их разброса, то есть оценки дисперсии, этот этап анализа вызывает некоторые трудности, поскольку количество релеватных примеров в НКРЯ не позволяет получить статистически надежные результаты. Целесообразно, однако, испробовать потенциальную применимость этого метода на небольшом однородном фрагменте корпуса, что и будет сделано в разделе 2.1.

2.1. Case study: Переводные эквиваленты дискурсивного *нет* в английских переводах рассказов Набокова

В качестве подходящего материала для тестирования предложенного в [Бунтман и др. 2014] метода здесь предлагаются перечисленные в разделе 1.3 рассказы Владимира Набокова. Соответствующие параллельные тексты представляют собой обозримый фрагмент корпуса, однородный как с точки зрения авторского стиля, так и выбранных переводчиком стратегий (все рассказы переведены на английский Дмитрием Набоковым).

В этих рассказах отмечены лишь пять случаев употребления дискурсивного *нет*, причем релевантные контексты обнаружились лишь в двух из одиннадцати проанализированных рассказов: в «Подлеце» и в «Звонке». Тем более, показательно, что при их переводе использованы пять разных английских эквивалентов. Приведем эти примеры.

(1) — Понимаю,— сказал Митюшин, косясь на шахматную доску, над которой нависла, шевеля пальцами, рука Гнушке.— *Нет*, ты слушай меня,— с тоской воскликнул Антон Петрович. (В. Набоков. Подлец)

"Second what?" said Mityushin absently, glancing askance at the chessboard, over which Gnushke's hand hung, its fingers wriggling. "*No*, you listen to me," Anton Petrovich exclaimed with anguish in his voice.

(2) — *Нет*, ты слушай! Не будем больше пить. Это серьезно, серьезно. (В. Набоков. Подлец)

"You just listen! Let us not drink any more. This is serious, very serious."

(3) *Нет*, конечно, Берга полагается убить, но сначала нужно было хорошенько все продумать и, если выбирать секундантов, то уж, во всяком случае, порядочных людей,<…>. (В. Набоков. Подлец)

Oh, of course Berg must be killed; only the matter ought to have been carefully thought out first, and, if it had come to choosing seconds, they should in any case have been gentlemen.

(4) Представь себе,— господин Берг тоже струсил. *Нет*, не тоже струсил, а просто: струсил. (В. Набоков. Подлец)

Just imagine Mr. Berg also lost his nerve. *Well*, not also, but lost his nerve anyhow.

(5) — Да ты разве не слушаешь, мама? — *Нет*, что ты,— я слушаю, я слушаю… (В. Набоков. Звонок)

"Aren't you listening, Mother?", "*Why*, yes I am."

Примеры показывают, что в качестве эквивалента русского *нет* могут выступать по крайней мере следующие английские слова — частицы, местоимения и междометия: *no* (в примере (1), *oh* (3), *well* (4), *why* (5). В примере (2) *нет* не имеет лексического соответствия, то есть в данном случае мы имеем дело с 0-эквивалентностью. Такое разнообразие способов перевода позволяет считать гипотезу, что русское *нет* в функции дискурсивной частицы является лингвоспецифичной лексемой, достаточно обоснованной.

2.2. Режимы употребления дискурсивного *нет* и их немецкие и английские соответствия

Задача данного раздела — выявить некоторые из основных типов употреблений слова *нет* в качестве дискурсивной частицы и посмотреть на материале параллельного корпуса, какими способами это слово переводится на немецкий и английский языки. Интересно также проследить, зависит ли способ перевода от того или

иного режима употребления *нет*. При этом используется не только корпус, в котором русский язык является исходным, но данные в направлении от английского или немецкого языков к русскому. Причина не только в том, что англо-русский и немецко-русский корпусы превосходят по объему русско-английский и русско-немецкий, но и в том, что при «инвертированном» подходе инвентарь возможных соответствий существенно расширяется.

Довольно типичны для русского языка случаи, когда *нет* выражает реакцию на невысказанное[1]. Пример такого рода — в связи с другими языковыми особенностями — приводит в своем учебнике по русской грамматике А. Тимберлейк: — *Что мы тогда пели?* ***Нет****, это вспомнить невозможно* [Timberlake 2004: 465]. Здесь *нет* — реакция на опущенную часть вопроса *ты помнишь?* В принципе за «невысказанным» могут скрываться разные аспекты ситуации. Это могут быть предполагаемые установки собеседника, собственные мысли или невербальные действия других людей. Ср. (6–8).

(6) Он стал на колени возле кровати, приподнял ее голову с подушки и прижал свои губы к ее холодеющим губам; она крепко обвила его шею дрожащими руками, будто в этом поцелуе хотела передать ему свою душу… *Нет*, она хорошо сделала, что умерла: ну, что бы с ней сталось, если б Григорий Александрович ее покинул? А это бы случилось, рано или поздно… (М. Ю. Лермонтов. Герой нашего времени)

Er kniete neben ihr Bett, er nahm sanft ihr Haupt mit dem Kissen auf und drückte einen Kuß auf ihre erkaltenden Lippen. In demselben Augenblick schlang sie fest ihre zitternden Arme um seinen Hals, als wollte sie ihre Seele in diesem Kusse übergeben… *Ach*, es war ein Glück für sie, daß sie starb! Was wäre aus ihr geworden, wenn Petschorin sie verlassen hätte? Und früher oder später wäre das ja doch geschehen.[2]

(7) Он оглянулся назад и по сторонам: точное море вокруг него. «*Нет*, лучше и не глядеть», подумал и шел, закрыв глаза, и когда открыл их, чтобы узнать, близко ли конец площади, увидел вдруг, что

[1] См. подробнее [Добровольский, Левонтина 2009а; 2009б], а также понятие “reaction to the unexpressed”, введенное в [Dobrovol’skij, Levontina 2009; 2012]. О связях «между высказанным и подразумеваемым» см., в частности, [Кобозева 1991: 155].

[2] В этом контексте идея реакции на невысказанное сочетается с фатической функцией *нет*.

перед ним стоят почти перед носом какие-то люди с усами, какие именно, уж этого он не мог даже различить. (Н. В. Гоголь. Шинель)

Er sah sich um und zurück — das Meer lag um ihn. „Besser nicht umsehen," dachte er und ging mit geschlossenen Augen weiter, und als er sie wieder öffnete, um zu sehen, wo er denn wäre, sah er vor seiner Nase Leute stehen mit Bärten; mehr konnte er nicht mehr unterscheiden.

(8) "What we'll really do," she said, "is get you ready for your friend Dr. Valentini." (E. Hemingway. A Farewell to Arms)

— *Нет*, сейчас мы займемся другим делом,— сказала она,— нужно приготовить тебя к приходу твоего друга доктора Валентини.

В некоторых контекстах «реакции на невысказанное» ввод в высказывание *нет* маркирует несобственно прямую речь. Являясь в принципе диалогической единицей, *нет* указывает на то, что перед нами свободный косвенный дискурс; ср. (9). В работах Е. В. Падучевой (ср. в первую очередь [1996: 303 ff.]) такие единицы называются первичными (или жесткими) эгоцентриками.

(9) Es war unbedingt nötig, daß K. selbst eingriff. (F. Kafka. Der Prozess)

Нет, К. непременно должен был сам вмешаться.

Частным случаем «реакции на невысказанное» является также употребление частицы *нет* для понижения иллокутивной силы собственного высказывания, как это имеет место в контексте (10).

(10) — Знаешь ли что? — сказал я ему,— я пари держу, что она не знает, что ты юнкер; она думает, что ты разжалованный… — Может быть! Какое мне дело! — сказал он рассеянно.— *Нет*, я только так это говорю… (М. Ю. Лермонтов. Герой нашего времени)

„Höre mal," sagte ich zu ihm, „ich wette, sie weiß nicht einmal, daß du Fähndrich bist; sie glaubt, du seist degradirt…" „Mag sein. Was liegt mir daran!" sagte er zerstreut. „*Nun*, ich meinte nur so…"

Ср. также пример (11) из «Мертвых душ», где *нет* используется для понижения значимости вопроса. Чичиков боится, что Собакевич заподозрит его в чрезмерной заинтересованности в информации о Плюшкине, и спешит заранее снять с себя возможные подозрения. Контекст (11) приводится как в английском, так и в немецком переводе, что на сегодняшнем уровне развития параллельного

корпуса НКРЯ стало возможным для многих произведений русской классической литературы. Интересно в этом примере, что в английском переводе частице *нет* соответствует форма *quite so*, то есть лексическая единица, значение которой при изолированном назывании прямо противоположно русскому *нет*. А в немецком переводе употреблена параллельная русскому *нет* форма *nein*, что, однако, не противоречит базовой идее о лингвоспецифичности этой дискурсивной частицы. Среди многочисленных разнообразных способов перевода слова *нет* в этой функции на другие языки встречаются и буквальные соответствия, ср. также (1) выше.

(11) — Извинительней сходить в какое-нибудь непристойное место, чем к нему. — *Нет*, я спросил не для каких-либо, а потому только, что интересуюсь познанием всякого рода мест, — отвечал на это Чичиков. (Н. В. Гоголь. Мертвые души)

"A man had far better go to hell than to Plushkin's." "*Quite so*," responded Chichikov, "my only reason for asking you is that it interests me to become acquainted with any and every sort of locality."

„Es ist verzeihlicher, irgendeinen unanständigen Ort aufzusuchen als ihn." „*Nein*, ich habe nicht aus einer bestimmten Absicht heraus gefragt, sondern nur, weil ich mich für Orte jeglicher Art interessiere", antwortete Tschitschikow darauf.

Еще одна базовая функция *нет* в качестве дискурсивной частицы — это усиление высказывания, подчеркивание важности сообщения, которое вводит эта частица; ср. (12). В этой функции *нет* часто сочетается с формами *в самом деле* или *правда*. В английском языке аналогичную функцию часто выполняет выражения *I tell you* (12), *really* (13) или *rather*, а в немецком — частица *ja* (14), союзы и союзные наречия *aber, doch*, *sondern* (15).

(12) *I tell you* this war is a bad thing. (E. Hemingway. A Farewell to Arms)

Нет, в самом деле, скверная штука война.

(13) "He was *really* a lovely horse," Catherine said. (E. Hemingway. A Farewell to Arms)

— *Нет*, правда, чудесная лошадь, — сказала Кэтрин.

(14) „*Ja*, es ist brillant, dass du gekommen bist", sagte er, und seine gemächliche Stimme war bewegt! (Th. Mann. Der Zauberberg)

— *Нет*, это чудесно, что ты приехал, — заявил он, и в его обычно спокойном голосе прозвучало тайное волнение.

(15) Das ist aber keine bloß äußerliche, *sondern* eine wirkliche, bindende Bürgschaft. (F. Kafka. Der Prozess)

И это не какое-нибудь пустяковое, формальное поручительство, *нет*, это поручительство настоящее, ко всему обязывающее.

Часто идея усиления сочетается с фатической функцией; ср. (16), а также (12–14) выше.

(16) *Ach*, ich wollte nur ein paar Worte mit ihr reden. (F. Kafka. Der Prozess)

Нет, мне просто хотелось сказать ей несколько слов.

Частным случаем усиления высказывания, подчеркивания его важности можно считать употребление слова *нет* для удвоения отрицания. В немецком языке *nein* — как и в английском *no* — для таких контекстов нетипично. Ср. следующие примеры из НКРЯ.

(17) **а.** Selbst dort zu arbeiten und zu unterhandeln, hat sich wenigstens für mich als ganz unmöglich erwiesen. (F. Kafka. Der Prozess)

Действовать там самому, самому вести переговоры — *нет*, мне это оказалось совершенно не под силу.

б. „Das Warten ist nicht nutzlos“, sagte der Kaufmann, „nutzlos ist nur das selbständige Eingreifen.“ (F. Kafka. Der Prozess)

— *Нет*, ожидание небесполезно, — сказал коммерсант, — бесполезны только попытки самому вмешаться.

в. „Mit dieser Meinung stimme ich nicht überein“, sagte K. kopfschüttelnd, „denn wenn man sich ihr anschließt, muß man alles, was der Türhüter sagt, für wahr halten.“ (F. Kafka. Der Prozess)

— *Нет*, с этим мнением я никак не согласен, — сказал К. и покачал головой, — если так думать, значит, надо принимать за правду все, что говорит привратник.

г. Soll man mir nachsagen dürfen, daß ich am Anfang des Prozesses ihn beenden wollte und jetzt, an seinem Ende, ihn wieder beginnen will? Ich will nicht, daß man das sagt. (F. Kafka. Der Prozess)

Неужто про меня потом скажут, что в начале процесса я стремился его окончить, а теперь, в конце, — начать сначала? *Нет*, не желаю, чтобы так говорили!

(18) **а.** *Нет*, ты не держи меня; как честный человек, поеду. (Н. В. Гоголь. Мертвые души)

Do not detain me; as an honourable man I say that I must go.

б. — *Нет*, брат, ты не ругай меня фетюком, — отвечал зять, — я ей жизнью обязан. (Н. В. Гоголь. Мертвые души)

"Do not insult me with the term Thetuk," retorted the brother-in-law, "To her I owe my life."

в. — *Нет*, это все не то, это все выдумки, это все. — Здесь Собакевич даже сердито покачал головою. (Н. В. Гоголь. Мертвые души)

"Such devices are no good at all". Sobakevitch shook his head wrathfully.

г. — Да бабу, Елизавету Воробья, еще и букву ъ поставили на конце. — *Нет*, никакого Воробья я не приписывал, — сказал Собакевич и отошел к другим гостям. (Н. В. Гоголь. Мертвые души)

"The woman 'Elizabet' Vorobei — 'Elizabet', not 'Elizabeta'?" "I added no such name," replied Sobakevitch, and straightway joined the other guests.

Дискурсивное *нет* используется в русском языке также как средство ввода самокомментария (19), причем в случаях, подобных (20), речь идет о своего рода самокоррекции с элементами усиления высказывания. Ни *nein*, ни *no* в таких случаях нетипичны.

(19) Und für wen führe ich diesen schweren Kampf? Für — es ist fast lächerlich, es auszusprechen — für Block. (F. Kafka. Der Prozess)

Вот ради кого я веду эту трудную борьбу! Ради — *нет*, просто смешно сказать — ради Блока!

(20) Мало сказать: он служил ревностно, *нет*, он служил с любовью. (Н. В. Гоголь. Шинель)

Akaki Akakiewitsch diente mit Eifer, *doch das ist noch nicht das Wort*: er diente mit Liebe.

Обращение к параллельным корпусам позволило выявить еще один тип контекстов, в которых русскому *нет* в немецком языке соответствует не *nein*, а какое-либо другое выражение. Это так называемое приимперативное *нет*; ср. (21).

(21) „Ich muß schon gehen." „Halt, warte!" (Th. Mann. Der Zauberberg)

— Мне пора. — *Нет*, подожди, постой!

В примере (21) из «Волшебной горы» Томаса Манна (пер.: В. Станевич) при императиве *warte!* вводящая реплику частица *nein* в немецком оригинале отсутствует. Ее добавление в русском переводе делает реплику более естественной. Ср. явно менее узуально приемлемый диалог: — *Мне пора.* — *Подожди, постой!* В принципе, вместо оригинальной фразы *Halt, warte!*, можно было сказать что-то вроде *Warte mal* или *Warte doch*. А дословно соответствующий русской реплике вариант *?Nein, warte!* представляется, скорее, неудачным. Употребление вводящей реплику частицы *нет* в русском языке явно предпочтительно. А в случаях повторного побуждения употребление *нет* становится почти обязательным. Ср. контекст (22).

(22) „Kommen Sie mit“, sagte K., „zeigen Sie mir den Weg, ich werde ihn verfehlen, es sind hier so viele Wege.“ „Es ist der einzige Weg“, sagte der Gerichtsdiener nun schon vorwurfsvoll, „ich kann nicht wieder mit Ihnen zurückgehen, ich muß doch meine Meldung vorbringen und habe schon viel Zeit durch Sie versäumt.“ „Kommen Sie mit!“ wiederholte K. jetzt schärfer, als habe er endlich den Gerichtsdiener auf einer Unwahrheit ertappt. (F. Kafka. Der Prozess)

— Пойдемте со мной, — сказал К, — покажите мне дорогу, не то я запутаюсь, здесь столько входов и выходов. — Нет, это единственный выход, — уже с упреком сказал служитель, — а вернуться с вами я не могу, мне еще надо передать поручение, я и так потерял с вами уйму времени. — *Нет*, пойдемте! — уже резче сказал К, словно наконец уличил служителя во лжи.

Вообще, повторное побуждение в русском языке маркируется практически обязательно, хотя и не всегда с помощью частицы *нет*. Часто такими маркерами выступают частицы *ну*, *да*, *же* и др.; см. [Левонтина 1999]. В контексте (22) мы имеем дело с эксплицитным отказом. После полученного отказа говорящий повторяет свое требование, не желая признавать права адресата на отказ. Именно в таких случаях повторное требование начинается с *нет*, что нетипично для немецкого или английского языков, причем использование *no* в английском вряд ли допустимо вообще. Ср. анализ этого и подобных примеров в [Добровольский, Левонтина 2009б]. В немецком и английском языке повторное побуждение либо никак не маркируется (ср. примеры 22 и 23), либо выделяется с помощью

таких форм как *nevertheless* (24), *doch*, в отдельных случаях — скорее под влиянием оригинала — с помощью *nein* (24).

(23) “You can’t come in now,” one of the nurses said. “Yes I can,” I said. “You can’t come in yet.” (E. Hemingway. A Farewell to Arms)

— Сейчас нельзя,— сказала одна из сестер.— Можно,— сказал я.— *Нет*, еще нельзя.

(24) — Да зачем, я и так вижу: доброй породы! — отвечал Чичиков.— *Нет*, возьми-ка нарочно, пощупай уши! (Н. В. Гоголь. Мертвые души)

“Why should I? — without doing that, I can see that he is well-bred.” “*Nevertheless*, catch hold of his ears and feel them.”

„Warum denn, ich seh es doch auch so: eine gute Rasse!“ antwortete Tschitschikow.— „*Nein*, faß sie nur an, befühle mal die Ohren!“

Последний рассматриваемый здесь тип контекстов, в которых русскому *нет* в немецком и английском языках соответствуют другие выражения,— это конструкции различных типов. Покажем это на примерах. Так, русскому *да нет* в немецком часто соответствует *nein, nein* (25), а русскому *ну нет* — как *bloß nicht*, так и *nein, nein* (26).

(25) „*Nein, nein*, das muß Haut und Fleisch sein, was im Innern wächst. (E. Kästner. Pünktchen und Anton)

— *Да нет*, это что-то из мяса и кожи.

(26) **а.** „*Bloß nicht*!“ rief er. (E. Kästner. Pünktchen und Anton)

— *Ну нет*! — воскликнул он,— только не это!

б. *Ну нет* <…> теперь с Петровичем нельзя толковать <…> (Н. В. Гоголь. Шинель)

Nein, nein, heute kann niemand mit Petrowitsch reden.

В высокой степени идиоматична русская конструкция [*нет-нет да и* V] с ее вариантами [*нет-нет да* V], [*нет-нет и* V]. Ее немецкими аналогами оказываются самые разнообразные наречия и адвербиальные фраземы в сочетании с соответствующим глаголом. Ср. *manchmal* (27), *hie und da* (28), *zwischendurch* (29), *bisweilen* (30).

(27) Kein Und, kein Bindewörtchen darf außenbleiben, und von allen Inversionen, die mir *manchmal* entfahren, ist er ein Todfeind; wenn man seinen Period nicht nach der hergebrachten Melodie heraborgelt,

so versteht er gar nichts drin. (J. W. Goethe. Die Leiden des jungen Werther)

Ни одного «и», ни одного союза он тебе не уступит и яро ополчается против инверсий, которые *нет-нет да* проскользнут у меня. Фразу ему надо строить на строго определенный лад, иначе он ничего не поймет.

(28) Ihre Augen, die sie in dem Bewußtsein, den beobachtenden Blicken der ganzen Stadt ausgesetzt zu sein, meistens gesenkt hielt, konnten es sich *hie und da* nicht versagen, über die Menge hinzuschweifen, in der sie auch Julchen Möllendorpf, geborene Hagenström, und ihren Gatten gewahrte… (Th. Mann. Buddenbrooks)

Глаза г-жи Перманедер, хоть она и потупляла их, в сознании, что взгляды „всего города“ устремлены на нее, *нет-нет и* обегали толпу, в которой она среди прочих заметила Юльхен Меллендорф, урожденную Хагенштрем, с супругом…

(29) Gleichzeitig aber vergaß er nicht, sich *zwischendurch* daran zu erinnern, daß viel besser als verzweifelte Entschlüsse ruhige und ruhigste Überlegung sei. (F. Kafka. Die Verwandlung)

Одновременно, однако, он не забывал *нет-нет да* напомнить себе, что от спокойного размышления толку гораздо больше, чем от порывов отчаяния.

(30) Alle waren ein wenig verweint; Grete drückte *bisweilen* ihr Gesicht an den Arm des Vaters. (F. Kafka. Die Verwandlung)

У всех были немного заплаканные глаза; Грета *нет-нет да* прижималась лицом к плечу отца.

Русскому выражению *но нет* в английском языке неожиданным образом часто соответствует *yes*. Его перевод на немецкий язык также оказывается идиоматичным. Ср. (31).

(31) Из-за нее две дамы, большие приятельницы и даже родственницы, перессорились совершенно, именно за то, что одна из них как-то манкировала контрвизитом. И уж как ни старались потом мужья и родственники примирить их, *но нет*, оказалось, что все можно сделать на свете, одного только нельзя: примирить двух дам, поссорившихся за манкировку визита. (Н. В. Гоголь. Мертвые души)

So sacred that on one occasion two closely related ladies who had also been closely attached friends were known to fall out with one another

over the mere fact of an omission to return a social call! *Yes*, in spite of the best efforts of husbands and kinsfolk to reconcile the antagonists, it became clear that, though all else in the world might conceivably be possible, never could the hatchet be buried between ladies who had quarrelled over a neglected visit.

Wegen einer Visitenkarte hatten sich einmal zwei Damen überworfen, die gute Freundinnen und sogar miteinander verwandt waren, und zwar, weil eine von ihnen der anderen keinen Gegenbesuch gemacht hatte. Und wie sehr sich dann auch die Männer und Verwandten darum bemühten, sie wieder miteinander auszusöhnen — *nichts half*, es zeigte sich, daß alles auf der Welt zu vollbringen ist, nur eines nicht: Zwei Damen miteinander auszusöhnen, die sich wegen eines ausgebliebenen Gegenbesuchs überworfen haben.

Рассмотренные в разделе 2.2 контексты иллюстрируют свойства русского *нет*, которые отличают его от английского аналога *no* и от немецкого *nein*. С одной стороны, это ориентация на внешний контекст (то есть на аспекты ситуации, внешние по отношению к пропозитивному содержанию высказывания), соотнесенность с разными слоями смысла высказывания, в том числе с предполагаемыми следствиями и импликатурами дискурса. С другой — это способность *нет* участвовать в ряде лингвоспецифических конструкций, которые, в принципе, не переводимы на другие языки поэлементно, а должны рассматриваться как структуры, близкие по своему статусу к фразеологии. Анализ этих и подобных примеров позволяет сделать вывод, что русское *нет* может быть связано как с диктумом, так и с модусом высказывания, в то время как ангийское *no* и немецкое *nein* в целом более тесно связано с диктумом, с пропозицией.

Есть все основания считать, что в случае дискурсивного *нет* мы имеем дело с лингвоспецифичной единицей русского языка. Такие слова представляют исключительный интерес для двуязычной лексикографии, поскольку они не имеют в других языках стандартных лексических эквивалентов. Значение дискурсивной частицы *нет* распределяется в английском языке между такими единицами, как *well, really, rather, however, nevertheless, but, why, oh, no, yes, quite so*, а в немецком — *nun, aber, doch, sondern, ach, nein, ja, bitte* и др.

От хорошего русско-английского и русско-немецкого словаря вряд ли следует ожидать фиксации всех этих и других потенци-

ально возможных лексических коррелятов дискурсивного *нет*. Это и не нужно, поскольку выбор некоторых из встречающихся в параллельных корпусах соответствий основывается в большей степени на субъективном решении переводчика, чем на их семантической близости исходному выражению. Вряд ли реалистично также ожидать от словаря, что в нем будут эксплицитно описаны контекстные условия, соблюдение которых требуется для употребления каждого из предлагаемых эквивалентов.

Понятно, что степень подробности лексикографического представления и выделение конкретных режимов употребления дискурсивного *нет* определяется типом словаря. В любом случае полезно указать на различные функции русского *нет* в тех случаях, когда это влияет на способ перевода этого слова. Так, появление в качестве английских эквивалентов русского *нет* слов и выражений *really, rather, I tell you* и немецких *aber, doch, sondern* естественно, если частица *нет* употребляется в функции усиления высказывания. А употребление *нет* в функции повторного побуждения, как правило, не имеет лексического соответствия ни в немецком, ни в английском языке, то есть такое *нет* чаще всего должно оставаться непереведенным. Фатическое *нет* часто передается на немецкий с помощью междометия *ach*, а на английский — *oh*.

Существующие русско-английские и русско-немецкие словари еще далеки от решения этой задачи. Так, в [Цвиллинг 2003] у *нет* выделяются всего два значения: «обычное» *нет = nein*, то есть *нет* как слово-предложение (в этом же значении даны употребления типа *ты придешь или нет?*) и *нет*-предикатив (*у меня нет времени*). В заромбовой части приводится пример на употребление конструкции [*нет-нет да и* V], а также выражения *сойти на нет*; *нет, так нет* и *чего здесь только нет*. В [РАС] эти два значения выделяются как омонимы, причем в рамках **нет I** (*нет* как «отрицание») под цифрой 2 описывается один из режимов употребления дискурсивного *нет*: «в начале реплики — с оттенком возражения, удивления». В качестве единственного английского эквивалента предлагается *but*: *нет, вы его не знаете! but you don't know him!*; *нет, почему вы так думаете? but why do you think so?* «Новый большой русско-английский словарь», который во многих отношениях существенно превосходит аналогичные издания, в описании слова *нет* сохраняет те же принципы; ср. [Ермолович, Красавина 2006].

Очевидно, что для лексикографов, работающих над составлением новых двуязычных словарей, здесь намечается широкое поле деятельности. Причем это относится не только к дискурсивному *нет*, но и к большинству лингвоспецифичных лексем.

3. Немецкие показатели пространственного дейксиса: элементы *her* и *hin* и их русские соответствия

Проблему лексической лингвоспецифичности интересно решать не только в направлении от русского языка к иностранному, но и в обратном направлении. Так, в немецком языке обнаруживается достаточно много лексических единиц, перевод которых на русский сопряжен с определенными трудностями, причем эти трудности мотивированы особенностями специфичной для немецкого языка концептуализации фрагментов действительности. Часто эти концептуальные особенности оказываются регулярным образом лексикализованы, то есть фиксация определенных понятийных различий пронизывает всю лексическую систему. Ниже речь пойдет об одном из таких случаев.

В данном разделе будет показано, как работа с параллельным корпусом может помочь в выявлении русских соответствий определенного немецкого слова, обладающего нетривиальными особенностями. Когда у некоторого слова в специфических употреблениях нет стандартных аналогов в других языках, может создаться впечатление, что подобные смыслы непереводимы в принципе. Корпус показывает, что способы передачи таких смыслов все-таки существуют, и позволяет эти способы описать. В качестве материала мы берем такие лексические единицы немецкого языка, про которые принято считать, что они не имеют русских аналогов, и пытаемся с помощью контекстов из параллельного корпуса выявить возможные способы передачи соответствующих смыслов.

В немецком языке есть огромное количество слов с пространственно-дейктическими элементами *her* (≈ ‘сюда’) и *hin* (≈ ‘туда’), выражающими идею направления движения.[1] Слова *hin* и *her* употребляются также и самостоятельно в различных комбинациях с глаголами и адвербиальными конструкциями для выражения про-

[1] С типологической точки зрения *her* и *hin* — это показатели глагольной ориентации; ср. [Плунгян 2011]. Немецкие показатели глагольной ориентации обсуждаются также в [Добровольский 2013].

странственных отношений между участниками ситуации. В русском языке нет таких системно организованных (то есть почти грамматических) средств для выражения соответствующих дейктических смыслов. Так, и глагол *hereinführen*, и глагол *hineinführen* переводятся на русский язык с помощью глаголов *вводить / ввозить*, так что идея ‘вводить/ ввозить сюда, то есть по направлению к говорящему’ vs. ‘вводить/ ввозить туда, то есть по направлению от говорящего’[1] в русском языке часто вообще не выражается.

В работе [Добровольский, Падучева 2008] было показано, что проблема состоит не только в отсутствии у этих немецких слов стандартных русских аналогов, но и в том, что комментарии к немецким *her* и *hin*, апеллирующие к говорящему, искажают реальную картину. Так, человек, желающий войти в некоторое помещение, (в стандартной ситуации) спросит, постучав, *Darf ich herein?* (то есть *Можно мне войти?*, сокращенная форма от *Darf ich hereinkommen?* ‘Можно мне сюда войти?’), а не *Darf ich hinein?* (букв. ≈ «Можно мне туда войти?»). Понятно, что предполагаемое перемещение в пространстве должно совершаться не по направлению к говорящему, а от его исходного местонахождения по направлению к адресату. Дело в том, что в таких случаях говорящий принимает точку зрения адресата, как бы ставя себя на его место.

Еще более сложно ведут себя такие дейктические элементы в гипотаксисе и нарративе. Ср. функционирование вопросительных местоименных наречий, в состав которых входят *her* или *hin*. Значение данных дейктических элементов (и соответственно их перевод на русский язык) меняется в зависимости от того, употребляется ли это наречие в прямом вопросе или в качестве относительного наречия в гипотаксисе. В гипотаксисе субъектом дейксиса является не говорящий, а подлежащее главного предложения. В [Добровольский, Падучева 2008] приводится такой пример: *Alle schauten entsetzt auf den Felsen, **woherunter** der Bergsteiger gestürzt war — Все в ужасе посмотрели на скалу, откуда <с которой> сорвался альпинист.* Выбор наречия *woherunter* (в отличие от *wohinunter*) мотивирован в этом случае местонахождением людей, смотрящих на скалу (*alle*). Контекст ясно показывает, что они стоят внизу и смотрят на скалу

[1] Точнее ‘по направлению к говорящему или к лицу, точку зрения которого принимает говорящий’ vs. ‘по направлению от говорящего или от лица, точку зрения которого принимает говорящий’. Ср. [Добровольский, Падучева 2008].

снизу вверх. Иными словами, *hin-* и *her*-слова показывают, где находится наблюдатель, а не говорящий. *Woherunter* — это как бы «где-сюда-низ», а *wohinunter* — «где-туда-низ». Соответственно, если наблюдатель наверху, он скажет «где-туда-низ», то есть *wohinunter*, по-русски что-то вроде *куда (вниз)*, а если он внизу — *woherunter* — «где-сюда-низ», то есть *откуда (вниз)*.

В качестве примера интерпретации дейктических слов в нарративном режиме мы использовали контексты (32) и (33) из НКРЯ.

(32) Prosper Alpanus senkte sich *herab* zu dem Jüngling, an dessen Seite er Platz nahm, während die Libelle aufflog in ein Gebüsch und in den Gesang einstimmte, der durch den ganzen Wald tönte. (E. T. A. Hoffmann. Klein Zaches genannt Zinnober)

Проспер Альпанус спустился к юноше и сел подле него, стрекоза упорхнула в кусты, вторя пению, наполнявшему весь лес.

Здесь в принципе возможно и *hinab*. Но в этом случае субъектом дейксиса был бы Проспер Альпанус. А в перспективе повествования, выбранной автором, субъект дейксиса — юноша. Как видно из перевода, русский язык не располагает стандартными средствами, позволяющими выразить эти различия: ‘спустился сюда’ vs. ‘спустился туда’. Более точный перевод был бы *Проспер Альпанус спустился сюда к юноше.*

(33) Er schaute *hinauf* und erblickte staunend Prosper Alpanus, der auf einem wunderbaren Insekt, das einer in den herrlichsten Farben prunkenden Libelle nicht unähnlich, daherschwebte. (E. T. A. Hoffmann. Klein Zaches genannt Zinnober)

Он поднял глаза и с изумлением увидел Проспера Альпануса, летевшего к нему на каком-то диковинном насекомом, не лишенном сходства с великолепной, сверкающей всеми красками стрекозой.

Hinauf в предложении (33) — которое в тексте предшествует предложению (32) — развертывает событие в противоположной перспективе. Субъект восприятия здесь — тот же юноша. Но в одном случае движение направлено к нему, потому что Проспер Альпанус спустился к нему вниз, а в другом — от него. В последнем случае решающим фактором для выбора дейктического элемента является направление взгляда юноши снизу вверх. Итак, субъект дейксиса в нарративном режиме — это один из персонажей, глазами

которого описываются события в конкретном контексте. Подобные сдвиги в осмыслении пространственных отношений еще более усложняют поиск адекватных русских соответствий.

Как параллельный корпус может использоваться для исследования особенностей значения и употребления дейктических слов с элементами *hin* и *her*, а также для выявления их контекстно-зависимых русских соответствий? Попытаемся ответить на этот вопрос на материале адвербиальных выражений *herein* и *hinein*, которые употребляются и как самостоятельные наречия, и в составе сложных глаголов.

Обратимся к примерам из НКРЯ. В первую очередь, примеры показывают, что, помимо самых простых и однозначных выражений с элементами *туда* для *hin* и *сюда* для *her*, существует целый ряд способов передать направление движения относительно точки отсчета — говорящего или наблюдателя, которые используются далеко не во всех контекстах, а только там, где это коммуникативно необходимо. Для *herein* это, например, такие предложные группы, как *на нас* (34)[1]. А для *hinein* это могут быть, среди прочего, глаголы, более или менее четко указывающие на направление движения от наблюдателя (35). В примере (36) глагол *войти* дейктически нейтрален: войти можно как куда-то, где находится наблюдатель (именно эта ситуация имеет место в данном контексте), так и в какое-то помещение прочь от наблюдателя.

(34) Sie weht *herein* mit dem Geruch der Erde, der Blumen und dem Funkeln der Sterne. (E. M. Remarque. Der schwarze Obelisk)

Она дышит *на нас* запахом земли, цветов и мерцанием звезд.

(35) Die gigantische Maschine *hob sich* mit Dr. h.c. Isaak Kohler in den Nachthimmel hinein, heulend, brüllend, ab nach Australien, als ich aus meinem Volkswagen sprang, den Revolver entsichert. (F. Dürrenmatt. Justiz)

Когда я выскочил из своего «фольксвагена», сняв револьвер с предохранителя, гигантская машина, уносящая доктора honoris causa Исаака Колера, с ревом и воем *взмыла* в ночное небо и взяла курс на Австралию.

[1] Подобный перевод возможен только в нарративе с диегетическим повествователем (как это имеет место в «Черном обелиске» Ремарака) или в свободном косвенном дискурсе.

(36) Dann ging die Tür auf, und Antons Mutter *kam herein*. (E. Kästner. Pünktchen und Anton)

Вдруг дверь открылась и в детскую *вошла* мать Антона.

Примеры (37–41) показывают, как выбор дейктического элемента в нарративе позволяет установить, кто из персонажей является в данный момент субъектом перспективы повествования, а также в каком месте располагается в каждом конкретном эпизоде точка наблюдения. В русских переводах это чаще всего видно только из контекста.

(37) <…> da gab ihm der Vater von hinten einen jetzt wahrhaftig erlösenden starken Stoß, und er *flog*, heftig blutend, weit in sein Zimmer *hinein*. (F. Kafka. Die Verwandlung)

И тогда отец с силой дал ему сзади поистине спасительного теперь пинка, и Грегор, обливаясь кровью, *влетел* в свою комнату.

(38) „Was denn?“ fragte Pünktchen harmlos. Gottfried Klepperbein vertraf ihr drohend den Weg. „Das weißt du schon ganz gut, stell dich nicht so dumm, mein Herzblatt!“ Pünktchen wollte gern ins Haus, aber er *ließ* sie nicht *hinein*. (E. Kästner. Pünktchen und Anton)

— Это что же? — простодушно осведомилась Кнопка. Готфрид Клеппербейн грозно загородил ей дорогу. — Сама знаешь, не строй из себя дурочку, дорогуша! Кнопка хотела войти в дом, но он не *пускал* ее.

(39) Aber es *fiel* so viel Ruß *herein*, dass das Mädchen die Luke gleich wieder zuziehen und mit ihrem Taschentuch die Hände K.s vom Ruß reinigen mußte, denn K. war zu müde, um das selbst zu besorgen. (F. Kafka. Der Prozess)

Но *посыпалось* столько сажи, что девушке пришлось тут же закрыть люк и смахнуть сажу с рук К. своим носовым платком, потому что сам он слишком ослабел.

(40) Wir wohnten in der Nähe des Ehrenfelder Bahnhofs, blickten vom Fenster unseres Zimmers aus auf die rote Backsteinmauer des Bahndamms, Braunkohlenzüge *fuhren* voll in die Stadt *herein*, leer aus ihr *hinaus*, ein trostreicher Anblick, ein herzbewegendes Geräusch, ich mußte immer an die ausgeglichene Vermögenslage zu Hause denken. (H. Böll. Ansichten eines Clowns)

Пансион находился неподалеку от Эренфельдской товарной станции; из окон нашей комнаты была видна красная кирпичная

стена и товарные составы, в город *шли* платформы с бурым углем, из города *отправлялся* порожняк — отрадная картина, приятный перестук колес: как тут не подумать, что имущественное положение семьи Шниров вполне надежно.

(41) Aber was weiß er schon? Beweisen kann er nichts. Trotzdem kann er mit mir spielen, wie er will. Der Arzt *kommt herein*. Es ist nicht der Direktor; es ist der behandelnde Arzt. (E. M. Remarque. Der schwarze Obelisk)

Но какая цена этому знанию? Доказать он ничего не может. И все-таки он играет со мною, как ему вздумается. *Входит* врач. Это не директор — это лечащий врач.

Так, в примере (37) субъект дейксиса — Грегор, в (38) — Кнопка. Причем важно, что не Кнопка, а другой персонаж является в данном случае субъектом действия и, соответственно, подлежащим в предложении. Аналогично и в (41): субъект действия и подлежащее — врач. Но не он является субъектом дейксиса. Иначе было бы не *kommt herein*, а *kommt hinein*. Выбор дейктического элемента ясно показывает, что наблюдатель находится внутри — в той комнате, в которую входит врач. А значит, субъект наблюдения и дейксиса (как и в примере 40) — диегетический повествователь, то есть персонаж, ведущий повествование от 1-го лица. В примере (40) комбинация дейктических адвербиалов *herein* и *hinaus* однозначно демонстрирует, что наблюдатель — он же диегетический повествователь — находится в городе, а не за его пределами, что коррелирует с указанием места расположения пансиона, в котором живут герои.

Ориентация немецкого пространственного дейксиса в нарративе на наблюдателя подтверждается выбором в русских переводах таких глаголов, как *появиться, показаться, явиться, доноситься*. Известно, что смысл ‘наблюдатель’ входит в значение этих глаголов, а указание на фигуру наблюдателя является неотъемлемой частью их толкования.

(42) <…> und es kam immer ein Punkt, wo er mich wie ein heulender Hund ansah, und meistens *kam* gerade dann Sabine *herein*, schaute ihn verbittert an, weil sie wieder schwanger war. (H. Böll. Ansichten eines Clowns)

<…> а потом вдруг наступала минута, когда он смотрел на меня глазами затравленного зверя; тут в комнате обычно *появлялась* Са-

бина и смотрела на Карла с видом жертвы: она опять ходила беременная.

(43) Wieder *kam* der junge Mann mit dem Schnurrbart *herein*, senkte einmal das Kinn auf die Brust und setzte sich mit sorgenvollverschlossener Miene. (Th. Mann. Der Zauberberg)

Снова *явился* молодой человек с усами, опустил подбородок на грудь и уселся с озабоченным и замкнутым видом.

(44) Nach Osten und Süden standen die Fenster des Turmgemachs offen, aus Garten und Feld *drangen* linde, süße Gerüche aller Art, aus der Landschaft unbestimmte Geräusche zu ihm *herein*, wie die kommende Frühe sie aus der Weite und Nähe zu bringen pflegt. (A. Schnitzler. Casanovas Heimfahrt)

Окна комнаты, выходившие на восток и на юг, были раскрыты. Из сада и с полей *доносились* нежные и сладкие ароматы, повсюду, вдалеке и вблизи, слышались неясные шорохи, которые обычно приносит с собой наступающее утро.

(45) Hie und da *spähten* ein paar Schutzleute mit grünlich schillernden Federbüschen *herein* und gingen in langsamem, schwerem Schritt wieder weiter. (G. Meyrink. Der Golem)

Время от времени *показывались* городовые с зеленоватыми и блестящими перьями на шляпах и медленными тяжелыми шагами шли дальше.

В основе семантики слова *hinein* (в его пространственном значении[1]), наряду с указанием на направление прочь от говорящего или от лица, точку зрения которого принимает говорящий, лежит идея перемещения из более открытого в более замкнутое пространство. Интересны контексты типа (46) и (47), в которых речь идет о ментальной (духовной или интеллектуальной) сфере. Употребление *hinein* в таких контекстах показывает, что в модели мира, зафиксированной в немецком языке, эта сфера мыслится как замкнутое пространство, причем отдельное от субъекта. Иными словами, даже если это внутренний мир человека, он описывается в языке как пространство, существующее отдельно от этого человека.

(46) In dies Priestertum, in diesen Hochmut, in diese Geistigkeit

[1] У этого слова есть еще и временное значение; ср. словарную статью из НБНС, приведенную ниже.

hinein hatte sein Ich sich verkrochen, dort saß es fest und wuchs, während er es mit Fasten und Buße zu töten meinte. (H. Hesse. Siddhartha)

В это-то священничество, в это-то высокомерие и духовность и заползло его Я, там оно засело крепко и росло, в то время как он воображал, что умерщвляет его постом и покаянием.

(47) Mit beginnendem Schlaf versenkte er sich tiefer und tiefer in sich *hinein* und hielt triumphalen Einzug in seiner inneren Festung, auf der er sich ein geruchliches Siegesfest erträumte, eine gigantische Orgie mit Weihrauchqualm und Myrrhendampf, zu Ehren seiner selbst. (P. Süskind. Das Parfum)

Засыпая, он все глубже погружался в самого себя и совершал триумфальный въезд в свою внутреннюю крепость, где он праздновал в мечтах некую победу обоняния, гигантскую оргию с густым дымом ладана и парами мирры в свою честь.

Рассмотренные здесь особенности немецких лексических единиц с дейктическим значением должны учитываться в их лексикографическом представлении. В первую очередь это касается необходимости переориентации субъекта дейксиса с говорящего (как это традиционно делалось в словарях) на наблюдателя и сходных с ним участников ситуации, которым говорящий как бы делегирует свои дейктические полномочия. В «Новом большом немецко-русском словаре» мы попытались учесть эти особенности следующим образом:

herein *adv dem* сюда; внутрь (*при обозначении перемещения из более открытого в более замкнутое пространство по направлению к говорящему или лицу, точку зрения которого принимает говорящий*); ~! входи (те)!; (nur) immer ~! входи (те) быстрей [смелей]!; hier ~! вход здесь! ◊ ~, **wenns kein Schneider ist!** *шутл. разг.* ≈ входи (те), коль не шутите!

herein= *отд. первый частотный элемент гл., указывает на движение внутрь по направлению к говорящему или лицу, точку зрения которого принимает говорящий (тж. перен.)*: hereinkommen входить; hereinlassen впустить; hereinfallen попасть впросак

hinein *adv dem* **1**. в, внутрь; вглубь; вплоть до (*при обозначении перемещения из более открытого в более замкнутое пространство по направлению от говорящего или лица, точку зрения которого принимает говорящий*); ins Zimmer ~ в комнату; ins Leere ~ в пустоту; nun aber ~ ins Bett (mit euch)! а ну, в кровать! **2**. *при обозначении времени*

до, вплоть до; bis in die Gegenwart ~ вплоть до наших дней; bis (tief) in die Nacht ~ до поздней ночи ◊ **ins Blaue** ~ *разг.* наобум, без четкой цели; **bis ins Herz [Mark, Innerste]** ~ *высок.* до глубины души

hinein= *отд. первый частотный элемент гл., указывает на движение внутрь по направлению от говорящего или лица, точку зрения которого принимает говорящий (тж. перен.)*: hineingehen входить; hineinblicken заглядывать (*напр. в помещение*); sich hineindenken вдумываться; hineinlegen привносить (*личное отношение в какое-л. дело, напр. чувство в исполнение музыкального произведения*)

* * *

Завершая эту работу, напомним, что ее целью было не столько описание отдельных сложных случаев, представляющих серьезные трудности как для контрастивной лексикологии и теории лингвоспецифичности, так и для двуязычной лексикографии, сколько поиски ответа на вопрос, насколько корпусы параллельных текстов могут оказаться полезным инструментом в решении подобных задач. Ответ на этот вопрос представляется однозначным. Предпринятый в статье анализ отдельных лексем, не имеющих в других языках стандартных эквивалентов, показал, что обращение к параллельным корпусам помогает выявить как нетривиальные семантические и дискурсивные особенности подобных слов, так и особенности их кросс-лингвистических функциональных соответствий.

Литература

Бунтман Н. В., Зализняк Анна А., Зацман И. М., Кружков М. Г., Лощилова Е. Ю., Сичинава Д. В. Информационные технологии корпусных исследований: принципы построения кросслингвистических баз данных // Информатика и ее применения. Т. 8, вып. 2, 2014. С. 98–110.

Гришина Е. А. Да в русском устном диалоге // Russian Linguistics. Vol. 35. 2011. P. 169–207.

Добровольский Д. О. Корпус параллельных текстов в исследовании культурно-специфичной лексики // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 383–401.

Добровольский Д. О. Беседы о немецком слове. М.: Языки славянской культуры, 2013.

Добровольский Д. О., Кретов А. А., Шаров С. Корпус параллельных текстов: архитектура и возможности использования // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005. С. 263–296.

Добровольский Д. О., Левонтина И. Б. 500 способов сказать «нет» (русско-немецкие соответствия) // Логический анализ языка. Ассерция и негация/Отв. ред. Н. Д. Арутюнова. М.: Индрик, 2009 а. С. 400–410.

Добровольский Д. О., Левонтина И. Б. Русское *нет*, немецкое *nein*, английское *no*: сопоставительное исследование семантики на базе параллельных копусов // Компьютерная лингвистика и интеллектуальные технологии. Труды междунар. конференции «Диалог 2009». М.: РГГУ, 2009 б. С. 97–101.

Добровольский Д. О., Падучева Е. В. Дейксис в отсутствие говорящего: о семантике немецких дейктических элементов *hin* и *her* // Компьютерная лингвистика и интеллектуальные технологии. Труды междунар. конференции «Диалог 2008». М.: РГГУ, 2008. С. 140–146.

Добрушина Е. Р. Что значит *нет*? // Дискурсивные слова русского языка: контекстное варьирование и семантическое единство. М.: Азбуковник, 2003. С. 146–193.

Йокояма О. Б. К метатеории перевода: перевод как дискурс // Человек о языке — язык о человеке. Сб. статей памяти Н. Ю. Шведовой. М.: Азбуковник, 2012. С. 152–159.

Кобозева И. М. Проблемы описания частиц в исследованиях 80-х гг. // Прагматика и семантика. Сб. научно-аналитических обзоров. М.: Изд-во АН СССР, 1991. С. 147–173.

Левонтина И. Б. Стратегии уговаривания: частицы в повторных просьбах // Язык. Культура, Гуманитарное знание. Научное наследие Г. О. Винокура и современность. М.: Научный мир, 1999.

Падучева Е. В. Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. М.: ШЯРК, 1996.

Плунгян В. А. Введение в грамматическую семантику. Грамматические значения и грамматические системы языков мира. М.: РГГУ, 2011.

Atkins S., Fillmore C. J., Johnson C. R. Lexicographic relevance: Selecting information from corpus evidence // International Journal of Lexicography. Vol. 16. 2003. No. 3. P. 251–280.

Cysouw M., Wälchli B. (eds.). Parallel texts. Using translational equivalents in linguistic typology. Sprachtypologie & Universalienforschung 60.2. (Special Issue). 2007.

Dobrovol'skij D., Levontina I. YES and NO: Universal Ideas in Language Specific Configurations // Meaning-Text-Theory. 4th International Conference, Montreal 2009. Proceedings. Ed. by David Beck, Kim Gerdes, Jasmina Milićević and Alain Polguère. URL: http://meaningtext.ilpga.fr/mtt2009/14-DDobrovolskijILevontina.pdf.

Dobrovol'skij D., Levontina I. Russian *NET* vs. German *NEIN* 'NO': a semiotic approach // Russian Linguistics. Vol. 36. 2012. P. 213–219.

Pope E. N. Questions and answers in English. The Hague, Paris: Mouton, 1976.

Timberlake A. Reference grammar of Russian. West Nyack: Cambridge University Press, 2004.

Словари

Ермолович, Красавина 2006 — *Ермолович Д. И., Красавина Т. М.* Новый большой русско-английский словарь. Изд-е 2-е, испр. М.: Русский язык — Медиа, 2006.

НБНС — Новый большой немецко-русский словарь: В 3 т./Под рук. Д. О. Добровольского. М.: Астрель, 2008–2010.

РАС — Русско-английский словарь/Под общим рук. А. И. Смирницкого / Изд-е 9-е, стереотип. под рук. О. С. Ахмановой. М.: Советская энциклопедия, 1971.

Цвиллинг 2003 — *Цвиллинг М. Я.* Русско-немецкий словарь. М.: Русский язык — Медиа, 2003.

D. O. Dobrovol'skij
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
dm-dbrv@yandex.ru

THE CORPUS OF PARALLEL TEXTS AND CONTRASTIVE LEXICOLOGY

The aim of the present article is to attempt to determine the extent to which corpora of parallel texts can serve as a useful instrument for solving problems of contrastive linguistics; i.e., for describing peculiarities in the functioning of Russian forms compared with those of another language. The study is based on empirical data taken from the Russian-English/English-Russian and Russian-German/German-Russian parallel corpora of the RNC. We focus on lexical phenomena of these pairs of languages, particularly cross-linguistic equivalence connected with the language-specific nature of certain lexemes and the means for representing them in bilingual dictionaries. The materials of the parallel corpora give rise to questions as to how cross-linguistic lexical correlates should be represented lexicographically, especially if these correspondences are not equivalent in a strict sense. Our analysis of lexical units that lack standard equivalents in other languages indicates that parallel corpora can help identify both the non-trivial semantic and discursive properties of such words and the peculiarities of their cross-linguistic functional correspondences.

Key words: parallel corpus, lexical semantics, bilingual lexicography, linguo-specific factors, Russian, English, German

References

Atkins S., Fillmore C. J., Johnson C. R. Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography*. Vol. 16. 2003, no. 3, pp. 251–280.

Buntman N. V., Zaliznjak Anna A., Zacman I. M., Kruzhkov M. G., Loshhilova E.Ju., Sichinava D. V. [Informational technology in corpus-based studies: towards a cross-linguistic database]. *Informatika i ee*

primeneniya [Informatics and its Applications]. Vol. 8, issue 2, 2014, pp. 98–110. (In Russ.)

Cysouw M., Wälchli B. (eds.) *Parallel texts. Using translational equivalents in linguistic typology. Sprachtypologie & Universalienforschung* 60.2. (Special Issue). 2007.

Dobrovol'skij D. O. [The corpus of parallel texts and culture-specific lexemes]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy.* [Russian National Corpus: 2006–2008. New Results and Prospects]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 383–401. (In Russ.)

Dobrovol'skij D. O. *Besedy o nemetskom slove* [Studies on German Lexis]. Moscow, Yazyki slavyanskoi kul'tury Publ., 2013.

Dobrovol'skij D. O., Kretov A. A., Sharov S. [The corpus of parallel texts: architecture and potential of using]. *Natsional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy* [Russian National Corpus: 2003–2005. Results and Prospects]. Moscow, Indrik Publ., 2005, pp. 263–296. (In Russ.)

Dobrovol'skij D., Levontina I. Russian *NET* vs. German *NEIN* 'NO': a semiotic approach. *Russian Linguistics*, 2012, Vol. 36, pp. 213–219.

Dobrovol'skij D., Levontina I. YES and NO: Universal Ideas in Language Specific Configurations. *Meaning-Text-Theory.* 4th International Conference, Montreal 2009. Proceedings. David Beck, Kim Gerdes, Jasmina Milićević and Alain Polguère (eds.). Available at: URL: http://meaningtext.ilpga.fr/mtt2009/14-DDobrovolskijILevontina.pdf.

Dobrovol'skij D. O., Levontina I. B. [500 ways to say "no" (Russian-German correlations)]. *Logicheskii analiz yazyka. Assertsiya i negatsiya* [Logical Analysis of Language. Assertion and Negation]. N. D. Arutyunova (ed.). Moscow, Indrik Publ., 2009a, pp. 400–410. (In Russ.)

Dobrovol'skij D. O., Levontina I. B. [Russian *net*, German *nein*, English *no*: contrastive semantic analysis based on parallel corpora]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunar. konferentsii "Dialog 2009"* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2009"]. Moscow, RSUH Publ., 2009b, pp. 97–101. (In Russ.)

Dobrovol'skij D. O., Paducheva E. V. [Deixis without the speaker: on semantics of German deictic elements *hin* and *her*]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunar. konferentsii "Dialog 2008"* [Computational Linguistics and Intellectual Technolo-

gies. Proceedings of International Conference “Dialog–2008”]. Moscow, RSUH Publ., 2008, pp. 140–146. (In Russ.)

Dobrushina E. R. [What does *net* mean?]. *Diskursivnye slova russkogo yazyka: kontekstnoe var'irovanie i semanticheskoe edinstvo* [Russian Discursive Words: Contextual Variation and Semantic Unity]. Moscow, Azbukovnik Publ., 2003, pp. 146–193. (In Russ.)

Grishina E. A. [*Da* in the Russian spoken dialogue]. *Russian Linguistics*, 2011, Vol. 35, pp. 169–207. (In Russ.)

Iokoyama O. B. [Towards the metatheory of translation: translation as a discourse]. *Chelovek o yazyke — yazyk o cheloveke. Sb. statei pamyati N.Yu. Shvedovoi* [Man about language — language about man. In memoriam of N.Yu. Shvedova]. Moscow, Azbukovnik Publ., 2012, pp. 152–159. (In Russ.)

Kobozeva I. M. [Particle studies in the 80s]. *Pragmatika i semantika. Sb. nauchno-analiticheskikh obzorov* [Pragmatics and semantics. Analytical survey]. Moscow, Izd. AN SSSR, 1991, pp. 147–173. (In Russ.)

Levontina I. B. [Strategies of persuasion: particles in repeated request]. *Yazyk. Kul'tura, Gumanitarnoe znanie. Nauchnoe nasledie G. O. Vinokura i sovremennost’* [Language. Culture. Humanities]. Moscow, Nauchnyi mir Publ., 1999. (In Russ.)

Paducheva E. V. *Semanticheskie issledovaniya: Semantika vremeni i vida v russkom yazyke. Semantika narrativa* [Semantic Studies: Semantics of Tense and Aspect in Russian. Narrative Semantics]. Moscow, Shkola “Yazyki russkoi kul'tury” Publ., 1996.

Plungyan V. A. *Vvedenie v grammaticheskuyu semantiku. Grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira* [Introduction to the Semantics of Grammar. Grammatical Meanings and Grammatical Systems in the Languages of the World]. Moscow, RSUH Publ., 2011.

Pope E. N. *Questions and answers in English.* The Hague, Paris, Mouton, 1976.

Timberlake A. *Reference grammar of Russian.* West Nyack, Cambridge University Press, 2004.

Dictionaries

Ermolovich D. I., Krasavina T. M. *Novyi bol'shoi russko-angliiskii slovar'* [A New Comprehensive Russian-English Dictionary]. 2nd ed. Moscow, Russkii yazyk — Media Publ., 2006.

NBNS — *Novyi bol'shoi nemetsko-russkii slovar'* [A New Comprehensive German-Russian Dictionary]. In 3 volumes. D. O. Dobrovol'skij (ed.). Moscow, Astrel' Publ., 2008–2010.

RAS — *Russko-angliiskii slovar'* [Russian-English Dictionary]. A. I. Smirnitskii (gen. ed.), 9th ed. O. S. Akhmanova (ed.). Moscow, Sovetskaya entsiklopediya Publ., 1971.

Tsvilling M.Ya. *Russko-nemetskii slovar'* [Russian-German Dictionary]. Moscow, Russkii yazyk — Media Publ., 2003.

Б. В. Орехов
Национальный исследовательский университет
«Высшая школа экономики»
(Россия, Москва)
nevmenandr@yandex.ru

ЕЩЕ РАЗ ОБ ИССЛЕДОВАТЕЛЬСКОМ ПОТЕНЦИАЛЕ ПОЭТИЧЕСКОГО КОРПУСА: МЕТР, ЛЕКСИКА, ФОРМУЛА

Статья продолжает начатую в публикациях других исследователей демонстрацию возможностей поэтического подкорпуса в составе Национального корпуса русского языка. Речь идёт о том, какие вопросы, связанные с историей русской поэзии, можно решать с помощью корпуса. В первой части статьи проводится пилотное исследование тяготения лексических единиц к определенным поэтическим метрам. Исследование проводится с применением коллокструкционного анализа. В качестве тестового материала используются «метапоэтические термины», то есть такие, которые традиционно используются при описании процесса создания поэтического произведения. Почти все они оказываются тяготеющими к ямбу. Исключение составляет слово «хорей», появляющееся чаще в хореических строках. Вторая часть статьи реконструирует поэтическую традицию, в которой находится стихотворение Ф. И. Тютчева «Последний катаклизм». Стихотворение рассматривается как составленное из ряда поэтических формул, повторяемых текстовых фрагментов, каждая из которых имеет свою историю, происхождение и частотность в русской поэзии. Уяснение этих аспектов для каждой формулы помогает полнее представить генезис всего стихотворения и встроить его в модель восприятия читателя-современника.

Ключевые слова: корпусная лингвистика, поэзия, русская поэзия, поэтический метр, поэтическая лексика.

Эта статья продолжает начатую в публикациях [Гришина и др. 2009], [Сичинава 2012] демонстрацию исследовательских возможностей поэтического подкорпуса в составе Национального корпуса русского языка. Если в упомянутых статьях в равной степени отмечаются способы использования поэтического подкорпуса для решения как чисто лингвистических задач (например, диахроническое описание акцентных вариантов), так и связанных с собственно историей русской поэзии, то мы сосредоточимся на втором круге вопросов, не претендуя при этом охватить его целиком.

Поэтический корпус предоставляет уникальную возможность изучения взаимодействий метра и различных лингвистических явлений (например, поведения лексических единиц) в стихе. Эта возможность появляется благодаря тому, что к привычным лингвистическим разметкам (морфологической, семантической, метатекстовой) добавлена специальная метрическая разметка, то есть каждой стихотворной строке приписана помета, определяющая метр, стопность и клаузулу стиха. Поисковые механизмы корпуса позволяют задавать пользовательские подкорпуса, включающие тексты, состоящие из строк, отвечающих требуемым характеристикам.

1. Метр и лексика

Получив такой инструмент, исследователь оказывается способен поставить вопрос о наличии неочевидных закономерностей взаимного тяготения лексики и метра. М. Л. Гаспаров оформил в виде последовательной теории «семантического ореола метра» высказывавшиеся ранее специалистами по русскому стиху идеи о системной связи между темой (и даже набором конкретных образов) стихотворения и избираемым для текста метром [Гаспаров 2013]. Так как тема текста в конечном счете формируется конкретными словами, гипотеза о том, что слово и метр могут формировать исторически обусловленную связь, выглядит логичной, а корпус предоставляет возможности для ее проверки.

М. Л. Гаспаров подчеркивал, что наличие семантического ореола не означает строгой последовательности в его реализации в тексте. Иными словами, классический пример тематического тяготения — тема дороги/пути в стихотворениях, написанных пятистопным хореем, может отсутствовать и, напротив, обнаруживать себя в сти-

хотворениях, написанных другим метром. Таким образом, исследователю приходится иметь дело с нечеткими границами и критериями анализируемого явления.

Однако такое нечеткое на первый взгляд понятие как «тяготение» вполне может быть осмыслено в терминах современной статистики. Задачи, связанные с нахождением взаимного влияния ряда факторов, скажем, в квантитативной лингвистике, решаются с помощью коллострукционного анализа, являющегося расширением традиционного критерия Фишера (подробнее см. Gries, Stefanowitsch 2004; Stefanowitsch, Gries 2003; Кузнецова, Велейшикова 2010)[1]. Эта статистическая методика позволяет устанавливать наличие «притяжения», «отталкивания» или индифферентности двух явлений (например, слов) в случае их совместной встречаемости в выборке.

Собрав материал в поэтическом подкорпусе, мы можем построить матрицу частотности лексем в строках определенного метра и вычислить зависимость или независимость слова от метра в русской поэзии. Ручная работа такого рода была бы чересчур объемной и ее выполнимость за разумное время без корпуса можно поставить под сомнение.

Для демонстрации метода мы избрали двадцать лексем, которые условно можно отнести к метапоэтическим терминам, то есть к классу слов, так или иначе являющихся метонимическими заместителями процесса поэтического творчества: *поэт, лира, поэзия, стих, муза, песнь, вдохновение, перо, слагать, сонет, пиит, элегия, ода, ямб, хорей, гомер, алкей/алцей, тетрадь, строка, строчка*.

Приведем ряд контекстов с этими лексемами в поэтическом корпусе:

Поэт-мечта дремотного Гиганта,
Спит Исполин и слышит звоны **лир**.
[К. Д. Бальмонт. «Задуман безошибочным Творцом…» [Два сонета к Италии, 2] (1892–1935)]

Гордяся им, я посвящаю вам
Игривый звук заздравного **сонета**,

[1] Автор благодарит Ю. Л. Кузнецову за консультации в этом вопросе.

Моей души усердный фимиам.
[Н. М. Языков. «На праздник ваш принес я два привета…» (1829–1833)]

А рядом здесь, под миртовою кущей,
Еще звучит, полна любовной неги,
Гармония Тибулловых **элегий**…
[Д. С. Мережковский. Тибур (1891)]

Жизнь сердца для него единственный предмет
Всех песней пламенных всех томных **вдохновений**;
Жизнь сердца кончится, — в молчаньи и смиреньи
Он укрывается… о, да!
[Е. П. Ростопчина. И он поэт (1840.05.09)]

Интерпретация лексемы *тетрадь* как метапоэтического термина может быть воспринята с сомнением современным читателем, однако контексты, извлеченные из поэтического корпуса, надежно подсказывают, что в стихотворном тексте слово *тетрадь* употребляется прежде всего как место для фиксации творческой (прежде всего, поэтической) активности человека.

Нам вчера
В редакцию товарищ прокурора
Прислал **тетрадь** рифмованного вздора.
[Д. С. Мережковский. Вера (1890)]

Я медленно закрыл **тетрадь**,
Перо подальше отложил
И начал терпеливо ждать,
Чтоб мрачный гость заговорил…
[С. Я. Надсон. Песни Мефистофеля (1886)]

А подле — с пестрыми каймами
Тетради песен и стихов
Счастливой Персии певцов…
[И. И. Козлов. Абидосская невеста (1826)]

Так как нас интересует плоскость семантики стиха, мы попытаемся провести анализ функционирования в поэтическом тексте именно лексем, а не конкретных словоформ. Таким образом, за пределами нашего исследования остаются все просодические особенности слов, проявляющие себя при словоизменении. Так, наращение слогов (*сонет — сонета*), сокращение длины слова (*муза — муз*) или смещение ударения (*стих — стихов*) при склонении в нашей выборке нивелируются. Это особенно важно осознавать, так как интуитивно может показаться, что слова ямбической структуры (с ударением на второй слог) должны тяготеть к ямбу, а слова хореической структуры (с ударением на первый слог) должны тяготеть к хорею. Однако это было бы заблуждением: граница полноударных слов в русском стихе не обязательно совпадает с границами стопы. В следующем примере двусложная словоформа *поэт* занимает только часть трёхсложной стопы анапеста:

Но, деляся с толпой этой тайной души сокровенной,
Изменяет себе, изменяет ли тайне **поэт**…
[О. Н. Чюмина. «Говорят, что порой, совлекая бесстрашно покровы…» (1893)]

В следующем примере словоформа *поэт* оказывается на границе двух дактилических стоп:

Я — не **поэт**, но с любовью глубокой
Музе-шалунье, как эхо, служу…
[Л. Н. Трефолев. Пляска весны (1892.01.20)]

Извлеченные из корпуса данные не нормализованы по авторам, так что остаются основания предполагать, что обнаруженные закономерности обусловлены не традицией русской поэзии, а индивидуальными стилевыми предпочтениями авторов, которые способны влиять на распределение. В то же время корпус объемом в 9,6 млн. словоупотреблений позволяет надеяться на сбалансированность исследованного материала и сглаженность влияния случайных факторов.

В ходе исследования нами последовательно были заданы подкорпуса для каждого из пяти метров русской силлабо-тоники. В каж-

дом из подкорпусов мы провели поиск интересующей нас лексемы. Результаты предварительных замеров приведены в Таблице 1.

Лексема	Я	Х	Д	Ам	Ан	Всего
поэт	3413	575	99	246	181	5603
лира	1440	255	37	66	37	2078
поэзия	707	62	19	40	27	1027
стих	4396	621	84	202	161	6758
муза	2048	243	70	56	55	2797
песнь	3051	1196	325	385	339	6577
вдохновение	210	28	4	9	8	294
перо	965	265	22	57	71	1808
слагать	247	35	10	24	12	381
сонет	487	28	2	14	9	565
пиит	242	13	10	1	7	292
элегия	289	26	4	10	7	352
ода	552	100	10	4	15	758
ямб	88	10	1	11	4	144
хорей	14	20	0	3	2	50
гомер	147	17	4	4	5	238
алкей/алцей	5	0	0	0	0	10
тетрадь	197	31	8	10	14	313
строка	617	118	24	56	44	1131
строчка	132	57	7	27	17	388
подкорпус	5205366	1386732	225491	415453	402174	9671137

Таблица 1. Частотность лексем в подкорпусах метров

Особенность полученных данных в бросающемся в глаза превалировании ямбического подкорпуса над остальными подкорпусами метров. Действительно, ямб — самый распространенный метр русской поэзии. Из-за разницы в объемах корпусов, подчас отличающихся по числу словоупотреблений на порядок, определить тяготение «на глаз» было бы крайне трудно.

Результаты подсчетов с использованием методики коллострукционного анализа приведены в Таблице 2. Плюсом обозначено тя-

готение слова к метру, минусом отталкивание, «0» маркирует нейтральность.

Лексема	Я	Х	Д	Ам	Ан
поэт	+	-	-	0	-
лира	+	-	0	-	-
поэзия	+	-	0	0	-
стих	+	-	-	-	-
муза	+	-	0	-	-
песнь	-	+	+	+	+
вдохновение	+	-	0	0	0
перо	0	0	-	-	0
слагать	+	-	0	0	0
сонет	+	-	-	-	-
пиит	+	-	0	-	0
элегия	+	-	0	0	-
ода	+	0	0	-	-
ямб	0	-	0	0	0
хорей	-	+	0	0	0
гомер	+	-	0	0	0
алкей/алцей	0	0	0	0	0
тетрадь	+	-	0	0	0
строка	0	-	0	0	0
строчка	-	0	0	+	0

Таблица 2. Притяжения, отталкивания и нейтральность лексем и метров по данным коллострукционного анализа

Действительно, ряд метапоэтических терминов обнаружил свое тяготение к определенным метрам. Более всего попавшие в выборку лексемы тяготеют к ямбу: 13 из 20 слов, скорее, следует ожидать увидеть в ямбических строках. При этом статистический курьез в том, что сама лексема *ямб* к ямбу как метру нейтральна, «отталкивается» от хореических стихов. Слово *хорей*, напротив, несмотря на ямбическую просодическую структуру, тяготеет к хореическим строкам и с наименьшей вероятностью встретится в ямбах.

Розы крови, рея,
Дали красный нимб…
Ровный бег **хорея**
Подарил Олимп.
[Е. И. Дмитриева. Хореи (1915)]

Любопытно и обнаруженное тяготение лексемы *строчка* к амфибрахию (независимо от стопности):

Альбом мой наполнен: всё врезано, вклеено;
Хотела еще бы хоть **строчку** от Мея, но…
[Л. А. Мей. (В альбом жене) (1858)]

Достойная **строчек**
О мраморе рук,
Вы платья комочек
Вздымаете вдруг,
[А. П. Ладинский. Как север (1930)]

Не сразу, не всем и не громко
должна я об этом сказать,
то **строчкой** мальчишески ломкой
то просто поглядом в глаза.
[О. Ф. Берггольц. «Мне многое в мире открыто…» (1928)]

Отдельного разбирательства требует случай слова *песнь*, отталкиваемого от ямба, но притягиваемого всеми остальными размерами. Можно выдвинуть еще нуждающуюся в уточнениях гипотезу, что хорей и разнообразные трехсложные размеры в большей степени закреплены за песенным дискурсом, чем «книжный» ямб (мы сознаем, что это утверждение найдет достаточное количество контрпримеров) и это могло повлиять на распределение лексемы в соответствующих подкорпусах.

2. Традиция и формулы

Одной из важнейших составляющих научной стратегии в гуманитарной сфере является установление генезиса описываемого яв-

ления в аспекте его взаимодействия с традицией. В филологии эта стратегия реализует себя в поиске «интертекстуальных параллелей» или источников культурно значимого текста. Поэтический корпус в этом смысле предоставляет наилучший инструмент для такого рода изысканий (см., в частности, Орехов 2013, Орехов 2014). Возможно, в будущем в корпусе может появиться разметка, автоматически маркирующая воспроизводимые фрагменты текста.

Мы называем «параллели» формулами (в духе Кавелти 1996), чтобы не делать различий между сознательным интертекстом и фразеологизованными ритмико-синтаксическими клише, так как в науке о литературе нет надежных способов установления «сознательности» использования автором того или иного приема.

Так как поэтический корпус содержит большинство значимых стихотворных текстов со второй трети XVIII века по середину XX века, поисковые механизмы позволяют легко определять степень распространенности той или иной конструкции, а также ее происхождение и ассоциированность с тем или иным авторским стилем.

Для примера можно, оттолкнувшись от стихотворения Ф. И. Тютчева «Последний катаклизм», проследить какая часть этого небольшого текста является формульной, а какие сочетания слов в контексте русской поэзии представляются оригинальными. При поиске в корпусе мы ограничили время создания попавших в выдачу текстов датой создания «Последнего катаклизма», 1829-м годом.

ПОСЛЕДНИЙ КАТАКЛИЗМ
Когда пробьет последний час природы,
Состав частей разрушится земных:
Всё зримое опять покроют воды,
И божий лик изобразится в них!
<1829>

Слово «катаклизм» до Тютчева в стихах не встречается. Более того, не упоминается оно и в текстах основного корпуса примерно до середины XIX века, по всей видимости, выступая в качестве экзотического «ученого» слова, привлекаемого поэтами барочного мироощущения (см. об этом Шайтанов 1998). Зато формула «когда пробьет» в инициальной позиции стиха не оригинальна и встречается до Тютчева, как минимум, однажды:

Когда пробьет печальный час
Полночной тишины
И звезды трепетно горят,
Туман кругом луны, —
[И. И. Козлов. Ирландская мелодия (1824)]

Тебя умерший не вспомянет,
Когда пробьет кончины час;
Безмолвен ад, — во тьме не станет
Звучать Тебе хвалебный глас.
[Д. П. Ознобишин. Псалом 6 (1820–1840)]

Сочетание «последний час» весьма частотно: до Тютчева оно использовалось не менее 44 раз. А вот конструкция «последний N природы» встречается в корпусе до Тютчева однажды в тексте автора, по признанию литературоведов, сильно повлиявшего на его творчество, что заставляет подозревать в сочетании «последний час природы» сознательное заимствование (однако доказать это, как нам представляется, невозможно):

Представь **последний** день **природы**,
Что пролилася звезд река…
[Г. Р. Державин. На взятие Измаила (1790–1791)]

Сочетание «состав частей» до Тютчева в русской поэзии не фигурирует (разумеется, по данным корпуса), а «части земные» употреблены в оде В. Майкова, однако, в другом значении (у Майкова 'части света', у Тютчева — 'составляющие материального мира'):

Посреди морской пучины
И во всех **частях земных**,
Где сражаются срацины…
[В. И. Майков. Ода на день брачного сочетания цесаревича великого князя Павла Петровича и великия княгини Наталии Алексеевны (1773)]

Конструкция «все зримое» до Тютчева в корпусе не встречается, а сочетание лексем «покрыть» и «вода» в одном контексте сравни-

тельно частотно. Особенное пристрастие к нему имел также повлиявший на стиль Тютчева М. В. Ломоносов: из 12 удовлетворявших запросу контекстов, которые нашлись в корпусе, 5 принадлежат этому автору.

И что надуты **вод** громады
Текли **покрыть** пространны грады,
Сравнять хребты гор с влажным дном.
[М. В. Ломоносов. Ода на день восшествия на всероссийский престол ее величества государыни императрицы Елисаветы Петровны самодержицы всероссийския 1746 года (1746)]

Покрыв водами высоты,
На легких облаках восходишь,
Крилами ветров шум наводишь,
Когда на них летаешь ты.
[М. В. Ломоносов. Преложение псалма 103 (1743–1749)]

Сочетание определяемого «лик» с определением «божий» в поэтическом корпусе до Тютчева встречается лишь однажды с инвертированным порядком в тексте М. М. Хераскова:

На лики божии взираю беспристрастно,
Что в чем-нибудь грешна, сие мне кажет ясно.
[М. М. Херасков. Венецианская монахиня (1758)]

Глагол «изобразиться» в корпусе сравнительно частотный (16 вхождений), в одном контексте с водами, в которых и происходит «изображение» он встречается у того же М. М. Хераскова:

Какою горестью душа твоя разилась,
Когда судьба твоя тебе **изобразилась**,
Когда взглянула ты ко брегу шумных вод,
Где вкруг твоих судов стесняется народ!
[М. М. Херасков. Россиада. Поэма эпическая. (1771–1779)]

Но в целом это, по всей видимости, устаревающий глагол, характерный, в основном, для поэтического языка XVIII века. После

Тютчева можно наблюдать еще 4 вхождения лексемы *изобразиться* в корпусе и ее полное исчезновение для поэзии XX века.

Таким образом, по контрасту с «Проблеском», в языковом смысле настроенном на сентиментальную поэтику В. А. Жуковского (подробнее см. об этом Орехов 2014), «Последний катаклизм» ориентирован на язык одической поэзии XVIII века, что выражается в избираемых Тютчевым формулах, призванных вызвать у читателя-современника соответствующий круг культурных ассоциаций.

Разумеется, исследовательский потенциал поэтического корпуса не ограничивается названными направлениями. Корпус представляет собой научный инструмент, идеально реализующий идею «поиск как язык», которая отражает два важнейших положения:

1. На языке можно порождать неопределённо большое множество заранее не заданных высказываний;

2. Создатель языка заранее не знает, что на нём будет выражено.

Аналогичным образом и поисковые инструменты в идеале должны быть достаточно гибкими и разнообразными, чтобы не ограничивать пользователя заранее придуманными создателем ресурса исследовательскими вопросами. Поэтический корпус полностью соответствует этому требованию и может быть использован в разнообразных лингвистических, стиховедческих и историко-культурных разысканиях.

Литература

Гаспаров М. Л. Метр и смысл. Об одном из механизмов культурной памяти. М.: Фортуна ЭЛ, 2013.

Гришина Е. А., Корчагин К. М., Плунгян В. А., Сичинава Д. В. Поэтический корпус в рамках НКРЯ: общая структура и перспективы использования // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 71–113.

Кавелти Дж. Изучение литературных формул // Новое литературное обозрение. 1996. № 22. С. 33–64.

Кузнецова Ю. Л., Велейшикова Т. В. Современные корпусные исследования языка: новые подходы // Вопросы языкознания. 2010. № 6. С. 108–124.

Орехов Б В. «Не Аполлон, но я кую сей цепи звенья…»: поздние стихи Батюшкова в свете корпусных данных // Корпусный анализ русского стиха. М.: Издательский центр «Азбуковник», 2013. С. 157–171.

Орехов Б. В. «Проблеск» Ф. И. Тютчева в ретроспективе Корпуса. Очерк корпусной поэтики // Корпусный анализ русского стиха: Сборник научных статей. Вып. 2. М.: Издательский центр «Азбуковник», 2014. С. 304–318.

Сичинава Д. В. Поэтический подкорпус Национального корпуса русского языка: несколько примеров поиска стиховедческой информации // Славянский стих. М.: Рукописные памятники Древней Руси, 2012. С. 482–491.

Шайтанов И. О. Уравнение с двумя неизвестными (Поэты-метафизики Джон Донн и Иосиф Бродский) // Вопросы литературы. 1998. № 6. С. 3–40.

Stefanowitsch A., Gries S. Th. Co-varying collexemes in the into-causative // Language, culture, and mind. Stanford, 2004. P. 225–236

Stefanowitsch A., Gries S.Th. Collostructions: investigating the interaction between words and constructions // International journal of corpus linguistics. 2003. 8 (2). P. 209–243

B. V. Orekhov
National Research University Higher School of Economics
(Russia, Moscow)
nevmenandr@yandex.ru

ONCE AGAIN ABOUT THE RESEARCH POTENTIAL OF THE POETIC CORPUS: METER, VOCABULARY, FORMULA

The article continues the trend of other researchers' publications that demonstrate the opportunities of the poetic subcorpus of the Russian National corpus. The question is, what issues related to the history of Russian poetry can be solved with the help of the corpus. In the first part of the article there is a pilot study of attraction of lexical units to poetic meters. The study is conducted with the use of collostructional analysis. As the test material we use "metapoetics terms" — those which are traditionally used when describing the process of creating poetry.

Almost all of them are attracted to iambic verse. The only exception is the word “trochee” that appears more frequently in trochaic lines. The second part of the article reconstructs the poetic tradition, which a poem by Fyodor Tyutchev “Last cataclysm” follows. The poem is considered as composed of a set of poetic formulas (repeated text fragments), each of them with its own history, origin and frequency in Russian poetry. Clarifying these aspects for each formula helps to represent the genesis of the whole poem and embed it in the model of perception of the contemporary reader.

Key words: corpus linguistics, poetry, Russian poets, poetic meter, poetic vocabulary

References

Cawelti J. [The study of literary formulas] *Novoe literaturnoe obozrenie* [New literary review]. Issue 22. Moscow, 1996. pp. 33–64. (In Russ.)

Gasparov M. L. *Metr i smysl. Ob odnom iz mehanizmov kul'turnoi pam'ati* [Meter and Meaning. One of the mechanisms of cultural memory]. Moscow, Fortuna EL Publ., 2013.

Grishina E. A., Korchagin K. M., Plungyan V. A., Sichinava D. V. [Poetic corpus in the RNC: General structure and prospects of use]. *Natsional'nyi korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy* [Russian National Corpus: 2006–2008. New Results and Prospects]. St. Petersburg, Nestor-Istoriya Publ., 2009, pp. 71–113.

Kuznetsova Yu. L., Veleyshnikova T. V. [Modern corpus-based language studies: new approaches]. *Voprosy Yazykoznaniya*, 2010, Issue 6, Moscow, pp.108–124. (In Russ.)

Orekhov B. V. ["Not Apollo, but I make this chain links…": late Batiushkov poems sub specie of corpus data]. *Korpusnyi analiz russkogo stikha* [Russian Verse corpus analysis]. Moscow, Azbukovnik Publ, 2013. pp. 157–171. (In Russ.)

Orekhov B. V. ["Glimpse" of F. I. Tyutchev in Corpus Retrospect. Essay on the Corpus poetics] *Korpusnyi analiz russkogo stikha* [Russian Verse corpus analysis]. Issue 2. Moscow, Azbukovnik Publ, 2014. pp. 304–318. (In Russ.)

Sichinava D. V. [Poetic subcorpus of the National corpus of the Russian language: a few examples of verse study relevant information].

Slavjanskii stikh [Slavic verse]. Moscow, Old Russian Manuscripts Publ., 2012, pp. 482–491. (In Russ.)

Shaytanov I. O. [Equation with two unknowns (the metaphysical Poets John Donne and Joseph Brodsky)]. *Voprosy literatury* [Questions of Literature]. Issue 6. Moscow, 1998. pp. 3–40. (In Russ.)

Stefanowitsch, A., Gries, S. Th. Co-varying collexemes in the into-causative. *Language, culture, and mind. Stanford*, 2004, pp. 225–236

Stefanowitsch, A., Gries, S.Th. Collostructions: investigating the interaction between words and constructions. *International journal of corpus linguistics*. 2003. 8 (2), pp. 209–243

[1]О. Н. Ляшевская, [2]Е. В. Кашкин
[1,2]Институт русского языка им. В. В. Виноградова РАН
[1]Национальный исследовательский университет
«Высшая школа экономики»
[1,2](Россия, Москва)
[1]olesar@yandex.ru, [2]egorkashkin@rambler.ru

ТИПЫ ИНФОРМАЦИИ О ЛЕКСИЧЕСКИХ КОНСТРУКЦИЯХ В СИСТЕМЕ ФРЕЙМБАНК

Система Russian FrameBank (ФреймБанк) является «дочерним» ресурсом Национального корпуса русского языка, который призван решать несколько задач. С одной стороны, ФреймБанк описывает семантику различных типов ситуаций через формализованные схемы — фреймы. Разметка ФреймБанка позволяет увидеть, как распределяется в тексте фреймовая информация: какие участники ситуации выражены, какие нет; в каком порядке выражаются участники и предикат; каким образом семантика ситуации распределяется по синтаксической структуре; с какими другими компонентами коммуникативной структуры (дискурсивными маркерами, отрицанием и т. п.) взаимодействуют элементы фрейма; как все это взаимодействует с контекстом, в частности, что происходит, если участники в двух фреймах перекрываются и т. п.

Помимо словарного ресурса (словарь лексических конструкций включает глагольные, именные и адвербиальные модели управления, устойчивые идиомы и т. п.), ФреймБанк предлагает разметку реализации лексических конструкций в тексте. Поэтому с помощью этого ресурса можно проследить, как часто и какими морфосинтаксическими средствами выражаются элементы фрейма, какие отклонения от стандарта допускаются; какие лексико-семантические

ограничения преодолеваются и с каким развитием значения это связано. Подобные ФреймБанку ресурсы, содержащие информацию об интерфейсе морфосинтаксиса и тематической семантики, могут использоваться для тренировки систем искусственного интеллекта, обучая их извлекать информацию о событиях, стоящих за текстом, с учетом всех различий в оформлении участников (и даже невыраженность участников).

С другой стороны, ФреймБанк является инструментом для изучения конструкционных свойств русской лексики. Здесь можно узнать, какие морфологические, синтаксические, семантические и лексические ограничения на выражение участников задает та или иная конструкция; получить иллюстрации глагольного и именного управления, а также периферийных идиоматизованных конструкций из текстов НКРЯ. Вместе с тем, ФреймБанк позволяет взглянуть на лексические конструкции как на мостик, соединяющий друг с другом разные лексические значения предиката; лексические значения семантически близких слов — но одновременно и увидеть, как другие значения противопоставляются с помощью конструкций. С помощью ФреймБанка мы можем проследить, как лексикон предикатов русского языка накладывается на сеть фреймов и как структурные лексико-семантические отношения типа полисемии, синонимии, антонимии и т. п. отражаются в лексических конструкциях.

Эта статья описывает структуру разметки словаря, банка примеров, графа фреймов и конструкций нашего ресурса.[1]

Ключевые слова: ФреймБанк, FrameNet, русский язык, грамматика конструкций, корпусная лингвистика, лексикография, конструкция, глагол, морфосинтаксис, семантическая роль, фрейм, полисемия.

[1] Проект выполнялся в рамках Программы фундаментальных исследований ОИФН РАН (2009-2011 гг.), направление «Лингвистические аспекты исследования текста», и Программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» (2011, 2012-2014 г.). Основные разработчики – О.Н. Ляшевская, Ю.Л. Кузнецова, Е.В. Кашкин. В разметке принимают участие студенты и сотрудники МГУ, НИУ ВШЭ, РГГУ и других организаций.

1. О системе ФреймБанк

ФреймБанк — это русскоязычный аналог ресурса FrameNet (http://framenet.icsi.berkeley.edu), разработанного под руководством Ч. Филлмора для английского языка ([Fillmore et al. 2003], [Johnson et al. ER]). Как и FrameNet, ФреймБанк является общедоступным онлайн-ресурсом. Основная задача ФреймБанка — это выявление в текстах естественного языка элементов фреймов, а именно, идентификация участников ситуаций, обозначаемых предикатами (глаголами, существительными, прилагательными и т. д.), и разметка способа их выражения — вне зависимости от того, связаны ли обозначающие участников единицы с предикатом синтаксически или нет. В результате разметки, ФреймБанк представляет информацию:

• о лексических конструкциях русского языка и системе фреймов;
• о семантико-синтаксическом интерфейсе в более общем смысле;
• о многозначности предикатных лексем и о том, как система значений связана с конструкционным потенциалом лексем.

1.1. Лексические конструкции

Лингвистический смысл разметки состоит в том, что мы можем на корпусных данных составить каталог лексических конструкций, т. е. таких морфосинтаксических паттернов языка, где как минимум один элемент — предикатное (целевое) слово — фиксирован лексически, а остальные представляют собой «слоты», в которых могут быть представлены разные единицы языка, связанные, однако, схожим способом оформления и определенной семантической общностью. Интерпретация отношений между участниками, представленными в каждом элементе конструкции, также должна быть сходной:

Мама **жарила** картошечку.
Дом, безусловно, ваш, но картошку **жарил** я.
Шашлыки наши разбойнички научились **жарить**, научатся и аманатов брать.

Рис. 1. Элементы конструкции *Мама жарила картошку*.

Среди ночи она **жарила** ему яичницу на электроплитке.
Жарил себе мясцо расхристанный, сердобольный мужик — на плитке, будто творил чудо.
... как раз в тот момент, когда Игрек **жарил** себе на времянке лепешки.

Рис. 2. Элементы конструкции *Она жарила ему яичницу на электроплитке*.

Конструкции, представленные на Рис. 1 и 2, отличаются количеством слотов. Конструкция, представленная на Рис. 3, формально похожа на Рис. 1, но отличается интерпретацией отношений между участниками, а также тем, что участник — прямое дополнение относится к иному лексико-семантическому классу и является лексически фиксированным[1] (*спирт*).

[1] Термин «лексически фиксированный элемент», конечно, до определенной степени условен. Хотя в текстах НКРЯ не удается найти других лексических вариантов для участника *спирт*, однако словообразовательные и близкие лексические расширения вполне ожидаемы в текстах электронной коммуникации и устной речи, ср. *Давно, в молодости, я работал в НовоУренгое .... там тундра....и за коротанием Белых ночей, после Вахты, мы жарили спиртягу чистую и курили беломор*... (http://www.musicforums.ru/vocal_arc/1203676105.html); *сиди жарь самогон и молись*... (newsland.com›news/detail/id/1190108/). Представляется, что полный запрет на вариативность полнозначной лексики шел бы вразрез с пониманием креативной и игровой составляющей языка.

Я читаю стихи проституткам И с бандитами **жарю** спирт. 'распивать крепкие алкогольные напитки (об уголовниках или алкоголиках)'

Рис. 3. Элементы конструкции *Я с бандитами жарю спирт* 'распивать крепкие алкогольные напитки (об уголовниках или алкоголиках)'.

Банк лексических конструкций включает:

• модели управления глаголов[1] (см. 1–2) и предикативов (см. 3–4);

• морфосинтаксические конструкции предикатных имен существительных (включают не только управление, но и атрибутивные связи и другие синтаксические комплексы отношений, в т. ч. контроль связи между именем-предикатом и аргументом с помощью глагола-лексической функции, см. (5–6); реляционные предметные имена типа *пациент*, *корпус* также считаются предикатными);

• морфосинтаксические конструкции прилагательных, наречий, вводных оборотов, разнообразных сирконстантных элементов[2] (см. 7–14);

• конструкции малого синтаксиса (фраземы, идиоматические конструкции) — в них, как правило, лексической константой является не только предикатное (целевое) слово, но и другие элементы, и наблюдается высокая некомпозиционность значения (см. Рис. 4).

(1) Snom жарить.V Sacc
Мама ***жарила*** *картошечку*.
(2) Snom жарить.V Sacc на + Sloc Sdat
Игрек ***жарил*** *себе на времянке лепешки*.

[1] Обратим внимание, что в корпусе обнаруживаются и достаточно частотные нестандартные случаи аргументно-предикатных конструкций типа *собрал силы и встал*; *собрав все свои силы, он заиграл*; они являются расширением стандартных моделей управления типа *собрал силы, чтобы встать* и *собрал последние силы для прыжка* [Апресян 2010].

[2] С синтаксической точки зрения прилагательные в атрибутивной позиции, наречия и др. элементы в этой группе не являются предикатами, однако они по-прежнему обладают способностью формировать вокруг себя морфосинтаксический паттерн.

(3) Sdat нельзя.PRAEDIC Vinf

Вам ***нельзя*** *ходить пешком*

(4) PRгде + Sx холодно.PRAEDIC

В доме ***было холодно****, мокро и пахло гарью.*

(5) Snom [прийти.SUPPORT к +] соглашение.Sdat о + Sloc

Без особых сложностей обе стороны пришли к ***соглашению*** *о создании в каждой из столиц специальных миссий.*

(6) соглашение.S между + Sins {A | CL}

С тех пор между ними мирное ***соглашение****: вне дома главная Наташа.*

(7) верный.A S

Это же ***верный*** *человек, это повод для перевода на более ответственную работу.*

(8) верный.A Sdat S

... скоро в МВД не останется ***верных*** *делу профессионалов.*

(9) Snom верный.Abrev Sdat

Большинство пенсионеров по-прежнему ***верны*** *Зюганову.*

(10) верный.A S

Ну что ты, кто решится? За такую штуку ***верных*** *десять лет дадут.*

(11) верно.ADV + V

Руководство корпорации смогло ***верно*** *оценить ситуацию.*

(12) верно.ADV + A

Знатная, богатая вдова… — И ***верно*** *прекрасная? — спросил Честон.*

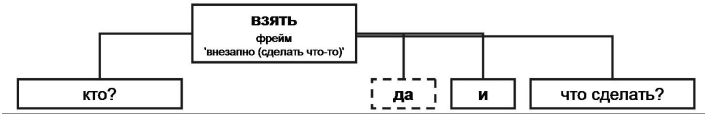
(13) верно.PRAEDIC + CL

— Кушать хотите, ***верно****? — спросил милый и близорукий.*

(14) себя.SPROdat V

Спит ***себе*** *человек и спит, может, он устал с дороги.*

...а старики пускай ***себе*** *дерутся, коли им это весело.*



Он сделал последний контрольный ход (допустим — 40-й), перевел часы, а флажок-предатель **взял** да и рухнул.
А полковник **возьми** да и откажись наотрез.

Рис. 4. Элементы конструкции *Взял и запил*.

Итак, лексическая конструкция — это ключевая лексема (лексическая константа; иногда их может быть несколько) и некоторый набор слотов. Количество слотов, их возможное морфологическое оформление, их возможное лексическое наполнение, тип синтаксической связи между элементами, интерпретация конструкции — вот те составляющие, которые в комплексе отличают лексические конструкции друг от друга. Конструкции могут иметь одинаковое морфосинтаксическое оформление, но разную интерпретацию, и наоборот, сходную интерпретацию и разное морфосинтаксическое оформление.

Лексические конструкции входят в иерархию вложенных классов конструкций. Различные генерализации связывают конкретные цепочки словоформ в тексте (ср. *хочу пить*) с лексическими конструкциями (ср. *хочу Vinf*, *хотеть Vinf*) и далее с конструкциями более абстрактного уровня (например, с инфинитивной конструкцией с волитивным глаголом; с общей инфинитивной конструкцией). Генерализации предполагают обобщение плана выражения конструкций и обобщение плана содержания, когда индивидуальные черты конструкций стираются — см. [Dabrowska, Tomasello 2008], [Dabrowska 2004], [Tomasello 2003] об усвоении конструкций носителями языка через генерализации.

Генерализации возможны и «по горизонтали», в частности, на уровне лексических конструкций. Конструкции одного предиката, различающиеся количеством элементов и их оформлением, также могут объединяться в классы, см. (1) и (2). Основанием для объединения является то общее в их семантике, что представляет один фрейм.

1.2. Фреймы

Фрейм (в лингвистическом понимании) — типовая ситуация, в которой заданы участники и некоторое отношение между ними (например, действие, состояние, свойство, связывающее их). В языке фрейм ассоциируется с определенным предикатом (а чаще с несколькими близкими по смыслу предикатами), при котором выражаются участники фрейма (аргументы).

Например, фрейм «обработка при помощи высокой температуры» состоит из следующих единиц:

Продукт
Источник тепла
Контейнер

В типовой ситуации Продукт, находясь в Контейнере, подвергается нагреванию от Источника тепла. Скорее всего, в результате этого процесса Продукт претерпевает изменения (становится теплым и съедобным). Иногда в тексте могут быть выражены некоторые дополнительные элементы, участвующие в данном фрейме. Например, это может быть Температура, ср. 120 °C в примере (17) или Образ действия, ср. *медленно* в примере (18).

Фрейм «обработка пищи при помощи высокой температуры» представляют в русском языке такие глаголы, как *вариться, жариться, готовиться, томиться* и др. (ср. англ. *bake, boil, cook, fry, stew* и др.).

(15) На толстых столах бурно кипели, отражая нестерпимое черноморское солнце, кривые самовары, и **жарилась** на сковородах украинская колбаса.

(16) После тщательного перемешивания кашу ставят **томиться** в печь или духовку.

(17) Благодаря герметически закрывающейся крышке пища **готовится** при 120 °C.

(18) Конина **варилась** медленно, жрать хотелось невыносимо.

Само английское название frame отсылает к «рамке», или интерпретации некоторой ситуации под определенным углом. Фреймы «покупка» и «продажа» являются частным случаем фрейма «товар-

но-денежная транзакция», представляя разную перспективу на ситуацию (со стороны покупателя или со стороны продавца). Аналогично, можно выделить фрейм «эмоция в перспективе экспериенцера» (ср. *я волнуюсь из-за экзаменов*) и «эмоция в перспективе причины, ее вызывающей» (ср. *предстоящий экзамен волнует меня*). Описанный выше примеры фрейм «обработка пищи при помощи высокой температуры» представляет изменение Продукта как «процесс в себе», а введение в рассмотрение человека, целенаправленно проводящего эту обработку, дает новый каузативный фрейм (ср. *Папа жарит сосиски*).

Важно, что фреймы связаны друг с другом в сеть. Такой фрейм, как «обработка пищи при помощи высокой температуры», будет частным случаем фрейма «обработка пищи», а тот, в свою очередь, частным случаем фрейма «физическое воздействие».

Фреймы могут быть связаны и как части сценария. Например, фреймы «получение работы», «работа (в должности)» и «увольнение с работы» являются подсценариями фрейма «наемная работа».

Примечание. Фрейм в понимании инженерии знаний — более общий способ представления окружающего мира, связанный с его членением на типовые ситуации и их сценарии. Например, всю совокупность отзывов о гостиницах можно представить в виде единого сценария, где говорящий описывает свой приезд, оформление на ресепшен, вид номера (большой, маленький), впечатления о кровати, санузле и т. п., понравился ли завтрак, отношение персонала, выезд из отеля и прочее. Специалисты по кластеризизации новостей и другим прикладным информационным задачам также обращают внимание на «модус», который задают фреймы в отношении одного и того же объективного события, ср., например, *Президент поздравил дзюдоистов с серебром на чемпионате мира* и *Дзюдоисты упустили золотые медали*. В информационных войнах симметричные события могут представляться «несимметрично», ср. *Израиль выпустил 30 ракет по территории сектора Газа* и *На израильской территории разорвались новые ракеты* (каузатор события, террористическая организация «Хамас», не упомянута).

В понимании инженерии знаний фрейм не ассоциируется напрямую с выражающими его предикатами естественного языка, однако можно видеть, что фрейм прибытия в отель можно соотнести с лингвистическими фреймами «кто приехал куда», «поздно (прибыть)», «кто выдал ключи кому», «быстро (выдать ключи)» и т. д.

Как уже было сказано, семантику лексической конструкции можно соотнести с тем или иным фреймом, а еще точнее, фрейм отражает прототипическое значение конструкции [Hamm 2007], результат генерализации индивидуальных значений лексических конструкций. Встает вопрос, где проходят границы между фреймами, насколько различия в перспективе должны быть учтены — и на этот вопрос нет четкого ответа.

Модус фрейма — это часть коммуникативной организации текста, а для нее важен и порядок представления элементов (порядок слов), и их статус (полная номинация, анафорическая отсылка к предшествующему контексту, опущение), и способ оформления (как морфосинтаксический, так и интонационный), и многое другое. Если вернуться к лексическим конструкциям, то можно вспомнить утверждение современной теории грамматики конструкций, что любое различие во внешнем оформлении конструкции влечет различие в семантике, ср. такие близкие примеры как *рассказывать о даче* и *рассказывать про дачу*. Получается, что классификация фреймов не имеет дна.

Тут, однако, вступают в силу практические соображения. Во FrameNet, а вслед за ним и во ФреймБанке фреймами нижнего уровня признаются обобщенные, типовые конструкты, примерно соответствующие одному лексическому значению слова в словаре. Можно указать следующие соглашения:

1. Если смена «перспективы наблюдения» и морфосинтаксического оформления влечет за собой смену лексического якоря, т. е. предиката (ср. *купить — продать, уволиться — уволить, преступление — халатность* и т. п.), то фреймы считаются разными. В противном случае говорят о повышении/понижении коммуникативного ранга участников внутри фрейма.

2. Каузативный и некаузативный фреймы в обязательном порядке признаются разными, хотя и связанными фреймами. Однако пассивная и активная диатезы, безличное, неопределенно-личное и обобщенно-личное употребление и другие регулярные средства мены коммуникативного статуса участников реализуют один фрейм.

3. Интерпретация функций (ролей) участников при отождествлении нескольких описываемых ситуаций как представляющих один фрейм не имеет противоречий.

4. Один и тот же набор ключевых участников может представлять разные фреймы. Различие происходит из интерпретации отношений между участниками и интерпретации сценария фрейма, которые задаются предикатом (предикатами).

1.3. Семантико-синтаксический интерфейс текста

Имея массив размеченных примеров, мы можем показать семантические и синтаксические комбинаторные возможности предикатных слов. Лексические конструкции задают ограничения на количество элементов в конструкции, морфологический и синтаксический способ их выражения, лексические способы выражения элементов, порядок слов, интонацию и т. д. Можно проследить, с помощью каких лексических конструкций выражается значение фрейма, как способ выражения участника зависит от способа выражения предикатного слова, насколько регулярно выражается тот или иной элемент (например, элемент *да* в конструкции *взял да и V* факультативен, ср. *взял да и выпил*, *взял и выпил*) и от каких факторов это зависит. Интересно проследить плотность выражения участников в тексте — всегда ли участники выражаются при предикате или упоминаются в более широком контексте, связаны ли предикат и участник синтаксически (возможно, например, они связаны через слово-контролер, ср. *разбойнички научились жарить*), каков синтаксический ранг выражаемого элемента — высокий (например, Субъект или Объект) или низкий (например, периферийная предложная группа), каков порядок упоминания элементов в тексте и т. д. — то есть интересно проанализировать все параметры информационной (коммуникативной) структуры текста. С помощью размеченных данных можно проанализировать дистрибуцию слов тех или иных лексических групп, выражающих участников — с какими предикатами они встречаются чаще всего, когда они более всего подвержены эллипсису и т. п. Наконец, ФреймБанк представляет сведения о дистрибуции сирконстантов, т. е. обстоятельствах места, времени и др. относительно разных классов предикатов и конструкций.

1.4. Многозначность предикатных лексем и их конструкционный потенциал

Фрейму как семантической единице может соответствовать несколько предикатных лексем естественного языка. Вместе с тем, многие предикатные лексемы соотносятся с более чем одним фреймом, поскольку у них может быть более одного значения, могут иметься метафорические расширения значений, лексемы могут входить в разные идиоматизованные, фразеологически связанные употребления. ФреймБанк показывает, как сеть значений предикатных слов накладывается на сеть фреймов. Во-первых, фрейм ассоциируется с набором лексем, употребленных в определенном лексическом значении. Во-вторых, он ассоциируется с набором лексических конструкций каждой из этих лексем. Конструкции представляют то или иное подмножество обозначаемых участников фрейма, с одной стороны, и формальный способ их выражения, с другой.

Таким образом, ФреймБанк позволяет ответить на вопросы:

• Похожи ли конструкции разных предикатных слов, представляющих фрейм?

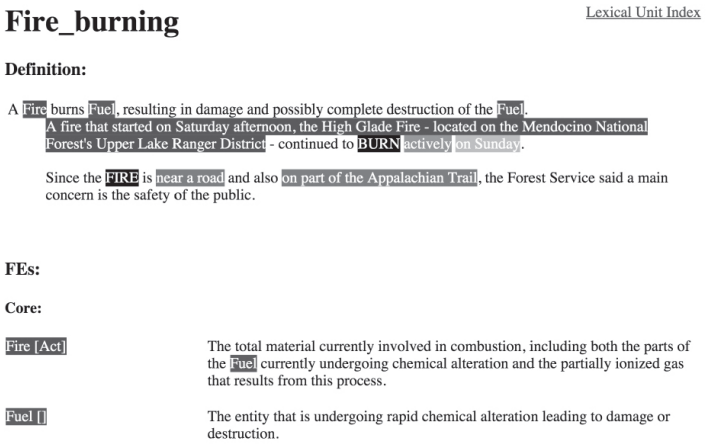
• Похожи ли конструкции предикатного слова, употребленного в разных значениях, т. е. конструкции, представляющие разные фреймы?

• Каким образом устроен механизм развития конструкционных возможностей предикатной лексики?

2. ФреймБанк и FrameNet

FrameBank относится к семейству FrameNet-ориентированных ресурсов. Первая, наиболее известная и полная система FrameNet была создана для английского языка, в дальнейшем появились версии FrameNet для японского, китайского, бразильского португальского, шведского, испанского и немецкого языков (см. https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages), которые клонируют полностью или частично структуру и формат английской системы. Кроме того, к FrameNet-ориентированным ресурсам можно отнести современные корпусно-ориентированные словари валентностей, глагольные компоненты WordNet-ов, а также

корпуса с глубокой разметкой, отражающей аргументную структуру, кореференцию, дискурсивные стратегии и мн. др. (ср., например, тектограмматику в Prague Dependency Treebank, тестовые корпуса для машинного перевода с Compreno-разметкой ABBYY и др.).



Fire_burning

Lexical Unit Index

Definition:

A Fire burns Fuel, resulting in damage and possibly complete destruction of the Fuel.
A fire that started on Saturday afternoon, the High Glade Fire - located on the Mendocino National Forest's Upper Lake Ranger District - continued to BURN actively on Sunday.

Since the FIRE is near a road and also on part of the Appalachian Trail, the Forest Service said a main concern is the safety of the public.

FEs:

Core:

Fire [Act]	The total material currently involved in combustion, including both the parts of the Fuel currently undergoing chemical alteration and the partially ionized gas that results from this process.
Fuel []	The entity that is undergoing rapid chemical alteration leading to damage or destruction.

Рис. 5. Фрагмент представления фрейма Fire_burning ‘гореть в огне’ в системе FrameNet (определение, пример и ядерные фреймовые элементы).

FrameNet, как следует из его названия, строится вокруг сети фреймов — типовых ситуаций с известным набором участников и расписанными ролями. Различаются фреймы высокого уровня (например, посессивного отношения, купли–продажи), промежуточного уровня (например, купля–продажа в перспективе продавца VS в перспективе покупателя) и самого частного уровня (например, фрейм, кодируемый глаголом *to shop* и существительным *shopping*). FrameNet связывает фреймы с набором предикатов, которые могут обозначать эти ситуации, и с набором средств выражения участников в тексте. В традиции лексической семантики фрейм соответствует одному из значений слова, а кодирование участников — конструкциям, строящимся вокруг лексически фиксированного центра, например, моделям управления глагола (или группы глаголов).

Структурно FrameNet состоит из следующих разделов:

• Словарь фреймов — схематичное толкование ситуации, в ко-

торой представлены ее участники, называемые по их роли (= фреймовые элементы), см. Рис. 5. Каждый фрейм снабжается иллюстративными примерами, а также списком глаголов и других предикатных слов, ассоциированных с этим фреймом. Например, к фрейму Fire_burning приписаны глаголы *blaze, burn* и *rage*, а также существительные *backfire, blaze, bonfire, fire, flames, hot spot, inferno* и др.

• Граф фреймов — надстройка над словарем, в которой фреймы связаны по нескольким типам связей, см. Рис. 6.

• База лексических единиц, представляющих фреймы — статистика способов формального выражения участников фрейма, см. Рис. 7 и 8.

• Аннотации — примеры, в которых размечены группы и клаузы, выражающие участников, см. Рис. 9.

• Конструкции малого синтаксиса типа *the more… the better…* и грамматические конструкции типа *four dollars an hour* выделены в отдельный блок "Constructicon", см. [Fillmore et al. 2012].

Помимо основной части, строящейся от фрейма к примерам, во FrameNet был реализован проект сплошной аннотации текстов.

Frame-frame Relations:

Inherits from:
Is Inherited by: Emergency_fire
Perspective on:
Is Perspectivized in:
Uses:
Is Used by:
Subframe of:
Has Subframe(s):
Precedes: Fire_going_out, Putting_out_fire
Is Preceded by: Catching_fire, Setting_fire
Is Inchoative of:
Is Causative of:
See also:

Рис. 6. Связи фрейма Fire_burning с другими фреймами в графе фреймов.

Frame Element	Number Annotated	Realization(s)
Depictive	(1)	PP[in].Dep (1)
Duration	(1)	PP[for].Dep (1)
Fire	(17)	NP.Ext (15) CNI.-- (1) DNI.-- (1)
Fuel	(8)	NP.Obj (5) INI.-- (1) NP.Dep (1) NP.Ext (1)
Manner	(2)	AVP.Dep (2)
Place	(10)	2nd.-- (1) PP[on].Dep (2) PP[throughout].Dep (2) N.Dep (1) PP[at].Dep (1) PP[in].Dep (3)
Subregion	(1)	PP[to].Dep (1)
Time	(2)	AJP.Dep (1) PP[into].Dep (1)

Рис. 7. Паттерны формальной реализации участников фрейма Fire_burning (фрагмент).

2 TOTAL	Fire	Fuel	Place	
(1)	NP Ext	NP Obj	PP[on] Dep	
(1)	NP Ext	NP Obj	PP[throughout] Dep	
1 TOTAL	Fire	Fuel	Time	
(1)	CNI --	NP Obj	AJP Dep	
1 TOTAL	Fire	Manner	Place	
(1)	NP Ext	AVP Dep	N Dep	
5 TOTAL	Fire	Place		
(1)	NP Ext	PP[at] Dep		
(2)	NP Ext	PP[in] Dep		
(1)	NP Ext	PP[on] Dep		
(1)	NP Ext	PP[throughout] Dep		
1 TOTAL	Fire	Place	Time	
(1)	NP Ext	PP[in] Dep	PP[into] Dep	

Рис. 8. Поверхностные реализации валентных паттернов фрейма Fire_burning (фрагмент)

5. [$_{\text{Fire}}$The first , a 8,000 hectare blaze near Blackheath] , [$_{\text{Fire}}$which] has been *BURNING*^{Target} [$_{\text{Duration}}$for nine days] broke containment lines at 11 PM AEDT on Monday .
6. [$_{\text{Fire}}$Wildfires in California] continued to spread on Tuesday , having *BURNED*^{Target} [$_{\text{Fuel}}$more than 140,000 acres of land] [$_{\text{Place}}$throughout the state] and caused an estimated US$13 million worth of damage .
7. [$_{\text{Fire}}$A second fire] has begun *BURNING*^{Target} [$_{\text{Place}}$on a 5 km (3 mile) front , along the Putty Road near Wiseman 's Ferry] .
8. [$_{\text{Fire}}$The blaze] started *BURNING*^{Target} last Wednesday north of Los Angeles .
9. [$_{\text{Fire}}$Fire] *BURNS*^{Target} [$_{\text{Place}}$at Barangaroo construction site , Sydney , Australia]

Рис. 9. Разметка примеров, представляющих фрейм Fire_burning.

Разработчики системы ФреймБанк не стали идти по пути клонирования FrameNet, поскольку тогда мы были бы связаны и с идеологическими, и с формальными ограничениями, накладываемыми этой системой. На наш взгляд, архитектура системы FrameNet идеологически отстает от активно развивающейся теории грамматики конструкций. Кроме того, эта архитектура не позволила бы отразить все особенности конструкций во флективных языках, к каковым относится и русский. Но что самое важное, система клонирования FrameNet-ов строится на сомнительном принципе универсальности фреймов во всех языках мира[1]. Нам представляется гораздо более ценным связать фреймы во ФреймБанке с фреймами в системах FrameNet, VerbNet, WordNet и др., не обязательно во взаимно-однозначном соответствии, а не строить систему от фреймов английского языка.

Во ФреймБанке основной единицей являются конструкции индивидуальных глаголов, а не обобщенные фреймы. Тем самым, подчеркивается своеобразие конструкций каждого глагола, уникальность всякой комбинации формы и значения как конструкции. Принимаются во внимание индивидуальные особенности конструкций у синонимов и даже диатетические вариации при одной лексеме. Как результат, русский ФреймБанк ориентирован в большей мере

[1] Этот принцип не раз высказывался создателями английского FrameNet Ч. Филлмором и М. Петрук (персональная коммуникация); см. также [Fillmore, Atkins 2000]; [Burchardt et al. 2009]; [Ohara 2009]; [Subirats 2009] и др. Х. Боас предлагает компромиссное решение, при котором существует универсальный костяк и отдельные культурно-специфичные фреймы, такие как англ. *to date, sugar daddy*, фин. *saunoa* ‘ходить в сауну’ [Boas 2011]. Однако, как показывает лексическая типология, даже самые ядерные фреймы, такие как *думать, считать, знать*; *стоять, сидеть, лежать*; *кидать* и *бросать*; *резать* и *разбивать*; *видеть* и другие фреймы восприятия не находят полного соответствия в языках мира, см. [Рахилина, Резникова 2013] и приводимые там ссылки.

на исследования лексических конструкций в духе Грамматики Конструкций (Ч. Филлмор, А. Голдберг и др.) и Московской семантической школы (Ю. Д. Апресян, Е. В. Падучева и др.), нежели на теорию Frame Semantics ([Fillmore 1977b, 1982]; [Hamm 2007]), лежащую в основе системы FrameNet.

Отрицая идею универсальности набора фреймов в языках мира, FrameBank направлен на анализ:

• русской лексической системы, структуры лексико-семантических групп и полисемии;

• парадигматических отношений между значениями многозначных слов — прежде всего в том, как они отражаются в системе связанных с этими значениями лексических конструкций;

• лексико-семантических ограничений на слоты конструкций;

• грамматических особенностей русского языка (порядок слов, падежи, согласование и т. п.).

ФреймБанк спроектирован с учетом традиций отечественной лексической семантики и специфики русского языка, где информация о предложно-падежной реализации управления предикатов и поверхностно-синтаксических свойствах других конструкций имеет особую ценность. Эта компьютерная система должна решать не только задачи словаря (ср. электронную систему «Лексикограф», см. [Падучева 2004]; [Кустова и др. 2005]; изданные на бумаге словари [Апресян, Палл 1982], [Сазонова 2008], лексикографические проекты [Азарова и др. 2004], [Апресян 2008] и др.), но и представлять аннотированный корпусный материал. Этот ресурс реализует принципы гибридных систем, в которых авторитетные лексикографы видят будущее словарей ([Atkins 1992], [Kilgarriff et al. 2006]): словарь в выходом в корпус. С одной стороны, это словарь — но с возможностью расширения иллюстраций за счет поиска в корпусе. С другой стороны, это корпус — но с экспертным отбором примеров. Наконец, это лексически ориентированный ресурс, в котором выборка примеров строится с опорой на конкретные лексемы, однако в каждом предложении разметке подлежат все предикатные слова и связанные с ними конструкции.

3. Словарь лексических конструкций

3.1. Шаблон конструкции

Словарь лексических конструкций представляет каждую конструкцию как шаблон (см. пример на Рис. 10), в котором указаны:

а) морфосинтаксические характеристики элементов конструкции;

б) синтаксический ранг участника;

в) экспликация (роль) участника;

г) лексико-семантические ограничения на заполнение слота конструкции;

д) статус участника: обязательный или факультативный;

е) буква, маркирующая участника в кратком паттерне вида *Y-а занесло из Z в W*.

Схема: Sacc Vimpers {ADV / PRкуда + Sx} {ADV / PRоткуда + Sx}
Пример: Рыбку занесло из речки в соленый океан.

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
Y	Sacc	пациенс перемещения	Объект	-
-	занести	-	Предикат	-
Z	{ADV / PRкуда + Sx}	конечная точка	Периферия	-
W	{ADV / PRоткуда + Sx}	начальная точка	Периферия	-

Рис. 10. Шаблон одной из конструкций глагола *занести* в значении перемещения

Например, для глагола *резать* выделено около 30 шаблонов конструкций (обозначаемых для наглядности ярлыками-примерами), среди которых есть кластер шаблонов *Продавщица* ***режет*** *сыр, Она* ***режет*** *хлеб на тонкие куски, Василий начал* ***резать*** *хлеб длинными ломтями, Портной* ***режет*** *ткань большими ножницами*, реализующих с помощью различных моделей управления значение глагола ‘разделять объект на части давлением острого инструмента’, кластер шаблонов *Старик* ***резал*** *четки из кипариса* и *Он* ***резал*** *деревянные ложки простым ножом*, соответствующих значению ‘изготавливать что-л. посредством резьбы’, а также кластер *В боку* ***режет*** и *У него в желудке* ***резало***, где глагол *резать* описывает определенный тип болевого ощущения.

Морфосинтаксические признаки задаются сочетанием частеречных и грамматических признаков, в основном в нотации НКРЯ, ср. Snom (имя существительное в именительном падеже), SPROdat (местоимение-существительное в дательном падеже), Ains (имя прилагательное в творительном падеже), ADV (наречие), Vinf (глагол в инфинитиве), CL (клауза). Предложно-падежные группы задаются следующим образом: *за* + Sacc (предлог *за*, управляющий именем в винительном падеже), *за* + Sins (предлог *за*, управляющий именем в творительном падеже. Вслед за [Апресян, Палл 1982] выделяются группы пространственных предлогов: PRгде + Sx (местное значение, предлог отвечает на вопрос «где?»), PRкуда + Sx (направительное значение, предлог отвечает на вопрос «куда?»), PRоткуда + Sx (направительное значение, предлог отвечает на вопрос «откуда?»), PRтраект + Sx (значение траектории, x везде обозначает падеж, который требует тот или иной предлог). Помета PR + Sx обозначает более общий вариант оформления предложной группы, чаще всего со значением времени или способа, ср. в конструкции *кончить работу к вечеру, до десяти часов утра, в восемь часов, задолго до звонка, за пять минут до конца урока* и т. д. Варианты поверхностной реализации обобщенных предложно-падежных групп представлены в Таблице 1, в порядке убывания частотности (по данным Фрейм-Банка). Возможные варианты оформления участников указываются в фигурных скобках, ср. {ADV/PRкуда + Sx}, {Vinf/что + CL}.

Группа предлогов	Варианты поверхностной реализации предложных групп
PRгде + Sx	в + Sloc, на + Sloc, за + Sins, у + Sgen, над + Sins, по + Sdat, под + Sins, среди + Sgen, между + Sins, перед + Sins, вокруг + Sgen, возле + Sgen, около + Sgen, подле + Sgen, меж + Sgen, внутри + Sgen, кругом + Sgen, вне + Sgen, при + Sloc
PRкуда + Sx	в + Sacc, на + Sacc, к + Sdat, под + Sacc, за + Sacc, навстречу + Sdat, до + Sgen, за + Sins, кругом + Sgen, на + уровень + Sacc, наперерез + Sdat, внутрь + Sgen
PRоткуда + Sx	из + Sgen, с + Sgen, от + Sgen, из-за + Sgen, из-под + Sgen
PRтраект + Sx	по + Sdat, через + Sacc, в + Sacc, между + Sins, под + Sins, за + Sins, вокруг + Sgen, мимо + Sgen, сквозь + Sacc

PR + Sx	в + Sloc, на + Sacc, по + Sdat, в + Sacc, до + Sgen, с + Sins, без + Sgen, прежде + Sgen, для + Sgen, после + Sgen, вокруг + Sgen, с + Sgen по + Sacc, при + Sloc

Таблица 1. Варианты поверхностной реализации обобщенных предложных групп по данным разметки примеров во ФреймБанке.

В нотации предикатов глаголов могут быть указаны грамматические ограничения, накладываемые данной конструкцией на форму предиката, ср. Vimpers (глагол в безличном употреблении), Vpass (глагол в пассиве[1]), Vpartcp.pass (глагол в форме пассивного причастия), PRAEDIC (предикат), Abrev (краткое прилагательное) и т. п.

В целом такой способ представления поверхностно-синтаксического участников отражает традиции грамматики зависимостей (ср. практику ТКС, Апресян, Палл 1982, системы «Лексикограф» и др.), однако оформление участников может обозначаться и в нотации, принятой в грамматике составляющих. В последнем случае частеречные пометы заменяются на NP (именная группа), VP (глагольная группа), PP (предложная группа), ADVP (наречная группа) и др.; помета CL (клауза) остается без изменений.

Экспликация — это адаптация понятия «семантическая роль» к разным типам предикатов. Как показал опыт развития FrameNet, традиционный список семантических ролей неизбежно разрастается при расширении круга размеченных глаголов и увеличении тематического разнообразия описываемых ими ситуаций, а применительно к именам существительным и прилагательным инвентарь ролей вообще не разработан. В силу этих обстоятельств было принято соглашение, что в поле «Экспликация» может быть указана семантическая роль или стоять иное пояснение, помогающее отождествить участника, например, в модели управления имени *слуга* (ср. *слуга Петра*) участник Петр может быть обозначен как «хозяин» или «тот, кому служат»[2]. Инвентарь ролей мы обсуждаем ниже, см. раздел 5.

В поле синтаксического ранга участника содержатся пометы

[1] Указывается только для конструкций, в которых употребление предиката в активном залоге невозможно.

[2] Технологически это устроено так, что аннотатор, вводящий в словарь новую конструкцию, может поставить любую помету, а затем редактор унифицирует пометы согласно базовому списку ролей.

Предикат, Субъект, Объект (для прямого объекта), Периферия (все прочие падежные, предложные, инфинитивные, наречные и т. п. группы), Клауза (для клауз с союзным и бессоюзным присоединением), Отрицание (для конструкций с обязательно встроенным отрицанием) и др. Для атрибутивных конструкций введены пометы Атрибут (для предиката) и Контролер согласования (ср. *верный друг*).

Лексические и лексико-семантические ограничения на заполнение слотов конструкции записываются по возможности в стандарте тегов семантической разметки НКРЯ, ср. inan/неодушевленный, abstr/абстрактный, hum/лицо, org/организация, food/пища, famn/фамилия, persn/имя и т. п., см. поле «Сем. ограничения» на Рис. 11. В случае жестких лексических ограничений здесь может быть перечислен список конкретных лексем русского языка, допустимых в данной конструкции (ср. конструкцию типа *батюшка служит молебен*, в которой объектная валентность может быть заполнена именами *служба, молебен, панихида, обедня*).

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
X	Snom	агенс	Субъект	лицо
-	служить	-	Предикат	-
Y	{ADV / PRгде + Sx}	место	Периферия	организация

Рис. 11. Глагол *служить*, конструкция *Дядя служил в министерстве юстиции*.

По традиции представления моделей управления в русской лексикологии и лексикографии каждый участник имеет буквенное «имя» типа X, Y, Z. Лексические константы обозначаются прочерком. Так же как и в системе «Лексикограф», во ФреймБанке буквы могут использоваться для отождествления участников в конструкциях одного фрейма, в конструкциях разных фреймов и в конструкциях разных глаголов.

Если участник факультативен, то его буквенное имя ставится в квадратные скобки: [Z], [R], [-] и т. п. В целях более точного описания русских конструкций в словаре мы по возможности «распарили» конструкции с факультативными участниками: теперь они пред-

ставляются как две и более связанных конструкции с обязательным набором участников. Факультативные участники оставлены лишь в зоне лексических констант, прежде всего, это факультативные союзы, частицы и т. п.

В пределах словарной статьи одного глагола или другого предиката все лексические конструкции получают свое условное название (например, «служить в министерстве», «служить царю» и т. д.) и связаны между собой отношениями «синтаксической деривации» [Падучева 2004].

Способ представления моделей управления и других лексических конструкций в словаре заимствован в основных чертах из системы «Лексикограф» [Падучева 2004]. Пилотная версия глагольной части словаря (инвентарь моделей управления и их характеристики) базировалась на словаре «Русский глагол — Венгерский глагол» [Апресян, Палл 1982], в котором собрана информация о более чем 1300 глаголах. В дальнейшем список конструкций редактировался, сокращался и расширялся с учетом информации из других источников («Лексикограф», толковые словари, ТКС, НОСС, синтаксический корпус НКРЯ, RussNet, материалы проекта дизамбигуации глагольных значений, см. [Толдова и др. 2008], и др.), но прежде всего за счет новых эмпирических данных из НКРЯ. По мере обработки примеров будут формироваться словарные входы имен существительных и других частей речи.

Источниками конструкций «малого синтаксиса» в словаре являются словарь МАС, [Грамматика 1980], [Шведова 1960], [Золотова 1988], лингвистические описания отдельных конструкций, а главное, конструкции, обнаруженные при разметке предложений.

3.2. Состав конструкций

В идеале, конструкции в словаре должны покрывать паттерны всех слов в корпусе, имеющих конструкционный потенциал. Рассмотрим для примера начало романа М. А. Булгакова:

(19) *Однажды весною, в час небывало жаркого заката, в Москве, на Патриарших прудах, появились два гражданина. Первый из них, одетый в летнюю серенькую пару, был маленького роста, упитан, лыс, свою приличную шляпу пирожком нёс в руке, а на хорошо выбритом лице его помещались сверхъестественных размеров очки в чёр-*

ной роговой оправе. Второй — плечистый, рыжеватый, вихрастый молодой человек в заломленной на затылок клетчатой кепке — был в ковбойке, жёваных белых брюках и в чёрных тапочках.

Первое предложение включает следующие базовые конструкции:
глагольная:
[два гражданина]$_{\text{Тема}}$ появились [в Москве, на Патриарших прудах]$_{\text{Место}}$
адъективная:
жаркий [закат]$_{\text{Носитель признака}}$
наречная:
небывало [жаркий]$_{\text{Признак}}$
именная:
закат (подчиненное имя не выражено, ср. *закат солнца*)
[Патриарший] пруд (конструкция наименования географического объекта)
конструкция с числительным:
два [гражданина]
сирконстантные конструкции (frame-setting constructions):
[однажды] Р
[весною, в час небывало жаркого заката] Р

Второе предложение включает конструкции:
глагольные:
[Первый из них] одет [в летнюю серенькую пару]
[Первый из них] нёс [свою приличную шляпу пирожком] [в руке]
[на… лице] помещались [очки…]
адъективные:
[Первый…] упитан
[Первый…] лыс
приличная [шляпа]
черная [оправа], серенькая [пара]
роговая [оправа]
сверхъестественный [размер]
выбритый [лицо]
свой [шляпа]
его [лицо]
наречная:

хорошо [выбритый]$_{\text{Признак}}$
именные:
[Первый…] [маленького] роста
[очки] [сверхъестественных] размеров
[шляпа] пирожком
очки [в… оправе]
лицо [его]
конструкция с числительным:
первый [из них]

Не перечисляя всех конструкций в третьем предложении, отметим интересные конструкции «одежды», такие как *человек в кепке* и *второй был в ковбойке* (со вспомогательным предикатом).

Все указанные конструкции включаются в словарь системы как лексические конструкции, т. е. они привязаны к индивидуальным лексемам. В некоторых случаях единственной лексической константой в них будет предлог, ср. *человек в кепке, первый из граждан*.[1] Вследствие принципа вложенности конструкций, несколько лексических конструкций могут образовывать конструкцию более абстрактного уровня, ср. параметрическую конструкцию типа *человек высокого роста*, *движения небольшой амплитуды*, инструментальную конструкцию формы *шляпа пирожком*, *нос уточкой*, конструкцию заполнения типа *заполнить шкаф книгами*, *залить сад электричеством*. О представлении таких конструкций речь пойдет во второй части раздела 3.3.

Кроме того, для объяснения поведения лексических конструкций и их взаимодействия в тексте мы должны включить в словарь и другие грамматические конструкции, не привязанные к конкретным лексемам, например, конструкцию аппозитивного прилагательного или причастия, конструкцию соположения атрибутов типа A A A… S, дитранзитивную конструкцию (ср. *испекла Пете пирог*), а также более общие синтаксические конструкции, такие как атрибутивная конструкция с полным прилагательным и предикативная конструкция с кратким прилагательным; активная и пассивная глагольные

[1] Конструкции с числительными задаются для всего частеречного класса или разряда, т. к. предполагается, что индивидуальные лексемы не вносят особого вклада в свойства конструкции, ср. NUMpauc + Sgen.sg (*два, три, четыре гражданина*).

конструкции, безличная, неопределенно-личная, обобщенно-личная императивная; сочинение и подчинение, перестановки порядка слов, конструкции повтора, компаративная (*гиены трусливее зайцев*), генитив при отрицании и т. д.

Словник словаря имеет две версии: целевой (все единицы, для которых целенаправленно собирались данные) и полный (все предикаты, которые встретились в предложениях). Словарь также делится на частеречные разделы: глаголы, абстрактные и предметные имена, прилагательные, наречия и проч.

3.3. Глагольные и неглагольные лексические конструкции

На Рис. 10 и 11 были представлены лексические конструкции глагольного управления. Здесь мы представим некоторые особенности синтаксических фразем и конструкций имен, наречий.

Конструкции глагольных фразем, как правило, включают другие лексические константы, помимо глагольного предиката, ср. шаблоны

Snom знать свое + место
Snom знать Sacc как свои + пять + пальцев
так и знай, CL
то и знай, CL
CL {надо/пора} и честь знать
бог знает {кто/что/какой/ADVPRO}
{бог/кто} {его/ее} знает Rel + CL

Схема: CL [CONJ] Snom V [себе] Vindic
Пример: А он знай себе играет на гармошке.

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
P	CL	фоновая ситуация	Клауза	-
W	CONJ	противопоставление	Союз	-
X	Snom	агенс	Несобственный	одушевленный
-	знай	-	Предикат	-
-	себе	-	Несобственный	-
Y	Vindic	содержание действия	Несобственный	-

Рис. 12. Представление конструкции *А он знай себе играет на гармошке*.

На Рис. 12 представлен шаблон идиомы типа *знай себе играет*, которая состоит из 6 элементов (два элемента, союз и дативное *себе*,— факультативные). Клауза, представляющая фоновую ситуацию, может быть выражена в преконтексте, за пределами предложения, содержащего элемент *знай*, ср. Рис. 12 и пример (20). Элементы Агенс и Содержание действия, как и *себе*, синтаксически не подчинены предикату *знать*; их ранг — Несобственный. Вместе с тем, Агенс является синтаксическим субъектом участника в индикативе Содержание действия; это вложенная предикация, на которую фразема, однако, накладывает определенные ограничения порядка слов. Возможности диатетических преобразований вложенной предикации (пассив, императив и т. п.) также ограничены. Фразема накладывает ограничение и на форму глагола *знать* — только форма императива 2 л. ед. ч. Вся эта информация сохраняется в словаре.

(20) *Ну а Пермяков с Пашутиным — ангелы небесные. Никого они не могли насиловать.* ***Знай работали себе*** *на местном предприятии — из тех, что еще не закрылись.*

У целого класса глаголов лексические конструкции включают, помимо собственно аргументно-предикатных структур (моделей управления) и фразем, конструкции, в которых эти глаголы выступают в роли лексических функций ([Mel'čuk 2004]; [Апресян 2004с, 2008]; [Апресян и др. 2010]). Шаблон ЛФ-конструкций, как правило, строится по образцу моделей управления, ср. «Snom оказать Sacc (*оказать влияние*)», но конструкции представлены в словаре как сами по себе, так и в составе комбинантных конструкций с предикатными именами (об этом ниже в этом разделе).

Если описание конструкций глаголов поддержано мощной традицией, как в плане теоретической проработки, так и в смысле достаточно полных словарных ресурсов, то этого нельзя сказать про конструкции имен. Хотя при разработке структуры субстантивного шаблона мы опирались на разработки словарей ТКС, НОСС, [Iordanskaja, Paperno 1996], базы имен системы «Лексикограф» [Красильщик, Рахилина 1992], мы не располагали готовым ресурсом, который можно было бы использовать в качестве ядра субстантивной части словаря лексических конструкций. Словарь имен строится по принципу bottom-up, т. е. конструкции заносятся в сло-

варь по мере их аннотирования в тексте. Как результат, фактор типичных, часто встречаемых в корпусе паттернов, безусловно, влиял на некоторые наши решения.

Словарь ФреймБанка включает конструкции трех типов:

• конструкции реляционных имен (ср. *муж Маши, край стола, командир подлодки, аналог НХЛ*);

• конструкции собственно предикатных слов (включая девербативы типа *прием, вывод, заключение*, деадъективы типа *красота, горечь* и собственно именные предикаты типа *мир (русской деревни), путь (к совершенству), (народовольческий) террор)*;

• частотные шаблоны предметных слов, например, названия организаций типа *Донская дивизия* и других именованных сущностей ([Лукашевич 2011]; [Толдова и др. 2013] и проч.).

Образец конструкции реляционного имени представлен на Рис. 13. Конструкция *Корпус Дохтурова* представляет фрейм с участниками Элемент (военнослужащие, которые служат в корпусе; вооружение, которое имеется в корпусе), Агенс (командующий корпусом), Цель (какие задачи выполняет корпус). Уточним, что поскольку *корпус* в этом значении не является предикатным словом, роли определяются относительно встроенных функциональных предикатов [Рахилина 2000], т. е. Агенс командует корпусом, Элемент несет военную службу в корпусе и т. п.

Схема: S Sgen
Пример: Я держу там корпус Криденера.

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
-	корпус	-	Предикат	-
Y	Sgen	агенс	Периферия	лицо

Рис. 13 Лексическая конструкция *Корпус Дохтурова.*

Помимо данной генитивной конструкции, у имени *корпус* в этом фрейме представлены лексические конструкции *корпус под командованием Тухачевского* (с предложным управлением), *корпус добровольцев* (еще одна генитивная, но с другим участником), *танковый корпус, экспедиционный корпус* (атрибутивные, с разными участниками) и др., а также их комбинации, ср. *конный корпус Гая*.

У собственно предикатных имен типа *вывод, красота, путь* мы отмечаем в словаре их управление (и другие типы синтаксической связи: атрибутивный, примыкание и т. д.). Но, помимо этого, нельзя оставить незамеченным их употребление с так называемыми light verbs или лексическими функциями типа *прийти к выводу, сделать вывод, напрашивается вывод* и т. п., см. Таблицу 2. Во-первых, известна лексическая избирательность как лексических функций, подчиняющих предикатное имя, так и самих предикатных имен в отношении подчиняющих их элементов. Во-вторых, с точки зрения синтаксической структуры (и разных подходов к ее построению), не всегда можно дать однозначный ответ, какому узлу, ЛФ-глаголу или целевому предикатному имени подчиняется актант во фразах типа *оказать влияние на начинающего поэта*. В-третьих, и, пожалуй, самое важное, оформление участников в ЛФ-конструкции может отличаться от управления предикатного имени per se и от управления ЛФ-глагола. Например, имя *заслон* управляет дативом только в конструкции *(по)ставить заслон пьянству*.[1] Кроме того, даже если способы оформления актантов у имени в независимой позиции и в ЛФ-конструкции совпадают, выбор того или иного способа оформления зависит от выбора лексической функции (количественная тенденция, например, глагол *напрашиваться* скорее связан с бессоюзным присоединением клаузы, ср. *напрашивается вывод*: CL, глагол *прийти (к)* скорее связан с союзным присоединением клаузы, ср. *прийти к выводу, что* CL, а при глаголе *делать* участник Содержание мысли скорее не выражен).

сделать (вывод)	28%
прийти (к выводу)	19%
делать (вывод)	12%
напрашиваться (вывод)	6%
приходить (к выводу)	4%
отсюда (вывод)	2%
правило (вывода)	1%
подтвердить (вывод)	1%
делать (вывод каким-л.)	1%
торопиться (с выводами)	<1%

Таблица 2. Наиболее частотные элементы синтаксической структуры, подчиняющие предикатное имя *вывод*. (По данным случайной выборки из НКРЯ, 150 примеров.)

[1] Об этой конструкции см. [Кузнецова, Ляшевская 2009], а также [Антошина, Ляшевская 2010].

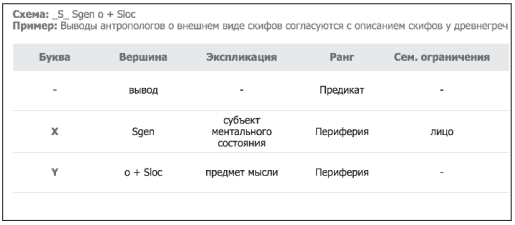


Схема: _S_ Sgen o + Sloc
Пример: Выводы антропологов о внешнем виде скифов согласуются с описанием скифов у древнегреч

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
-	вывод	-	Предикат	-
X	Sgen	субъект ментального состояния	Периферия	лицо
Y	o + Sloc	предмет мысли	Периферия	-

Рис. 14. Конструкция *вывод Sgen o + Sloc* в словаре лексических конструкций.

Схема: Snom прийти к + выводу {что + CL / «CL» / CL}
Пример: Я пришел к выводу, что война проиграна.

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
X	Snom	субъект ментального состояния	Субъект	лицо
-	прийти / приходить	-	Предикат-ЛФ	-
-	к + выводу	-	Периферия (Предикат)	-
Y	{что + CL / «CL» / CL}	содержание мысли	Клауза	пропозиция

Рис. 15. Конструкция *Snom прийти к выводу, что CL* в словаре лексических конструкций.

Пример разметки двух конструкций с именем *вывод* (в значении результата ментальной деятельности) приведен на Рис. 14 и 15. ЛФ-контролер получает синтаксический ранг Предикат-ЛФ (ср. *прийти, сделать, напрашиваться* и т. п.) или Предлог-ЛФ (ср. *согласно выводам Sgen, CL*). Предполагается, что Субъект подчиняется именно этому участнику. Целевое предикатное имя помечается как Объект (Предикат), Субъект (Предикат), Периферия (Предикат): отсылка к Предикату в скобках является условным способом обозначить целевое слово (семантически предикат).

Конструкции целевых прилагательных включают, в стандартном случае, их употребление в атрибутивной позиции (в полной форме) и в предикативной позиции (в краткой и полной форме), а также употребление с различными служебными связками[1] (лексическими функциями), ср. *вечер оказался испорчен / испорченным*. Помимо прилагательного, элементами конструкции является именной элемент — Контролер согласования (в атрибутивной конструкции) или

[1] Следует уточнить, что конструкции со связкой *быть* с именами существительными, прилагательными, предикативами в словаре отдельно не выделяются.

Субъект (в предикативной конструкции), а также другие элементы, находящиеся под синтаксическим управлением (если есть). Примеры двух конструкций с прилагательным *вынужденный* приведены на Рис. 16–17.

Схема: Snom _A_ Vinf
Пример: Я вынужден признать поражение.

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
X	Snom	пациенс социального отношения	Субъект	лицо / организация
-	вынужденный	-	Предикат	-
W	Vinf	содержание действия	Периферия	-

Рис. 16. Предикативная конструкция *Я вынужден признать поражение* в словаре лексических конструкций.

Схема: _A_ S
Пример: Вынужденное молчание.

Буква	Вершина	Экспликация	Ранг	Сем. ограничения
-	вынужденный	-	Атрибут	-
X	S	содержание действия	Контролер согласования	абстрактный

Рис. 17. Атрибутивная конструкция *Вынужденное молчание* в словаре лексических конструкций.

3.4. Общая информация о лексической конструкции

Помимо шаблона элементов, в словаре конструкций отражена следующая общая информация о конструкции:

- метка конструкции
- имя конструкции
- пример
- признаки
- синопсис
- комментарий
- служебная информация о статусе разметки конструкции в системе.

Метка конструкции представляет собой, как правило, лексему с индексом, например, *волноваться 2.2* (первая цифра — порядковый номер значения, в основном соответствует нумерации в словаре МАС, вторая цифра — порядковый номер лексической конструкции).

Имя конструкции — типичный пример, отражающий самые яркие особенности оформления конструкции и тип участников, ср. *Она волновалась из-за того, что от дочери долго не было известий* и *Мать волновалась за сына*.

Поле признаков содержит сведения о парном по виду глаголе для данного значения (например, СВ *взять*, СВ *написать*, НВ *оценивать*, СВ нет, а также информацию об ограничениях на оформление лексических констант).

Синопсис содержит пояснения о структуре значений, например, 1. ‘испытывать эмоцию’ или 2. ‘находиться в колебательном движении’ для глагола *волноваться*.

Словарь лексических конструкций снабжен указателем целевых лексем (лексических констант), а надстройкой над ним являются граф семантических ролей (см. п. 5) и граф конструкций и фреймов (frame grapher), связывающие конструкции по семантической и формальной близости (см. п. 6).

4. Аннотированный корпус примеров: реализации конструкций

4.1. Информация о реализации участников

Информация о моделях управления и других свойствах содержится в двух частях системы: в словаре (шаблоны лексических конструкций) и при конкретных лексических элементах в предложении (реализация конструкции в тексте, далее кратко «реализация»).

В корпусной части ресурса FrameBank представлено приблизительно по 100 примеров из НКРЯ (в дальнейшем количество примеров планируется расширить) на каждый глагол. При примере сохраняется вся информация из корпуса: метатекстовая, лексико-грамматическая, лексико-семантическая. Разметчик должен соотнести пример с нужным шаблоном и его контекстным вариантом (при деепричастии, пассивном причастии и т. п.), определить группы, соответствующие элементам конструкции (а также их синтаксические

и семантические вершины), несовпадения в морфосинтаксическом оформлении и ограничениях на лексико-семантическую сочетаемость. Отдельно размечаются сирконстанты и модальные слова, которых шаблон, естественно, не охватывает (см., например, выше Рис. 2). Если нужный шаблон в системе отсутствует, разметчик пополняет словарь конструкций.

В общих чертах структура таблицы, в которой размечается реализация конструкции в тексте, сходна с шаблоном лексической конструкции в словаре, но имеет дополнительную часть. Как и в словаре, в шаблоне реализации перечисляются элементы конструкции, как постоянные, так и переменные, и указаны их лексические, грамматические и семантические ограничения. Дополнительными полями в шаблоне конструкции являются:

- заполнение группы
- заполнение семантической вершины группы
- тип реализации

В поле «Заполнение вершины» и «Заполнение группы» заносятся сегменты размечаемого предложения, соответствующие элементам конструкции. Семантическая вершина размечается для того, чтобы проследить вариации заполнения элементов с точки зрения лексико-семантических классов (как работают семантические ограничения). В простом случае семантическая вершина группы совпадает с ее синтаксической вершиной (ср. *маки* в *красные маки*, *волновался* в *очень волновался*). В случае квантификации размечается элемент, подчиненный «прозрачному» оператору, ср., например, *две чашки молока* (семантическая вершина: *молоко*), *по две капли* (в поле «Заполнение вершины» записывается разрывная группа: *по / капли*). Помимо словесных цепочек, в системе сохраняется информация о координатах групп и вершин в предложении, которую затем можно использовать для вычисления порядка слов и расстояний между словами.

Информация в полях «Вершина», «Группа», «Экспликация», «Ранг», «Лексико-семантические ограничения» исходно берется из шаблона лексической конструкции в словаре. Далее в этих полях фиксируется информация обо всех отклонениях от стандартной реализации по морфологическому, синтаксическому, семантическому и лексическому оформлению.

В поле «Реализация» (в первой строке над полями «Заполнение вершины» и «Заполнение группы») отмечаются стандартный и нестандартные типы реализации конструкции (сочинение, контроль, отрицание, императив, пассив, деепричастие и др., см. п. 4.2).

Пример шаблона конструкции представлен на Рис. 18.

	Ранг	Морфосинтаксис	Экспликация	Сем. класс
свести	Предикат	Vimpers		
Y	Объект	Sacc	часть субъекта физиологического состояния	часть тела
X	Периферия	от + Sgen	причина	абстрактный

Рис. 18. Паттерн конструкции *Пальцы свело от холода* в словаре лексических конструкций.

	Ранг	Морфосинтаксис	Экспликация	Сем. класс	Причина альтернации	Выражение элемента
X	Субъект	Snom	причина	абстрактный		
	Периферия	Sins	=	=	конструкция с пассивным причастием	*простудой*
свести	Предикат	V				
	Предикат-Атрибут	V.partcp. pass.full. acc				*сведенные*
Y	Объект	Sacc	Часть субъекта физиологического состояния	часть тела		
	Контролер согласования	Sacc	=	=	конструкция с пассивным причастием	*пальцы*

Рис. 19. Аннотация конструкции *Судорога свела пальцы* в примере *… она подсела к печи, сведенные$_Y$ простудой$_X$ пальцы$_Y$ засовывала в самый огонь — грела.*

Рис. 19 иллюстрирует разметку конкретного примера, соотнесен-

ного с определенным шаблоном. Для каждого элемента конструкции первая строка отражает информацию из словаря, вторая — аннотацию в примере.

Кроме основных элементов лексической конструкции, размечаются элементы, отсутствующие в словаре конструкции:

• сирконстанты (периферийные элементы фрейма);

• модальные слова (отрицание и другие частицы)[1];

• синтаксические контролеры управления (глаголы или другие предикаты, управляющие целевым словом, через которые идет синтаксическая связь между целевым словом и Несобственным участником, ср. *Папа хотел купить*; о С-командовании см. [Тестелец 2001: 120–122]).

Иллюстрация разметки такого примера представлена на Рис. 20.

-- Вот он, один из этих штукарей из Варьете ,-- послышался грозный голос над онемевшим бухгалтером . И тут же Василия Степановича арестовали.

Имя	Вершина	Группа	Заполнение вершины	Заполнение группы	Экспликация	Ранг	Семантические ограничения
X	Snom	NPnom	неопределе личная конструкци	неопределе личная конструкци	агенс	Субъект	лицо / организация
Реализация	-	-			агенс	Не выражен	-
-	арестовать	арестовать			-	Предикат	-
Реализация	арестовать	N/A	арестовали	N/A	-	Предикат	-
Y	Sacc	NPacc	стандартны	стандартны	пациенс	Объект	лицо
Реализация	Sacc	NPacc	Василия	Василия Степанович	пациенс	Объект	лицо

№	Тип	Группа
1	сирконстант	тут же

Рис 20. Разметка реализации конструкции в примере *И тут же Василия Степановича арестовали* с учетом сирконстантного элемента (вид в системе онлайн-разметки).

Каждый пример содержит метатекстовую разметку (заполняется автоматически по данным НКРЯ): автор, его пол и год рождения, время создания текста, жанровая принадлежность. Размечаются характеристики предложения: длина, место в тексте (начало, середина, конец), пунктуационные знаки, а также его синтаксическая сложность:

• простое предложение;

• сложносочиненное предложение;

[1] Модальные глаголы, осуществляющие функцию синтаксического контроля.

• сложноподчиненное предложение;
• другое[1].

Назначение такой разметки — помочь пользователю в отборе примеров, в частности, по критериям времени создания, тематики и языковой сложности (readability), см. Рис. 21.

Более 70 процентов участников рынка будут вынуждены искать средства для увеличения означенной суммы. [Стань сильней или умри (2002) // «Финансовая Россия», 2002.09.19].
ссылка: Стань сильней или умри (2002) // «Финансовая Россия», 2002.09.19
автор: —
пол: —
год рождения: —
время создания текста: 2002
функциональная сфера: публицистика
тема текста: бизнес, коммерция, экономика, финансы
тип текста: статья
длина предложения (слов): 13
длина предложения (символов): 101
место предложения в тексте: середина
пунктуационные знаки: нет
сложность предложения: простое предложение
слов с частотным рангом 1–500: 6
слов с частотным рангом 501–2500: 5
слов с частотным рангом 2501–10000: 1
слов с частотным рангом 10001–50000: 1
слов с частотным рангом 50001+: 0

Рис. 21. Метатекстовая информация и информации о предложении.

Каждое слово в предложении также получает свой паспорт (Рис. 22): поля паспорта заполняются в том числе с использованием данных лексико-грамматической и лексико-синтаксической разметки НКРЯ, а также Грамматического словаря [Зализняк 1977]. В дальнейшем эта информация может пополняться и редактиро-

[1] Эта достаточно примитивная иерархия может быть использована преподавателями русского языка и разработчиками компьютерных систем и компонентов. Категория «другое» включает предложения с причастными и деепричастными оборотами, с прямой речью, с сочетанием разных типов сочинения и подчинения и т. п.

ваться. Предусмотрена возможность посмотреть информацию о слове в толковых словарях (интернет-версии словарей МАС и Ожегова-Шведовой), а также получить справку об употреблении слова в НКРЯ (основной и синтаксический корпус).

(21) Он служил в конном корпусе Гая. [*Анатолий Рыбаков. Тяжелый песок* (1975–1977)]

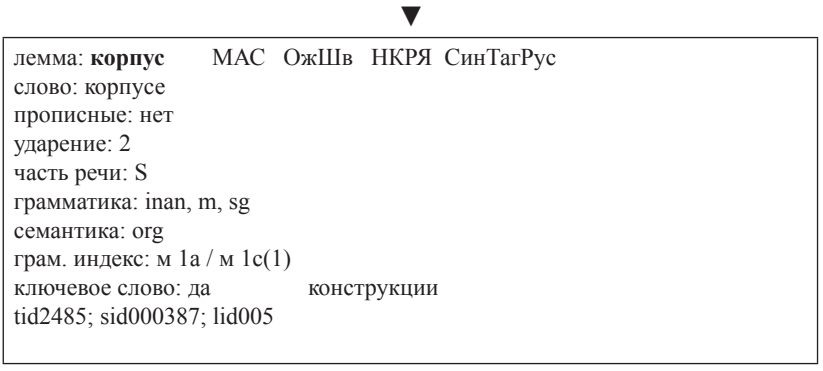


Рис. 22. Пример предложения и паспорт слова *корпус*.

4.2. Типичная нестандартная реализация элементов

В словаре представлены лишь так называемые «базовые» конструкции. Все варианты поверхностной реализации, которые можно предсказать по правилам русской грамматики, привязываются к соответствующим базовым конструкциям, а примеру приписывается информация о варианте нестандартной реализации. Перечислим случаи, в которых элемент либо выражен нестандартно, либо вообще не выражен в пределах предложения: участник выражен другой частью речи, но в той же грамматической форме; нестандартная форма (падеж, предлог) участника, не связанная с грамматическими конструкциями, перечисленными ниже; изменение лексико-семантического класса; сочинение (участники синтаксически зависят от предиката, сочиненного с данным); генитив субъекта или объекта при отрицании; пассив; причастие в атрибутивной (полной) форме; контроль субъекта или объекта другим деепричастием; инфинитивом; *сам, один, весь* и *оба* при деепричастии, инфинитиве, причастии в случае контроля; дативный субъект при инфинитиве;

императив; другие случаи несобственных участников; субстантивированное употребление причастий: предикат обозначает самого участника. Далее рассмотрим эти случаи более подробно.

1. Простая частеречная замена, ср. *Им дают разные задачи но они **решают** только простые* (эллипсис; элемент выражен не существительным, а прилагательным; ср. также замену существительного на местоимение-существительное и релятивизатор *который*).

2. Изменение формы (падежа, предлога), не связанное с грамматической конструкцией: участник оформлен нестандартным падежом или предложной группой, не представленной в словаре (например, в случае замены группы NPgen на NPacc, ср. *то, что так алкало его воображение*; квантификации группы, ср. *Пришло около тысячи человек*; *Стадион вмещает до пяти тысяч зрителей* и т. п.).

3. Изменение лексико-семантического класса участника. Например, в реализации участник обозначает «животное», а в словаре указано лексико-семантическое ограничение «лицо».

4. Сочинение. Элемент является Субъектом (или Объектом) другого глагола, который связан с данным отношением сочинения, ср. *Иван кричал и **волновался***[1]).

5. Генитив субъекта или объекта при отрицании. Участник, который в стандартном случае был бы синтаксическим Субъектом или Объектом, выражен родительным падежом из-за отрицания при глаголе (ср. *фонарей не горело*, *не знал слов*). То же для пассива (ср. *Ни одной суетной человеческой эмоции не **допущено** в эту храмовую постройку*).

6. Пассив: творительный субъекта. Участник, который в стандартном случае был бы Субъектом, выражен творительным падежом в пассивной конструкции (предикат — краткое страдательное причастие, полное страдательное причастие или форма на *-ся*, ср. *Чьи права **затронуты** постановлением суда*; *жюри, **возглавляемое** Сергеем Шуваловым*; *Метеориты **бомбардируются** космическими лучами*).

7. Пассив: именительный объекта. Участник, который в стандартном случае был бы Объектом, выражен именительным падежом

[1] Тег «сочинение» ставится на второй, третий и т. п. предикат в сочиненной группе, так как именно они являются несобственными в синтаксическом представлении НКРЯ (СинТагРус).

в пассивной конструкции (предикат — краткое страдательное причастие или форма на *-ся*, ср. *чьи права* ***затронуты*** *постановлением суда*, *Метеориты* ***бомбардируются*** *космическими лучами*)[1].

8. Пассив: семантический объект как контролер согласования при полном пассивном причастии (*у Ивана, исключенного из комсомола*; *жюри, возглавляемое Сергеем Шуваловым*).

9. Семантический субъект как контролер согласования при полном активном причастии (ср. *у Ивана, волновавшегося за меня*). Если предикат стоит между другими элементами группы субъекта (что вполне вероятно, т. к. он синтаксически входит в группу и зависит от ее вершины), ср. *этой огромной алчущей волной*, то поле «Заполнение группы» размечается как разрывная группа: этой огромной / волной. Зависимые от целевого предиката заполняются в своих строках. Если контролирующий предикат стоит в причастной форме и субстантивируется, т. е. обозначает Субъекта (*приходящий ко мне алкать*), то дополнительно ставится комментарий «контроль: субстантивация».

10. Несобственный семантический субъект при деепричастии (*Иван, волнуясь за меня*).

11. Дативный субъект при инфинитиве. В дативной конструкции семантический субъект выражается дательным падежом как при контроле предикатива или другого дативного предиката, так и без него, ср. *Нечего вам гулять по ночам*; *Где ему за мной угнаться*; *Мне выйти? Мне и встать перед ним на колени!*

12. Императив: несобственный и стандартный семантический субъект. Субъект при императиве может быть выражен как обращение, ср. *Витя,* ***иди*** *сюда*. Если Субъект при императиве выражен стандартно (NPnom), ср. *И ты приходи*; *Сам купи!*; *А тот поезжай в Германию и живи там еще 20 лет*, в т. ч. если обращение стоит при стандартном Субъекте, ср. *А ты, Маша, не бойся!*, реализация признается стандартной. Если Субъект при императиве опущен, но называется в контексте, см. след. п.

13. Участник не выражен зависимым при целевом предикате, но тем не менее упомянут в других частях предложения, ср. *А был бы кто пошустрей — подобрал бы кошелек*; *Нищие духом — не потому, что не имеют его, а потому, что* ***страждут*** *и алчут*.

[1] О разметке сложных форм пассива (*был обнаружен*, *может быть обнаружен*) см. раздел 4.4.

Это свойственно предикатам-инфинитивам и императивам, ср. *Одно тебе остается, чтобы избежать несчастья*; *довести товар до потребителя для отечественного производителя по-прежнему остаётся сверхзадачей*; *Если я тебе не понравлюсь, пройди мимо*.

14. Вводная конструкция: целевой предикат вводит предложение или группу (ср. *Допустим, эту кличку придумал отец*; *… трудно было сравнивать, допустим, с Кубанскими*).

15. Субстантивированное употребление причастий: предикат обозначает самого участника. При субстантивации Субъект активного причастия и Объект пассивного причастия обозначаются самим предикатом, ср. *Блаженны* ***алчущие*** *правды*; *Один из* ***собравшихся*** *встал*; *Нехорошо пользоваться* ***ворованным***. Однако, если Субъект имеет атрибутивные зависимые (ср. *всё собранное*; *эти алчущие богатства*), то они указываются в полях участника-Субъекта «Заполнение вершины», «Заполнение группы», «Вершина», «Группа»; «Реализация» — «субстантивация»; «Ранг» — «Атрибут».

4.3 Опущение и предупомянутость участника

Участник может быть не выражен в следующих случаях:

1. Опущение семантического субъекта при пассиве (*сотрудник был уволен*), деепричастии (*Подъезжая к станции, я потерял шляпу*), императиве (*не волнуйся*).

2. Участники — говорящие. Если участником является говорящий или адресат, и он не выражен при целевом глаголе (*захожу вчера в магазин*; *знаете как оно бывает?*).

3. Безличная конструкция. Если пример представляет безличное употребление (ср. *Пещеру завалило* — глагол в форме ср. рода прош. времени или 3-го лица ед. числа наст.-буд. времени), то на него заводится отдельная новая конструкция, ср. *Вавилову не разрешено было поехать в Эдинбург, хотя он был избран президентом конгресса*; *Делай, как велено*; *Курить не разрешается*.

4. Неопределенно-личная конструкция (*Ему разрешили уехать за границу*): глагол во мн. числе прош. времени или в 3-м лице мн. числа наст.-буд. времени.

5. Обобщенно-личная конструкция 2-го лица. Ср. *Ударишь лошадь кнутом — а она начнет лягаться* (глагол обычно во 2-м лице ед. числа).

6. Причастие как термин. В терминологическом употреблении причастие обычно теряет зависимые, ср. *Особенности родингитов в целом определяются спецификой вмещающей рамы* (причастие по сути адъективируется).

7. Контроль имплицитный, при котором целевой глагол в инфинитве подчинен другому предикату, но при этом в контролирующем предикате участник не выражен (в силу конструкционных свойств контролирующего предиката или грамматической конструкции, в которую он попадает), ср. *Идите пить чай* (участник: кто пьет чай); *Трудно установить, что стало причиной аварии* (участник: кто установил причину).

Если группа неполная (ср. *Четыре человека остались в комнате, а пятый уехал*), то все размечается так же, как и при полной группе (в данном примере реализация — «стандартный»). Различные другие случаи эллипсиса классифицируются как «сочинение», «преконтекст», «постконтекст» и т. п. Если ни один из случаев не подходит, ставится помета «эллипсис».

4.4 Сирконстанты, модальные и грамматические элементы, элементы контроля

Сирконстантные участники представляются в разметке предложений как неядерные участники, не предсказываемые в словаре [Падучева 2004: 74]. Информация о них хранится в особой зоне дополнительных элементов; в той же зоне представлены модальные элементы (например, отрицание, показатель косвенного вопроса *ли*, вводные дискурсивные обороты, в сферу действия которых попадает глагольный фрейм и т. д.), грамматические показатели аналитических форм целевого глагола (показатели сослагательного наклонения, форм будущего и прошедшего времени) и, наконец, предикаты контроля, т. е. предикаты-посредники, которые в предложении подчиняют целевое слово и которым синтаксически переподчиняются актанты целевого слова (например, *должно*, переподчиняющее второй актант глагола достигнуть в предложении *И его научное решение должно быть достигнуто на основе анализа*).

Если предикат стоит в форме сложного будущего времени с глаголом-связкой *быть*, ср. *Но мы не будем их ограничивать*, то это

размечается не как контроль, а как стандартная реализация (в поле «Заполнение вершины» предиката заносится *будем ограничивать*). Остальные связки со значением будущего (*станет, начнет*) трактуются как контроль.

Аналогично (добавлением в поле «Заполнение вершины») размечается глагол-связка *быть* в сложном пассиве, при предикатах-прилагательных и предикативах и некоторых др. аналитических формах (ср. *был объявлен, будет начата, (может) быть начата*; *был замечательный, будет веселее*; *было холодно*; *был выпивши*). Примеры разметки модальных конструкций: *Может быть понята — может* добавляется в поле «Модальные слова», *быть понята* — в поле «Заполнение вершины» (контролирующий предикат *может* управляет инфинитивной группой); *Любоваться было нечем — было нечем* добавляется в поле «Модальные слова», *любоваться* — в поле «Заполнение вершины» (контролирующий предикатив *нечем* управляет инфинитивом и имеет форму сложного прошедшего с глаголом-связкой *быть*).

Сам, один, весь и *оба* при деепричастии, инфинитиве, причастии (контроль). Если в ситуации контроля есть плавающие определители *сам, один, весь* и *оба*, которые отрываются от своей группы и выражаются как стандартный Субъект (в им. падеже) при целевом субъекте в форме деепричастия, инфинитива, причастия, ср. *Он всё видел и всё слышал, сам оставаясь незаметным*; *Я не дурак, чтобы сам себя губить за собственные деньги*; *Бакунин решил сам переломить свою участь*. Если те же определители *сам, один, весь* и *оба* стоят в форме датива, ср. *Он говорил, что нельзя ничего узнать, самому не попробовав*, то реализация — «деепричастный датив» или «инфинитивный датив».

4.3. Формирование банка предложений

Список «целевых» глаголов (2500 единиц) формируется по признаку их частотности, а также разнообразия и вариативности управления. На первом этапе на каждую лексическую единицу собирается по 100 предложений НКРЯ — случайным образом. На втором этапе выборка может быть доведена до 200 примеров, в том числе и с целью балансировки выдачи.

Выборка должна быть сбалансирована по следующим показателям:

1) метатекстовые признаки:

время создания текста (после 1950 г., первая половина XX в., XIX в., XVIII в.);

(в идеале) не более 1 примера из 1 автора;

жанр (художественная литература, публицистика, прочая нехудожественная литература, устная речь);

2) характеристики предложения:

длина;

сложность (простое, в т. ч. осложненное знаками препинания, сложносочиненное, сложноподчиненное, парцелляты);

место предложения в тексте;

3) место предиката в предложении:

начало — середина — конец.

Для особо частотных глаголов выборка может быть дополнительно сбалансирована относительно контекстных маркеров, а именно, слов открытых лексических классов, которые наиболее часто встречаются в предложении с данным глаголом, ср. *речь идет, подписать… договор* и т. д. Если их частотность в выборке будет не слишком велика, это послужит некоторой гарантией тематического разнообразия иллюстраций.

Банк предложений включает отобранные предложения и их ближайший контекст (как правило, 3 предложения справа и слева от них в тексте).

5. Семантические роли в системе ФреймБанк[1]

Инвентарь семантических ролей может строиться на основании самых разных подходов и теоретических установок, см., например, [Fillmore 1968, 1977, 1982], [Dowty 1991], [Лютикова и др. 2006: 17–22], [Плунгян 2011: 160–165], а также существующие списки ролей для русского языка ([Апресян 1974/1995: 125–126], [Апресян и др. 2010: 370–377], [Падучева 2004: 587–588], «Лексикограф»). Объем этого инвентаря тоже может существенно разниться — во многом в зависимости от конкретных нужд его использования — ср., с одной стороны, небольшой набор «глубинных падежей» в [Fillmore 1971], с другой стороны, потенциально неограниченный инвентарь,

[1] Первый вариант данного раздела был опубликован в [Кашкин, Ляшевская 2013].

используемый в системе FrameNet (содержащий, в частности, семантические роли Agriculturist, Colonists, Electricity, Examiner, Hair, Hunter, Rope и т. п.).

В системе FrameBank инвентарь семантических ролей строится согласно следующим принципам:

— инвентарь должен быть иерархически организован с целью создания более гибкого инструмента поиска и кластеризации: при желании, его можно свести к 5–10 проторолям, в других случаях он может быть расширен до нескольких десятков и даже сотен ярлыков;

— интерпретация первого и второго аргумента в большей мере зависит от семантики предиката, нежели трактовка третьих, четвертых и т. д. аргументов типа Инструмента, Траектории и т. д.;

— инвентарь коррелирует с семантической классификацией глагольной лексики[1]: в частности, это означает, что традиционные очень широко понимаемые роли Агенса и Пациенса должны в разных группах получать разные ярлыки;

— объем роли строится по принципу прототипа и периферии: например, прототипом Пациенса является участник, претерпевающий изменение под физическим воздействием контролирующего ситуацию Агенса, периферийные случаи (пациенс нефизического процесса, пациенс, не претерпевающий изменения, пациенс, который создается в результате физического действия и проч.) получают собственные ярлыки (ср. Тема, Результат) и считаются частным случаем роли Пациенса.

За основу для составления списка был взят инвентарь семантических ролей, приведенный в [Апресян и др. 2010: 370–377]. Практическая работа с имеющимися в системе шаблонами конструкций потребовала, однако, внесения в этот инвентарь ряда изменений. В первую очередь эти изменения обусловлены тем, что ряд содержащихся в списке Ю. Д. Апресяна семантических ролей объединяет достаточно разнородные семантические сущности. Если роль соот-

[1] Поскольку FrameBank является «дочерним» ресурсом НКРЯ, с надстроенным слоем разметки и интегрированным словарем, он ориентирован на систему глагольных классов Основного корпуса [Kustova et al. 2009], с учетом их дополнения и расширения. Вместе с тем, сам принцип иерархического выделения ролей может быть связан с любыми другими лексическими классификациями.

носилась с несколькими семантическими классами глаголов, то она разделялась нами на несколько — например, это коснулось роли Экспериенцера, которой в нашей разметке соответствуют семантические роли Субъект восприятия (*видеть, слышать*), Субъект ментального состояния (*думать, понимать*), Субъект психологического состояния (*бояться, любить*), Субъект физиологического ощущения (*болеть, колоть в боку*), Субъект физиологической реакции (*трястись от холода, тошнить*) и Субъект психологической реакции (*смеяться, краснеть от стыда*). Роль Агенса была сохранена для ядерных агентивных контекстов, но в дополнение к ней в список были включены экспликации Говорящий, Субъект поведения (*лениться, медлить*), Субъект социального отношения (*дружить, помириться*), Субъект перемещения (последняя экспликация используется для всех, не только агентивных, одноместных глаголов перемещения, коррелируя тем самым с их выделением в особый класс; агентивность глагола в этом случае однозначно устанавливается по одушевленности субъекта).

В результате для разметки шаблонов конструкций был использован список из 91 базовых экспликаций, классифицированный по принципу семантической близости на несколько блоков: блок Агенса, блок посессивных ролей, блок Пациенса, блок Адресата, блок Экспериенцера, блок Инструмента, блок обстоятельственных характеристик (внутри которого далее выделяются признаковые характеристики, временны́е характеристики, параметры, причинно-следственные характеристики, пространственные характеристики, а также группа источников и ресурсов как метафорическое расширение пространственных фреймов). При выделении обстоятельственных участников как актантов предиката мы в первую очередь руководствуемся принципом смыслоразличительности актанта в толковании (см. [Апресян 1974/1995: 119–133], [Падучева 2004: 72–79], [Апресян (ред.) 2010: 365–370]). Кроме того, мы учитываем здесь и некоторые формальные критерии. Так, в конструкции *Больной провел ночь беспокойно* выделяется обязательный участник ночь (NPacc) с ролью Срок, поскольку он является метафорическим заполнением той же самой валентности прямого объекта, которая уже существует у глагола *провести* в его исходных употреблениях (ср. *провести лошадь через реку, провести кабель* и пр.). В некоторых случаях употребление в конструкции вспомогательного участника

накладывает импликации на интерпретацию глагола: например, глагол *лечь / ложиться*, употребленный в конструкции с наречием или с предложной группой, выражающими роль Время (*Мы ложимся* ***очень рано***; *Он лёг* ***в два часа ночи***), дефолтно интерпретируется как ‘лечь спать / ложиться спать’. В этом и подобных случаях мы также включаем обстоятельственного участника в модель управления предиката.

Семантические роли и их блоки образуют единый граф, что позволяет выбирать между разными уровнями дробности поиска, релевантными для конкретного теоретического исследования или прикладной задачи (например, найти как все шаблоны конструкций, в которых реализуются семантические роли из Блока Агенса, так и все шаблоны конструкций, где есть участник с ролью Говорящего). Здесь следует отметить, что классификация семантических ролей потенциально может быть и более дробной: например, для глагола *помочь* экспликации «Тот, кто помогает» и «Тот, кому помогают» будут более точны, чем существующие Агенс и Контрагент. Вместе с тем, во-первых, ярлыки с настолько узкой областью применения будут редко востребованы пользователями; во-вторых, при необходимости они могут быть автоматически порождены по определенной схеме. Поэтому рабочий статус во FrameBank получает базовый инвентарь из 91 роли. В то же время, и теоретические предпосылки построения этого инвентаря, и техническое устройство системы допускают дальнейшее дробление любого из узлов графа ролей в зависимости от конкретной исследовательской задачи.

Отметим, что для целого ряда шаблонов конструкций оказалось невозможным приписать участнику ровно одну семантическую роль, поскольку имело место сочетание семантики двух различных ролей. В этих случаях в разметку вводились двойные семантические роли (ср. здесь [Апресян 1974/1995: 140] об отдельных примерах синкретичного выражения валентностей). Так, в примерах *«Вот это фокус», — удивился* ***он****, «Вот еще!» — сердится* ***брат****,* ***Пассажиры*** *долго возмущались, что поезд опоздал на два часа* как утверждается, что субъект произносит некоторое высказывание, так и характеризуется эмоциональное состояние субъекта. По этой причине одушевленный участник получает здесь сдвоенную роль Говорящий — Субъект психологического состояния. В контекстах *обрабатывать детали* ***на станке***, *кормить ребенка* ***с ложечки***,

ловить птиц ***сетями*** речь идет об инструменте совершения действия, но одновременно этот инструмент имеет локативные свойства, поэтому в данном случае использовалась экспликация Инструмент — Место. Конструкции типа *Пехотинцы строились* ***клином***, *Мальчик завязал шнурки двойным* ***узлом***, *Кассир сложил ассигнации* ***стопками*** описывают результат (то, что получилось в результате построения) и одновременно способ совершения действия, и в таких случаях в разметку вводилась двойная экспликация Результат — Способ. В целом, сдвоенные роли присутствуют в конструкциях, где либо участник выражается морфосинтаксически нестандартно (ср. *на станке*), либо предикат относится к нескольким лексическим классам (ср. *удивился* — эмоциональное психологическое состояние и говорение), либо сам аргумент семантически связан с разными классами объектов (ср. *двойной узел* как объект, который можно создать, либо как определенный способ завязывания веревки).

Инвентарь семантических ролей

1. БЛОК АГЕНСА

Агенс: активный (прототипически одушевленный) участник ситуации, целенаправленно изменяющий что-л. во внешнем мире. Как правило, экспликация используется для глаголов физического воздействия (***Хозяин*** *часто бьет собаку*), еды и питья (***Водитель*** *выпил стакан молока*), обработки (***Рабочий*** *намазал деревяшку клеем*), создания нового объекта (***Женщины*** *варили обед*), каузации перемещения (***Милиционер*** *вывел нарушителя*), в случае деятельности, направленной на измененение чего-л. (***Мы*** *будем воевать до победного конца*).

Каузатор: участник ситуации (прототипически одушевленный), порождающий какое-л. действие без намерения, обычно при глаголах физического воздействия или каузации перемещения (*Отойдите, чтобы* ***я*** *Вас не задел*; ***Вы*** *испортите аппарат неумелым обращением*; *Осторожно, смотри, куда* ***ты*** *льешь вино*).

Эффектор: неодушевленный участник ситуации, который может рассматриваться как активный источник энергии, вызывающий изменения — обычно при глаголах физического воздействия (***Ветром***

сорвало крышу; ***Колеса автомобиля*** *давили саранчу*; *Уши горят* ***на морозе***).

Потенциальный эффектор: участник, который мог бы стать Эффектором ситуации, но не стал им — при глаголах с семантикой предотвращения (*Козырёк хорошо защищал глаза* ***от солнца***).

Контрагент: активный (прототипически одушевленный) участник агентивной ситуации, который выполняет роль, отличную от роли Агенса (*Петр яростно боролся* ***с новеньким мальчиком***; *Я не могу справиться* ***с лошадью***; *Мать кормила* ***малыша***). Экспликация также используется для маркирования «собеседника» (не только слушающего, но и участвующего в разговоре) при глаголах речи (*Учитель разговаривал* ***с родителями***; *Продавец торгуется* ***с покупателем***). При глаголах перемещения обозначает второго участника ситуации преследования или обгона (*Собака гонится* ***за кошкой***). В коммуникативной структуре Контрагент имеет статус ниже, чем Агенс.

Потенциальный контрагент: участник, который мог бы стать Контрагентом ситуации, но не стал им — при глаголах с семантикой предотвращения (*Лесник спас девочку* ***от медведя***).

Субъект перемещения: совершающий перемещение участник ситуации, описываемой одноместным глаголом перемещения (***Мальчик*** *бегает по двору*. ***Поезд*** *едет в Новосибирск*. ***Уголек*** *описал круг*. ***Книга*** *упала с полки*).

Говорящий: активный участник при глаголах речи (***Мама*** *сказала, что мы поедем в Новосибирск*; ***Он*** *клянется в своей невиновности*; ***Ты*** *мне солгал*).

Субъект социального отношения: активный участник при глаголах социального отношения (***Маша*** *дружит с Васей*; ***Мой отец*** *женился по любви*; ***Командир*** *уволил офицера в запас*).

Контрагент социального отношения: активный участник ситуации социального отношения, выполняющий роль, отличную от роли Субъекта социального отношения (*Маша дружит* ***с Васей***; *Принц женился* ***на бедной девушке***).

Субъект поведения: активный участник при глаголах поведения, в отличие от Агенса не оказывающий явного воздействия на других участников ситуации (***Ученики*** *в начале четверти ленятся*. ***Котята*** *возятся в углу*. *Почему* ***вы*** *медлите с отъездом?*).

2. ГРУППА ПОСЕССИВНЫХ РОЛЕЙ

Посессор: участник ситуации, являющийся обладателем какого-л. предмета либо непредметной сущности (***Он*** *владел прачечной*; *Не знаю, откуда взялись* ***у меня*** *силы*). Также используется при глаголах, описывающих движения частей тела (*Рот* ***у рыбы*** *то открывался*, *то закрывался, она постепенно заваливалась на бок*; *Пушистый хвост* ***у лисы*** *вилял из стороны в сторону*)

Конечный посессор: участник при глаголе смены посессора, вступающий во владение каким-л. предметом (***Воры*** *украли у бабушки кошелек. Мальчик дал* ***другу*** *карандаш*).

Исходный посессор: участник при глаголе смены посессора, утрачивающий владение каким-л. предметом (*Воры украли* ***у бабушки*** *кошелек*. ***Мальчик*** *дал другу карандаш*).

3. БЛОК ПАЦИЕНСА

Пациенс: участник ситуации, претерпевающий самопроизвольно или под воздействием другого участника изменения — как правило, при переходных глаголах физического воздействия (*Девочка режет* ***хлеб***), обработки (***Дыру*** *надо заклеить*), еды и питья (*Собака ела* ***сахар***), смены посессора (Мальчишки *воровали* ***яблоки*** *у соседа*), при непереходных глаголах изменения состояния (***Вода*** *быстро застывает*), уничтожения (***Ваза*** *разбилась*).

Часть пациенса в фокусе внимания: часть Пациенса, непосредственно изменяемая в указанном выше классе ситуаций. Роль используется в тех паттернах, где пациенс уже эксплицитно выражен другой именной группой (*Старик бил слугу* ***по голове***. *Парикмахер бреет клиенту* ***бороду***).

Потенциальный пациенс: участник, который мог бы стать Пациенсом ситуации, но не стал им — при глаголах с семантикой предотвращения (*Лесник спас* ***девочку*** *от медведя*).

Пациенс перемещения: объект при глаголах каузации перемещения (*Мальчик кидал* ***камни*** *в ведро*; *Немцы вели* ***пленных*** *на расстрел*).

Пациенс социального отношения: пассивный участник при глаголах социального отношения (*Отец выдал* ***дочь*** *замуж*; ***Васю*** *отчислили из университета*).

Тема: пассивный участник ситуации, который никак не меняется

в ходе ее развития — как правило, при бытийных глаголах (*Летом в Прибалтике бывают* ***чудные дни***), глаголах возникновения (*Вышел* ***новый номер газеты***), позиции (***Лампа*** *висит над столом*), «фиктивного движения» (***Тропинка*** *вдруг повернула налево*), обладания (*В те времена Россия еще не владела* ***Сибирью***), контакта (*Чайки едва касались* ***воды***).

Часть темы в фокусе внимания: та часть Темы, о которой идет речь в указанном выше классе ситуаций. Роль используется в тех паттернах, где Тема уже эксплицитно выражена другой именной группой (*Пиджак вытерт* ***на локтях***; *Мужчина стоял* ***спиной*** *к окну*).

Точка контакта: при глаголах контакта — та точка, в которой один объект касается другого (*Рыба забилась* ***о пристань***; *Лошадь занесла его головой* ***о камень***; *Он укрепил палку* ***за верхний конец***).

Стимул: источник зрительной, слуховой и т. п. информации для Субъекта восприятия (*Путешественники видят* ***берег***; *Я слушаю* ***красивую песню***).

Предмет высказывания: то, о чем говорится в ситуации речи (*Дипломаты беседовали* ***о политике***; *Он называл* ***ее*** *очаровательной*).

Содержание высказывания: то, что говорится в ситуации речи (*Он говорит* ***правду***. *Он кричит,* ***что его никто не понимает***).

Предмет мысли: то, о чем думает, о чем знает какие-л. факты и т. п. Субъект ментального состояния (*Я думаю* ***о прошлом***; ***Игоря*** *все считают гениальным*).

Содержание мысли: то, что думает, что знает и т. п. Субъект ментального состояния (*Я думаю*, ***что он врет***; *Я знаю* ***о победе Пети в конкурсе***).

Ситуация в фокусе: пропозиция, о свойствах которой делается утверждение в конструкции (*Он взял и* ***убежал***. *Стоит подняться на эту горку —* ***и виден будет наш дом***).

Содержание действия: разновидность Ситуации в фокусе — то основное, к чему сводится существо действия — при глаголах, описывающих самостоятельное протекание действия (***Удары*** *падали на спину раба*; *Остается* ***проститься с Вами***), глаголах деятельности (*Они вели* ***напряженную работу***; *Он играет* ***вальс***), фазовых глаголах (*Ее муж начал* ***пить***; *Прекратите* ***шум***).

Потенциальная угроза: нежелательная ситуация или нежелательный объект, воздействие которого предотвращается — обоб-

щение ролей Потенциального агенса, Потенциального эффектора, Потенциальной ситуации (*Мы боремся* ***против хищнического лова рыбы*** / ***против браконьеров***).

Запах: основное содержание при глаголах запаха (*В комнате воняло* ***тухлой капустой***. *Чай отдаёт* ***рыбой***).

Часть субъекта физиологического ощущения: часть тела, в которой локализуется испытываемое субъектом физиологическое ощущение (*У меня колет* ***в боку***; ***Глаза*** *стали ему отказывать*).

Часть субъекта физиологической реакции: часть тела, в которой проявляется физиологическая реакция (*У Ирины тревожно билось* ***сердце***; *Он моргал* ***воспаленными глазами***).

Результат: объект, создаваемый в результате выполнения действия (*Она умеет бить* ***масло***. *Рабочие копают* ***канаву***. *Он разломал кусок хлеба* ***на 4 части***).

Цель: планируемый результат действия либо объект, которым субъект предполагает завладеть в результате совершения действия, часто в результате перемещения (*Комитет собирает людей* ***на митинг***; *Ты идешь* ***за билетами****?*).

Предел: ситуация (или метонимически связанный с ней объект), по достижении которой завершается действие (*Защитники дрались* ***до последнего патрона***; *Мы довели беседу* ***до конца***).

4. БЛОК АДРЕСАТА

Адресат: получатель информации в ситуации речи (*Он сообщил* ***мне****, что не приедет*; *Благодарю* ***Вас***).

Аудитория: множественный участник — зрители или слушатели какого-л. лица — как правило, при глаголах, описывающих публичное выступление (*Вы должны открыть* ***перед коллективом****, что вынудило вас совершить преступление*; *Ректор обратился* ***к собравшимся*** *с приветственной речью*).

Бенефициант: участник, интересы которого непосредственно затрагивает ситуация и который получает от этого пользу, при этом действие не направлено непосредственно на него (***Население*** *выигрывает от снижения цен*; ***Батраки*** *работали на хозяина*) — часто дативный участник при глаголах различных семантических классов (*Дедушка открыл* ***бабушке*** *дверь*; *Он хотел сшить* ***жене*** *модное пальто*).

Малефициант: участник, интересы которого непосредственно затрагивает ситуация и которому это приносит вред, при этом действие не направлено непосредственно на него (*Он заработал проблем* ***на свою голову***).

5. БЛОК ЭКСПЕРИЕНЦЕРА

Субъект восприятия: активный участник при глаголах восприятия (***Дедушка*** *услышал крик.* ***Ты*** *видел этот фильм?*).

Субъект ментального состояния: активный участник при глаголах ментального состояния (***Он*** *понимает, что надо делать*; ***Мы*** *думаем о будущем*).

Субъект психологического состояния: активный участник при глаголах психологического состояния (***Ребенок*** *боится собак*; ***Петя*** *любит Машу*).

Субъект физиологического ощущения: одушевленный участник ситуации физиологического ощущения, испытывающий данное ощущение (***У меня*** *колет в боку*; *Кашель не отпускал* ***старика***).

Субъект физиологической реакции: одушевленный участник ситуации физиологической реакции, обусловленной физиологическими причинами (***Меня*** *тошнит*).

Субъект психологической реакции: одушевленный участник ситуации физиологической реакции, обусловленной психологическими причинами (***Девочка*** *смеется*).

6. БЛОК ИНСТРУМЕНТА

Инструмент: участник ситуации, не являющийся частью Агенса, прототипически используемый Агенсом для воздействия на Пациенс и не расходуемый в ходе этого воздействия (*Фашисты били партизана* ***кнутом***. *Защитники Брестской крепости били* ***из зенитных орудий***. ***Пила*** *плохо пилит*).

Квазиинструмент: участник ситуации, являющийся частью Агенса, прототипически используемый Агенсом для воздействия на Пациенс и не расходуемый в ходе этого воздействия (*Муж бил жену* ***кулаками***. *Он зажал* ***двумя руками*** *челюсти собаки*). Многие фреймы имеют общий слот для участника-инструмента и участника-квазиинструмента.

Часть контрагента в фокусе внимания: часть тела, используемая Контрагентом в ситуации (*Командир пожал бойцу* ***руку***).

Часть субъекта восприятия: часть тела (или метафорически отождествляемая с ней сущность), используемая Субъектом восприятия для осуществления ситуации (*Я видел это* ***своими глазами***; *Если* ***глаза*** *меня не обманывают, это Иванов*).

Часть субъекта ментального состояния: часть тела (или метафорически отождествляемая с ней сущность), используемая Субъектом ментального состояния для осуществления ситуации (*Он держит* ***в голове*** *много информации*; ***Перед его мысленным взором*** *встают картины прошлого*).

Часть субъекта психологического состояния: часть тела (или метафорически отождествляемая с ней сущность), используемая Субъектом психологического состояния для осуществления ситуации (*Надежда шевельнулась у него* ***в душе***; *Он несет* ***в сердце*** *чувство любви*).

Рабочая часть: часть тела живого существа или часть предмета, используемая для поддержания данного пространственного положения — обычно при глаголах позиции, опоры и перемещения (*Мой отец плавал* ***на спине***. *Дом стоял на* ***бетонном фундаменте***).

Способ: спецификация действия — то, как или с помощью каких приемов оно осуществляется (*Командир влиял на бойцов* ***личным примером***; *Рыцарь стоял* ***в горделивой позе***; ***По-птичьи*** *кликали, звали сестрицу*).

Средство: участник ситуации, прототипически используемый Агенсом для воздействия на Пациенс и расходуемый в ходе этого воздействия (*Артиллеристы били по блиндажу* ***минами***; *Котлеты лучше жарить* ***на постном масле***).

7. БЛОК ОБСТОЯТЕЛЬСТВЕННЫХ ХАРАКТЕРИСТИК

7.1 Признаковые характеристики

Признак: тот аспект объекта, по которому он отличается от других похожих объектов (*Вещи можно ассоциировать* ***по смежности***. *Он признал свой старый дом* ***по высокой трубе***. ***Мертвенная бледность*** *указывала на его сильнейшее волнение*).

Признак действия: тот аспект действия, по которому оно отли-

чается от других аналогичных действий (*Он* ***красиво*** *пишет*; *Ружье* ***точно*** *стреляет*).

Эталон: та сущность, с которой происходит сравнение (*Печи спорили в белизне* ***со стенами***. *Автобус ходит* ***по расписанию***).

Статус: та работа или то использование, для которых предназначен человек или предмет — обычно при бытийных глаголах или глаголах социального отношения (*Его приглашают в институт* ***заместителем директора***. *Земля представляет собой* ***сфероид***).

7.2 Временны́е характеристики

Момент: точка во времени, в которой локализуется ситуация (*Часы бьют* ***одиннадцать***; *Пароход уходит* ***в восемь***).

Срок: отрезок времени, в котором локализуется ситуация (*Солдат растянул сухой паёк* ***на неделю***; *Учитель отвел* ***20 минут*** *на вопросы*).

Время: заранее никак не ограниченная длительность действия (***Время*** *бежит*; *Отец вставал* ***рано***).

Возраст: количество времени с момента появления на свет существа или предмета (*Маше исполнилось* ***5 лет***).

7.3 Параметры

Параметр: количественный или качественный параметр — обычно при глаголах, описывающих изменение состояния (*Самолет набирает* ***высоту***; ***Температура*** *поднялась до 40 градусов*).

Значение параметра: выражаемое количественно или качественно значение параметра (*Термометр показывает* ***минус сорок***; *Машинист* ***немного*** *уменьшил скорость локомотива*).

Вес: вес человека или неодушевленного объекта (*Мешок весит* ***10 кг***; *Ее муж сбросил* ***5 кг***).

Скорость: скорость выполнения действия (*Судно делает* ***12 узлов***).

Цена: цена объекта — как правило, при глаголах смены посессора или иных глаголах из сферы товарно-денежных отношений (*Мы купили дачу* ***за миллион рублей***; *Билет стоил* ***тысячу рублей***).

7.4 Причинно-следственные характеристики

Причина: участник, который является причиной ситуации или обусловливает возникновение состояния — часто при глаголах психологического состояния (*Мы волнуемся, **что он не отвечает на письма***; ***Твоё долгое молчание** беспокоит меня*; *Он тоскует **по друзьям***), изменения состояния (*Щеки впали **от бессонницы***), физиологического состояния и физиологической реакции (*Я начал кашлять **от холода***).

Свойство причины: то в Причине, что непосредственно вызывает данное состояние или положение вещей (*Он бесит нас **своим нахальством**. Она нравилась ему **своей непосредственностью***).

Мотивировка: действие, состояние, свойство или положение вещей, объясняющее, почему агентивный участник предпринял данное действие, но не являющееся его причиной (*Сержант бил солдат **за неподчинение***; *Полицейский арестовал его **за торговлю оружием***). Экспликация используется также в паттернах с другими ролями из Блока Агенса, ср. *Специалисты критиковали статью **за необоснованность выводов*** (субъект — Говорящий), *Командир отметил бойца в своем приказе **за отвагу и находчивость*** (субъект — Субъект социального отношения).

7.5 Пространственные характеристики

Конечная точка: точка, в которой заканчивается перемещение участника (*Самолет летит **в Париж**. Рыбаки толкали лодку **в воду***).

Конечное состояние: состояние, в которое входит участник экспериенциальной ситуации (*Мальчика тянуло **ко сну***; *Поэт погрузился **в мечты***).

Направление: сторона, в которую устремлено перемещение — при глаголах перемещения, либо сторона, куда ориентирован неподвижный объект — при глаголах позиции (*Он отвернулся **к своему соседу справа***; *Концы усов висят **вниз***).

Траектория: линия или смежный с ней участок пространства, где происходит перемещение (*Ребята бегали **через поле***; *Ветер несет **по дороге** желтые листья*).

Место: участок пространства, где локализуется ситуация — как

правило, описываемая глаголами позиции или перемещения (*Влюбленные ходят* ***у реки***; *Книга лежит* ***на столе***).

Начальная точка: точка, в которой начинается перемещение участника (*Учитель уронил* ***книгу*** *из рук*; *Мотай* ***отсюда****!*).

Расстояние: расстояние, на которое переместился Субъект перемещения или Пациенс перемещения (*Рота отошла* ***на два километра***; *Подай доску* ***на метр влево***).

Опора: поверхность объекта, на которой удерживается расположенный на ней другой объект — при глаголах позиции, перемещения, обработки (*Он медленно шел по дорожке, опираясь* ***на плечо сына***; *Я глажу брюки* ***на доске***).

Уровень: горизонтальная линия или плоскость, являющаяся границей чего-л. — при глаголах перемещения или позиции (*Вася* ***по колено*** *провалился в сугроб*; *Мой сын отпустил волосы* ***до плеч***).

Сфера: область деятельности, в рамках которой осуществляется ситуация — часто при агентивных глаголах (*Этот прибор используют* ***в медицине***), глаголах ментального состояния (*Мой друг математик, но хорошо понимает* ***в музыке***).

Среда: вещество или окружающая обстановка, где осуществляется ситуация (*Она мылась* ***в теплой воде***; ***Из мрака*** *выступил сад*). Часто используется при глаголах восприятия (*Даже* ***в густом тумане*** *можно было различить контуры собора*; *Он не рассмотрел* ***в темноте*** *приближающегося человека*).

Социальная среда: социум, где осуществляется ситуация (*Постоянно вращаясь* ***среди актеров****, я стал привыкать к их языку*; *Они считали своим долгом сеять знания* ***в народе***; *Он происходил* ***из древнего и знатного рода***) — часто при агентивных глаголах, глаголах социального отношения, глаголах психологического состояния.

Множество: группа однородных объектов, из которой по определенному признаку выделяется какая-либо часть (***Отряд*** *разбился на группы*; ***Шествие*** *завершали три монаха*) — часто при глаголах выбора (***Из массы сложных вопросов*** *он выделил два основных*).

7.6 Группа источников и ресурсов

Источник звука: при глаголах звука — объект, непосредственно создающий звук (*Завидев неприятеля, дежурный бил* ***в барабан***. ***Гром*** *гремит*).

Источник света: при глаголах света — объект, непосредственно создающий свет (***Луна*** *льет серебристый свет*; ***Роса*** *блестела на солнце*).

Источник запаха: при глаголах запаха — объект, непосредственно создающий запах (***От бомжа*** *воняет*; ***Цветы*** *льют аромат на поля*).

Источник вещества: при глаголах возникновения или обработки — объект, из которого непосредственно выделяется вещество (***Из растений*** *выделяется кислород*; *Туземцы научили его извлекать сок* ***из плодов манго***).

Источник: экспликация выбирается в том случае, если семантические ограничения на заполнение паттерна конструкции включают несколько типов перечисленных выше источников (*Шорох исходил* ***из кустов*** / ***От котлов*** *исходил аппетитный запах* / ***От свечи*** *исходит свет*; *Это идёт* ***от английской народной баллады***).

Носитель информации: объект, содержащий текст / информацию или предназначенный для того, чтобы содержать текст / информацию — обычно при глаголах речи или ментального состояния (***В аннотации*** *говорится: «Книга предназначена для широкого круга читателей»*; *Я знаю об этом* ***из надежного источника***; *Ученики списали задание* ***с доски***).

Ресурс: материал, источник получения чего-л — при глаголах создания нового объекта (*Лодку сделали* ***из дерева***; *Пастухи шьют сапоги* ***из овечьей кожи***), либо при глаголах, описывающих деятельность, требующую ресурса для своего поддержания (*Чтоб ты жил* ***на одну зарплату***! ***На свои деньги*** *я могу покупать, что хочу*).

Таблица 3 представляет список наиболее частотных ролей. Показано количество лексических конструкций, в которых присутствует эта роль, а также количество предикатов, у которых есть конструкции с таким участником. Граф семантических ролей доступен на Рис. 23.

Экспликация	Кол-во констр. в словаре	Пример конструкции	Кол-во предикатов в словаре
агенс	4787	*Продавщица режет сыр.*	1824
пациенс	3086	*Продавщица режет сыр.*	1498
тема	1591	*На берегу лежал человек.*	1004
субъект перемещения	1520	*Мы выезжаем в Москву.*	515
говорящий	1304	*«Годовой план мы выполним досрочно», – объявил директор.*	749
пациенс перемещения	1049	*Мальчик водил слепого за руку.*	358
конечная точка	921	*Мы выезжаем в Москву.*	657
место	903	*На берегу лежал человек.*	738
содержание высказывания	776	*Он сообщил, что работает над книгой.*	454
эффектор	643	*Ветер треплет знамена.*	565
субъект психологического состояния	643	*Он тоскует по друзьям.*	526
содержание мысли	637	*Мы считали его опасным человеком.*	438
содержание действия	634	*Потрудитесь встать, пожалуйста!*	526
результат	633	*Достиг я высшей власти.*	445
причина	616	*Командира беспокоило, если разведчики долго не возвращались.*	501
субъект ментального состояния	614	*Павел точно знал адрес больницы.*	475
субъект социального отношения	520	*Я не ссорюсь с друзьями.*	550
начальная точка	485	*Дождь лился из туч.*	421
способ	479	*Он выразил неудовольствие одним взглядом.*	477
контрагент	446	*Я прощаюсь с тобой.*	453
адресат	421	*Так велел ему князь.*	384
цель	377	*Отряд остановился передохнуть.*	292
конечный посессор	361	*Товарищ одолжил мне книгу.*	284
стимул	303	*Капитан видит человека за бортом.*	231
исходный посессор	292	*Товарищ одолжил мне книгу.*	227
предмет мысли	284	*Мы считали его опасным человеком.*	311

Экспликация	Кол-во констр. в словаре	Пример конструкции	Кол-во предикатов в словаре
пациенс социального отношения	283	*Он воспитывает трех сыновей.*	380
субъект поведения	258	*Он работает спустя рукава.*	341
предмет высказывания	230	*Автор отразил в своих очерках настроение рабочих.*	219
сфера	228	*По физике он учился неплохо.*	267
субъект восприятия	225	*Перед ним открылся вид на море.*	155
ситуация в фокусе	218	*Стоит подняться на эту горку, и виден будет наш дом.*	286
агенс – субъект ментального состояния	189	*Следователь вывел, кто будет следующей жертвой*	132
каузатор	188	*Он невольно выдал себя.*	169
место – пациенс	185	*Беседка заросла репейником.*	148
подвергающаяся воздействию часть пациенса	178	*Тарас тыкал ее сапогом прямо в лицо.*	187
траектория	164	*Сани ехали по глухой лесной дороге.*	138
инструмент / квазиинструмент	158	*Она коснулась земли зонтиком / руками.*	178
агенс / каузатор	149	*Горничная <споткнулась и> разбила тарелку на мелкие кусочки.*	123
статус	148	*Он выступил в роли верховного арбитра в этой ситуации.*	170
признак	144	*Ваш завод отстает по всем показателям.*	193
контрагент социального отношения	141	*Он ладил со всеми.*	209
посессор	136	*Новый препарат обладает интересным свойством.*	145
источник звука	128	*Звонарь бил в сторожевой колокол.*	58
инструмент	124	*Подписывайтесь ручкой, карандашом нельзя.*	112
инструмент – место	123	*Мы выехали домой поездом.*	136
срок	121	*Его посадили на три года.*	122

Экспликация	Кол-во констр. в словаре	Пример конструкции	Кол-во предикатов в словаре
субъект физиологической реакции	120	*Ленка взглянула на него и, не удержавшись, прыснула.*	112
часть субъекта физиологического ощущения	120	*Болит нога.*	58
признак действия	119	*Земля хорошо родит.*	149
мотивировка	108	*Он оправдывался перед гостями за свое опоздание.*	151
субъект физиологического ощущения	108	*Спазм сжал женщине горло.*	126
средство	104	*Девушка перевязывает письма ленточкой.*	133
потенциальная угроза	86	*Он избежал ареста.*	57
эталон	85	*Иван подражал отцу в манере говорить.*	98
агенс – субъект восприятия	74	*Боец целится в противника.*	43
бенефициант	70	*Время работает на нас.*	89
параметр	68	*Такое впечатление, будто микрофон усиливает шум.*	56
конечная точка – сфера	66	*Мой коллега перешел в другой институт.*	94
направление	64	*Вершины гор уходят за облака.*	77
ресурс	63	*Каменщики сложили из кирпича высокую стену.*	78
цена	63	*Коляска стоила пятьсот рублей золотом.*	59
значение параметра	60	*Водитель понижает скорость до сорока километров в час.*	65
опора	56	*Я поскользнулся, но удержался за косяк.*	58
квазиинструмент	55	*Руки влюбленных встретились.*	75
источник света	52	*Мелькают огоньки.*	30
часть субъекта восприятия	50	*Я видел все это собственными глазами.*	56
потенциальный пациенс	50	*Наши войска удержали город.*	41
расстояние	49	*Планку подняли еще на пять сантиметров.*	55

Экспликация	Кол-во констр. в словаре	Пример конструкции	Кол-во предикатов в словаре
часть субъекта психологического состояния	48	*Душа у него тает, когда он смотрит на свою маленькую дочь.*	59
часть субъекта физиологической реакции	40	*Уши у него покраснели.*	29
результат – способ	37	*Адмирал вытянул корабли в кильватерную колонну.*	38
субъект перемещения метаф.	37	*Дорога уходила вдаль.*	35
множество	36	*Пятнадцать союзных республик образуют Советский Союз.*	37
пациенс / тема	36	*Отец ударил кулаком по столу.*	29
субъект психологического состояния / часть субъекта психологического состояния	35	*Иногда он вдруг загорался / его глаза вдруг загорались.*	41
носитель информации	34	*В книге говорится о последних открытиях в области физики.*	44
субъект психологического состояния / субъект физиологического ощущения	33	*Заботливый уход вернул больному жизнь.*	48
свойство причины	33	*Работа увлекла его своими перспективами.*	47
предмет высказывания / содержание высказывания	33	*Старик жаловался на боли в области сердца.*	39
место / траектория	33	*По дорогам / степи плыли караваны верблюдов.*	31
ресурс метаф.	32	*Умственное воспитание складывается из следующих элементов.*	44
субъект психологической реакции	30	*Девочка заплакала.*	64
момент времени	30	*Теплоход отходит ровно в полночь.*	50
пациенс / результат	30	*Клены распускают почки.*	29

Экспликация	Кол-во констр. в словаре	Пример конструкции	Кол-во предикатов в словаре
субъект ментального состояния / субъект психологического состояния	29	*Эх, дал я маху.*	43
часть субъекта ментального состояния	28	*Голова не соображает.*	43
часть темы в фокусе внимания	26	*Православные церкви обращены алтарем на восток.*	40
предел	25	*Его работа приближается к идеалу.*	37
начальная точка – сфера	24	*Докладчик отошел в сторону от своей основной темы.*	37
конечное состояние	24	*Она дошла до слез.*	33
конечная точка – пациенс / конечная точка – тема	24	*Они бросались шишками друг в друга.*	21
говорящий – субъект психологического состояния	24	*«Вот это фокус», – удивился он.*	18
социальная среда	23	*Постоянно вращаясь среди актеров, я стал привыкать к их языку.*	33
среда	22	*Он не различил в темноте приближающегося человека.*	28
агенс – субъект психологического состояния	21	*Зря ты себя так казнишь.*	28
аудитория	20	*Думаете, мне было приятно краснеть перед товарищами?*	20
контрагент – субъект ментального состояния	20	*Учитель дважды в неделю занимается с отстающими.*	17

Таблица 3. Наиболее частотные семантические роли в системе «ФреймБанк».

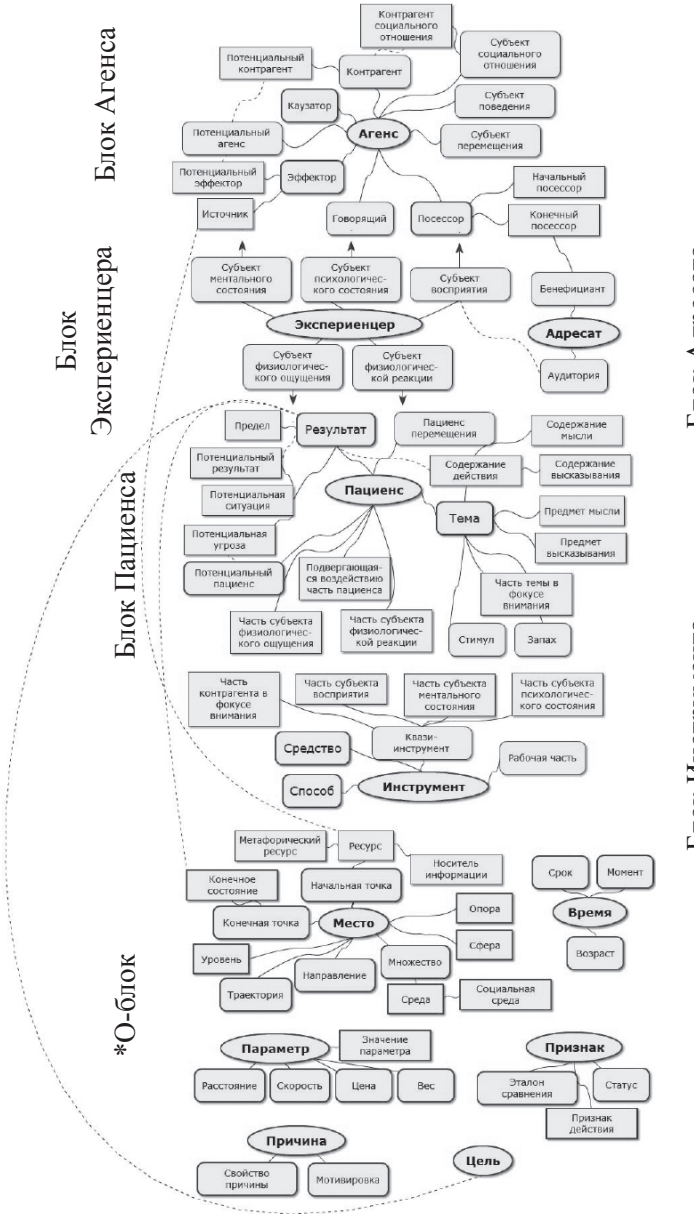


Рис. 23. Иерархия семантических ролей системы «ФреймБанк».

6. Граф конструкций в системе ФреймБанк

В этом разделе мы сформулируем принципы построения графа конструкций и фреймов, которые связывают отдельные словарные входы для лексических конструкций в систему. Собственно, граф фреймов состоит из:

— входов для фреймов, где перечисляются все возможные участники фрейма; фрейм может относиться к конкретной лексической единице, группе лексических единиц и далее к более абстрактному классу;

— связей между этими входами, показывающих их таксономическую организацию (генерализацию и специализацию фреймов);

— связей между разными фреймами одной лексической единицы;

— соответствий между фреймом и всеми лексическими конструкциями, которые к нему относятся.

Граф конструкций — это более глубокая проработка связей между конструкциями, относящимися к одному фрейму, к фреймам одной лексической единицы и к фреймам разных лексических единиц. Граф конструкций позволяет проследить, насколько системно наследуются свойства конструкций от своего прототипического класса к конструкциям конкретных лексем, в каких точках наследование блокируется и как наследование/изменение в морфосинтаксическом оформлении конструкций отражается в семантических переходах на всем множестве предикатной лексики.

6.1. Типы переходов между конструкцями

При разметке системы FrameBank для каждого из глаголов строится семантическая сеть, иллюстрирующая направления и типы переходов между всеми конструкциями этого глагола. Прежде всего, выделяются внутрифреймовые и межфреймовые связи. Связи между конструкциями, относящимися к одному значению глагола и, соответственно, к одному фрейму, маркируют введение нового участника, замену участника при переносе фокуса внимания, мену коммуникативного статуса и морфосинтаксического оформления и т. п. явления. Межфреймовые связи соединяют конструкции, относящиеся к разным значениям глагола.

Кроме того, различаются связи, маркирующие семантический переход, и формальные связи, маркирующие изменение или на-

следование формального паттерна конструкции. Примером семантического перехода является метафора — например, при переходе от конструкции *Гром* ***гремит*** к конструкции *Директор* ***гремел****, обличая пороки*. Вместе с тем, для конструкции *Директор гремел…* устанавливается отношение точного формального наследования (Snom V) как с конструкцией *Гром гремит*, так и с конструкцией *Друзья говорили до утра*, обозначающей прототипическую речевую ситуацию.

В работе [Кашкин, Ляшевская 2013] была изложена классификация переходов, учитывающая различные варианты соответствий между конструкциями, относящимися к фреймам одного глагола (в смысле плана выражения и плана содержания конструкций). Кратко перечислим эти классы:

А1. Мена морфосинтаксического оформления участника (без смены иерархии участников);

А2. Мена статуса участников, диатетический сдвиг;

А3. Невыражение участника, относящегося к известному классу;

А4. Невыражение участника, дейктически или ситуативно известного;

А5. Невыражение участника, неопределенного (неважного) в ситуации;

А6. Добавление участника;

А7. Мена участников (перенос фокуса с одного участника на другого);

Б. Добавление неядерного участника ситуации: в производную конструкцию эксплицитно добавляется участник, не предусмотренный прототипом фрейма;

В1. Специализация значения фрейма, связанная с невыражением одного из участников;

В2. Идиоматизация значения, связанная с введением в конструкцию новых лексических констант;

Г1. Метонимия: смежный участник (замена участника на другого, смежного, в пределах одного фрейма);

Г2. Метонимия: перераспределение коммуникативных акцентов между участникмами ситуации (при диатетическом сдвиге, А2);

Г3. Метонимия: сдвиг домена;

Г4. Метонимия: смежный класс (используется при переходе глагола в смежный с исходным тематический класс);

Д. Метафора (смена таксономического класса какого-либо из участников ситуации, сопровождающаяся сдвигом значения глагола);

Е. Ребрендинг (переход лексемы в другой таксономический класс, основанный на механизме импликатуры: результат семантического перехода является следствием или выводом из исходного значения);

Ж. Другие, более далекие и менее прозрачные переходы, включая выветривание значений у глаголов, выступающих в роли лексической функции.

Иногда поведение конструкций может быть объяснено только через заимствование свойств конструкций других предикатов. Ниже мы покажем подходы к установлению связей между конструкциями для двух таких сложных случаев.

6.2. Конструкция *брать левее* и ее место в графе конструкции

Конструкция *брать (взять) влево / вправо / вверх / вниз / левее / правее / выше / ниже*, ср. (22) — (25), интересна с нескольких точек зрения. Так, она имеет формальный паттерн *Snom V ADV. DIR*, который не используется в других фреймах глагола *брать*[1]. Появление нового участника — Направление / Конечная точка — сигнализирует о переходе глагола в новый для него фрейм физического перемещения (в случае собственно перемещения — левее или правее, выше или ниже от исходной траектории, в случае стрельбы — траектория выстрела пойдет выше, ниже и т. п. относительно исходной и т. д.). Доноры формального паттерна — конструкции типа *идти / ехать* и т. п. + направительная группа {ADV/PRкуда + Sx} и *повернуть* + {ADV/PRкуда + Sx} – причем, заметим, что *повернуть* в контексте типа *влево / левее* обозначает не только ‘сделать поворот’, но и ‘повернув, продолжать движение так, что траектория проходит левее исходной’.

(22) **—** *Капитан, а вы не сбились с дороги? — ежеминутно слышался голос командира дивизиона. — По-моему, надо* ***брать левее***… [А. А. Игнатьев. Пятьдесят лет в строю. Кн. 1–2 (1947–1953)];

[1] Кроме как *бери выше*, которая является метафорическим расширением от данной и вполне объяснимо наследует формальные свойства.

(23) *Река резко* ***брала влево*** <…> [Михаил Тарковский. Кондромо // «Октябрь», 2003];

(24) *Разговор в ритме брасса: — Как сносит. — Надо* ***брать выше****. — Куда? — На ту иву. — Ладно*. [И. Грекова. На испытаниях (1967)];

(25) *Братец Сашенька выстрелил, но не попал:* ***взял выше****!* [А. Ф. Писемский. Люди сороковых годов (1869)].

Однако обратим внимание на два идиоматических свойства этой конструкции. Во-первых, не совсем очевидно, почему центром конструкции, обозначающей перемещение, становится посессивный предикат *брать / взять* (а также их производные типа *забирать*). Во-вторых, имеет свои особенности и оформление участника Направление/Конечная точка: допустима только адвербиальная реализация, но не предложная группа (ср.??*Он взял на Ленинский проспект*). Можно было бы просто констатировать, что перед нами — конструкция-идиома[1], но хотелось бы более последовательно исследовать связи этой конструкции в системе других лексических конструкций русского языка.

В первую очередь, обратим внимание на то, что для нового фрейма автономного перемещения, который по определению требует непереходной конструкции, выбирается такой прототипически переходный глагол, как *брать*. Можно предположить, что между конструкциями *Snom V ADV. DIR* и *Snom V Sacc* должно быть промежуточное звено вида

Snom V Sacc ADV. DIR,

и действительно, его можно обнаружить среди других конструкций с глаголом *брать*.

[1] Вполне вероятно, что конструкция появилась как заимствование из западноевропейских языков, ср. первые примеры в НКРЯ, датируемые концом XVIII – началом XIX вв., причем с возможностью предложной реализации: *Ежели бы граф Мелин не* ***брал влево*** *и следовал точному повелению, тоб о злодее был совершенно известен.* [Военно-походный журнал командира карательного корпуса подполковника Михельсона И.И. (1774)]; *Сначала виден с Артиллерийской бухты дом Снаксарева. Мы* ***берем вправо, в заставу****. Под-вечер гуляю.* [А.С. Грибоедов. Отдельные заметки (1822-1825)]. Однако исследование истории конструкции не входит в задачи нашего синхронного описания.

А. Контексты с функциональным объектом, в т. ч. с генитивной модификацией:

(26) *Лоцман, склонясь вперед, легонько стал* ***брать руль налево***. [С. Т. Григорьев. Красный бакен (1923)];

(27) *Шофер, наверно, не такой опытный, как другие, чуть-чуть* ***взял руля не туда****, заднее колесо заскользило по балке, и грузовик провалился, повиснув карданом на одной из балок и, по счастью, зацепившись передними колесами за другую.* [К. Симонов. Живые и мертвые (1955–1959)];

(28) *Я* ***чуть влево беру руля*** *и, знаете, как метлой смел балласт с флагманской пироги, со второй, с третьей. [А. Некрасов. Приключения капитана Врунгеля* (1960–1980)].

(29) *Несмотря на действительно очень близкое расстояние, Якубович* ***взял прицел чуть левее и выше****, чем следовало, и пуля вместо ноги попала в левую кисть Грибоедова.* [http://diletant.ru/blogs/39739/6056].

По импликации, взять в руки руль, штурвал, вожжи, ружье, лопату, лом — значит поехать, начать стрелять, копать и т. п. Если движение или другая деятельность, включающая движение, уже имеет место, то повернув инструмент, человек меняет его траекторию, например, копает левее / правее исходного места[1]. Таким образом, использование функционального объекта вынуждает возникновение смежных фреймов, отсюда мотивационная связь в цепочке «посессивность» — «изменение траектории движения».

Б. Конструкция вида *брать курс / направление* {ADV/PRкуда + Sx}:

(30) *Залил полный бак и* ***взял курс в Вольск****.* [Героиновый тюремщик // «Криминальная хроника», 2003.07.24];

(31) *Самолет выровнялся, и Сергей увидел, что Бардыкин* ***бе-***

[1] Ср. также примеры, описывающие управление самолетом: *Лейтенант быстро* ***взял штурвал на себя —*** *«ястребок» резко взмыл кверху*. [С.Н. Сергеев-Ценский. В снегах (1941)], хотя, в отличие от *ружья*, в них нет прямого соответствия между направлением движения штурвала и транспорта. Имя *лопата* требует несколько другой конструкции Snom V Sins ADV, ср. *взять лопатой левее*.

***рет курс чуть левее холмов**, чтобы их можно было видеть сверху и немного сбоку.* [Олег Куваев. Территория (1970–1975)];

(32) *Ехать пора, — сказал я и 415-й танк с гордо задранной пушкой **взял направление в сторону «Аминовки»**.* [С. Погодаев. Письмо другу об афганской войне (2004)];

(33) *Грэй **взял направление к открытому морю**, затем стал держаться левого берега*. [А. С. Грин. Алые паруса (1922)].

Помимо этих двух конструкций, близких нашей конструкции формально и относящихся непосредственно к глаголу *брать*, укажем конструкции глагола *держать*, в которых прослеживается параллельное развитие свойств.

В. Конструкции вида а) *держать руль левее* (34), б) *держать путь на запад* (35–36) и в) *держать корабль против ветра* (37–38):

(34) *Стартуй и **держи руль немного левее*** [www.drive2.ru/l/3445510];

(35) *Хорошо. **Держим курс в парк**.* [Максим Милованов. Естественный отбор (2000)];

(36) *Мы тогда **держали курс в SW четверти***. [Ф. Ф. Беллинсгаузен. Двукратные изыскания в Южном Ледовитом океане… (1831)];

(37) *Нам оставалось одно **— держать судно против ветра** и ждать улучшения погоды*. [А. С. Новиков-Прибой. «Коммунист» в походе (1924)];

(38) *Рулевой упорно смотрел на картушку, стараясь **держать корабль на заданном курсе***. [А. С. Новиков-Прибой. «Коммунист» в походе (1924)].

Брать левее и *держать левее* отличаются фазой: последнее означает ‘продолжать двигаться по траектории, находящейся левее исходной (т. е. той, которая признана точкой отсчета)’. Вследствие большей статичности фрейма, с этим глаголом допустимы как направительная конструкция Snom V Sacc {ADV/PRкуда + Sx}, так и локативная конструкция Snom V Sacc {ADV/PRгде + Sx} (ср. здесь *эскадра движется в море*, *продолжать движение в первой*

четверти). Однако, если мы рассмотрим пары *взять яблоко в руку* : *держать яблоко в руке* и *взять левее* : *держать левее*, то нужно признать, что направительная конструкция была бы менее ожидаема[1], если бы не фрейм движения, из которого «копируются» данные конструкционные свойства.

Объединяет *брать левее* и *держать левее* и то, что они подразумевают контролируемость события, ср. неконтролируемое *пуля пошла* / *лом пошел левей*.

Таким образом, мы можем установить связи между конструкцией *взять левее* и конструкциями:

• *идти левее, повернуть левее* (донор конструкции, полное формальное и фреймовое соответствие);

• *брать руль левее, брать курс левее* (неполное формальное соответствие, поддерживающее, однако, паттерн нашей конструкции; эти конструкции ближе к паттернам *брать* в основных посессивных значениях);

• *держать левее* (копирование свойств «заимствованной» конструкции далее к фрейму, отличающемуся фазой).

6.3. Связь конструкций на примере конструкции вида Xnom V перед Yins у глаголов социального отношения

В данном разделе мы обсудим глаголы социального отношения, выступающие в конструкции «Xnom V перед Yins». Эта конструкция оказывается достаточно продуктивной и возможной для большого числа глаголов, ср. в частности *гордиться, дрожать, краснеть, ломаться, оправдываться, ползать, пресмыкаться, унижаться*. Сразу же нетрудно заметить, что здесь возможны как глаголы с исходной семантикой социального отношения (ср. *унижаться*), так и глаголы, у которых подобное значение вторично (ср. *ползать*). Мы остановимся на последнем классе и рассмотрим четыре глагола, выступающие в конструкции с предлогом *перед*: *ползать, пресмыкаться, дрожать* и *ломаться*. Нас будет интересовать, во-первых, одинаков ли механизм сдвига в класс социального отношения для этих глаголов (исходно относящихся к разным семантическим классам); во-вторых, какие сходства и различия в модели сдвига мы мо-

[1] Ср. другие идиомы *брать под контроль* : *держать под контролем, брать на мушку* : *держать на мушке* и т.п.

жем увидеть при сопоставительном исследовании данных глаголов, возможность чего и предоставляет «ФреймБанк» как общая база данных глагольных конструкций.

Ползать, пресмыкаться

Обсудим для начала семантические сдвиги двух глаголов перемещения — *ползать* и *пресмыкаться*.

Лексема *пресмыкаться* исходно описывает перемещение субъекта «ползком, волоча брюхо по земле» [МАС]. Субъектом перемещения здесь может быть только животное, как в (39) и (40), ср. также название биологического класса пресмыкающихся.

(39) *Из пресмыкающих в степи приметить нам только случилось ехидн, которые перед глазами нашими везде* ***пресмыкались****: однако, ни скоту, ни людям никакой пакости от них не слышно*. [Иван Лепехин. Дневные записки (1768–1769)];

(40) ***Пресмыкаясь*** *по песку, выполз из воды налим*. [Юрий Коваль. Листобой (1972)].

Переносно этот глагол сдвигается в интересующую нас область социального отношения (значение по МАС — ‘раболепствовать, угодничать, унижаться’), как в (41) и (42) — причем этот сдвиг, по всей видимости, достаточно старый, ср. пример (41) из текста 1789 г.

(41) *При дворе султанском гордыня имеет престол свой в сердцах одних безобразных евнухов, кои, быв отвержены от всего, что питает дух истинного величия, и не имея ни супруг, ни детей, естественно свирепы и враждуют непрестанно противу прочих чинов и служителей, кои с сокрушением всегда пред ними* ***пресмыкаются****; но когда воспримут какую власть, стараются уподобиться им в коварстве и злости и изрыгают обыкновенно лютость свою на безгласных граждан, пока не возымеют случая распространить оной далее.* [П. А. Левашов. Цареградские письма (1789)];

(42) *Он не умел* ***пресмыкаться*** *перед начальством, не умел просить и любой приказ руководства считал законом.* [Николай Леонов. Лекарство от жизни (2001)].

Данный сдвиг очевидным образом анализируется как метафорический: меняется семантический класс субъекта глагола (животное → человек), а исходная семантика перемещения в «сниженном» пространственном положении переосмысляется как нахождение в сниженном социальном положении.

Глагол *ползать*, исходная семантика которого очень близка глаголу *пресмыкаться*, развивает и аналогичное значение из класса социального отношения, как в (43) и (44). Как указано в [Кузнецова 1963/2010: 203], это значение зародилось в XVIII в., но о первичности его vs. аналогичного значения *пресмыкаться* сведений нет.

(43) *Как бы ни были велики наши исторические несчастия и крушения, мы призваны самостоятельно быть, а не* ***ползать*** *перед другими; творить, а не заимствовать; обращаться к Богу, а не подражать соседям; искать русского видения, русских содержаний и русской формы, а не ходить в кусочки, собирая на мнимую бедность.* [И. А. Ильин. О русской идее (1948)];

(44) *Они зарубили первую диссертацию Бабаева — инфаркт, а потом* ***ползали*** *перед ним на коленях — защищайтесь, ну защищайтесь же, жена про него сказала: «За внешней сдержанностью — бешеный темперамент.* [Александр Терехов. Бабаев (2003)].

На первый взгляд, здесь мы имеем такой же метафорический сдвиг, как и в случае глагола *пресмыкаться*. На самом же деле ситуация оказывается несколько сложнее. В НКРЯ нередки контексты, описывающие позу унизительного, попрошайнического ползания человека, ср. (45) — (46), и именно они могут дополнительно поддерживать развитие такого переноса. Изначально значение ‘унижаться’ могло возникнуть здесь как метонимический результат действия, буквально описываемого глаголом *ползать*, а затем уже мог произойти полный отрыв от исходной семантики и сохранение только семантики результата. Метонимический сдвиг, следующий модели ДЕЙСТВИЕ → РЕЗУЛЬТАТ, хорошо известен в семантической теории, см. краткий обзор в [Рахилина и др. 2010 а: 408–409] и различные примеры такого сдвига в [Падучева 2004] и [Кустова 2004]. В нашем случае, однако, не вполне стандартен уход глагола *ползать*

из исходного семантического класса, типичный уже для метафорического, а не для метонимического переноса.

(45) *Постоянно нуждаясь в деньгах, он выманивал их у Карла, причем* ***ползал*** *перед ним, целуя ему ноги, уверяя в своей любви и преданности.* [П. П. Каратыгин. Временщики и фаворитки 16, 17 и 18 столетий. Книга первая (1870)];

(46) *Утром он, как обычно,* ***ползал*** *перед женой на коленях, хватал ее за ноги, умолял простить и обещал выбросить ружье на помойку или продать.* [Владимир Войнович. Жизнь и необычайные приключения солдата Ивана Чонкина (1969–1975)].

Дрожать

Исходный класс употреблений глагола *дрожать* охватывает ситуации колебания с маленькой амплитудой, в том числе ситуацию определенной физиологической реакции человека. В качестве причины этой физиологической реакции часто выступает страх, что наглядно демонстрируют и данные НКРЯ, где содержится 284 вхождения сочетания *дрожать от страха*. Иллюстрацией могут послужить примеры (47) — (48), где семантика глагола *дрожать* очевидно относится к области физиологической реакции:

(47) *Меня волновало тогда только одно: если нас выведут на расстрел до восхода солнца, не подумают ли наши убийцы, что я* ***дрожу*** *от страха, а не от холода?* [Анастас Микоян. Так было (1971–1974)];

(48) *Бледный,* ***дрожащий*** *от страха Зимовской встал, сделал шаг из-за стола и зашатался.* [Г. М. Марков. Строговы. Кн. 1 (1936–1948)].

В результате глагол *дрожать* сдвигается в обозначение собственно эмоции страха (49–50), причем в НКРЯ такие примеры фиксируются начиная уже с XVIII в.

(49) *Но вместо ее все вельможи* ***дрожали*** *перед Бироном.* [М. М. Щербатов. О повреждении нравов в России (1786–1787)];

(50) *Больше всех* ***дрожали*** *перед своим председателем сами министры*. [А. А. Игнатьев. Пятьдесят лет в строю. Кн. 5 (1947–1953)].

В [Апресян В. Ю., Апресян Ю. Д. 1993] этот и некоторые анало-

гичные случаи анализируются исходя из идеи о сходстве языковой категоризации страха и холода (ср. приводимые в указанной статье примеры *сжаться от страха / от холода*, *зубы стучат от страха / от холода* и нек. др.) и о наличии в языке метафорической модели «ХОЛОД → СТРАХ». Этот анализ очень логично вытекает из представленного авторами эмпирического материала, но, в то же время, мы наблюдаем здесь случай, во многом похожий на рассмотренный выше глагол *ползать*. Имеются две метонимически связанные ситуации: состояние (страх) и его результат (дрожь), и глагол, исходно описывающий одну из них, сначала начинает передавать одновременно обе ситуации, а затем — только вторую ситуацию. Различие, однако, состоит в том, что в случае *ползать* глагол действия сдвигается в семантику результата, а в случае *дрожать*, напротив, глагол результата переносится в обозначение его причины.

Ломаться

Глагол *ломаться* имеет два переносных значения, реализуемых в современном русском языке в конструкции *ломаться перед Yins*. Первое из них определяется в МАС как ‘манерничать, жеманиться’, см. (51) — (52), второе — как ‘упрямиться, не соглашаться на что-л., не имея на то основательных причин’, см. (53) — (54).

(51) *Залез под потолок, копошится там — и вдруг замечаю, как девочки стали* ***ломаться*** *перед ним, кокетничать.* [Михаил Шишкин. Венерин волос (2004) // «Знамя», 2005];

(52) *Перед работником инспекции по делам несовершеннолетних Эльвирой Кириенко позирует, бравирует, рисуется и* ***ломается*** *юная хулиганка и бездельница.* [В. Пронин. Давайте разберемся… (Правовая тема на ленинградском экране) // «Человек и закон», 1978];

(53) *Товарищ Сталин подозвал всех ребят из охраны и, как они ни* ***ломались****, заставил их усесться на ковер и принять участие в этом обеде под открытым небом.* [Фазиль Искандер. Сандро из Чегема (Книга 1) (1989)].[1]

[1] В (53) контрагент социального отношения, обозначаемый обычно предложной группой *перед* + *Yins*, не выражен эксплицитно в силу дейктической определенности соответствующего референта.

(54) ***Поломавшись*** *немного перед Мефистофелем, Фауст быстро подписывает условие и, видимо, страшно рад.* [М. О. Меньшиков. Немецкая душа (1915.01.01)]

Можно предположить, что данные употребления развились из исходного для *ломаться* значения деформации через посредство еще одного (редкого, но всё же встречающегося в НКРЯ), в котором *ломаться* описывает кривляния шута, как в (55).

(55) *Сначала мулла прочитает им что-то из Корана; потом дарят молодых и всех их родственников, едят, пьют бузу; потом начинается джигитовка, и всегда один какой-нибудь оборвыш, засаленный, на скверной хромой лошадёнке,* ***ломается****, паясничает, смешит честную компанию; потом, когда смеркнется, в кунацкой начинается, по-нашему сказать, бал.* [М. Ю. Лермонтов. Герой нашего времени (1839–1841)].

Употребление из (55) связано с «социальными» употреблениями, по-видимому, прозрачным метафорическим сдвигом: неестественное поведение человека описывается той же лексемой, что и принятие им неестественных поз. Более интересен здесь переход от исходного значения деформации ко всему кластеру значений, описывающих действия человека. С одной стороны, здесь можно видеть метафору по внешнему сходству между ломающимся на части (и в этот момент изменяющим форму) предметом и принимающим неестественное положение (и, далее, неестественно ведущим себя) человеком. С другой стороны, рассматриваемый сдвиг не сводится просто к смене семантического класса субъекта и вытекающей из него смены значения глагола, как происходит при стандартной метафоре, но влечет за собой, как мы сейчас увидим, еще и изменение ряда морфосинтаксических характеристик лексемы.

Во-первых, глагол *ломаться* при семантическом сдвиге в область человеческой деятельности приобретает идею контролируемости, отсутствующую у него в исходном значении деформации. В случае употреблений типа (55) это свойство очевидно из контекста. В социальных употреблениях оно тоже сохраняется, что формально проявляется, например, в сочетаемости предиката с обстоятельствами цели (56–58), невозможной в контекстах типа *ветки ломаются*.

(56) — *Я вас не понимаю: для чего вы* ***ломаетесь****?* [Андрей Белый. На рубеже двух столетий (1929)];

(57) — *Он принялся снова терпеливо нашептывать ей на ухо, убежденный, что она слышит и понимает его, но* ***ломается****, чтобы, по бабьей привычке, набить себе цену.* [А. А. Фадеев. Разгром (1925–1926)];

(58) *Саня поначалу считал, что Пашка не знает, куда девать себя от успехов, так внезапно свалившихся на его голову, а потому нарочно* ***ломается*** *и распускает жалость с тоской.* [Виктор Курочкин. На войне как на войне (1965)].

Во-вторых, описываемый сдвиг сопровождается сменой аспектуальных свойств глагола *ломаться*. Исходно этот глагол обозначает моментальное действие, поэтому его формы несовершенного вида не могут иметь длительной интепретации, выражая вместо этого либо семантику глагольной множественности (59), либо генерическое значение (60), либо свою стандартную пунктивную семантику в нарративных употреблениях настоящего времени (61–62).

(59) *Сразу вспоминается старая притча про веник, который переломить не удается, хотя все его прутики по отдельности* ***ломаются*** *без труда.* [И. Демонис. Во все лопатки // «Наука и жизнь», 2007];

(60) *Надписи на воске недолговечны. Папирус со временем темнел и* ***ломался****. А когда хитрые египтяне ввели запрет на вывоз папируса за границу, греки вспомнили, что в Персии для письма издревле использовался дефтер — высушенные шкуры животных.* [Алёна Палажченко. История бумажного листа // «Наука и жизнь», 2009];

(61) *Но вдруг мокрое ее лицо проясняется решимостью, она вцепляется в брючную пряжку, щелкает ноготь,* ***ломается****, он помогает ей…* [Александр Иличевский. Облако // «Октябрь», 2008];

(62) *Так подходит он к водосточной трубе, ползет по ней: «Сейчас полежу на крыше, погреюсь на солнышке». И тут труба* ***ломается****. Летит кот вниз и думает: «Сейчас, как обычно, на лапы приземлюсь».* [Владимир Шахиджанян. 1001 вопрос про ЭТО (№№ 501–1001) (1999)].

Напротив, в рассматриваемых переносных употреблениях с субъектом-человеком глагол *ломаться* сдвигается из аспектуального класса мгновенных событий в класс непредельных процессов, ср. примеры (63) — (64), иллюстрирующие его сочетаемость с обстоятельствами длительности, невозможную в прямом значении деформации:

(63) *Алена изобразила некоторое удивление, но* ***ломалась*** *недолго*. [Екатерина Романова, Николай Романов. Дамы-козыри (2002)];

(64) *Ни-ни, не желает. Долго* ***ломался****, наконец смилостивился.* **—** *С тобой не пойду, с мамой-Мусей.* [В. С. Новицкая. Безмятежные годы (1912)].

Указанное изменение аспектуальных свойств предиката *ломаться* коррелирует и с его семантическим изменением при переносе в область человеческой деятельности. Если исходно он обозначает точечный результат некоторого воздействия на предмет, то в семантику его переносного употребления входит уже не результат, а длящийся сам по себе процесс. Тем самым, можно говорить о том, что к метафорическому сдвигу *ломаться* здесь добавляется еще и близкий к метонимическому сдвиг значения с результата на процесс.

Таким образом, мы определили употребления глагола *ломаться* в классе социального отношения как метафору от значения ‘кривляться, паясничать’, которое, в свою очередь, связано с исходным значением деформации — на первый взгляд, тоже метафорически. При этом, однако, в употреблениях, касающихся человеческой деятельности, глагол *ломаться*, во-первых, приобретает семантику контролируемости, во-вторых, существенно меняет свои аспектуальные свойства, в-третьих, этот переход осложнен дополнительным квазиметонимическим сдвигом с результата на процесс. В результате трудно говорить о проявлении в данном случае метафоры как переноса «по сходству» или, в терминах [Fauconnier 1985], проекции и наложения одного домена на другой — поскольку не кажется прозрачным «сходство» между, с одной стороны, пунктивным, неконтролируемо происходящим событием и, с другой стороны, контролируемым непредельным процессом. Такие переносы, при которых лексема подвергается существенной перестройке семантики и грамматических свойств, уже описывались в литературе

под названием «ребрендинг» — см., например, [Брицын и др. (ред.) 2009], [Рахилина и др. 2010 б] о предикатах боли, [Рахилина и др. 2010 а] о русских качественных прилагательных, ср. также [Резникова 2014] о нестандартных семантических сдвигах в русском интернет-сленге. В данном случае мы наблюдаем пример похожего явления.

Итак, анализ сдвигов в область социального отношения, развиваемых глаголами *ползать, пресмыкаться, дрожать* и *ломаться*, показывает, что во всех этих случаях в основе семантического изменения лежит метафора, но чаще всего механизм сдвига оказывается более сложным — причем по-разному сложным для разных глаголов. Так, для глаголов *ползать* и *дрожать* имеются прозрачные метонимические модели, поддерживающие развитие такой полисемии, а для глагола *ломаться* — еще и существенные морфосинтаксические изменения, сопровождающие его уход из базовой семантической зоны. В практическом плане эти данные демонстрируют пользу изучения глагольной полисемии на основе общей базы данных, объединяющей широкий массив русской глагольной лексики. В теоретическом плане они указывают на необходимость более дробной классификации моделей семантических переходов по сравнению с набором традиционно предлагаемых для этого ярлыков.

* * *

Граф лексических конструкций содержит сведения о системной организации предикатной лексики, о ее многозначности в той части, которая связана с формальными конструкционными свойствами и с переходами между разными значениями. Поисковая система ФреймБанка, разрабатываемая в настоящее время, позволит отфильтровать, например, все глаголы, в конструкциях которых участник выражается определенным образом, посмотреть способы выражения заданной семантической роли и т. п. В целом это даст возможность представить систему наследования конструкций (продолжая опыт [Апресян 1967]) в разных зонах глагольной и именной лексики. Кроме того, интересны формально неполные и семантически непрозрачные переходы — как кажется, ФреймБанк даст стимул к изучению не только «точечных случаев», но и их системы в целом.

7. База сирконстантных участников

База сирконстантных (неядерных) участников, выраженных при предикатно-аргументных конструкциях глаголов, является очень интересным приложением к ФреймБанку. Постоянно пополняемая в процессе разметки реализации конструкций в предложении, база позволяет судить о том, как часто участник с той или иной ролью выражается с некоторым классом лексических конструкций или с некоторым тематическим классом глагола. Эта база дает богатый материал для проверки гипотезы о так называемом «свободном» присоединении сирконстантных участников, а точнее, показывает, что дистрибуция этих участников далеко не свободна.

В Таблице 4 приводятся данные по сочетаемости 14 самых частотных типов сирконстантов и 13 лексико-семантических зон фреймов.

Зона фрейма	время	место	степень	образ действия	узуальность	причина	длительность	одновременность	последовательность	цель	точность	частота	сравнение	скорость
движение	426	172	133	219	148	93	107	62	58	64	26	38	74	82
речь	120	131	44	139	68	56	31	18	24	18	22	38	11	9
физическое воздействие	70	89	44	80	36	28	33	20	21	19	4	24	12	7
эмоция	69	44	224	21	41	39	21	11	11	8	29	3	8	
ментальный	79	79	40	36	31	28	32	18	17	12	21	5	6	
социальное взаимодействие	71	64	58	41	43	35	14	17	9	11	13	10	6	4
начало существования	90	104	5	24	27	29	8	28	17	18	4	13	5	2
посессивный	62	69	10	20	12	14	7	14	2	25	4	5	2	3
психическая сфера	33	22	71	8	16	22	14	16	6	5	9	2	3	6
звук	17	42	2	16	9	2	2	5	6	2	7	2	2	
физиология	23	21	18	12	10	6	6	4	5		2	2		2
изменение состояния или признака	17	8	12	15	15	10	8	6	5	4		6	4	4
прекращение существования	19	6	10	2	10	8	7	8	10		4	5	1	8
SUM	1096	851	671	633	486	370	290	227	191	184	145	153	134	127

Таблица 4. Частотность сирконстантов в различных фреймах.

Гораздо выше математического ожидания сочетание фреймов эмоций и психической сферы с сирконстантами степени; движения с сирконстантами времени, скорости и сравнения; речи с сирконстантами образа действия и места; начала существования и посессивных с сирконстантами места.

Реже, чем ожидалось бы, сочетаются фреймы движения и сирконстанты места, степени, причины и точности; фреймы речи и сирконстанты степени и времени; фреймы физического воздействия с сирконстантами времени и т. д. Интересно, что сирконстанты со значением узуальности и частоты, одновременности и последовательности почти не обнаруживают преференций по совместной встречаемости с фреймами разных тематических зон.

8. Заключение

Проект ФреймБанк стартовал в 2011 году при поддержке программы «Корпусная лингвистика». Сейчас он вошел в «средний возраст» — это значит, он растет вглубь и вширь. Построено ядро словаря лексических конструкций, отработана технология разметки реализации конструкций в тексте. В стадии конструирования находится сеть фреймов и граф лексических конструкций. В планах — нанесение слоя синтаксической разметки, а также разработка поисковой системы для диагностики расхождений между аннотациями в словаре и в тексте, для диагностики взаимосвязей между различными элементами конструкций на разных уровнях разметки — морфологической, синтаксической, лексической, фреймовой и т. д.

Наиболее проработанной является часть, связанная с глагольными конструкциями (ядро ФреймБанка составляют 2,5 тысячи частотных русских глаголов и ассоциированных с ними конструкций и корпусных примеров), однако конструкции предикативов, прилагательных, наречий, предикатных имен существительных не менее интересны и ждут своего представления во ФреймБанке.

Также пока не реализована сплошная полнотекстовая framenet-разметка (full-text annotation) корпуса текстов, которая дала бы возможность изучать дистрибуцию синтаксических и семантических элементов на уровне предложения, абзаца и текста.

Наконец, пока что вся разметка системы производится вручную — вместе с тем, становится все яснее, как можно автоматизировать те или иные задачи. Очевидный прогресс в развитии систем автоматического синтаксического анализа русского языка ([Nivre, Boguslavsky, Iomdin 2008], [RU-EVAL 2012]), систем распознавания семантических ролей ([Kuznetsov 2013], [Smirnov et al. 2014]), систем извлечения знаний в целом дает основания надеяться

на создание компонента автоматической разметки во ФреймБанке в будущем.

Литература

Азарова И. В., Синопальникова А. А., Яворская М. В. Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2004». М.: РГГУ, 2004. С. 542–547.

Антошина С. А., Ляшевская О. Н. Именные модели управления с точки зрения грамматики конструкций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог-2010». М.: РГГУ, 2010. С. 7–12.

Апресян В. Ю., Апресян Ю. Д. Метафора в семантическом представлении эмоций // Вопросы языкознания. 1993. № 3. С. 27–35.

Апресян Ю. Д. Лексическая семантика (синонимические средства языка). М.: Наука, 1974. (2-е изд.: Избранные труды, том I. Лексическая семантика. М.: Языки русской культуры, 1995.)

Апресян Ю. Д. (ред.) Новый объяснительный словарь синонимов русского языка. М.-Вена: Языки славянской культуры, Wiener Slawisitscher Almanach, 2004.

Апресян Ю. Д. О проекте активного словаря (АС) русского языка // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог-2008». М.: РГГУ, 2008. С. 23–31.

Апресян Ю. Д. О семантической мотивированности лексических функций — коллокатов // Вопросы языкознания. 2008. № 8. С. 3–33.

Апресян Ю. Д. О семантической непустоте и мотивированности глагольных лексических функций // Вопросы языкознания. 2004. № 1. С. 3–30.

Апресян Ю. Д. Экспериментальное исследование семантики русского глагола. М.: Наука, 1967.

Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Санников В. З. Теоретические проблемы русского синтаксиса: взаимодействие грамматики и словаря. М.: Языки славянских культур, 2010.

Апресян Ю. Д., Палл Э. Русский глагол — венгерский глагол. Управление и сочетаемость. Будапешт: Tankyonvkiado, 1982.

Брицын В. М., Рахилина Е. В., Резникова Т. И., Яворская Г. М. (ред.) Концепт БОЛЬ в типологическом освещении. Киев: Видавничий Дім Дмитра Бураго, 2009.

Евгеньева А. П. (ред.) Словарь русского языка в 4-х тт. (МАС). М.: Русский язык, 1999.

Зализняк А. А. Грамматический словарь русского языка. М.: Русский язык, 1977.

Золотова Г. А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М.: Наука, 1988.

Кашкин Е. В., Ляшевская О. Н. Семантические роли и сеть конструкций в системе FrameBank // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог-2013». Т. 1: Основная программа конференции. М.: Изд-во РГГУ, 2013. С. 297–311.

Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2012». Раздел III. Доклады, представленные участниками тестирования систем синтаксического анализа (RU-EVAL 2012). С. 77–131.

Красильщик И. С., Рахилина Е. В. Предметные имена в системе «Лексикограф» // НТИ. Сер. 2. 1992. № 9. С. 24–31.

Кузнецова А. И. Динамика русских глаголов движения с XI по XX век. Воронеж: Изд-во ВГУ, 2010.

Кузнецова А. И. Смысловые отношения и их исторические изменения в лексико-семантической группе глаголов движения русского языка (с XI по XX в.) Дисс. … канд. филолог. наук. М., 1963.

Кузнецова Ю. Л., Ляшевская О. Н. Конструкции и трансформации. Рукопись. 2009.

Кустова Г. И. Типы производных значений и механизмы языкового расширения. М.: Языки славянской культуры, 2004.

Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М: Индрик, 2005. С. 155–174.

Лексикограф. Электронный ресурс: http://lexicograph.ruslang.ru.

Лукашевич Н. В. Тезаурусы в задачах информационного поиска. М.: Издательство МГУ, 2011.

Лютикова Е. А., Татевосов С. Г., Иванов М. Ю., Пазельская А. Г.,

Шлуинский А. Б. Структура события и семантика глагола в карачаево-балкарском языке. М.: ИМЛИ РАН, 2006.

Мельчук И. А., Жолковский А. К. Толково-комбинаторный словарь современного русского языка. Вена: Wiener Slavistischer Almanach, 1984.

Падучева Е. В. Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004.

Плунгян В. А. Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира. М.: РГГУ, 2011.

Рахилина Е. В. Когнитивный анализ предметных имен: семантика и сочетаемость. М.: Русские словари, 2000 (2-е изд. М.: Азбуковник, 2010).

Рахилина Е. В., Резникова Т. И. Фреймовый подход к лексической типологии // Вопросы языкознания. 2013. № 2. С. 3–31.

Рахилина Е. В., Резникова Т. И., Карпова О. С. Семантические переходы в атрибутивных конструкциях: метафора, метонимия и ребрендинг // Лингвистика конструкций / Отв. ред. Е. В. Рахилина. М.: Азбуковник, 2010. С. 398–455.

Рахилина Е. В., Резникова Т. И., Бонч-Осмоловская А. А. Типология преобразования конструкций: предикаты боли // Лингвистика конструкций / Отв. ред. Е. В. Рахилина. М.: Азбуковник, 2010. С. 456–540.

Резникова Т. И. Слог у автора неплох: реально доставляет: о нестандартных семантических переходах в глагольной лексике (по данным интернета) // Современный русский язык в Интернете / Отв. ред Е. В. Рахилина. М.: Языки славянских культур, 2014. С. 169–180.

Сазонова И. К. Толково-грамматический словарь русских причастий. М.: АСТ-Пресс Книга, 2008.

Тестелец Я. Г. Введение в общий синтаксис, М.: РГГУ, 2011.

Толдова С. Ю., Кустова Г. И., Ляшевская О. Н. Семантические фильтры для разрешения многозначности в национальном корпусе русского языка: глаголы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог-2008». М.: РГГУ, 2008. С. 522–529.

Толдова С. Ю., Брыкина М. М., Файнвейц А. В. Проблема идентификации именованных сущностей при их автоматическом извлечении // Актуальные инновационные исследования: наука и прак-

тика. 2013. № 1. Электронная публикация: http://elibrary.ru/item.asp?id=19010697

Шведова Н. Ю. Очерки по синтаксису русской разговорной речи. М.: Изд-во АН СССР, 1960.

Шведова Н. Ю. Русская грамматика. М.: Наука, 1980.

Atkins B. T. Tools for computer-aided corpus lexicography: the Hector project // F. Kiefer, G. Kiss, J. Pajsz (eds.) Papers in Computational Lexicography: Complex'92. Budapest: Hungarian Academy of Sciences, 1992. Pp. 1–60.

Boas H. C. Constructing parallel lexicon fragments based on English FrameNet entries: Semantic and syntactic issues // H. Hedeland, T. Schmidt, K. Woerner (eds.). Multilingual Resources and Multilingual Applications. GSCL 2011. Hamburg: Universität Hamburg, 2011. Pp. 9–18.

Burchardt A., Erk K., Frank A., Kowalski A., Pado S., Pinkal M. Using FrameNet for the semantic analysis of German: annotation, representation, and automation // Boas H. C. (ed.). Multilingual FrameNets in Computational Lexicography: Methods and Applications. Berlin, New York: Mouton de Gruyter, 2009. Pp. 209–244.

Dabrowska E. Language, Mind and Brain. Edinburgh: Edinburgh University Press, 2004.

Dabrowska E., Tomasello M. Rapid learning of an abstract language-specific category: Polish children's acquisition of the instrumental construction // Journal of Child Language. 2008. Vol. 35 (3). Pp. 533–558.

Dowty D. R. Thematic proto roles and argument selection // Language. 1991. Vol. 67. Pp. 547–619.

Fauconnier G. Mental spaces: Aspects of meaning construction in natural languages. Cambridge, MA: MIT Press, 1985.

Fillmore Ch. J. The Case for Case // E. Bach, R. T. Harms (eds.). Universals in Linguistic Theory. New York: Holt, Rinehart and Winston, 1968. Pp. 1–88.

Fillmore Ch. J. Topics in lexical semantics // Cole R. W. (ed.), Current issues in linguistic theory. Bloomington: Indiana University Press, 1977. Pp. 76–138.

Fillmore Ch. J. Frame semantics // Linguistics in the morning calm. Seoul: Hanshin Publishing Co, 1982. Pp. 111–137.

Fillmore Ch. J., Johnson C. R., Petruck M. R. L. Background to FrameNet. International Journal of Lexicography. 2003. Vol. 16 (3). Pp. 235–250.

Fillmore Ch. J., Atkins B. T. S. Describing polysemy: The case of 'crawl' // Ravin Y., Laecock C. (eds.). Polysemy. Oxford: Oxford University Press, 2000. Pp. 91–110.

Fillmore Ch. J., Lee-Goldman R. R., Rhomieux R. The FrameNet constructicon // H. C. Boas, I Sag (eds.) Sign-based construction grammar. Stanford: CSLI, 2012. Pp. 309–372.

Hamm F. Frame Semantics. 2007. Электронная публикация:
http://www.uni-stuttgart.de/linguistik/sfb732/files/hamm_FrameSemantics.pdf

Iordanskaja L., Paperno S. A. Russian-English collocational dictionary of the human body. Columbus / Ohio: Slavica Publishers, 1996.

Johnson C., Fillmore Ch. J., Petruck M, Baker C., Ellsworth M., Ruppenhofer J., Wood E. FrameNet: Theory and Practice. Electronic resource, available at: http://www.icsi.berkeley.edu/framenet.

Kilgarriff A., Rundell M., Uì Dhonnchadha E. Efficient Corpus Creation for Lexicography: building the New Corpus for Ireland // Language Resources and Evaluation. 2006. Vol. 40. Pp. 127–152.

Kustova G. I., Lashevskaja O. N., Paducheva E. V., Rakhilina E. V. Verb taxonomy: from theoretical lexical semantics to practice of corpus tagging // B. Lewandowska, K. Dziwirek (eds.). Cognitive Corpus Linguistics Studies. Frankfurt: Peter Lang, 2009. Pp. 41–56.

Mel'čuk I. Actants in semantics and syntax I: actants in syntax // Linguistics. 2004. № 42–1. Pp. 1–66.

Nivre J., Bogulavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian // Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Brighton: One digital, 2008. Pp. 641–648.

Ohara K. Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru // Boas H. C. (ed.), Multilingual FrameNets in Computational Lexicography: Methods and Applications. Berlin, New York: Mouton de Gruyter, 2009. Pp. 163–182.

Subirats C. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon // Boas H. C. (ed.). Multilingual FrameNets in Computational Lexicography: Methods and Applications. Berlin, New York: Mouton de Gruyter, 2009. Pp. 135–162.

Tomasello M. Constructing a language: a usage-based theory of child language acquisition. Cambridge, MA: Harvard University Press, 2003.

[1]O. N. Lyashevskaya, [2]E. V. Kashkin
[1,2]V. V. Vinogradov Russian Language Institute of RAS
[1]National Research University Higher School of Economics
[1,2](Russia, Moscow)
[1]olesar@yandex.ru, [2]egorkashkin@rambler.ru

TYPES OF INFORMATION ABOUT LEXICAL CONSTRUCTIONS IN RUSSIAN FRAMEBANK

Russian FrameBank is an online resource based on the data from the Russian National Corpus. It includes a dictionary of lexical constructions and a corpus of their uses tagged with a FrameNet-like annotation scheme.

On the one hand, FrameBank links the semantics of lexical units to formalized conceptual structures – frames. The users can observe how the slots of a frame are arranged in a text: which participants are overt or covert; what is the order of a predicate and its participants; what other sentence elements (discourse markers, negation etc.) interact with the participants of the target frame, etc. FrameBank also provides information on the morpho-syntactic features of frame elements in a discourse (e. g. typical morpho-syntactic patterns, deviations from the prototype, etc.). Such data on the interface between morpho-syntax and frame semantics may be useful for developing systems of artificial intelligence aimed at retrieving information from texts with due account for morpho-syntactic variation and possible omission of participants.

On the other hand, FrameBank is a tool for studying the constructional properties of the Russian lexicon. It makes it possible to find out what morpho-syntactic, semantic or lexical restrictions are imposed by a particular construction on its participants. The illustrations of verbal and nominal government are also available in FrameBank, as well as the list of idioms which are present in the bank of corpus examples.

In addition, FrameBank represents lexical constructions as a bridge between different meanings of a predicate (or between the meanings of synonyms), at the same time highlighting constructional oppositions between lexical meanings. The database makes it possible to trace how the lexicon of Russian predicates is overlaid with the network of frames and how the structural relations in lexical semantics (polysemy,

synonymy, antonymy, etc.) are reflected in the network of lexical constructions.

This article describes the structure of the dictionary, the principles of annotating corpus examples, the graph of semantic roles and the graph of lexical constructions in FrameBank.

Key words: FrameBank, FrameNet, Russian, Construction Grammar, corpus linguistics, lexicography, construction, verb, morpho-syntax, semantic role, frame, polysemy.

References

Antoshina S. A., Lyashevskaya O. N. [Nominal case patterns from the viewpoint of Construction Grammar]. *Komp'yuternaya lingvistika i intellektual'nye tehnologii: trudy mezhdunarodnoi konferentsii "Dialog–2010"* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2004»]. Moscow, RSUH Publ., 2010, pp. 7-12. (In Russ.)

Apresjan Yu. D. *Eksperimental'noe issledovanie semantiki russkogo glagola* [An experimental study on the semantics of Russian verb]. Moscow, Nauka Publ., 1967.

Apresjan Yu. D. *Leksicheskaya semantika (sinonimicheskie sredstva yazyka)* [Lexical semantics (synonymical tools of a language)]. Moscow, Nauka Publ., 1974. / 2nd edition: Izbrannye trudy. Tom I. Leksicheskaya semantika [Selected papers. Vol. 1. Lexical semantics]. Moscow, Yazyki russkoi kul'tury Publ., 1995.

Apresjan Yu. D. (Ed.) *Novyi obyasnit'el'nyi slovar' sinonimov russkogo yazyka* [A new explanatory dictionary of Russian synonyms]. Moscow-Wien, Yazyki slavyanskoi kul'tury Publ, Wiener Slawisitscher Almanach, 2004.

Apresjan Yu. D. [On the semantic non-emptiness and motivation of verbal lexical functions]. *Voprosy yazykoznaniya*. 2004, no. 1, pp. 3–30. (In Russ.)

Apresjan Yu. D. [On creating an active dictionary of Russian]. *Komp'yuternaya lingvistika i intellektual'nye tehnologii: Trudy mezhdunarodnoi konferentsii "Dialog–2008"* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2008"]. Moscow, RSUH Publ., 2008, pp. 23–31. (In Russ.)

Apresjan Yu. D. [On the semantic motivation of lexical functions – collocates]. *Voprosy yazykoznaniya*, 2008, no. 8, pp. 3–33. (In Russ.)

Apresjan Yu. D., Boguslavskij I. M., Iomdin L. L., Sannikov V. Z. *Teoreticheskie problemy russkogo sintaksisa: vzaimodeystvie grammatiki i slovarya* [The theoretical issues of Russian syntax: interaction between grammar and lexicon]. Moscow, Yazyki slavyanskikh kul'tur Publ., 2010.

Apresjan Yu. D., Pall Je. *Russkii glagol – vengerskii glagol. Upravlenie i sochetaemost'* [Russian verb – Hungarian verb. Government and combinability]. Budapest, Tankyonvkiado Publ., 1982.

Apresjan V. Yu., Apresjan Yu. D. [Metaphor in the semantic representation of emotions]. *Voprosy yazykoznaniya*, 1993, no. 3, pp. 27–35. (In Russ.)

Azarova I. V., Sinopal'nikova A. A., Yavorskaya M. V. [Principles of building a Wordnet-thesaurus RussNet]. *Komp'yuternaya lingvistika i intellektual'nye tehnologii: trudy mezhdunarodnoi konferentsii "Dialog–2004"* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2004"]. Moscow, RSUH Publ., 2004, pp. 542–547. (In Russ.)

Atkins B. T. Tools for computer-aided corpus lexicography: the Hector project // *Papers in Computational Lexicography: Complex'92*. F. Kiefer, G. Kiss, J. Pajsz (eds.). Budapest, Hungarian Academy of Sciences, 1992, pp. 1–60.

Boas H. C. Constructing parallel lexicon fragments based on English FrameNet entries: Semantic and syntactic issues. *Multilingual Resources and Multilingual Applications. GSCL 2011*. H. Hedeland, T. Schmidt, K. Woerner (eds.). Hamburg, Universität Hamburg, 2011. Pp. 9–18.

Britsyn V. M., Rakhilina E. V., Reznikova T. I., Yavorskaja G. M. *Koncept BOL" v tipologicheskom osveshchenii* [The concept of PAIN in a cross-linguistic perspective]. Kiev, Vidavnichii Dim Dmitra Burago Publ., 2009.

Burchardt A., Erk K., Frank A., Kowalski A., Pado S., Pinkal M. Using FrameNet for the semantic analysis of German: annotation, representation, and automation. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. H.C. Boas (ed.). Berlin, New York, Mouton de Gruyter, 2009, pp. 209–244.

Dabrowska E. *Language, Mind and Brain*. Edinburgh, Edinburgh University Press, 2004.

Dabrowska E., Tomasello M. Rapid learning of an abstract language-specific category: Polish children's acquisition of the instrumental construction. *Journal of Child Language*, 2008, Vol. 35(3), pp. 533–558.

Dowty D. R. Thematic proto roles and argument selection. *Language.* 1991, vol. 67, pp. 547–619.

Evgen'eva A. P. (Ed.) *Slovar' russkogo yazyka v 4-x tt*. [The dictionary of Russian in 4 volumes]. Moscow, Russkii yazyk Publ., 1999.

Fauconnier G. *Mental spaces: Aspects of meaning construction in natural languages.* Cambridge, MA: MIT Press, 1985.

Fillmore Ch. J. The Case for Case. *Universals in Linguistic Theory.* E. Bach, R.T. Harms (eds.). New York: Holt, Rinehart and Winston, 1968, pp. 1–88.

Fillmore Ch. J. *Topics in lexical semantics. Current issues in linguistic theory.* R. W. Cole (ed.). Bloomington, Indiana University Press, 1977, pp. 76–138.

Fillmore Ch. J. Frame semantics. *Linguistics in the morning calm.* Seoul: Hanshin Publishing Co, 1982, pp. 111–137.

Fillmore Ch. J., Johnson C. R., Petruck M. R. L. Background to FrameNet. *International Journal of Lexicography*. 2003, vol. 16(3), pp. 235-250.

Fillmore Ch. J., Atkins B. T. S. Describing polysemy: The case of 'crawl'. *Polysemy*. Y. Ravin, C. Laecock (eds.). Oxford: Oxford University Press, 2000, pp. 91–110.

Fillmore Ch. J., Lee-Goldman R. R., Rhomieux R. The FrameNet constructicon. *Sign-based construction grammar*. H. C. Boas, I Sag (eds.) Stanford: CSLI, 2012, pp. 309–372.

Hamm F. (2007), *Frame Semantics*, available at: http://www.uni-stuttgart.de/linguistik/sfb732/files/hamm_FrameSemantics.pdf (accessed on 20.06.2015)

Iordanskaya L., Paperno S. *A Russian-English collocational dictionary of the human body.* Columbus/Ohio: Slavica Publishers, 1996.

Johnson C., Fillmore Ch. J., Petruck M, Baker C., Ellsworth M., Ruppenhofer J., Wood E. FrameNet: Theory and Practice, available at: http://www.icsi.berkeley.edu/framenet. (accessed on 20.06.2015)

Kashkin E. V., Lyashevskaya O. N. [Semantic roles and construction network in FrameBank]. *Komp'yuternaya lingvistika i intellektual'nye tehnologii: po materialam ezhegodnoi Mezhdunarodnoi konferentsii "Dialog-2013". T. 1. Osnovnaya programma kon-*

ferentsii [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference “Dialog-2013”. Vol. 1. The main program of the conference]. Moscow, RSUH Publ., 2013, pp. 297–311. (In Russ.)

Kilgarriff A., Rundell M., Uì Dhonnchadha E. Efficient Corpus Creation for Lexicography: building the New Corpus for Ireland. *Language Resources and Evaluation*, 2006, Vol. 40, pp. 127–152.

Krasil'shhik I. S., Rakhilina E. V. [Names of objects in the Lexicographer database]. *NTI, ser. 2*, no. 9, 1992, pp. 24–31. (In Russ.)

Kustova G. I. *Tipy proizvodnyh znachenij i mehanizmy jazykovogo rasshirenija* [Types of derived meanings and patterns of linguistic extension]. Moscow, Yazyki slavyanskoi kul'tury Publ., 2004.

Kustova G. I., Lashevskaja O. N., Paducheva E. V., Rakhilina E. V. Verb taxonomy: from theoretical lexical semantics to practice of corpus tagging. *Cognitive Corpus Linguistics Studies*. B. Lewandowska, K. Dziwirek (eds.). Frankfurt, Peter Lang, 2009, pp. 41–56.

Kustova G. I., Lyashevskaya O. N., Paducheva E. V., Rakhilina E. V. [Semantic annotation in the Russian National Corpus: principles, problems, perspectives]. *Nacional'nyi korpus russkogo yazyka: 2003–2005. Rezul'taty i perspektivy*. [Russian National Corpus: 2003–2005. Results and perspectives]. Moscow, Indrik Publ, pp. 155–174. (In Russ.)

Kuznetsova A. I. *Smyslovye otnosheniya i ikh istoricheskie izmeneniya v leksiko-semanticheskoi gruppe glagolov dvizheniya russkogo yazyka (s XI po XX v.)* [Semantic relations and their historical changes in the domain of motion verbs in Russian (from XI till XX century)]. Phd diss. Moscow, MSU Publ., 1963.

Kuznetsova A. I. *Dinamika russkikh glagolov dvizheniya s XI po XX vek* [The evolution of Russian motion verbs from XI till XX century]. Voronezh, VSU Publ., 2010.

Kuznetsova Yu. L., Lyashevskaya O. N. *Konstruktsii i transformatsii* [Constructions and transformations]. 2009, ms.

Lexicographer, available at: http://lexicograph.ruslang.ru. (accessed on 20.06.2015)

Lyutikova E. A., Tatevosov S. G., Ivanov M. Ju., Pazel'skaja A. G., Shluinskij A. B. *Struktura sobytija i semantika glagola v karachaevo-balkarskom jazyke* [Event structure and the semantics of a verb in Karachay-Balkar]. Moscow, IMLI RAS Publ., 2006.

Lukashevich N. V. *Tezaurusy v zadachakh informatsionnogo poiska* [Thesauri in the tasks of information retrieval]. Moscow, MSU Publ., 2011.

Mel'čuk I. Actants in semantics and syntax I: actants in syntax. *Linguistics*. 2004, no. 42–1, pp. 1–66.

Mel'chuk I. A., Zholkovskij A. K. *Tolkovo-kombinatornyj slovar' sovremennogo russkogo jazyka* [Explanatory and combinatorial dictionary of modern Russian]. Vienna, Wiener Slavistischer Almanach Publ., 1984.

Nivre J., Bogulavsky I., Iomdin L. Parsing the SynTagRus Treebank of Russian. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Brighton: One digital, 2008, pp. 641–648.

Ohara K. Frame-based contrastive lexical semantics in Japanese FrameNet: The case of risk and kakeru. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. H.C. Boas (ed.). Berlin, New York: Mouton de Gruyter, 2009, pp. 163–182.

Paducheva E. V. *Dinamicheskie modeli v semantike leksiki* [Dynamic patterns in lexical semantics]. Moscow, Yazyki slavyanskoi kul'tury Publ., 2004.

Plungian V. A. *Vvedenie v grammaticheskuyu semantiku: grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira* [An introduction to grammatical semantics: grammatical meanings and grammatical systems in the World's languages]. Moscow, RSUH Publ., 2011.

Rakhilina E. V. *Kognitivnyi analiz predmetnykh imen: semantika i sochetaemost'* [Cognitive-based analysis of object names: semantics and combinability]. Moscow, Russkie slovari Publ., 2000. (2nd edition – Moscow, Azbukovnik Publ., 2010)

Rakhilina E. V., Reznikova T. I. [A frame-based approach to lexical typology]. *Voprosy yazykoznaniya*, 2013, no. 2, pp. 3–31. (In Russ.)

Rakhilina E. V., Reznikova T. I., Karpova O. S. [Semantic shifts in attributive constructions: metaphor, metonymy and rebranding]. E. V. Rakhilina (ed.). *Lingvistika konstruktsii* [Construction linguistics]. Moscow, Azbukovnik Publ., 2010, pp. 398–455. (In Russ.)

Rakhilina E. V., Reznikova T. I., Bonch-Osmolovskaya A. A. [The typology of constructional changes: predicates of pain]. E. V. Rakhilina

(ed.). *Lingvistika konstrukcii* [Construction linguistics]. Moscow, Azbukovnik Publ., 2010, pp. 456–540. (In Russ.)

Reznikova T. I. [*Slog u avtora neplokh: real'no dostavlyaet*: on nonstandard semantic shifts in the verbal lexicon (evidence from Internet texts)]. Rakhilina E. V. (ed.). *Sovremennyj russkii yazyk v Internete* [Modern Russian on the Internet]. Moscow, Yazyki slavjanskih kul'tur Publ., 2014, pp. 169–180. (In Russ.)

Sazonova I. K. *Tolkovo-grammaticheskii slovar' russkikh prichastii* [An explanatory and grammatical dictionary of Russian participles]. Moscow, AST-Press Kniga Publ., 2008.

Shvedova N. Yu. *Ocherki po sintaksisu russkoi razgovornoi rechi* [Essays on the syntax of Russian colloquial speech]. Moscow, Academy of Sciences of the USSR Publ., 1960.

Shvedova N. Yu. (Ed.) *Russkaya grammatika* [Russian grammar]. Moscow, Nauka Publ., 1980.

Subirats C. Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. *Multilingual FrameNets in Computational Lexicography: Methods and Applications.* H.C. Boas (ed.). Berlin, New York: Mouton de Gruyter, 2009, pp. 135–162.

Testelec Ja. G. *Vvedenie v obshhii sintaksis* [An introduction to the general theory of syntax]. Moscow, RSUH Publ., 2001.

Toldova S. Ju., Brykina M. M., Fainveits A. V. [The problem of named entity recognition in automatic extraction]. *Aktual'nye innovatsionnye issledovaniya: nauka i praktika* [Current innovative studies: theory and practice]. 2013, no. 1. Available at: http://elibrary.ru/item.asp?id=19010697 (In Russ.)

Toldova S. Yu., Kustova G. I., Lyashevskaya O. N. [Semantic filters for polysemy resolution in the Russian National Corpus]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoi konferentsii «Dialog–2008»* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2008"]. Moscow, RSUH Publ., 2008, pp. 522–529. (In Russ.)

Tomasello M. *Constructing a language: a usage-based theory of child language acquisition*. Cambridge, MA, Harvard University Press, 2003.

Zaliznyak A. A *Grammaticheskii slovar' russkogo yazyka* [A grammatical dictionary of Russian]. Moscow, Russkii Yazyk Publ., 1977.

Zolotova G. A. *Sintaksicheskii slovar'. Repertuar elementarnykh edinits russkogo sintaksisa* [A syntactic dictionary. The inventory of elementary units in Russian syntax]. Moscow, Nauka Publ., 1988.

Е. А. Гришина
Институт русского языка им. В.В. Виноградова РАН
(Россия, Москва)
rudi2007@yandex.ru

О РУССКОМ ЖЕСТИКУЛЯЦИОННОМ ОТРИЦАНИИ[1]

В статье исследуется соотношение слова и жеста в стандартных и наиболее частотных конструкциях отрицания предиката. В первом разделе рассматриваются фразы с отрицательной конструкцией «*не* + предикат» и анализируются закономерности, характерные для выбора отрицательных жестов «качать головой» и «отрезать рукой». Во втором разделе описаны конструкции с глубинным отрицанием — контексты, которые не содержат отрицательной конструкции на поверхностном уровне, но, тем не менее, сопровождаются отрицательными жестами. Был рассмотрен довольно большой массив данных, на котором установлены корреляции между лингвистическими признаками конструкций с наличием или отсутствием жестикуляционного отрицания. Получены некоторые статистические закономерности употребления основных отрицательных жестов, которым дана содержательная интерпретация.

Key words: жестикуляционное отрицание, отрицательные конструкции, конструкции с глубинным отрицанием

1. Введение

Исследование жестикуляционных способов выражения отрицания является довольно популярным и, пожалуй, одним из самых разработанных направлений в современной жестикуляционной лингвистике и уступает, очевидно, только исследованию способов

[1] Работа выполнена при поддержке РФФИ (гранты № 14-06-00245 и № 15-06-04334) и РГНФ (грант № 15-04-12018).

выражения согласия и подтверждения. Основные направления, в которых продвигаются соответствующие исследования, таковы.

1. Выражение отрицания в языках глухонемых (дескриптивные и типологические исследования, а также сравнение структуры отрицательных конструкций в языках глухонемых и в соответствующих «звуковых» языках)[1].

2. Возникновение отрицательных конструкций и отрицательных жестов в онтогенетическом[2] и филогенетическом[3] аспектах.

3. Состав и значение жестов отрицания в разных языках, а также сопоставительные межкультурные исследования жестикуляционного отрицания. В качестве необходимого компонента в эти работы входит анализ семантических, прагматических и ритмических взаимоотношений жестов отрицания с коммуникативной ситуацией и сопровождающей речью. Здесь прежде всего следует упомянуть работы Ивлин МакКлейв [McClave 2000, 2001], [McClave et al. 2008], Адама Кендона [Kendon 2002, 2004], Саймона Харрисона [Harrison 2009, 2010, 2013], Женевьевы Кальбрис [Calbris 2008, 2011].

Наше исследование задумывалось как работа, примыкающая к третьему направлению. Мы планируем на материале Мультимедийного русского корпуса (МУРКО) рассмотреть состав русских отрицательных жестов и исследовать отношения между основными отрицательными жестами и сопровождающей их речью. База данных включает в себя в общей сложности более тысячи двухсот контекстов, что позволяет полагаться на полученные статистические данные.

Сразу заметим, что в статье не рассматриваются отрицательные жесты, которые используются в качестве сопровождения или замены русских предложений *Да* и *Нет*, служащих положительной или отрицательной реакцией на высказывание собеседника. Основой для статьи послужили наиболее стандартные **отрицательные конструкции**, состоящие из сочетания частицы *не* и предиката, в подавляющем большинстве случаев — глагола (несколько замечаний

[1] См., в частности, среди многих других, [Veinberg and Wilbur 1990], [Coerts 1992], [Bergman 1995], [Nespor and Sandler 1999], [Sandler 1999], [McClave 2001], [Pfau 2002], [Pfau, Quer 2002], [Atkinson et al. 2004], [McKee 2006], [Pfau 2008].

[2] См., например, [Pea 1980], [Guidetti 2005], [Esteve-Gibert, Prieto 2014] и литературу при этих работах.

[3] См. [Pinfield 2013], [Schneider et al. 2010] и соответствующую литературу.

о противопоставительном отрицании *не X, а Y, не только X, но и Y*, а также о конструкциях с двойным отрицанием и о конструкциях с предикатными отрицаниями типа *нечего, некому, негде* и под. будут сделаны в отдельном Разделе 2.4 — поскольку основным предметом исследования в данной работе является соотношение слова и жеста в стандартных и наиболее частотных конструкциях отрицания предиката).

Статья состоит из двух основных разделов. В Разделе 2 анализируются закономерности, характерные для выбора отрицательных жестов во фразах с отрицательной конструкцией *не* + предикат. В Разделе 3 описаны основные типы контекстов, которые не содержат отрицательной конструкции на поверхностном уровне, но, тем не менее, сопровождаются отрицательными жестами, — такие контексты мы будем называть **конструкциями с глубинным отрицанием**.

Методика исследования заключается в следующем. В отрицательных конструкциях и в конструкциях с глубинным отрицанием мы выделяем те или иные лингвистические характеристики — семантика связанных с жестикуляционным отрицанием лексем, синтаксические структуры, морфологические компоненты значения, ритмические особенности (прежде всего, различение эмфатических и нейтральных конструкций). Каждому примеру из нашей базы данных, тем самым, присваивается некоторый набор лингвистических признаков, наличествующих в сопровождающем отрицательный жест (или его отсутствие) устном высказывании. Понятно, что при достаточно большой базе данных такая ее структура позволяет оперировать статистическими параметрами. С помощью критерия χ-квадрат мы определяем, какие лингвистические характеристики существенны для наличия / отсутствия жестикуляционного отрицания, а также для его качества (органа, которым оно осуществляется). На этом заканчивается, так сказать, фактический этап работы и начинается интерпретационный — когда мы делаем попытки предложить объяснения найденным закономерностям.

2. Жестикуляционное сопровождение отрицательных конструкций

2.1. ОСНОВНЫЕ И ФАКУЛЬТАТИВНЫЕ ОТРИЦАТЕЛЬНЫЕ ЖЕСТЫ

Состав отрицательных жестов, характерных для того или иного языка, в значительной степени зависит от того, что именно исследователь считает жестом, т. е. до какой степени подробности он доходит в членении на компоненты того или иного семантически нагруженного движения человеческого тела. В качестве примера можно привести движение руки/рук в горизонтальном плане параллельно земле, которое описывается как отрицательный жест во многих работах — у А. Кендона, Ж. Кальбрис, С. Харрисона и др.

Этот отрицательный жест

1) может осуществляться одной или двумя руками;

2) при осуществлении одной рукой рука может двигаться поперек всего тела говорящего (например, правая рука — из левой периферии через центр в правую периферию) или просто выходить из центральной зоны в периферийную (например, правая рука — из центральной зоны в правую периферию);

3) при осуществлении двумя руками руки могут двигаться изнутри наружу — из центра в периферийные зоны (правая — в правую периферию, левая — в левую),— или производить (каждая) поперечное движение относительно тела говорящего (левая — из правой периферии через центр в левую периферию, правая — из левой периферии через центр в правую периферию);

4) движение в горизонтальном плане может сочетаться с перемещением руки сверху вниз в вертикальном плане (т. е. исходный пункт движения отрицающей руки может быть расположен выше конечного пункта);

5) основной ориентацией ладони в этом жесте является ориентация ладони вниз, параллельно земле, однако ладонь может быть ориентирована также вверх и вертикально;

6) обычно этот жест осуществляется открытой ладонью, но ладонь может быть конфигурирована и иначе — например, в конфигурацию *кольцо, указательный палец, кулак, угол*;

7) жест может осуществляться однократно или многократно.

Возникает вопрос — различные комбинации перечисленных возможностей являются одним и тем же жестом или разными? Саймон Харрисон, например, вслед за Адамом Кендоном полагает, что из всей богатой совокупности комбинаций следует выделить основной признак — открытая ладонь, ориентированная вниз и движущаяся поперек тела говорящего, — и на основе этого признака сформировать семейство жестов (см. [Mueller 2004] о понятии gesture family) со значением предикативного (относящегося ко всей предикации) отрицания. Тем самым, с этой точки зрения, жесты, изображенные на рис. 1, заимствованном нами из работы [Harrison 2013] (отрицательный жест одной рукой, двумя руками из противоположных зон, двумя руками из центральной зоны), принадлежат к одному и тому же семейству жестов и могут описываться как один жест.

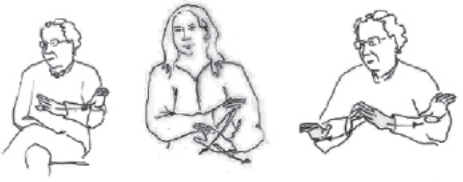


Рис. 1.

Мы не собираемся в данной статье подробно рассматривать понятие жеста в целом и отрицательных жестов в частности у разных авторов, поскольку задача сделать соответствующий обзор перед нами не стоит. Обозначим основные принципы, по которым мы вычленяли отрицательные жесты для нашего исследования.

Мы различаем основные и факультативные отрицательные жесты, вне зависимости от органа, которым они исполняются, и от способов их исполнения.

Основными отрицательными жестами мы называем жесты, которые могут использоваться в режиме пантомимы (т. е. без речевого сопровождения) в качестве отрицательного ответа на вопрос, и при этом передают исключительно значение ‘нет’, без добавочных семантических компонентов. Такими жестами в русской жестикуляционной системе являются отрицательное качание головой (однократное или многократное, то, что в англоязычной традиции назы-

вается headshake (s)), а также резкое поперечное (горизонтальное) движение рукой (с любой конфигурацией и ориентацией ладони, но чаще всего — с открытой ладонью, ориентированной вниз). При этом поперечное движение может осуществляться одной или двумя руками, ладонь может иметь, в принципе, любую конфигурацию (*кольцо, указательный палец, кулак, угол*), ладонь может быть ориентирована не только вниз, но и вверх, под углом и строго вертикально. Таким образом, для отрицательного жеста рукой в русской жестикуляционной системе смыслоразличительными признаками является горизонтальный компонент в движении и резкость (повышенная четкость начала и конца ударной фазы жеста, повышенная скорость осуществления).

Как будет показано ниже, основные отрицательные жесты могут осуществляться одновременно, т. е. говорящий может отрицательно качать головой (ниже условно обозначаем этот жест *качание головой*, КГ, ☹) и одновременно осуществлять основной отрицательный жест рукой, который мы условно называем *отрезать рукой* (обозначения ОР, ✂). При этом не происходит взаимного погашения этих жестов, т. е. одновременное исполнение двух отрицательных жестов ведет не к появлению положительного значения предикации (по типу *я не могу не сказать = я должен сказать*), а скорее напоминает использование головных и ручных отрицательных компонентов в языке глухонемых, где один компонент может исполняться головой (обязательный отрицательный жест), но при этом отрицаемый компонент фразы снабжается отрицательным аффиксом или лексемой, исполняемыми рукой (конфигурация *кольцо*, обозначающая в этом случае ‘ноль, ничто’, *указательный палец*, направленный на адресата, движение руки из центра на периферию в сочетании с рядом предикатов, например, *хотеть, знать*, и др. способы).

Факультативными отрицательными жестами мы считаем жесты, которые в режиме пантомимы передают не только семантический компонент ‘отрицание’, но и добавочные семантические компоненты; чаще всего ими являются некоторые классы предикатов. Наиболее известным таким отрицательным жестом является жест *стоп*: движение открытой ладони, развернутой на адресата или на некоторый предлагаемый собеседником объект. В режиме пантомимы этот жест обозначает ‘не нужно’, т. е. помимо чистого отрицательного компонента ‘нет’ включает семантический компо-

нент ‘необходимость, потребность’. Еще одним факультативным жестом отрицания является жест *развести руками*, который в режиме пантомимы означает что-то из ряда ‘не знаю’ или ‘не могу’. Жесты дистанцирования (*поднять брови, пожать плечами, отодвинуться, склонить голову набок* и ряд других) также являются факультативными отрицательными жестами и в режиме пантомимы обозначают что-то из ряда ‘не знаю’, ‘не понимаю’, ‘мне кажется это необычным=мне не кажется это обычным’, ‘я не интересуюсь затронутой тобой темой’, ‘я не хочу иметь с этим ничего общего’ и под.

Наше исследование посвящено основными отрицательным жестам в русской жестикуляционной системе — ☹ *качать головой* и ✂ *отрезать рукой*, факультативные отрицательные жесты если и будут затронуты, то лишь вскользь или как некоторое вспомогательное средство при анализе контекстов.

2.2. ЛИНГВИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ОСНОВНЫХ ОТРИЦАТЕЛЬНЫХ ЖЕСТОВ

Итак, предметом анализа в данном исследовании является соотношение основных отрицательных жестов КГ и ОР с речевыми контекстами, включающими отрицание на поверхностном уровне (*отрицательные конструкции*) и формально не имеющими отрицания на поверхностном уровне (*конструкции с глубинным отрицанием*)[1].

Отрицательные конструкции могут не сопровождаться отрицательным жестом (1)[2], сопровождаться жестом КГ (2), жестом ОР (3) и сочетанием жестов КГ и ОР (4).

(1) *Чтобы не говорить дальше загадками, я приведу примеры*

(2) *Дата его рождения* ☹{*не совпадает*} *с официальной датой*[3]

[1] В литературе мы встретили только одно замечание о том, чем различаются ручные и головные указательные жесты, – в работе [Jungheim 2004: 137], где были проанализированы жесты, сопровождающие речь говорящих, для которых японский язык является родным, и тех, кто изучает японский как второй язык: ручные жесты доминировали в случае отказа от предложения, а головные – в случае отказа от приглашения.

[2] Отсутствие отрицательного жеста при отрицательной конструкции далее обозначается разрядкой.

[3] Далее примеры оформлены следующим образом: в фигурных скобках заключена зона действия жеста, т.е. та часть высказывания, на которую приходится ударная часть жеста; перед скобкой словесно или с помощью условных знаков обозначается тип сопровождающего жеста/жестов.

(3) … *а я* ✂{*не поняла*}, *что же мне теперь делать*

(4) … *и многие люди, кстати сказать,* ☹✂{*до этого не доходят*}.

На первый взгляд кажется, что выбор отрицательного жеста (КГ, ОР, КГ+ОР) или выбор между наличием и отсутствием сопровождающего отрицательно жеста является абсолютно случайным, поскольку при анализе одиночных примеров не удается выявить никаких лингвистических оснований для выбора между имеющимися четырьмя возможностями. Если мы сравним следующие несколько контекстов (5)–(7), то нам будет трудно определить, какие именно факторы влияют на выбор жестикуляционного сопровождения, и есть ли вообще такие факторы:

(5) *а я* ☹✂{*не знал, сколько стоит*}

(6) *я тогда еще* ☹{*еврейского языка*} *не знал*[1]

(7) *да я не знал, товарищ капитан*

Как видим, в чрезвычайно сходных контекстах мы встречаемся с тремя возможностями — сочетание основных жестов (5), жест КГ (6) и отсутствие любого жеста (7).

Именно в таких случаях, когда не помогает лингвистическая интуиция, на помощь приходит статистика. При большом количестве примеров есть надежда вычленить те лингвистические характеристики контекстов, которые значимо влияют на выбор между четырьмя имеющимися возможностями жестикуляционного оформления отрицательных конструкций.

Приведем далее несколько таблиц, которые показывают распределение жестов между разными типами контекстов.

2.2.1. Коммуникативная дистанция

Табл. 1 показывает, как отрицательная жестикуляция распределяется в соответствии с такой характеристикой высказывания, как лицо предиката. Приведем несколько примеров для пояснения, что имеется в виду:

(8) **1 лицо**: *сюда я* ☹✂{не *пойду*}; **3 лицо**: *параллели и меридианы* ☹✂{тоже *не видны*}

[1] В этом примере мы сталкиваемся с ситуацией, хорошо известной в литературе по жестикуляционной лингвистике, когда при порождении речи ad hoc, т.е. неподготовленной речи, жест несколько опережает по времени семантическую зону высказывания, к которой он, собственно, относится (см. [Calbris 2011: 54–55, 59–62, 68–69]).

(9) **1 лицо**: *я уже* ☹{*давно это не слышу*}; **3 лицо**: *например* ☹{*фотоны сами с собой непосредственно не взаимодействуют*}

(10) **1 лицо**: *я не приношу новых слов*; **3 лицо**: *почему* [они] *не пускали калек и увечных?*

(11) **1 лицо**: [я] *места себе* ✂{*не находил*}; **3 лицо**: *они* ✂{*не слушают*}

Таблица 1.[1]

Лицо / **Отрицательный жест**	**1 лицо**	**3 лицо**
качать головой + отрезать рукой	8	21
качать головой	**58**	*59*
нет отрицательного жеста	*94*	229
отрезать рукой	*13*	44
χ^2=19,8, p=.0002, параметры связаны, распределения достоверны		

Из табл. 1 мы видим, что жест КГ 1) тяготеет к глаголам в форме первого лица или синтаксически связанным с субъектом в первом лице и 2) отталкивается от глаголов в форме третьего лица или связанных с субъектом в третьем лице. Отсутствие отрицательного жеста и жест ОР не характерны для первого лица. Таким образом, жест КГ противопоставлен паре ОР–отсутствие отрицательного жеста по параметру лица.

Как показывает анализ нашей базы данных, по отношению к себе говорящий гораздо чаще применяет глаголы «внутреннего состояния» — глаголы восприятия (*видеть, слышать*), ментальные (*думать, знать, понимать*), способности (*мочь, уметь*), эмоций (*любить, жалеть*),— чем по отношению к персонажам своей речи. И наоборот, по отношению к персонажам и предметам говорящий гораздо чаще употребляет глаголы существования (*быть, происходить, случаться*), чем по отношению к самому себе. Это хорошо видно из табл. 2.

[1] Далее в таблицах ячейки, где данные существенно отличаются от средних распределений в бóльшую сторону, оформляются полужирным шрифтом, в меньшую сторону — курсивом.

Таблица 2.

Лицо / Тип глагола	1 лицо	3 лицо
глаголы внутреннего состояния	**93**	*70*
глаголы существования	*15*	**101**
χ^2=55,6, p=8,8–14, параметры связаны, распределения достоверны		

Поскольку распределение типов предикатов достаточно сильно связано с лицом предиката, а лицо предиката связано с типом отрицательного жеста, то довольно естественно, что мы наблюдаем и связь типа предиката с типом отрицательного жеста, табл. 3.

Таблица 3.

Тип глагола / Тип глагола	Глаголы внутреннего состояния	Глаголы существования
качать головой	**60**	*17*
отрезать рукой	*13*	**18**
χ^2=13,06, p=.003, параметры связаны, распределения достоверны		

Примеры:

(12) *я* ☹{*не знаю*}, *какое у меня сегодня лицо*

(13) *без него магнитного поля* ✂{*не существовало бы*}

Интересно при этом, что одушевленность или неодушевленность субъекта не имеют никакой значимости при выборе сопровождающего отрицательную конструкцию жеста, что видно из табл. 4 (в таблице учитываются, естественно, данные только по третьему лицу, поскольку субъекты первого и второго лица по определению являются одушевленными).

Таблица 4.

Тип глагола / Тип глагола	неодушевленный	одушевленный
качать головой + отрезать рукой	8	9
качать головой	40	23
нет отрицательного жеста	124	105
отрезать рукой	25	19
χ^2=2,3, p=.5, параметры не связаны, распределения недостоверны		

Итак, можно видеть, что первым фактором, влияющим на выбор отрицательного жеста, является фактор коммуникативной дистанции между говорящим и субъектом высказывания. В случае, если эта дистанция нулевая, т. е. говорящий и субъект совпадают, отрицательная конструкция чаще всего сопровождается жестом КГ и не сопровождается жестом ОР или нулевым (отсутствующим) отрицательным жестом. Напротив, когда коммуникативная дистанция между говорящим и субъектом отрицаемого предиката максимальна, как в ситуации третьего лица, говорящий с пониженной частотой использует жест КГ.

2.2.2. Эпистемическая дистанция

Анализ материала показал, что работающим фактором, воздействующим на отрицательную жестикуляцию, является противопоставление высказываний **со снятой утвердительностью** (или, что то же, высказываний с нейтральной модальностью, см. [Падучева 2013: 32] и соответствующие ссылки[1]) и высказываний с отрицаемым предикатом в настоящем времени, которое предполагает синхронность речи и событийного ряда.

Отрицаниями со снятой утвердительностью являются отрицания, предикаты в которых, среди прочего, имеют форму сослагательного наклонения, являются компонентом чужой речи, передаваемой говорящим, формируют контексты неуверенности, предположительности, а также отрицания, где использованы вводные слова, дизъюнктивные конструкции, подчиненные предложения после условных и целевых союзов и др.[2] Для высказываний со сня-

[1] Понятие «снятая утвердительность» несколько парадоксально использовать по отношению к отрицаниям, поскольку отрицание и является первым кандидатом на снятую утвердительность. Материал, однако, показывает, что «типы неассертивности бывают разные» (Е.В. Падучева, устное сообщение), и внутри «широкой» снятой утвердительности, заданной отрицанием, существуют более узкие ее варианты, которые различают, в частности, степени уверенности говорящего в его собственном отрицании (см. также ниже, примечание на с. 568).

[2] О том, что высказывания со снятой утвердительностью существенно влияют на жестикуляцию, в частности, для них характерно движение руки или головы жестикулирующего субъекта вправо (в противоположность прошедшему времени, для которого характерна ориентация влево), мы уже писали в работе [Гришина 2013], не используя, впрочем, там данный термин, а пользуясь термином *ирреальное наклонение.*

той утвердительностью характерен уход говорящего от ответственности за содержание высказывания, в то время как для контекстов с настоящим временем предиката, не включающих маркеров снятой утвердительности, напротив, характерна полная ответственность говорящего за высказывание, поскольку в них говорящий одновременно является непосредственным наблюдателем отражаемого в предикации события и, соответственно, не может переложить ответственность за ложность или истинность предикации на постороннего наблюдателя.

Несколько примеров.

(14) **Снятая утвердительность**: *А поручителя как такового* ☹✂{*не найти*}; **Настоящее время**: *в режиме пантомимы указания открытой ладонью* ☹✂{не *используются*}

(15) **Снятая утвердительность**: *Мы* ***можем*** *создать условия, которых* ☹{*вообще не существует*} *в природе*; **Настоящее время**: *у нас здесь дороги* ☹{*не чистятся*}

(16) **Снятая утвердительность**: *как можно организовать,* ***чтобы*** *естественный отбор не выбраковал*; **Настоящее время**: *такие задачи не решаются в принципе*

(17) **Снятая утвердительность**: *мне* ***кажется****, что булгаковеды еще вплотную* ✂{*не подошли*} *к трактовке*; **Настоящее время**: ✂{*не хочет*} *он гореть в геенне огненной*

Приведем соответствующие распределения в табл. 5.

Таблица 5.

Модальность/Время Тип жеста	Снятая утвердительность	Настоящее время
качать головой + отрезать рукой	10	9
качать головой	*33*	**60**
нет отрицательного жеста	132	117
отрезать рукой	**27**	*16*
χ^2=11,6, p=.009, параметры связаны, распределения достоверны		

Как видим, отрицательные конструкции в контекстах со снятой утвердительностью и в контекстах с настоящим временем предиката в отношении сопровождающих отрицательных жестов различаются достаточно отчетливо: жест КГ характерен для настоящего времени, а жест ОР — для снятой утвердительности.

Примечание. Здесь нужно сделать одно очень важное отступление. В работе [Гришина 2014а] мы привели данные, которые описывали распределение русских указательных и иконических жестов, выровненных с глаголами, в пространстве по сагиттальной (вперед–назад), поперечной (вправо–влево) и вертикальной (вверх–вниз) осям[1]. В рассмотрение было включено существенное количество контекстов (порядка 3 тыс. жестоупотреблений), бóльшая часть из которых (порядка 95%) содержала положительные, а не отрицательные формы предикатов, так что наличие отрицательных конструкций практически не влияло на статистические распределения.

Табл. 6 содержит выдержки из упомянутой таблицы в [Гришина 2014а], и мы можем видеть, как именно положительные предикаты 1) в конструкциях со снятой утвердительностью, 2) в настоящем времени (без учета настоящего исторического, которое ведет себя как прошедшее время) и 3) в прошедшем времени распределены по поперечной оси.

Таблица 6.

Модальность/Время Направление	**Снятая утвердительность**	**Настоящее время**	**Прошедшее время**
налево	*148*	57	**160**
направо	**261**	64	*112*
χ^2=33,9, p=4,34–08, параметры связаны, распределения достоверны			

Как видим, в положительных контекстах наблюдается отчетливая асимметрия: жестикуляция при контекстах со снятой утвердительностью тяготеет вправо, при контекстах с прошедшим временем — влево, для настоящего времени данный параметр не существен. В работе [Гришина 2013] эта закономерность была истолкована следующим образом: направо от говорящего расположена зона доминантной (правой) руки, которая ассоциируется говорящим с ним самим. Этот факт в сочетании с «эпистемической ущербностью» контекстов со снятой утвердительностью, связанной с нежеланием говорящего нести полную ответственность за истинность собственного высказывания и с желанием дистанцироваться от содержания

[1] Приведенная в цитируемой работе таблица является расширенным вариантом аналогичной таблицы, опубликованной нами в статье [Гришина 2013], где рассматривались только указательные жесты.

высказывания, позволил нам интерпретировать данные результаты как связь правой зоны жестикуляционного пространства с **мнением** говорящего. Что касается предикатов в прошедшем времени, то они излагают информацию как уже полностью состоявшуюся, как **факт**, известный говорящему и сообщаемый им адресату. Для контекстов в настоящем времени противопоставление мнения и факта оказывается незначимым — настоящее узуальное вообще не имеет привязки к декартовым координатам, поскольку называет ситуацию без указания конкретного момента ее осуществления, а настоящее узуально-длительное располагается по вертикальной оси, ориентировано сверху вниз и, тем самым, привязано к той точке в пространстве, которое занимает говорящий в момент речи.

Теперь мы можем отобрать из нашей базы данных отрицательные конструкции, которые сопровождаются отрицательным жестом ОР, и посмотреть, как именно происходит распределение в пространстве жестикулирующей руки при осуществлении этого отрицательного жеста — в соответствии с теми же модальностями/временами, но при наличии отрицания (разумеется, контекстов существенно меньше, но основную тенденцию проследить удается, см. табл. 7).

Таблица 7.

Модальность/Время Направление	**Снятая утвердительность**	**Настоящее время**	**Прошедшее время**
налево	5	2	4
направо	20	13	14
χ^2=0,45, p=.8, параметры не связаны, распределения недостоверны			

Как видим, контексты со снятой утвердительностью и контексты в прошедшем времени теряют свою привязку к правой и левой зоне жестикуляционного пространства: этот параметр оказывается для них столь же несущественным, как и для настоящего времени: все они в равной степени тяготеют к правой зоне, характерной для контекстов со снятой утвердительностью. И в этом нет ничего удивительного, поскольку конструкция отрицания является одним из способов лишить утверждение его «утверждающей» природы и одним из операторов, который переводит высказывания в нейтральную модальность, т. е. снимает утвердительность (см. [Падучева 2013: 32)]). Таким образом, в случае отрицания предиката в контексте сня-

той утвердительности говорящий постулирует отсутствие события, реальность которого **и в положительном контексте** была под сомнением: говорящий занимал по отношению к событию существенную эпистемическую дистанцию, не будучи уверенным в утверждаемом; в случае прошедшего времени отрицательная конструкция переводит сообщение о некотором факте в разряд сообщений **об отсутствии этого факта**, что, очевидно, уже является не фактом, а таким же мнением говорящего, как и прочие конструкции со снятой утвердительностью; тот же самый процесс имеет место в отрицательных конструкциях в настоящем времени.

Теперь можно вернуться еще раз к табл. 5. В ней мы видим, что два **в равной степени** ирреальных контекста (отрицательная конструкция в настоящем времени и отрицательная конструкция в нейтральной модальности) сопровождаются **разными** отрицательными жестами: настоящее время — жестом КГ, контексты со снятой утвердительностью — жестом ОР. Выбор разных отрицательных жестов заставляет нас предполагать имеющиеся различия и в самих конструкциях.

Мы считаем, что различие между двумя типами контекстов связано с понятием **наблюдаемого отсутствия**, которое было подробно разобрано Е. В. Падучевой [Падучева 2013: 104–114] в связи с анализом употребления генитива в отрицательных конструкциях. Воспользовавшись предложенным направлением рассуждений и в высшей степени удачным термином, приложим их к нашей, весьма отличной от генитивных конструкций ситуации. Действительно, положительные контексты со снятой утвердительностью подают некоторую предикацию как то ли имевшую, то ли не имевшую места — говорящий не берет на себя полноту ответственности за выбор между этими возможностями. Привнесение в этот тип контекстов отрицания, в принципе, ничего не меняет — говорящий, опять же, не берет на себя ответственности за выбор между *истиной* и *ложью*. Напротив, положительные контексты в настоящем времени предполагают, что говорящий без обиняков утверждает прохождение некоторого события перед его глазами, т. е. утверждает, строго говоря, 1) событие имеет место, 2) я не только информирую об этом событии, но и подтверждаю его достоверность, поскольку воспринимаю его собственными органами чувств. Внесение в данную конструкцию отрицания отрицает, с нашей точки зрения, только первую позицию, т. е. событие

не имеет места, при этом второе положение под отрицание не попадает, и говорящий, тем самым, остается информатором-наблюдателем **не происходящего** в данный момент события.

Подводя итог — мы полагаем, что отрицательные конструкции со снятой утвердительностью эпистемически находятся на более далеком расстоянии от говорящего, чем отрицательные конструкции в настоящем времени.

2.2.3. Аксиологическая дистанция

Последний параметр, отмеченный нами как существенный для выбора отрицательного жеста, сопровождающего отрицательную конструкцию,— это использование разного рода интенсификаторов при отрицательных конструкциях (*абсолютно, в принципе, вообще, до конца, конечно, отнюдь, очень, резко, совершенно, совсем*). Несколько примеров:

(18) *оно* ***абсолютно*** *не нужно человеку*

(19) *такие задачи не решаются* ***в принципе***

(20) *мы можем создавать условия, которых* ☹{***вообще*** *не существует в природе*}

(21) *ведь общая теория относительности, она эту программу не довела* ***до конца***

(22) *хотя,* ***конечно****,* ☹{*в казино не хожу*}

(23) *это люди* ☹{***очень*** *не любят*}*, кстати*

(24) *вот тот идеальный эквивалент* ☹✂{***совершенно*** *не влезает*}

(25) *у меня дело, но я* ***совсем*** *не знаю, как войти*

Примеров, к сожалению, не очень много (их, по-видимому, следует собирать целенаправленно, ориентируясь на контрольные лексемы-интенсификаторы), но закономерность, тем не менее, просматривается достаточно хорошо (см. табл. 8).

Таблица 8.

Интенсификатор / **Тип жеста**	**Есть**	**Нет**
качать головой + отрезать рукой	1	32
качать головой	**10**	118
нет отрицательного жеста	*9*	373
отрезать рукой	3	62
χ^2=8,0, p=.04, параметры связаны, распределения достоверны		

Как видим, для отрицательного жеста, включающего в себя компонент ОР (чистый ОР или ОР в сочетании с КГ) данный параметр не значим; отрицательная конструкция с интенсификатором тяготеет к жесту КГ; для отрицательной конструкции с интенсификатором не характерно отсутствие жеста.

Еще более отчетливо эта закономерность будет видна, если мы возьмем для сравнения контексты без интенсификатора с предикатами в третьем лице, где предмет высказывания находится на максимальной коммуникативной дистанции от говорящего (см. п. 2.2.1 и табл. 9).

Таблица 9.

Интенсификатор / **Тип жеста**	**Есть**	**Нет**
качать головой + отрезать рукой	1	20
качать головой	**10**	50
нет отрицательного жеста	*9*	224
отрезать рукой	3	44
χ^2=13,2, p=.004, параметры связаны, распределения достоверны		

Приведенные данные выглядят в высшей степени согласованными с предшествующим изложением, поскольку наличие интенсификатора при отрицательной конструкции вносит в предикацию элемент оценки со стороны говорящего: таким образом говорящий не просто фиксирует ненаступление какого-либо события, но и оценивает степень этого ненаступления (*абсолютно, совершенно, очень* и под.), а также фиксирует уровень собственной уверенности в неизбежности ненаступления события (*конечно*).

И снова, как и в предыдущих случаях, минимальная дистанция между говорящим и его высказыванием, в данном случае аксиологическая, сопровождается жестом КГ, в то время как для безжестовых отрицательных конструкций наличие интенсификатора не характерно.

В заключение раздела следует уточнить, что три перечисленных параметра являются взаимно независимыми, что показывает сводная табл. 10.

Таблица 10.

Интенсификатор **Лицо**	**Есть**	**Нет**
1 лицо	5	167
3 лицо	15	344
χ^2=0,52, p=.47, параметры не связаны, распределения недостоверны		
Интенсификатор **Модальность/Время**	**Есть**	**Нет**
снятая утвердительность	4	201
настоящее время	11	95
χ^2= 3,35, p=.07, параметры не связаны, распределения недостоверны		
Лицо **Модальность/Время**	**1 лицо**	**3 лицо**
снятая утвердительность	51	130
настоящее время	70	129
χ^2=2,1, p=.14, параметры не связаны, распределения недостоверны		

Таким образом, как взаимно независимые мы можем расценивать и распределения жестов в соответствии с тремя параметрами – коммуникативной, эпистемической и аксиологической дистанцией между говорящим и его высказыванием.

2.3. ПРОМЕЖУТОЧНЫЕ ИТОГИ

Первая закономерность, которая отчетливо проявилась при анализе материала, заключается в том, что не удалось обнаружить никакого лингвистического параметра, который бы позволил противопоставить отсутствие отрицательного жеста при отрицатель-

ных конструкциях — и использование ручного жеста *отрезать рукой/руками* (ОР). Таким образом, отсутствие отрицательного жеста и жест ОР выступают в отрицательных конструкциях как единое жестикуляционное событие, противопоставленное в таком качестве жесту *качать головой* (КГ). Это, на наш взгляд, объясняет, почему при сочетании в высказывании двух основных отрицательных жестов — КГ и ОР — они не нейтрализуют друг друга: они относятся к разным зонам передачи смысла и при сочетании не отменяют, а дополняют друг друга.

Вторая закономерность — жест КГ используется говорящим в тех случаях, когда говорящий в максимальной степени несет ответственность за свое высказывание, т. е. в случаях, когда мы имеем дело с нулевой коммуникативной дистанцией (использование первого лица предиката), нулевой эпистемической дистанцией (использование настоящего времени) и нулевой аксиологической дистанцией (использование интенсификаторов). Напротив, отсутствие жестикуляционного сопровождения и жест ОР используются говорящим в случае максимальной дистанции между ним и его высказыванием (третье лицо, высказывания со снятой утвердительностью, отсутствие интенсификаторов).

Таким образом, примеры типа (26) и (27)– (28) находятся на противоположных концах предложенной шкалы:

(26) *хотя,* ***конечно****,* ☹{*в казино не хожу*}

vs.

(27) *почему вот зеленых млекопитающих не бывает, насколько я знаю?*

(28) *там нет соотнесенности, и* ✂{*не может быть в общем-то*} *соотнесенности*.

Чем можно объяснить именно такое распределение жестов между минимально и максимально дистанцированными от говорящего высказываниями?

Обратимся к противопоставлению, предложенному Д. Макниллом в [McNeill 1992], между жестами, отражающими точку зрения наблюдателя (OVPT, observer viewpoint gestures), и жестами, отражающими точку зрения персонажа речи (CVPT, character viewpoint gestures). Различие между ними заключается в следующем. Исполняя жесты, отражающие точку зрения персонажа, говорящий как бы мысленно совмещает свое тело с телом персонажа своей речи и же-

стикулирует так, как будто и является этим персонажем. Так, например, при CVPT-жесте, сопровождающем глагол *лететь*, говорящий может изображать руками крылья (птицы или самолета). Напротив, при исполнении жестов с точки зрения наблюдателя говорящий сохраняет дистанцию между своим телом и телом персонажа, между своим местоположением и местоположением персонажа, т. е. между своей точкой зрения и точкой зрения персонажа. В этом случае говорящий не вовлекает в жестикуляцию свое тело, а исполняет условный обобщенный жест рукой для схематичной передачи того или иного предиката: так, глагол *лететь* в OVPT-режиме будет, скорее всего, исполнен либо подъемом руки вверх, либо движением руки/головы в сагиттальном, поперечном и вертикальном направлении для передачи временных и модальных характеристик высказывания с данным предикатом.

Говоря в общем, при исполнении жестов наблюдателя сохраняется дистанция между говорящим и персонажем, а при исполнении жестов, отражающих точку зрения персонажа, данная дистанция исчезает (или сводится к минимуму).

Посмотрим на полученные результаты по отрицательной жестикуляции с этой точки зрения. *Качание головой* выступает как жест, отражающий точку зрения персонажа, поскольку в этом случае дистанция между говорящим и персонажем речи сокращается до нуля. В случае, если говорящий и персонаж являются одним и тем же лицом, это проявляется особенно отчетливо,— и частота использования жеста КГ в высказываниях с нулевыми дистанциями значимо возрастает. Напротив, отсутствие отрицательной жестикуляции и жест *отрезать рукой* функционируют как жесты (отсутствие жестов), отражающие точку зрения наблюдателя, когда говорящий, который в данном случае и выступает в качестве наблюдателя, дистанцирует себя от персонажей и содержания своей речи.

Если это так, то становится понятным, почему для нулевой дистанции был выбран жест КГ, а для ненулевой, большой дистанции (коммуникативной или эпистемической или аксиологической) были выбраны либо ручной жест ОР, либо отсутствие жеста. Жест *качать головой* может рассматриваться как жест, осуществляемый телом, а не рукой, а следовательно, он в наибольшей степени подходит для ситуации CPTV, т. е. для совпадения, так сказать, в одном теле, в одной пространственной зоне тела говорящего и тела персонажа.

Напротив, отсутствие отрицательного жеста сохраняет дистанцию между наблюдателем и персонажем, а использование в отрицательной жестикуляции руки/рук для осуществления жеста *отрезать рукой* — максимально удлиняет эту дистанцию, физически отчуждая содержание высказывания от говорящего, а именно — отодвигая жестикуляцию от тела говорящего на расстояние, задаваемое длиной руки. Таким образом, отсутствие жеста при отрицательных конструкциях может быть истолковано как характеристика высказывания как «не вполне своего», а использование жеста ОР — как характеристика высказывания как «чужого». При этом использование жеста КГ маркирует отрицательную конструкцию как «свою» для говорящего.

Примечание. В жестикуляционной литературе широко распространено понимание, что жестикуляция говорящим контролируется в существенно меньшей степени, чем речепорождение. Важным следствием из этого психолингвистического факта является возможность для жестикуляции сохранять свои семантические свойства в любых речевых контекстах, а следовательно, жестикуляция может корректировать содержание высказывания с семантической (в широком смысле слова) точки зрения. Как следствие, если нам удается правильно определить семантическую палитру того или иного жеста или семейства жестов, то мы можем использовать жестикуляцию как своеобразный лингвистический маркер того, что на самом деле хотел сказать говорящий, для чего у него не хватило слов или конструкций и что он добавил в высказывание с помощью жестов. Как это происходит?

Прежде всего, в тех случаях, когда жестикуляционная структура фразы противоречит ее лингвистической структуре, логично, исходя из сказанного, отдавать приоритет именно жестикуляционной структуре. В частности, интересен вопрос с будущим временем, которое может трактоваться и как вариант нейтральной модальности (снятой утвердительности), и как вариант уверенного утверждения. Выбор говорящим жестикуляционного сопровождения отрицательной конструкции в будущем времени является аргументом для одного из двух прочтений. Например, во фразе

(29) *Мы сейчас не будем обсуждать структуру, иначе б мы ушли в обсуждение «Апокалипсиса». Сюда я* ☹✂{не *пойду*}.

отрицательная конструкция сопровождается двойным отрицательным жестом — КГ и ОР. И если жест КГ хорошо коррелирует с пер-

вым лицом предиката, то жест ОР, сопровождающий синтагму в будущем времени *не пойду*, означает, если принять изложенное выше, что говорящий данной синтагмой выражает не твердое намерение не двигаться куда-либо в дальнейшей беседе, а утверждает, что он **не хотел бы** продолжать разговор на эту тему, т. е. на самом деле мы имеем здесь — с жестикуляционной точки зрения — высказывание со снятой утвердительностью.

Аналогичные проблемы возникают с трактовкой фраз, имеющих в качестве вводной части утверждение *я уверен*:

(30) *Но я уверен, что до этого* ☹{*не дойдет*}.

С точки зрения модальности такие фразы двусмысленны. С одной стороны, главное предложение передает уверенность говорящего в утверждении, которое вводится союзом *что*. С другой стороны, предложение *я уверен* не является прозрачным для отрицания, поскольку при добавлении отрицания последнее не затрагивает придаточного предложения:

(30 а) *Но я не уверен, что до этого не дойдет*.

Тем самым, фраза *до этого не дойдет* является фразой со снятой утвердительностью, т. е. говорящий в ней дистанцируется от утверждаемого содержания, а следовательно, не является в нем полностью уверенным. Жестикуляция, а именно жест КГ, который сопровождает отрицательную конструкцию в примере (30), показывает, что говорящий уверен в собственном высказывании, невзирая на то, что лингвистически мы имеем дело со снятой утвердительностью.

Аналогичным образом разрешается неоднозначность в безличных конструкциях.

(31) *И сбежал бы, а уже* ☹{*не можешь*} *без дозы*.

Во фразе (31) конструкция с обобщенно-личной формой глагола во 2 л. ед. ч. может трактоваться как сентенция (имеющая отношение к любому человеку), но с равной степенью вероятности может относиться и лично к говорящему. Употребление жеста КГ, который сокращает коммуникативную дистанцию между говорящим и содержанием высказывания, позволяет считать, что в этом случае мы имеем дело с предикатом в финитной форме (‘я бы сбежал, но уже не могу без дозы’).

А в предложениях (32)– (33)

(32) *Покуда пороховые погреба не возьмем, шведов нам* ✂{*не одолеть*}.

(33) *если мы искусственно <…> осушим,* ✂{*не будет у нас*} *океанов, то все равно останутся два максимума.*

отрицательные конструкции могут трактоваться семантически (как финитные конструкции с субъектом в первом лице *нам не одолеть = мы не одолеем, не будет у нас = мы не будем иметь*), но с равной вероятностью могут трактоваться и синтаксически, т. е. в полном соответствии с имеющейся безличной конструкцией, где подлежащего нет, а местоимение первого лица выступает как дополнение (в примере (33) при этом на место семантического подлежащего претендует в этом случае компонент *океаны*). Жест ОР, который сопровождает эти конструкции «в одиночестве», без жеста КГ, показывает, что для говорящего верна вторая трактовка.

2.4. ПРИМЕЧАНИЕ: ИНЫЕ ТИПЫ ОТРИЦАТЕЛЬНЫХ КОНСТРУКЦИЙ

В этом разделе будет сказано несколько слов о других отрицательных конструкциях, отличающихся от стандартных отрицательных конструкций *не* + предикат. Это противопоставительное отрицание, т. е. отрицание, выраженное конструкциями *не…, а…*; *не только…, но и…* (см. [Падучева 2013: 249–250], там же ссылка на [Богуславский 1985: 129–137]), предикативные конструкции с *не*-местоимениями (*нечего, некому, негде* и проч.) и конструкции с двойным отрицанием.

Отметим, прежде всего, что по сравнению со стандартной отрицательной конструкцией в жестикуляционном отношении перечисленные конструкции имеют ряд отклонений, которые можно увидеть из табл. 11.

Таблица 11

Тип конструкции **Тип жеста**	**противопоставительное отрицание**	**стандартное отрицание**	**двойное отрицание**	**предикативное отрицание**
КГ (изолированно или в сочетании с ОР)	*4*	161	**53**	*9*
ОР (изолированно или в сочетании с КГ)	3	95	21	*1*
нет жеста	**21**	382	*83*	**41**
χ^2=18,5, p=.005, параметры связаны, распределения достоверны				

Из табл. 11 можно видеть, что для противопоставительного и предикативного отрицания характерно безжестовое употребление и не характерно использование любой отрицательной жестикуляции. Для конструкций с двойным отрицанием характерно использование жестов головы и не характерно безжестовое функционирование.

Одинаковое поведение противопоставительного и предикативного отрицания в отношении плотности и качества сопровождающей жестикуляции можно интерпретировать следующим образом — обе эти конструкции не являются отрицанием в классическом смысле. Противопоставительное отрицание является в большей степени противопоставлением, чем отрицанием (что сказывается и на его лингвистических особенностях, см. ссылку выше); предикативное отрицание в большинстве случаев отрицает не предикат, а его актанты, поэтому и ведет себя иначе с жестикуляционной точки зрения, чем классическая отрицательная конструкция.

Что касается конструкции с двойным отрицанием, то здесь ситуация более сложная. Во-первых, если мы выведем из статистического анализа противопоставительное и предикатное отрицание, то увидим, что распределения жестикуляции в случае классической отрицательной конструкции и в случае двойного отрицания практически не различаются (см. табл. 12).

Таблица 12

Тип отрицания **Модальность/Время**	**Отрицательная конструкция**	**Конструкция с двойным отрицанием**
КГ (изолированно или в сочетании с ОР)	161	53
ОР (изолированно или в сочетании с КГ)	95	21
нет жеста	382	83
χ^2=4,65, p=.09, параметры не связаны, распределения недостоверны		

Во-вторых, выясняется, что все параметры, которые были значимы для жестикуляционного сопровождения отрицательной конструкции, не работают в случае конструкции с двойным отрицанием,— ни коммуникативная, ни эпистемическая дистанция

(см. табл. 13; использование интенсификаторов мы не анализировали, поскольку их количество в конструкциях с двойным отрицанием весьма невелико, и это, по-видимому, не случайно, см. ниже).

Таблица 13

Лицо / **Лицо**	**1 лицо**	**3 лицо**
качать головой+отрезать рукой	5	167
качать головой	15	344
нет жеста	18	69
отрезать рукой	0	14
χ^2=3,7, p=.3, параметры не связаны, распределения недостоверны		
Модальность/Время / **Тип жеста**	**Снятая утвердительность**	**Настоящее время**
качать головой+отрезать рукой	2	2
качать головой	15	21
нет жеста	20	36
отрезать рукой	1	9
χ^2=3,8, p=.29, параметры не связаны, распределения недостоверны		

Более подробный анализ материала показал, что на жестикуляционное сопровождение конструкций с двойным отрицанием влияет фактор совершенно иного плана — наличие и место расположения эмфазы относительно отрицания (см. табл. 14).

Примеры (выделенная с помощью эмфазы единица оформлена КАПИТЕЛЬЮ):

(34) **Эмфаза на местоимении**: *и больше ☹{НИКОГДА не возвращается}; поэтому ✂{не было НИКАКИХ} так сказать демонстраций*

(35) **Эмфаза на отрицательной конструкции, на иных членах предложения и фразы с отсутствием специальной эмфазы**: *я принял вид, что как бы ☹{и не НАХОДИЛ} ничего; никуда он не уедет, пока не убьет меня; приписывали слова, что атомная энергия никогда не будет полезна человеку.*

Таблица 14

Наличие жеста **Наличие и место эмфазы**	**Отрицательный жест КГ**	**Отрицательного жеста нет**
Эмфаза на местоимении	**22**	*13*
Эмфаза на глаголе + отсутствие эмфазы	*27*	**69**
χ^2=9,3, p=.002, параметры связаны, распределения достоверны		

Как видим, жестикуляция распределена крестообразно — для контекстов с эмфазой на местоимении характерен отрицательный жест *качать головой*, для остальных контекстов — отсутствие отрицательного жеста.

Сочетание трех фактов:

— нехарактерность для конструкций с двойным отрицанием жеста *отрезать рукой* и безжестового употребления (табл. 11);

— независимость конструкций с двойным отрицанием от параметра лица (коммуникативная дистанция) и параметра модальность/время (эпистемическая дистанция; табл. 13);

— связь жеста *качать головой* с эмфатически выделенным отрицательным местоимением (табл. 14)

позволяет предположить, что в конструкциях с двойным отрицанием отрицательное местоимение ведет себя как интенсификатор (в нормальном случае в стандартных отрицательных конструкциях именно на интенсификатор падает фразовая эмфаза, например *он ВООБЩЕ не умеет читать*), и именно наличие такого интенсификатора является определяющим для появления отрицательного жеста и для его качества (*качать головой*, как и в случае со стандартными интенсификаторами при обычных отрицательных конструкциях). Именно этим, с нашей точки зрения, объясняется сравнительная редкость использования обычных интенсификаторов в конструкциях с двойным отрицанием (например, *никто ВООБЩЕ не умеет читать*).

Если наше объяснение верно, то можно утверждать, что в конструкциях с двойным отрицанием с эмфазой на местоимении жестикуляционное сопровождение определяется нулевой аксиологической дистанцией между говорящим и его высказыванием, т. е. отрицательная конструкция такой структуры является — с жестикуляционной точки зрения — оценочным высказыванием.

3. Жестикуляционное сопровождение конструкций с глубинным отрицанием

3.1. ТИПЫ КОНТЕКСТОВ С ГЛУБИННЫМ ОТРИЦАНИЕМ

О том, что отрицательные жесты часто сопровождают высказывания, в которых нет отрицания, сообщали разные авторы. Так, в частности, А. Кендон в работах [Kendon 2002: 180], [Kendon 2004: 255–264] писал о том, что отрицательный жест может использоваться как дополнительное сообщение к основному сообщению, выраженному говорящим с помощью слов, например, тогда, когда говорящий с помощью жестикуляции выражает свое отношение к содержанию своего собственного высказывания, — негативное отношение к ситуации, которую он описывает, или сомнение в том, что он утверждает. Кроме того, отрицательные жесты могут проявлять подразумеваемое отрицание, которое не выражается на поверхностном уровне в высказывании, и тогда отрицание выражается только жестикуляционным способом, а не с помощью слов. Такая ситуация имеет место, например, в «высказываниях следующей структуры: ‘Я могу сделать X, но я сделаю Y’, где ‘несовершение X’ является импликацией, подразумеваемой говорящим; аналогичным образом, использование лексемы *единственный*, например, во фразе *Это единственная вещь, которую ты можешь сделать*, подразумевает невозможность иных линий поведения» [Kendon 2002: 180].

Отрицательные жесты могут быть использованы в контекстах, где постулируется тотальность некоторой предикации и отсутствие исключений; при использовании превосходной степени и утверждерний с интенсификатором — здесь также подразумевается тотальность и отсутствие исключений. Отрицательные жесты используются в случаях автокоррекции, когда говорящий посредством данного жеста корректирует и комментирует собственное высказывание.

Таким образом, тот факт, что отрицательный жест широко используется для сопровождения «положительных» высказываний, хорошо известен. У нас остаются, собственно, всего два вопроса, на которые мы планируем ответить в в данном разделе:

1) есть ли различия в употреблении жестов *качать головой* и *отрезать рукой* в конструкциях с глубинным отрицанием? и если есть, то с чем они могут быть связаны?

2) все ли возможные ситуации использования отрицательных жестов в конструкциях с глубинным отрицанием учтены на данный момент в работах по жестикуляционной лингвистике?

Начнем со второго вопроса. В работе [Гришина 2012], посвященной конфигурации указующей руки в русском автодейксисе (автоуказании говорящего на самого себя), мы провели различие между автоуказанием, связанным с содержанием высказывания (в высказывании так или иначе должна присутствовать сема ‘я’, а автоуказание производится указательным пальцем), и автоуказанием, связанным со скрытым перформативом [Ross 1970], который лежит в основе любого высказывания (‘обрати на меня внимание: я говорю тебе, я прошу тебя, я утверждаю для тебя, я обращаюсь к тебе’ и проч.) и осуществляется открытой ладонью. Таким образом, мы показали, что жестикуляция может относиться к двум пластам высказывания — собственно семантической, связанной лишь с содержанием высказывания, и с прагматической, которая включает, среди прочего, тип речевого акта, осуществляемый говорящим, его коммуникативные отношения со слушающим, отношение говорящего к своему собственному высказыванию и др.[1]

Представляется, что различение этих двух планов в любом устном высказывании может оказаться весьма полезным и при анализе отрицательных жестов, сопровождающих высказывания с глубинным отрицанием.

3.1.1. Тематическое глубинное отрицание

Тематическим глубинным отрицанием (ТГО) мы далее будем называть те случаи глубинного отрицания, которые не имеют никакого очевидного отношения к прагматическим характеристикам высказывания.

[1] Это согласуется с различением, проведенным А. Кендоном в книге [Kendon 2004], между двумя основными функциями, которые выполняются речевыми (co-speech) жестами, а именно: между сущностными (substantial) и прагматическими жестами. Первые добавляют нечто в содержание высказывания, в то время как вторые передают те или иные аспекты ситуации, в которую включено высказывание. Аналогичным образом, в работе [McClave 2000] показано, что, в отличие от ручных жестов, жесты головы в основном используются для передачи прагматических, а не чисто семантических значений. В работе [Goodwin 1981] утверждается, что ручные жесты оказываются более подходящими для выражения содержания пропозиции, чем жесты головы, поскольку последние менее свободны в своих чисто биомеханических свойствах.

ТГО являются лексемы (и конструкции) с **внутрисловным отрицанием** ([Падучева 2013: 226–234]), в толкование которых с необходимостью включен семантический компонент ‘отрицание’. К таковым на нашем материале можно отнести лексемы *анти-* (приставка), *блокировать, всё!* ‘больше не надо’, *выбросить, выносить за скобки, довольно!* ‘больше не надо’, *другой* ‘не этот’, *забыть, закончить, запрещать, зачеркнуть, исказить, исключить, исчезнуть, лишиться, надоесть, нейтрализовать, оппортунистический* ‘не имеющий строгих принципов’, *ориентировочный* ‘неточный’, *отдельно* ‘не смешивая с другим’, *отказаться, отложить* ‘не делать в настоящий момент’, *отмести, отрицать, отсекать, отсеять* ‘выбрать для того, чтобы выбросить’, *отсутствовать, ошибиться, перестать, потерять, прекратить, пресекать, пустой* ‘не имеющий чего-л.’, *разобрать* (самолет)*, сделать секретным* ‘недоступным для чужих’, *удалить, утратить, уходить, хватит!* ‘больше не надо’.

Сюда же примыкают конструкции с *вряд ли*, с предлогом *без*, идиома *пара пустяков* ‘не требует усилий’.

Крупными подгруппами в составе группы лексем с внутрисловным отрицанием являются конструкции с лексемой *только*, которая включает в толкование компонент ‘не существует’.

Аналогичным образом, поскольку в их толкование включен компонент ‘не существует’, как лексемы с внутрисловным отрицанием мы рассматриваем все контексты, где присутствует идея тотальности, отсутствия исключений. Естественно, главными в этой группе являются лексемы *весь, всё, всякий, всегда, каждый* и *любой*. В эту же группу входят интенсификаторы, которые описывают ***предельную*** степень проявленности признака (*совершенно, постоянно, абсолютно, стопроцентно, полностью, исключительно, чисто*).

В эту же группу входят и конструкции с превосходной степенью.

3.1.2. Прагматические глубинные отрицания

Прагматические глубинные отрицания (ПГО) представляют собой неоднородную группу, но при этом общим свойством всех контекстов с ПГО является повышенное участие в высказывании говорящего и необходимость — для понимания высказывания — обращаться к структуре коммуникативной ситуации, к отношению говорящего к слушающему и к его собственному высказыванию. Ниже мы перечислим основные группы контекстов с ПГО.

3.1.2.1. Автокоррекция: отношение говорящего к форме своего высказывания

В случае автокоррекции говорящий либо меняет общую стратегию построения своего высказывания, либо, не меняя стратегии, делает поправку, выбирая например, другую лексему (подробно об автокоррекции см. [Подлесская 2014]). Тем самым говорящий, не будучи удовлетворен первоначальным планом построения своего высказывания, отказывается от него и прибегает к новому плану. Вот этот момент **отказа** говорящего от первого, неудачного плана и маркируется с помощью отрицательного жеста (при этом на поверхностном уровне, в смысловом строении фразы, никакого отрицания, ни в виде явного отрицательного маркера, ни в виде ТГО может не быть).

Примеры:

он открыл, что существует только двести… ☹{*существует*} *двести тридцать и только двести тридцать способов идеального заполнения пространства*; *жуткий рассказ Саламова* ☹{[пауза]} *Шаламова*; *и вот наступает момент, когда он…* ☹{*а, да, так был задуман*} *роман*; *Толстой описывал ощущения свои…* ☹{*ощущения*}*… представления Пьера об Элен*; *а официально* ☹{[пауза]} *неофициально, то есть реально, он родился…*; *двумя сторонами…* ☹{[пауза]} *двумя способами, с двух сторон*; *еще очень много…* ☹{[пауза]} *еще очень долго времени пройдет, пока…*; *ну вот значит что…* ☹{*структура словарной статьи*} *здесь показана*; *чем че…* ☹{*четыре*} *миллиарда; исчезающее много…* ☹{*то есть мало*}; *но он мог…* ☹{*у него было два пути тогда*}

3.1.2.2. Дистанцирование и несогласие

Достаточно большой пласт контекстов с ПГО связан с выражением разных степеней и форм несогласия говорящего. И именно прагматический компонент несогласия определяет появление отрицательных жестов — при полном отсутствии семы отрицания в формальной и содержательной структуре фразы.

Несогласие может выражаться на **лексическом** уровне:

постоянное искушение восстать, ☹{*выступить*}; *когда я понял, что жизнь прошла, я* ☹{*взбунтовался*}; к*огда кто-то говорит,*

что мы отстали навсегда, ☹{*это вранье*}; *отец мне неоднократно говорил, что* ☹{*это обычное*} ***преувеличение*** *обычных журналюг*

на **синтаксическом** уровне:

☹{***но*** *я*} *студень варю иначе*; ☹{*дождик нужен для посева,* ***а я*** *в гараж*}; ***а*** *в дальнейшем* ☹{***все-таки*** *его надо открыть*}; ☹{*их* ***на самом деле*** *очень много*}; ☹{***а*** *таких контекстов очень много*}; *одни люди верят,* ***а другие*** *думают, что* ☹{*это он говорит нарошно*}; *некий выверт эволюции, который* ☹{*применительно к одноклеточным*} *работает,* ***а*** *для многоклеточных не имеет места* [в последнем примере отрицается не предшествующий, а последующий компонент]

на **интонационном** уровне (эмфаза выделена КАПИТЕЛЬЮ):

на которые ты вынужден отвечать, ☹{*ДАЖЕ ЕСЛИ ты занимаешься*} *механизмом…*; *поэтому будущее сегодня* ✂{*РЕАЛЬНО*}

на **текстовом** уровне:

так что *электромагнитное* ☹{*взаимодействие продолжает оставаться*} *здесь симметричным* [в предшествующем тексте озвучивалась чужая позиция, что электромагнитное взаимодействие в некотором эксперименте оказывается несимметричным; после ряда аргументов говорящий утверждает, что оно продолжает оставаться симметричным]; *это* ***в самом деле*** *каменистый материал,* ☹{*потому что*} *свойства кратеров одни и те же* [в предшествующем тексте озвучивалась чужая позиция, что в кратерах на Венере — не каменистый материал; после ряда аргументов говорящий делает противоположное утверждение]; ***так что*** ☹{*симметрия по-прежнему*} *сохраняется* [в предтексте озвучивалась идея, что в некотором событии симметрия не сохраняется; после ряда аргументов говорящий делает обратное утверждение]; *показал, что, используя комплементарность* ☹{***на самом деле***} *… можно строить достаточно длинные цепочки* [до данного эксперимента существовало мнение, что строить длинные цепочки нельзя].

В диалоговом режиме (который на нашем материале наблюдается чаще в речи кинематографа, но встречается и в некинематографической речи) возражение и несогласие часто встречаются по отношению к адресату. Самый простой случай — когда положительное (не имеющее в себе отрицательной компоненты) высказывание используется как отрицательный ответ на утверждение первого говорящего или как возражение говорящему:

[первый говорящий что-то утверждает по телефону, второй говорящий возражает:] *Ну почему же?* ☹{*Здесь нормально*}; — *Извиняюсь, надо бы протокол составить.*— *Надо. Но и* ☹{*Новый год надо встретить*}; [первый говорящий утверждает, что второй говорящий вложил деньги в чарты; второй говорящий возражает:] ☹{*Если бы я вложил*} *бабки в чарты, твои бы и в первую сотню не попали бы.* [первый говорящий спрашивает:] *Могут или пугают?* [второй говорящий отвечает:] ☹{*Могут.*} [в последнем случае интересно распределение ответов на дизъюнктивный вопрос между речевым и жестикуляционным каналом — на речевом уровне второй говорящий подтверждает одну из альтернатив (*могут*), а на жестикуляционном уровне возражает говорящему, который может выбрать неправильную альтернативу, вторую, *пугают*]

Отрицательным жестом в диалоговом режиме часто сопровождаются стандартные способы выразить возражение:

☹{*Осмелюсь заметить*}, *ваше величество*… [дальше идет возражение]; ☹{*Вы мне льстите*} [в смысле — вы неправы, характеризуя меня таким-то образом].

Отдельный блок возражений адресату представляют собой высказывания, в которых говорящий возражает на реально не прозвучавшую, но **предвосхищенную** (неважно, правильно или нет) реплику слушающего. В этом случае отрицательный жест сопровождает формально «положительное» высказывание говорящего, которое, с точки зрения говорящего, является, тем не менее, несогласием, возражением, опровержением и под., что и приводит к использованию отрицательного жеста. Примеры:

он сказал: «Ну ребят, ну это ☹{*очевидно*}! [говорящий полагает, что слушающие видят в некоторой ситуации те или иные сложности и неочевидности, и заранее опровергает такую трактовку]; *Она его поняла*, ☹{*всё естественно*} [говорящая полагает, что слушающая считает некоторую ситуацию неестественной, и заранее опровергает такую трактовку]; *Поппер, которого не приглашали…* ☹{*я заметил твой скептический взгляд*}, *я сейчас дам тебе высказаться* [говорящий замечает, что слушающий не согласен с его высказыванием, и заранее опровергает возможное предположение слушающего, что ему не дадут возразить]; ☹{*пусть смеются надо мной*}, *лишь бы войти* [говорящий ожидает со стороны слушающего сообщения, что над ним, говорящим, будут смеяться, и возражает, что согласен на эти на-

смешки]; ☹{*у меня все в порядке*} [говорящий предвосхищает вопрос слушающего «А у тебя тоже все плохо?» и заранее отвечает на него, опровергая это предположение]; ☹{*и колхоз у нас хороший*} [говорящая ожидает от слушающего сомнений в качестве колхоза и заранее опровергает эти сомнения]; ☹✂{*это нормально*}, *это мы все время так жестикулируем* [докладчик предполагает сомнения аудитории, что некоторое явление стандартно, и заранее опровергает это сомнение]; *в Москве емк…* ☹✂{*грубо, очень грубо*}, *емкость рынка это где-то…* [говорящий заранее предполагает сомнение в точности цифр, которые он планирует озвучить, и делает превентивную оговорку, заранее опровергая эти сомнения]; ☹{*я могу сказать, что Ленина там, Маркса*} *он действительно читал десятки раз* [говорящий заранее опровергает возможное сомнение аудитории в том, что Сталин многажды перечитывал работы Маркса и Ленина]; *кроме Евразии есть еще Африка, Австралия, Америка и* ☹{*Новая Гвинея*} [говорящий предполагает, что слушающий думает, что проблема исчерпывается одной Евразией, и напоминает ему, что, напротив, существуют еще важные с точки зрения обсуждаемой проблемы регионы]; *то он* ☹{*безусловно*} *имел бы в виду себя* [докладчик предвосхищает утверждение аудитории: может быть, он имел в виду не себя? и возражает на это предвосхищенное утверждение].

Итак, мы видим, что глубинное отрицание сопровождает речевой жанр возражения, опровержения, дискуссии с третьими лицами и со слушающим. Но, как показывает материал, говорящий может занимать позицию опровергающего и по отношению к своему собственному высказыванию. В этом случае мы имеем дело с ситуацией дистанцирования говорящего от содержания своего собственного высказывания или, иными словами, с ситуацией отчуждения высказывания от говорящего. В этих случаях в высказывании появляются маркеры дистанцирования, которые дают нам знать, что говорящий не желает нести за свое высказывание всю полноту ответственности. Маркеры дистанцирования очень часто сопровождаются ПГО, которое в этих случаях означает что-то типа ‘я не знаю точно’, ‘я не уверен’, ‘я не хочу иметь с этим ничего общего’ и под. Примеры:

есть ***как бы*** ☹{*локальное нарушение симметрии*}, *но симметрия при этом сохраняется*; ☹{*мне вот, например, очень* ***странно***} *вот такое каждый год увеличение бюджета*; ***наверное****, есть* ☹{***какой-то*** *оптимизм*}, *но…*; ☹{***вообще***}, ***в принципе****, мы рас-*

сказали всё; ☹{***вообще-то***}, и *брошенные, и незаселенные — они все пустые;* ***по некоторым таким*** *параметрам* ☹{***типа*** *личных местоимений*}; ☹{***каким-то*** *образом*}, *каким, пока мы* ***не знаем***; ***может****, она* ☹{***как-то***} *поднимается из недр земли*; *то,* ☹{***может быть***, *это красинокавказцы*}; ☹{*то есть* ***мы можем говорить***} *о некоторой евразийской семье*; *это можно делать,* ☹{*как это*} *делают,* ***видимо****, киты, или черепахи, или землекопы*; ☹{***по-моему****, Валерий все* ***достаточно*** *хорошо сформулировал*}; ☹{*это* ***кажется*** *очень* ***странным***}; *формируется* ☹{***достаточно своеобразная*** *картина*}; ***ну***… ☹{*до моего дома хватит*}, *а обратно нет*; *то есть они* ☹✂{[пауза]} *одинаковы и* ***вроде бы*** *и по актуальному значению*; ☹{*влюбляться*} *в* ***какого-то*** *графа Данилу*.

3.1.2.3. Оценочные высказывания

Говорящий может выражать с помощью ПГО оценку того, что передает его высказывание. Здесь мы сталкиваемся, очевидно, с оценками ‘дурно’ и ‘хорошо’. Структура пейоративных высказываний, оценивающих содержание дурно, устроена достаточно просто: посредством отрицательного жеста, сопровождающего ПГО, говорящий осуждает нечто, оценивает нечто как неправильное, неуместное, нехорошее, т. е. выражает оценки, включающие в себя компонент ‘плохое как антоним хорошего или нормального’. Примеры:

☹ {*глупый ты, Лешка*};
эти болезни ☹ {*идут и лечатся очень медленно и печально*};
☹ {*зима будет долгой*}, *но мы должны…*;
что-то вы ☹ {*все взъерепенились, я погляжу!*};
☹ {*тяжело*} *было;*
☹ {*ух и тяжело она нам досталась*} *в этом году!*;
это ☹ {*слишком длинный эксперимент*};
☹✂ {*в общем, тяжело*}; ☹ {*но это мало что дает*}

Гораздо сложнее и интересней устроена положительная оценка, выраженная с помощью отрицательного жеста в ПГО. Высказывания с положительной оценкой в общем случае устроены следующим образом: в семантическую структуру фразы входит некоторое утверждение, а над ним осуществляется оценочная надстройка ‘данное утверждение кажется говорящему настолько удивительным, что говорящий не верит своим органам чувств, задействованным

в восприятии утверждаемого факта’. Таким образом, компонентом, влекущим за собой появление отрицательного жеста, является ‘это удивительно, я не верю собственным глазам, ушам и под.’. Иногда лексемы с компонентом ‘вызывающий удивление’ появляются в таких контекстах и на поверхностном уровне:

это ☹{***удивительная***} *вещь*; *может быть, вот эти* ☹{*просторы*} ***необычные***

В группу ‘положительно окрашенное удивление’ попадают практически все восклицания:

он проснется, увидит, ☹{*ему приятно будет!*}; *если бы я умел петь* ☹{[пауза]}, *как бы я спел!*; ☹{*Васька!*} *Молоток!*; ☹{*Как это смешно и мило*}, *Семен!*; ☹{*Ой*}, *я смотрю, у вас вторичные половые признаки стали проявляться!*; ☹{[пауза]} *Хорош, подлец!*; ☹{*Хорош!*}; ☹{*Умница!*}; ☹{*Какая неожиданная и во всех смыслах приятная*} *для меня встреча!*; ☹{*Какая*}, *право, хорошенькая,* ☹{*симпатичная*} *… барышня!*; ☹{*Ой, хороший гусь получился!*}; ☹{*Ммм!* [междометие]}; *Там* ☹{*такое творится!*}; ☹{*Хорошо!*}

Заметим, что в восклицаниях с отрицательной оценкой мы не можем точно сказать, какой именно фактор влияет на появление отрицательного жеста, — семантический компонент ‘плохое, нехорошее’ или семантический компонент ‘удивительно, не верю’, и можем констатировать лишь наличие оценочного компонента в высказывании, который влияет на появление отрицательной жестикуляции:

☹{*Какой кошмар!*} ☹{*Какой кошмар!*}; *Ну и люди, знаете,* ☹{*это тоже!*}; *Ну и* ☹{*силен*} *ты врать!*

В группу ‘не верю своим органам чувств’ попадают оценочные конструкции с местоимениями *так, такой*:

☹{*Ты бы ушел от такой женщины?*}; *Земля, которая* ☹{*так ценилась*}, *что…*; *Если это … роботы, то почему* ☹{*они так похожи*} *на живых?*

Таким же образом устроены контексты, в которых использованы лексемы, окрашенные для говорящего в высшей степени положительно (ситуативно обусловленная положительная оценка), а также лексемы, включающие в себя объективно, вне зависимости от ситуации, интенсивную положительную оценку. Обычно и ситуативные, и словарные лексемы с положительной оценкой попадают в эмфазу. К этой же группе примыкают контексты с интенсификаторами, которые относятся к словам с положительной оценкой:

☹{*Вы НАСТОЯЩИЙ поэт!*}; ☹{*Vous parlez bien français!*}; ☹{*Я МЕЧТАЮ создать центр*}, *где бы можно было бы собрать людей…*; *В вас заложен дух* ☹{*ВОИТЕЛЬНИЦЫ!*}; *Я жертвую любовью.* ☹{*Для ИСКУССТВА*}; *Он* ☹{*ОБОЖАЛ читать*}; *А одна нога, которая в аварию попала,* ☹{*ОЧЕНЬ сексуальная*}; ☹{*ПОТРЯСАЮЩИЙ*} *учитель!*; *Ой,* ☹{*она у нас ЗАМЕЧАТЕЛЬНО готовит*}; ☹{*Я ОЧЕНЬ уважаю*} *искусство*; ☹{*А это КОЛОССАЛЬНОЕ*} *дело!*; ☹{*ПОРАЗИТЕЛЬНОЕ*} *сходство!*; ☹{*БЕСКОНЕЧНО благодарен вам*}, *любезнейший…*

Отдельную подгруппу составляют контексты с семой ‘много’, которая, если контекст не подразумевает иного, по-видимому, включает положительную оценку (в пограничных случаях — неотрицательную):

Забавно, что у меня, ☹{*например, куча*} *знакомых, кто в кризис ... сократил штат*; *Но кроме этого есть* ☹{*колоссальное количество*}…; *Ну,* ☹{*лет триста*} *я вам гарантирую*; *Это, конечно, предмет для* ☹{*огромного*} *количества спекуляций*; *корпус …* ☹✂{он *еще намного больше*}, *чем…*; ☹{*44 дня*} *жил в коммунизме абсолютном!*; *англичане в Австралии* ☹{*на большом*} *телескопе наблюдали*; *Мы знаем, что* ☹{*очень многие*} *органические соединения…*; *у каждого* ☹{*порядка там тысячи человек*} *в среднем*; *одно из самых впечатляющих событий … это* ☹{*65*} *миллионов лет назад…*; *одно и то же явление может иметь* ☹{*очень*} *много имен*

3.2. СОПУТСТВУЮЩИЕ ЖЕСТЫ

Отрицательные жесты в контекстах с ПГО часто сопровождаются жестами иной семантики, которые либо совпадают с отрицательными жестами по времени, либо непосредственно следуют после них. Интересно посмотреть, как соотносятся такие сопутствующие жесты с выделенными группами контекстов с ТГО и ПГО, и определить, есть ли в полученных распределениях какие-то закономерности.

Сопутствующие жесты можно разбить на следующие основные группы (жесты, встретившиеся всего несколько раз, мы не включаем в рассмотрение, поскольку их количество статистически незначимо).

Жесты дистанцирования: *поднять брови, склонить голову набок, махнуть рукой, отодвинуться от слушающего, двинуть голову назад, пожать плечами, развести руками*

Жест подтверждения: *кивок сверху вниз*

Эмфатические жесты: *двинуть головой вперед, нахмуриться, сморщиться*

Глазное поведение: *закрыть глаза*

Табл. 15 демонстрирует, как сопутствующие жесты распределены между выделенными в данном разделе типами глубинного отрицания.

Таблица 15

Сопутствующие жесты Группы контекстов	**дистанцирования**	**подтверждения**	**эмфатические**	**закрытые глаза**	**нет сопутствующих жестов**
Дистанцирование и несогласие	**45**	*2*	*7*	*6*	*23*
Положительная оценка (‘не верю своим органам чувств’)	10	4	11	**10**	17
Отрицательная оценка	5	3	**9**	4	10
Отрицание предвосхищенной реплики	12	**7**	6	4	14
Тематическое глубинное отрицание (ТГО)	12	8	20	8	**68**
Автокоррекция	0	3	0	**7**	3
χ^2=106,6, p=8,1–14, параметры связаны, распределения достоверны					

Как видим, данные в существенной степени распределены несимметрично, что косвенно подтверждает объективное существование границ между выделенными нами группами контекстов: все они сопровождаются отрицательными жестами (*качать головой* и/или *отрезать рукой*), но при этом сопутствующие жесты для разных групп контекстов различаются.

Довольно ожидаемо для группы контекстов с семантической доминантой ‘дистанцирование’ или ‘несогласие’ единственной характерной группой сопутствующих жестов является группа «дис-

танцирование», т. е. говорящий, осуществляя отрицательный жест, одновременно поднимает брови, пожимает плечами, разводит руками, склоняет голову набок и отодвигается от слушающего.

При этом, однако, группа контекстов с несогласием, которое выражается отрицанием предвосхищенной реплики адресата, имеет совершенно другой состав сопутствующих жестов — ни жесты дистанцирования, ни остальные жесты не показывают какого-либо пика в этой группе, демонстрируя не отличающиеся от средних распределения, и лишь группа жестов подтверждения, а именно, подтверждающий кивок, показывает заметный всплеск. Это означает, что, отрицая предвосхищенное утверждение адресата, говорящий сначала осуществляет отрицательный жест головой, а потом, практически сразу после отрицательного жеста, кивает:

и, ☹✄{безуслов} кивок{*но*}, *они выражают обман*
вот ☹{*ОНИ будут*} *в словаре* ✄ кивок{*ТОЧНО*}
☹{*пусть*} кивок{*смеются надо мной*}, *лишь бы войти*
☹{*и колхоз у нас*} кивок{*хороший*}
вместе с тем, герр Хеснер, я ☹{*прошу соблюдать условия нашего*} кивок{*договора*}

Этот кивок подтверждает достоверность и правильность утверждения, которое говорящий делает, отрицая предвосхищенную реплику слушающего.

Группа контекстов осуждения, где отрицательный жест сопровождает сему ‘плохо, нехорошо’, в качестве сопутствующих с повышенной частотой имеет эмфатические жесты, усиливающие отрицательную оценку содержания высказывания.

Очень важным является тот факт, что тематическое глубинное отрицание (внутрисловное отрицание, включая контексты с семой всеобщности) разительным образом отличается от контекстов с прагматическим глубинным отрицанием, — ТГО не имеет сопутствующих жестов и выражено в визуальном режиме чистыми отрицательными жестами.

Интересно, что две группы — положительная оценка с семой ‘не верить собственным органам чувств’ и автокоррекция — обе имеют пики на сопутствующем жесте *закрыть глаза*.

В нашей работе [Гришина 2011] среди прочих значений жеста *закрыть глаза* мы выделили два: 1) жест *закрыть глаза* как жест, выражающий внутреннюю сосредоточенность говорящего

на мыслительных, эмоциональных, оценочных процессах, происходящих в его внутреннем мире (жест *закрыть глаза* в данном случае выступает как способ оградить говорящего от внешних впечатлений, которые могут помешать ему сосредоточиться на его внутреннем мире); 2) жест *закрыть глаза* как жест, который символизирует выход говорящего из ситуации коммуникации, завершение неудачной, неправильной коммуникации; этот жест говорящий осуществляет, чтобы, открыв глаза, войти в коммуникацию как совершенно новый участник (в этом значении жест *закрыть глаза* с повышенной частотой сопровождает отрицательный жест *качать головой*).

Мы считаем, что в первом значении (жест сосредоточенности) *закрыть глаза* сопровождает глубинное отрицание положительной оценки (‘этого не может быть, не верю своим органам чувств’), тогда как во втором значении (‘выход из акта коммуникации’) жест *закрыть глаза* сопровождает ситуацию автокоррекции, где говорящий с помощью отрицательного жеста прерывает первую, неудачную попытку коммуникации и переходит к новой, более удачной.

3.3. РАСПРЕДЕЛЕНИЕ ОТРИЦАТЕЛЬНЫХ ЖЕСТОВ МЕЖДУ ГРУППАМИ КОНТЕКСТОВ

Теперь мы можем перейти к анализу распределения типов контекстов с глубинным отрицанием между разными отрицательными жестами — *качать головой* и/или *отрезать рукой*, см. табл. 16.

Таблица 16

Тип жеста **Тип контекста**	**качать головой + отрезать рукой**	**качать головой**	**отрезать рукой**
Прагматические глубинные отрицания (ПГО)	10	**203**	*3*
Внутрисловные отрицания	4	*39*	**30**
ТГО с семой ‘всеобщность’	**7**	*28*	**10**
χ^2=90,8, p=8,7–19, параметры связаны, распределения достоверны			

Итак, мы видим, что

1) для контекстов с прагматическим глубинным отрицанием характерен жест *качать головой* и не характерен жест *отрезать рукой*,

2) для контекстов с внутрисловным отрицанием, наоборот, характерен жест *отрезать рукой* и не характерен жест *качать головой*,

3) для контекстов с тематическим глубинным отрицанием, включающих в себя сему ‘всеобщность’ (*весь, всё, каждый, любой* и проч.) не характерен жест *качать головой*, но характерен жест *отрезать рукой* (как и для других тематических глубинных отрицаний), но при этом контексты всеобщности с повышенной по сравнению со средней частотой сопровождаются обоими жестами одновременно.

Последний пункт, по-видимому, объясняется тем, что контексты ТГО с семантическим компонентом всеобщности легко переосмысляются как оценочные, поскольку метафорический переход от семы ‘всё’ к семе ‘много’ для жестикуляции довольно стандартен[1], а как мы показали выше, контексты с семантическим компонентом ‘много’ образуют отдельную подгруппу в группе контекстов с положительной оценкой. Таким образом, контексты всеобщности с повышенной частотой сочетают в себе стандартную для группы ТГО жестикуляцию *отрезать рукой* — с оценочной жестикуляцией ПГО *качать головой*.

С учетом этой поправки, из табл. 16 с очевидностью следует, что жестикуляция головой и рукой в контекстах с глубинным отрицанием распределена крестообразно: за тематическим внутрисловным отрицанием, не зависящим от прагматического контекста, закреплена жестикуляция рукой, а за прагматическим глубинным отрицанием, тесно связанным с поведением говорящего внутри коммуникативного акта и с его отношениями со слушающим и с формой/содержанием его собственного высказывания, закреплена жестикуляция головой[2].

[1] В частности, он имеет место в круговых движениях, где круговые движения руками, имеющие значение всеобщности, легко переносятся и на контексты с семантическим компонентом ‘много’, см. [Гришина 2014б].

[2] Интересно, что полученные данные частично коррелируют с данными, полученными в ходе исследования развития отрицательных конструкций и отрицательных жестов у детей. Так, в работе [Pea 1980] показано, что у детей появляется снача-

4. Заключение

В нашей работе мы попытались ответить на три вопроса:

1) Есть ли какие-то закономерности в употреблении основных отрицательных жестов, *качать головой* и *отрезать рукой*, которые сопровождают стандартную отрицательную конструкцию *не* + предикат?

2) В **положительных** контекстах какого типа в русской жестикуляции употребляются **отрицательные** жесты (иными словами, каков набор конструкций с глубинным отрицанием в русском языке)?

3) Можно ли найти какие-то закономерности в употреблении основных отрицательных жестов, *качать головой* и *отрезать рукой*, в контекстах с глубинным отрицанием?

Был рассмотрен довольно большой массив данных и на нем получены некоторые статистические закономерности, которые были проинтерпретированы нами следующим образом.

I. В стандартных отрицательных конструкциях *не* + предикат на выбор сопровождающего отрицательного жеста влияет позиция говорящего относительно его высказывания. Если говорящий воспринимает отрицание как «свое», то он использует отрицательный жест *качать головой*. «Принадлежность» высказывания говорящему проявляется в том, что в таких высказываниях говорящий одновременно является основным действующим лицом (предикат имеет форму первого лица), либо является непосредственным наблюдателем отсутствия какого-либо события (используется настоящее время), либо оценивает степень ненаступления события (при отрицательных конструкциях используются разного рода интенсификаторы). В случае отчуждения говорящего от его собственного высказывания, т. е. в случае восприятия высказывания как «не вполне своего» (отсутствие отрицательного жеста) или как «чужого» (жест

ла эмоциональный отказ от предложения, затем – констатация исчезновения объекта наблюдения из поля зрения, и лишь затем, на самых поздних этапах, – логическое отрицание в парадигме истина/ложь, т.е. в ходе развития ребенка явно происходит восхождение от минимально абстрактного к максимально абстрактному отрицанию. И при этом, как указывает автор [op. cit.: 170], именно первый, наиболее чувственный и эмоциональный вид отрицания сопровождается жестом *качать головой*, в то время как отрицание истинности какого-либо утверждения сопровождается словесными формулировками с *no* и *not*, и практически никогда не сопровождается жестом *качать головой*.

отрезать рукой),— точка зрения говорящего и точка зрения наблюдателя ненаступления события не совпадают (используется предикат в третьем лице, высказывание имеет нейтральную модальность, т. е. его предикация не может быть оценена ни как истинная, ни как ложная, и говорящий никак не оценивает степень ненаступления события с помощью интенсификаторов). Таким образом, жест *качать головой* противопоставляет «свое» отрицание — «не вполне своим» или «чужим» отрицаниям, которые маркируются либо отсутствием отрицательного жеста, либо жестом отрезать рукой (в терминологии Д. Макнилла жест качать головой ведет себя как жест персонажа (CPVT), а жест *отрезать рукой* — как жест наблюдателя (OPVT)).

II. Положительные контексты, в которых достаточно регулярно используются отрицательные жесты, делятся на две большие группы. Первая группа, контексты с тематическим глубинным отрицанием (ТГО), характеризуются использованием лексем и конструкций, в толкование которых включено отрицание (предикаты, квантворные слова, конструкции и частицы, которые сохраняют компонент отрицания в своей семантической структуре вне зависимости от прагматических характеристик высказывания и акта коммуникации в целом). Вторая группа контекстов, прагматическое глубинное отрицание (ПГО), не имеет в себе постоянного семантического компонента ‘отрицание’ и приобретает его в зависимости от коммуникативной ситуации, от намерений говорящего, от речевого акта, который он совершает (ситуация автокоррекции — отказ от первоначальной стратегии построения высказывания,— ситуация возражения и дистанцирования, оценка события как отрицательного или положительного).

III. В контекстах с ТГО, не имеющих зависимости от коммуникативных и прагматических характеристик высказывания, используется жест *отрезать рукой*, в контекстах с ПГО — жест *качать головой*. Тем самым жест *качать головой* может быть описан как жест субъективный и ситуативный, связанный с позицией говорящего относительно коммуникативного и прагматического контекста высказывания, а жест *отрезать рукой* — как объективный, связанный с постоянными, словарными семантическими характеристиками используемых лексем и конструкций.

Литература

Atkinson et al. 2004 — *J. Atkinson, R. Campbell, J. Marshall, A. Thacker, B. Woll.* Understanding 'not': neuropsychological dissociations between hand and head markers of negation in BSL // Neuropsychologia, 42 (2004), p. 214–229

Bergman 1995 — *B. Bergman.* Manual and non-manual expression of negation in Swedish Sign Language // H. Bos, T. Schermer (Eds.). Sign language research 1994. Hamburg: Signum Press., p. 85–103

Calbris 2008 — *G. Calbris*. La tête de Nicolas Sarkozy, ou les fonctions des gestes de la tête durant l'énonciation // Mots. Les langage du politique, 86/2008. p. 99–118

Calbris 2011 — *G. Calbris*. Elements of Meaning in Gesture. Amsterdam/Philadelpjia, 2011

Coerts 1992 — *J. Coerts*. Non-manual grammatical markers: An analysis of interrogatives, negations and topicalisations in Sign Language of The Netherlands. Amsterdam: University of Amsterdam.

Esteve-Gibert, Prieto 2014 — *N. Esteve-Gibert, P. Prieto*. Infants temporally coordinate gesture-speech combinations before they produce their first words // Speech Communication. 2014, T. 57, p. 301–316.

Goodwin 1981 — *Ch. Goodwin*. Conversational organization: interaction between speakers and hearers. Academic Press, New York, 1981

Guidetti 2005 — *M. Guidetti.* Yes or no? How young French children combine gestures and speech to agree and refuse // Journal of Child Language, 32 (2005), p. 911–924

Harrison 2009 — *S. Harrison.* Grammar, Gesture, and Cognition: The Case of Negation in English. Ph. D. Dissertation, Université Michel de Montaigne Bordeaux 3

Harrison 2010 — *S. Harrison*. Evidence for node and scope of negation on coverbal gesture // Gesture, 10 (1), p. 29–51

Harrison 2013 — *S. Harrison*. The temporal coordination of negation gestures in relation to speech // Proceedings of TiGeR, 2013

Jungheim 2004 — *N. O. Junghaim*. Hand in Hand: A Comparison of Gestures Accompanying Japanese Native Speaker and JSL Learner Refusals // JALT Journal, Vol. 26, № 2, November, 2004, p. 127–146

Kendon 2002 — *A. Kendon*. Some uses of the head shake // Gesture 2 (2), 2002, p. 147–182

Kendon 2004 — *A. Kendon*. Gesture. Visible Action as Utterance. Cambridge Univ. Press, 2004

McClave 2000 — *E. Z. McClave*. Linguistic functions of head movements in the context of speech // Journal of Pragmatics, 32, p. 855–878

McClave 2001 — *E. Z. McClave, Evelyn Z.* The relationship between spontaneous gestures of the hearing and American Sign Language // Gesture 1, 2001, p. 51–72

McClave et al. 2008 — *E. Z. McClave, H. Kim, T. Tamer, M. Mileff*. Linguistic movements of the head in Arabic, Bulgarian, Korean, and African American vernacular English // Gesture, 7 (3), 2008, p. 343–390

McKee 2006 — *R. McKee*. Aspects of interrogatives and negation in New Zealand Sign Language // U. Zeshan (toim.) Interrogative and Negative Constructions in Sign Languages, Preston (UK), 2006, p. 70–90

McNeill 1992 — *D. McNeill*. Hand and Mind. Chicago, 1992

Mueller 2004 — *C. Mueller*. The Palm-Up-Open-Hand. A case of a gesture family? // C. Mueller, R. Posner (Eds.), The Semantics and pragmatics of everyday gesture. The Berlin conference. Berlin, p. 233–256

Nespor, Sandler 1999 — *M. Nespor, W. Sandler*. Prosody in Israeli Sign Language // Language and Speech, 42, 1999, p. 143–176.

Pea 1980 — *R. D. Pea*. The development of negation in early child language // The social foundation of language and thought. Essays in honor of Jerome S. Bruner. N. Y. –L., 1980, p. 156–186

Pfau 2002 — *R. Pfau*. Applying morphosyntactic and phonological readjustment rules in natural language negation. // R. P. Meier, Q. -P. D. Cormier (Eds.). Modality and structure in signed and spoken languages. Cambridge: Cambridge University Press., 2002, p. 263–295

Pfau 2008 — *R. Pfau*. The Grammar of Headshake: A Typological Perspective on German Sign Language Negation // Linguistics in Amsterdam. 1, 2008, p. 37–74

Pfau, Quer 2002 — *R. Pfau, J. Quer*. V-to-Neg raising and negative concord in three sign languages // Rivista di Grammatica Generativa, V. 27 (2002), p. 73–86

Pinfield 2013 –T. Pinfield. Gestural head movements in captive bonobos (Pan paniscus); use, function and evolutionary implications // Canopy Journal of the Primate Conservation, MSc Programme, 2013, v. 14, № 1

Ross 1970 — *Ross J. R.* On declarative sentences // Readings in English transformational grammar. Waltham (Mass.), 1970. — N 222. — P. 272

Sandler 1999 — *W. Sandler*. Prosody in two natural language modalities // Language and Speech, 42, 1999, p. 127–142

Schneider et al. — *C. Schneider, J. Call, K. Liebal.* Do bonobos say NO by shaking their head? // Primates, 2010, v. 51, №. 3, p. 199–202.

Veinberg, Wilbur 1990 — *S. C. Veinberg, R. B. Wilbur.* A linguistic analysis of the negative headshake in American Sign Language // Sign Language Studies, 68, p. 217–244

Богуславский 1985 — *И. М. Богуславский.* Исследования по синтаксической семантике. М., 1985

Гришина 2011 — *Е. А. Гришина.* Закрытые глаза: Жестикуляция, пунктуация, акцентуация // «Филология», № 2, 2011, с. 39–56

Гришина 2012 — *Е. А. Гришина*. Автодейксис: типы и значения // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып. 11 (18). — М.: Изд-во РГГУ, 2014, с. 173–186

Гришина 2013 — *Е. А. Гришина.* Темпоральные дейктические жесты // Известия РАН. Серия литературы и языка, 2013, том 72, № 1. С. 3–31

Гришина 2014 а — *Е. А. Гришина*. Вертикальная ось в жестикуляции: лингвистический аспект // Русский язык в научном освещении, 26, 2014. С. 42–89

Гришина 2014 б — *Е. А. Гришина*. Круги и колебания: семантика сложных траекторий в русской жестикуляции // Когнитивные подходы к языку: Сб. статей, М., 2014. С. 238–286

Падучева 2013 — *Е. В. Падучева.* Русское отрицательное предложение. М., 2013

Подлесская 2014 — *В. И. Подлесская.* «То есть, не убили, а зарезали саблей»: самоисправления говорящего в устных рассказах // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.). Вып. 13 (20). — М.: Изд-во РГГУ, 2014, с. 547–561

E. A. Grishina
Vinogradov Russian Language Institute
of the Russian Academy of Sciences
(Russia, Moscow)
rudi2007@yandex.ru

ON GESTURAL NEGATION IN THE RUSSIAN LANGUAGE

The article focuses on the topic of correlation between a word and a gesture in standard and the most frequent constructions with negation of a predicate. The article consists of two parts. The first part discusses negative constructions "ne [not] + predicate" and analyzes consistent patterns typical for choosing negative gestures. The second part describes constructions with deep negation, i.e. contexts without a negative construction on the surface, but accompanied with negative gestures. The research has been conducted using the Multimodal Russian Corpus (MURCO).

The research procedure is based on defining linguistic features in negative constructions and constructions with deep negation: a) semantics of lexical units related to gestural negation, b) syntactic structures, c) morphological components d) rhythmic features. Thus we assign every sample from the gestural database with a number of its linguistic characteristics in an utterance accompanying the gestures or the lack of such. With the help of statistical parameters, we determine which linguistic characteristics are essential for the presence or the lack of gestural negation and its quality (the body organ that performs it), and interpret all the statistical consistencies.

Key words: gestural negation, negative constructions, constructions with deep negation

References

Atkinson J., Campbell R., Marshall J., Thacker A., Woll B. Understanding 'not': neuropsychological dissociations between hand and head markers of negation in BSL. *Neuropsychologia*, 42 (2004), pp. 214–229.

Bergman B. Manual and non-manual expression of negation in

Swedish Sign Language. H. Bos, T. Schermer (Eds.). *Sign language research* 1994. Hamburg, Signum Press., pp. 85–103.

Boguslavskij I. M. *Issledovanija po sintaksicheskoj semantike* [Research on syntactic semantics]. Moscow, 1985.

Calbris G. La tête de Nicolas Sarkozy, ou les fonctions des gestes de la tête durant l'énonciation. *Mots. Les langage du politique*, 86/2008. pp. 99–118.

Calbris G. *Elements of Meaning in Gesture*. Amsterdam / Philadelpjia, 2011.

Coerts J. *Non-manual grammatical markers: An analysis of interrogatives, negations and topicalisations in Sign Language of The Netherlands*. Amsterdam: University of Amsterdam.

Esteve-Gibert N., Prieto P. Infants temporally coordinate gesture-speech combinations before they produce their first words. *Speech Communication*, 2014, v. 57, pp. 301–316.

Goodwin Ch. *Conversational organization: interaction between speakers and hearers*. Academic Press, New York, 1981

Grishina E.A. [Closed eyes: Gesticulation, punctuation, accentuation]. *Filologija* [Philology], 2011, no. 2, p. 39–56. (In Russ.)

Grishina E.A. [Autodeixis: types and meanings]. *Komp'juternaya lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii «Dialog»*. [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2012"]. Issue 11 (18). Moscow, RSUH Publ., 2012, pp. 173–186. (In Russ.)

Grishina E. A. [Temporal deictic gestures]. *Izvestija RAN. Serija literatury i jazyka*, 2013, v. 72, no. 1, pp. 3–31. (In Russ.)

Grishina E.A. [Vertical axis in gesticulation: linguistic aspect]. *Russkii yazyk v nauchnom osveshchenii*, 26, 2014, no. 26, pp. 42–89. (In Russ.)

Grishina E. A. [Circles and sways: semantics of complex trajectories in Russian gensticulation]. *Kognitivnye podhody k yazyku* [Cognitive approaches to Language], Moscow, 2014, pp. 238-286 (In Russ.)

Guidetti M. Yes or no? How young French children combine gestures and speech to agree and refuse. *Journal of Child Language*, 32 (2005), pp. 911–924.

Harrison S. *Grammar, Gesture, and Cognition: The Case of Negation in English*. Ph. D. Dissertation, Université Michel de Montaigne Bordeaux 3.

Harrison S. Evidence for node and scope of negation on coverbal gesture. *Gesture*, 10(1), pp. 29–51.

Harrison S. The temporal coordination of negation gestures in relation to speech. *Proceedings of TiGeR*, 2013.

Junghaim N. O. Hand in Hand: A Comparison of Gestures Accompanying Japanese Native Speaker and JSL Learner Refusals. *JALT Journal*, Vol. 26, no. 2, November, 2004, pp. 127–146.

Kendon A. Some uses of the head shake. *Gesture* 2 (2), 2002, pp. 147–182.

Kendon A. *Gesture. Visible Action as Utterance.* Cambridge Univ. Press, 2004.

McClave E. Z.. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32, pp. 855–878.

McClave E. Z., Evelyn Z. The relationship between spontaneous gestures of the hearing and American Sign Language. *Gesture* 1, 2001, pp. 51–72.

McClave E. Z., Kim H., Tamer T., Mileff M. Linguistic movements of the head in Arabic, Bulgarian, Korean, and African American vernacular English. *Gesture*, 7(3), 2008, pp. 343–390.

McKee R. Aspects of interrogatives and negation in New Zealand Sign Language. U. Zeshan (toim.) *Interrogative and Negative Constructions in Sign Languages*, Preston (UK), 2006, pp. 70–90.

McNeill D. *Hand and Mind*. Chicago, 1992

Mueller C. The Palm-Up-Open-Hand. A case of a gesture family? C. Mueller, R. Posner (Eds.), *The Semantics and pragmatics of everyday gesture*. The Berlin conference. Berlin, pp. 233–256.

Nespor M., Sandler W. Prosody in Israeli Sign Language. *Language and Speech*, 42, 1999, pp. 143–176.

Paducheva E.V. *Russkoe otricatel'noe predlozhenie* [Russian negative sentence]. Moscow, 2013.

Pea R. D. The development of negation in early child language. *The social foundation of language and thought*. *Essays in honor of Jerome S. Bruner*. N.Y.–L., 1980, pp. 156–186.

Pfau R. Applying morphosyntactic and phonological readjustment rules in natural language negation. R. P. Meier, Q.-P. D. Cormier (Eds.). *Modality and structure in signed and spoken languages*. Cambridge, Cambridge University Press., 2002, pp. 263–295.

Pfau R. The Grammar of Headshake: A Typological Perspective

on German Sign Language Negation. *Linguistics in Amsterdam*, no. 1, 2008, pp. 37–74.

Pfau R., Quer J. V-to-Neg raising and negative concord in three sign languages. *Rivista di Grammatica Generativa*, V. 27 (2002), pp. 73–86.

Pinfield T. Gestural head movements in captive bonobos (Pan paniscus); use, function and evolutionary implications. Canopy Journal of the Primate Conservation, MSc Programme, 2013, v. 14, no. 1.

Podlesskaya V.I. ["To est', ne ubili, a zarezali sablej": speaker's self-correction in oral retelling]. *Komp'yuternaya lingvistika i intellektual'nye tehnologii: Po materialam ezhegodnoi Mezhdunarodnoi konferentsii "Dialog"* [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference "Dialog–2014"]. Issue 13 (20). Moscow, RSUH Publ., 2014, pp. 547–561. (In Russ.)

Ross J. R. On declarative sentences. *Readings in English transformational grammar.* Waltham (Mass.), 1970, N 222, p. 272

Sandler W. Prosody in two natural language modalities. *Language and Speech*, 42, 1999, pp. 127–142.

Schneider C., Call J., Liebal K. Do bonobos say NO by shaking their head? *Primates*, 2010, v. 51, no. 3, pp. 199–202.

Veinberg S. C., Wilbur R. B. A linguistic analysis of the negative headshake in American Sign Language. *Sign Language Studies*, 68, pp. 217–244

А. А. Бонч-Осмоловская
Научно-исследовательский университет
«Высшая школа экономики»
(Россия, Москва)
abonch@gmail.com

КУЛЬТУРОМИКА В НАЦИОНАЛЬНОМ КОРПУСЕ РУССКОГО ЯЗЫКА, К ПОСТАНОВКЕ ЗАДАЧИ: ТРИ ВЕКА РУССКИХ ДОРОГ

Культуромика — новый метод исследования культурных и социальных процессов с помощью статистического анализа употребления слов и словосочетаний в огромных текстовых массивах. Исследования по культуромике выполняются на базе Google ngrams, хранилища всех сочетаний слов для каждого года длиной от 1 до 5 на 8 языках. Однако работа с большими данными обладает рядом проблем, решить которые предлагается, сузив объем данных, но существенно улучшив их по качеству. В статье вводится понятие medium data — средних данных, достаточных для того, чтобы строить с их помощью надежные статистические предсказания, но и допускающих, благодаря своему объему, качественную очистку, фильтрацию, объединение данных в семантические классы. В статье рассматриваются данные XVIII–XX веков, полученные как результат запроса «имя прилагательное + дорога». Весь массив данных был поделен на семантические классы, соответствующие одному из аспектов значений существительного *дорога*. В работе строятся наблюдения о том, какие из семантических классов ведут себя похожим образом, как меняется частотность вхождений семантических классов, а также их лексический состав с течением времени. Эти данные связываются с экстралингвистической реальностью. Фактически предпринимается попытка исследования того, как изменения

в общественной жизни находят отражение в изменениях языковой практики. Это исследование проводится с помощью данных Национального корпуса русского языка.

Ключевые слова: корпус, семантика, дорога, общественное развитие, культурное развитие, русский язык

Цель настоящей работы состоит в том, чтобы показать возможности использования Национального корпуса русского языка для исследования нелингвистических задач. Корпус, в данном случае, используется как упорядоченное, сбалансированное и размеченное хранилище культурной памяти в разные исторические периоды. Объектом для исследования и интерпретации может быть количественное изменение частотности употребления той или иной лексемы или лексического выражения, обозначающих общественно значимые понятия, явления и концепты, качественный анализ сопутствующих контекстов, вытеснение одной лексемы другими — словом, все виды взаимодействия между языковыми единицами, которые могут быть интерпретированы как признаки культурного слома или развития или же как фиксация актуальных экономических и политических процессов и их отражения в общественном сознании.

Принципиально важным и интересным является тот факт, что, рассматривая те или иные лексические изменения и интерпретируя их с помощью экстралингвистических факторов, мы опираемся в целом на корпус текстов, напрямую тематически не связанных с тем или иным фактором. Поясню эту мысль на следующем примере. Резкий взлет частотности употребления слова «женщина» в середине XIX века безусловно связан с первой волной эмансипации. Если мы сравним два подкорпуса текстов периода с 1820 по 1830 и с 1850 по 1860 и ограничим поиском по художественным текстам, т. е. романам, повестям, заведомо не содержащим в себе манифестов о «женском вопросе», мы увидим рост частотности лексемы «женщина» на 20% — с 524 ipm до 623 ipm соответственно. Такая динамика свидетельствует об изменении фокуса интереса писателей и читателей в эту эпоху, а стало быть и об изменениях в общественном сознании. Это изменение и его связь с известными нам из истории этапами общественного и культурного развития и является объектом интерпретации и исследования культуромики.

Данная работа не претендует на то, чтобы представить глубокое социокультурное исследование. Задача видится в другом: описать методологию формирования лингвистической выборки и подходы к лингвистической интерпретации данных для исследования нелингвистических объектов. Предлагаемый метод базируется на представлении о важности и недооцененности того, что можно было бы назвать medium data (средние данные), противопоставляя их big data (большим данным): данные, достаточные для того, чтобы делать статистически надежные предсказания, но в то же время позволяющие качественный содержательный анализ и надежную фильтрацию шума.

Culturomics, или культуромика — термин, обозначающий исследование культурных и социальных процессов с помощью статистического анализа употребления слов и словосочетаний в огромных текстовых массивах, был впервые введен в статье в журнале Science 2010 года, опубликованной большой группой ученых [Michel et al 2010]. Вдохновителями и руководителями работы по культуромике стали два исследователя-биотехнолога Жан-Батист Мишель и Эрез Либерман Айден. Имея серьезный опыт работы с большими данными, необходимых для изучения геномных последовательностей, Ж. Мишель и Э. Айден решили использовать гигантскую библиотеку Google Books для квантитативного культурного анализа. Сейчас в Google Books в электронном виде представлено более 5 млн книг (4% от того, что когда-либо было издано). Именно в результате этой работы появился сервис Google Ngram Viewer, позволяющий быстро сравнить частотность нескольких слов или словосочетаний и построить графики, отражающие диахронические изменения. Так, темами для культуромических исследований, становилось, например, сравнение популярности (частоты цитирования) величайших ученых прошлых веков — Галиллея, Дарвина, Фрейда, Эйнштейна, или демонстрация того, как постепенно уточнялись цифры после запятой числа, равняющегося скорости света. Одним из наиболее известных сюжетов стала визуальная демонстрация того, как работает цензура: упоминание Марка Шагала резко падает в немецких текстах периода третьего Рейха, и равномерно растет в текстах на английском языке этого же времени[1].

[1] См. более подробные примеры на странице Google Ngram Viewer https://books.google.com/ngrams/info.

В целом, можно сказать, что основная проблема культуромики как подхода является оборотной стороной его достоинств: использование огромного массива Google Books позволяет получить надежные статистические данные, но только для очень простых запросов. Мы можем запрашивать лишь n-граммы, а это значит, что предметом исследования могут становиться лишь атомарные смыслы или явления, но не отношения между ними, которые часто, особенно в языках со свободным порядком слов, выражаются с помощью лексем, находящихся на определенном расстоянии друг от друга.

Разметка Национального корпуса русского языка позволяет получить из существенно меньшего объема данных гораздо более качественную и чистую выборку. Это возможно, во-первых, благодаря существенно более высокому качеству исходных цифровых данных, в частотности минимизации ошибок распознавания в НКРЯ по сравнению с Google Books. Во-вторых, Google ngrams для русского языка посчитаны без учета лемматизации, в-третьих, в них не учтен переход со старой орфографии на новую, что осложняет исследования языковых изменений до 1917 года. Чрезвычайно важным преимуществом НКРЯ является то, что документы имеют существенно больший комплекс метаданных. В частности, в то время как Google Books позволяют фильтровать только по дате издания, документы НКРЯ снабжены также указателями на дату создания. Эта разница может быть весьма существенна, например, при изучении советской эпохи, поскольку далеко не все тексты могли пройти цензуру и быть изданными. Наконец, сложная лингвистическая разметка позволяет искать не по лексемам, а по конструкциям, лишь с отчасти заданными слотами. Эта возможность существенно расширяет потенциалы исследований по культуромике.

Таким образом, исследовательский подход может опираться на идею «средних данных», или medium data. Материалом в этом случае являются данные существенно меньшего объема, но выигрывающие в качестве. Ключевыми квалификационными параметрами для средних данных являются следующие характеристики полученной выборки.

1) Надежность источников данных, которые могут быть отфильтрованы относительно релевантности исследуемому вопросу.

2) Достаточность данных для достоверных статистических измерений.

3) Возможность дополнительной семантической разметки, позволяющей обобщить результаты выборки.

Использование medium data дает возможность работать не просто с выборкой, а с выборкой контекстов, которые впоследствии могут быть подвергнуты дополнительному анализу. Контексты дают возможность работы со смыслами, а не только со словами: объединять слова в группы по общему значению (семантические домены), разрешать полисемию. Иначе говоря, полученный в результате первоначального запроса график динамических изменений можно рассматривать не как результат, «диагноз», а как уточнение исследовательской задачи, «симптом».

В целом, порядок работы с данными выглядит примерно следующим образом.

1) Таймлайн, интересующий исследователя, делится на значимые периоды, например, по 50 лет. Для каждого пятидесятилетия выгружаются данные по запросу, которые проходят дополнительную очистку: лемматизация, дизамбигуация, возможный фильтр ненадежных или же смещающих фокус исследования источников, сокращение списка лексем с помощью обобщающих категорий или семантических доменов.

2) Обработанные данные представляются в виде графика, который отображает основные временные точки взлетов и падений.

3) Для анализа изменений употреблений, отражаемых на графике, используются контексты конкретных употреблений из выборки той или иной временной точки. Соотношение этих контекстов с общекультурными трендами этого времени ложится в основу выдвигаемых интерпретаций.

Ниже будет представлено пилотное исследование по культуромике, материалом которого будет сравнение частотных характеристик для разных временных периодов именной группы, состоящей из существительного *дорога* и согласованного с ним прилагательного.

1. Дорога как явление социальной жизни. Подход к исследованию.

Выбор существительного *дорога* для пилотного исследования по культуромике с помощью данных Национального корпуса русского языка мотивирован, с одной стороны, богатой семантикой этого существительного, с другой стороны, его семиотической

значимостью для культуры и общества. Дорога является одной из сложнейших культурных единиц. В мифо-ритуальном сознании перемещение, противостоящее обыденной жизни, сопряжено с комплексом этнических традиций, обычаев и норм [Щепанская 2003]. Использование образа дороги в литературе и живописи является способом ментального освоения пространства и времени, ключевым элементом «культурной географии» — представления в искусстве окружающего пространства и важнейших топосов (см. об этом подробно, например, в [Замятин 2003]). Чрезвычайно важно то, что образ дороги — предметный, артефактный, доступный для изображения или же для именного обозначения, непосредственно связан с действием — с семантикой движения, которое в свое очередь включает в себя компоненты пространства, времени и собственно движения, т. е. одновременного изменения времени и пространства.

В статье «Путь по дороге и бездорожью» Н. Д. Арутюнова рассматривает существительные *путь* и *дорога* как определенные символические комплексы, метафорические смыслы которых дают ключи к интерпретации сложнейших поэтических и прозаических текстов XIX-го и XX-го века [Арутюнова 1999]. Семиотические потенциалы *дороги* и *пути* непосредственно происходят из полисемии, заложенной в этих существительных: «Референтное значение слов *дорога* и *путь* вырисовывается на фоне когнитивного комплекса, объединяющего следующие сведения: 1) среда, в которой осуществляется движение, 2) контактная поверхность (информация о том, из чего сделана или где проложена дорога), 3) ее параметры, 4) отправной и конечный путь движения, 5) движущийся объект — человек, предмет, животное, 6) средство передвижения, 7) направление движения. Они могут быть распределены между такими опорными понятиями, как 1) статический пространственно-предметный денотат, 2) динамический объект, 3) движение в пространстве, 4) движение во времени» [Арутюнова 1999: 4–5].

Представляется, что для моделирования семантической структуры существительного *дорога* чрезвычайно полезным является использование понятия Наблюдателя (см., например, подробно о разных типах Наблюдателя в [Падучева 2006]). Характерно, что, хотя обычно роль Наблюдателя выделяется для описания особенностей нарративного употребления глаголов, так или иначе, включающих в себя перцептивный семантический компонент, в данном случае

она оказывается применима и для различения значений существительного, именно потому, что его семантика непосредственно связана с предикатом движения, требующим заданного ориентира. Заметим, что классический пример, иллюстрирующий необходимость введения роли Наблюдателя, как раз связан с существительным *дорога*: «На дороге появился всадник». Наблюдатель задает точку исходной перцепции возникновения фигуры всадника, однако объектом восприятия является не только сама передвигающаяся фигура, но и то пространство, по которому он передвигается.

Специфика дороги как объекта состоит в том, что она растянута в пространстве и ее физические границы, начало и конец, заведомо превосходят границы восприятия ее человеком в единственной точке пространства, в единственный момент времени. Восприятие дороги как цельного объекта возможно только с помощью ее схематического отображения — карты или вида с высоты, или же с помощью суммирования опыта движения по дороге, т. е. последовательного перемещения из одной части дороги в другие ее части в течение заданного временного интервала. Таким образом, можно говорить о трех позициях Наблюдателя, задающих спектр возможных значений *дороги*[1]:

1) дорога в вещественном смысле — полоса земли, приспособленная для перемещения;

2) дорога в топологическом смысле — вектор, связывающий две точки, путь следования;

3) дорога в физическом смысле перемещения — путешествие.

Каждое из этих значений характеризуется разным набором прилагательных-дескрипторов, активирующим в семантике лексемы те свойства, которые непосредственно доступны Наблюдателю в текущей позиции.

Итак, дорога в значении полосы земли предполагает Наблюдателя, находящегося на этой самой полосе. Это максимально приближенное восприятие дороги Наблюдателем. Дескрипторы описывают прежде всего физические свойства объекта (грязная, мокрая, ухабистая, асфальтовая) или же визуальные характеристики (красивая, белая, туманная).

Наблюдатель дороги в топологическом смысле — это, условно

[1] Ср. Словарь Ожегова, в котором это противопоставление проведено наиболее четко.

говоря, человек, склонившийся над картой. В данном случае, наблюдатель не видит дорогу как физический объект, но способен воспринимать свойства дороги в большом пространстве: ее начало и конец (начальный и конечные пункты), ее специфические свойства, по сравнению с другими дорогами — центральность или дополнительность, ее геометрические характеристики. Основными прилагательными, в данном случае, являются прилагательные направления (*Московская, Варшавская*), или же прилагательные, в которых заложено сравнение свойств дороги с другими дорогами этой местности (*окружная, центральная, объездная, боковая*), при этом сам охват сопоставления зависит от позиции наблюдателя, масштаба той карты, которая подразумевается.

Наконец, описание дороги как путешествия требует максимальной вовлеченности «субъекта сознания» [Падучева 2006], накапливающего опыт перемещения: кроме собственно временных характеристик (*долгая дорога*), прилагательные, использующиеся в данном случае, отражают не столько свойства дороги, сколько метонимический перенос состояния путешественника (*унылая дорога, утомительная дорога, веселая дорога*).

Метафорический перенос семантики дороги на человеческую жизнь или развивающийся во времени процесс, связан, как пишет Арутюнова, с заменой темпоральной модели жизни на пространственную. «Динамика жизненных событий выражается метафорой дороги» (Арутюнова 1999:11). При этом прилагательные могут быть атрибутами уже переносного значения (*жизненная дорога, трудовая дорога, сознательная дорога, бабья дорога*), или же участвовать в метафорическом переносе вместе со своей вершиной (*широкая дорога, кривая дорога, незнакомая дорога*).

Таким образом, сам выбор прилагательного может служить хорошим контекстом, классифицирующим основные значения лексемы *дорога*. Подобная классификация имеет прямое отношение к задаче исследования изменений в обществе с помощью Национального корпуса: общественное развитие непосредственно связано с мобильностью населения, которая, в свою очередь, зависит от целого ряда параметров: связанности между собой разных населенных пунктов, преодолимости этих расстояний, качества дорог, скорости и сложности путешествия. Разные параметры, определяющие динамику общественной мобильности,

связаны с разными значениями существительного дорога. Таким образом, наблюдая изменения частотности разных значений существительного *дорога*, мы можем сделать выводы (или же выдвинуть гипотезы) об изменении значимости различных параметров мобильности в разные периоды общественного развития.

Ниже будут представлены результаты корпусного исследования частотности существительного *дорога*, сопровождаемого прилагательным, в XIX и XX веках: в (2) будут описаны методы сбора данных и их обработки, в (3) будут представлены основные результаты анализа.

2. Методы сбора и обработки данных

В целом, метод сбора данных состоял из четырех этапов:

1) Выгрузка данных по выделенным периодам таймлайна и очистка полученной выборки от шума.

2) Разделение данных на семантические классы.

3) Статистический анализ близости полученных классов.

Рассмотрим более подробно основные задачи каждого из этапов и те решения, которые принимались в проблемных случаях.

2.1. Выгрузка и очистка данных

Для сравнения изменения частотности существительного *дорога* в сочетании с прилагательным были выбраны 8 временных периодов, по которым были составлены соответствующие подкорпуса. XVIII век был представлен целиком отдельным подкорпусом общим объемом более 3,7 млн словоформ, XIX век был поделен на два подкорпуса: с 1801 по 1850 год объемом 11,9 млн словоформ и с 1851 по 1900 год объемом 39,4 млн словоформ. В XX веке были выделены 5 периодов по 20 лет: с 1901 по 1920 (21 млн словоформ), с 1921 по 1940 (31,9 млн словоформ), с 1941 по 1960 (19 млн словоформ), с 1961 по 1980 (24,6 млн словоформ) и с 1981 по 2000 (40,7 млн словоформ). Ниже в таблице 1 представлены обобщенные сведения о каждом рассматриваемом периоде и нормализованной частоте рассматриваемого сочетания (слово1: прилагательное, слово2: дорога). Цифры в таблице отражают количество конструкций, полученных после очистки данных от шума.

	Период	Объем подкорпуса	Общая частота конструкции в ipm	Доля конструкции на период
1	XVIII век	3795391	62,25	9%
2	1801-1850	11917009	109,96	16%
3	1851-1900	39464065	97,22	14%
4	1901-1920	21049150	76,62	11%
5	1921-1940	31934949	93,87	14%
6	1941-1960	19026302	99,01	14%
7	1961-1980	24666067	90,67	13%
8	1981-2000	40741373	61,91	9%

Таблица 1. Объемы подкорпусов для каждого временного периода и общая частота конструкции «прилагательное+дорога»

Медиана значений частоты конструкции «прилагательное+дорога» в восьми временных периодах составляет 92,2 ipm. Из последнего столбца таблицы видно, что конструкция распределена более или менее равномерно, за исключением 1-го, 4-го, и 8-го периодов. По-видимому, меньшая представленность существительного *дорога* с определением в XVIII-м веке и в начале XX-го свидетельствует о несбалансированности корпуса. Действительно, можно предположить, что существительное *дорога* чаще встречается в художественных текстах. Для того, чтобы избежать омонимии с прилагательным *дорогой*, эта гипотеза была проверена с помощью косвенной формы *дороге*: для художественных текстов в целом по корпусу частотность этой формы составляет 145,6 ipm, а для нехудожественных текстов почти в два раза меньше — 88,39 ipm. Таким образом, можно предположить, что падение частотности конструкции может быть связано с недостаточной представленностью в корпусе художественной литературы в этот период. В таблице 2 показана доля художественных текстов в каждый из рассматриваемых периодов, соотнесенная с долей рассматриваемой конструкции на период от общей выборки. Эти два ряда имеют положительную корреляцию, коэффициент которой равен 0,62.

	Период	Доля художественных текстов	Доля конструкции на период
1	XVIII век	18%	9%
2	1801-1850	37%	16%
3	1851-1900	39%	14%
4	1901-1920	26%	11%
5	1921-1940	38%	14%
6	1941-1960	42%	14%
7	1961-1980	39%	13%
8	1981-2000	38%	9%

Таблица 2. Соотнесенность доли художественных текстов в подкорпусе с долей конструкции на период времени

Из таблицы 2 видно, что действительно в первом и четвертом периодах доля художественных текстов меньше, и доля рассматриваемых конструкций тоже меньше. Кроме этого, обращает на себя внимание самый последний временной период с 1981 по 2000 год. Доля художественных текстов уменьшилась в этом периоде всего лишь на 1% по сравнению с предыдущим, а конструкция стала реже на 4%. Таким образом, можно сделать предварительный вывод, что к концу XX-го века частотность конструкции снижается. Однако для того, чтобы предложить содержательную интерпретацию этому факту, необходимо более подробно рассмотреть то, какие из выделенных выше базовых значений *дороги* стали встречаться реже. К этому вопросу мы еще вернемся ниже.

Как уже было сказано, существительное *дорога* в большой части своей парадигмы омонимична кратким и полным формам весьма частотного прилагательного *дорогой*: в единственном числе мы имеем только две уникальные формы — *дороге* и *дорогу* (винительный падеж краткой формы прилагательного *дорогой* можно игнорировать как очень редкий и устаревший в рассматриваемых временных периодах), а во множественном числе таких форм имеется три — *дорогам, дорогами, дорогах*. Омонимия снималась полуавтоматическим способом. Сначала были установлены фильтры на согласование прилагательного и существительного по падежу и числу. Однако в ряде контекстов, таких как в предложении (1), омонимия снималась вручную:

1) *Ах, я и забыла сказать, что дедушка Федор Андреевич Толстой, уезжая за границу снимать свои катаракты, вместе с дачей наделал отцу моему еще много подарков: во-первых, он купил для него чудную верховую лошадь, жеребца арабской породы, Гектора; потом подарил еще ньюфаундлендскую собаку Зюлему, которая у него всегда жила при этой даче; еще для балкона* ***большой дорогой*** *телескоп и наконец шестивесельный морской катер, чтобы кататься в нем по Неве.* [М. Ф. Каменская. Воспоминания (1894)].

Всего после очистки данных от шума объем рабочей выборки составил 15845 предложений с конструкцией, включающей в себя существительное *дорога* с предшествующим ему согласованным прилагательным.

2.2. Деление конструкций на семантические классы.

Всего в выборке встретилось 1590 уникальных прилагательных. Безусловно, семантический анализ данных такого объема требует очень больших временных ресурсов. При этом в соответствии с законом Ципфа лишь три прилагательных имели частотность более 1000 вхождений, и лишь семнадцать прилагательных более 100 вхождений. Стандартный подход с использованием простого порога отсечения по частотности не соответствует поставленной задаче — во-первых, интересующие нас значения существительного *дорога* могут быть в основном отражены в конструкциях «хвоста» выборки, во-вторых, такой способ игнорирует диахроническую лексическую вариативность, вытеснение одного прилагательного другим. Поэтому было выбрано более трудоемкое решение — объединить близкие по значению прилагательные в семантические классы таким образом, чтобы прилагательные одного и того же класса описывали одни и те же свойства существительного *дорога*. Из этого следовала необходимость снятия неоднозначности прилагательных. Например, словосочетание прямая дорога может обозначать маршрут движения, как в (2):

2) *Возвращались* ***прямой дорогой*** *на Горелово, минуя полустанок.* [Борис Васильев. Дом, который построил Дед (1990–2000)].

Или же это же словосочетание может обозначать дорогу в переносном значении, как метафору саморазвивающегося процесса (3):

3) *Очевидно, что здесь избыток денег и рост цен ведут* ***прямой дорогой*** *к свертыванию производства.* [Юлий Ольсевич. Инфляция: лекарство или яд // «Огонек». № 14, 1991].

В итоге были выделены 19 семантических классов. Они неравнозначны по своему содержанию и широте охвата, прежде всего потому, что кроме собственно семантики учитывались также и экстралингвистические факторы, например, обозначение специфического явления или объекта, который может иметь свои собственные уникальные свойства и динамику частотности. Так, словосочетание *железная дорога* (вместе с несколькими синонимичными выражениями) было выделено в отдельный класс, несмотря на то, что также были выделены семантические классы «способ передвижения», «средство передвижения». Ниже в таблице 3 перечислены все выделенные семантические классы в алфавитном порядке, они соотнесены с тем значением *дороги*, которые они характеризуют, даны основания выделения класса и примеры прилагательных.

Таблица 3. Состав семантических классов

Семантический класс	**Тип значения**	**Основания выделения класса**	**Примеры самых частотных прилагательных класса**
Время перемещения	Дорога как полоса	Прилагательные времени года: проходимость дороги напрямую была связана с сезонностью	Зимняя, летняя, весенняя
Железная	Дорога как полоса, а также другие значения дороги, кроме метафорического	Устойчивое словосочетание, определяющее отдельный от собственно дороги, но связанный с ним объект	Железная, чугунная, конно-железная

Семантический класс	Тип значения	Основания выделения класса	Примеры самых частотных прилагательных класса
Локальная	Дорога как вектор	Прилагательные, соотносящие дорогу с локальными населенными пунктами или локальными точками	Проселочная, колхозная, городская
Маршрут	Дорога как перемещение	Прилагательные, характеризующие отношение субъекта к траектории пути	Знакомая, прямая, прежняя, ближайшая
Местность	Дорога как полоса	Прилагательные, характеризующие пространство, окружающее дорогу	Лесная, горная, полевая, степная
Метафора	Переносное значение	Прилагательные, расширяющие контекст переносного значения	Ложная, настоящая, новая, открытая
Направление	Дорога как вектор	Прилагательные, соотносящие дорогу с локусами, которые она связывает	Николаевская, Смоленская, Военно-грузинская
Перцептивные свойства	Дорога как полоса	Прилагательные, определяющие визуальные свойства дороги	Широкая, извилистая, узкая
Плохая	Дорога как полоса	Прилагательные отрицательной оценки, связанные с проходимостью дороги и качеством дорожного покрытия	Плохая, разбитая, грязная

Семантический класс	**Тип значения**	**Основания выделения класса**	**Примеры самых частотных прилагательных класса**
Путешествие	Дорогая как перемещение субъекта	Прилагательные, параметризующие ситуацию путешествия: его продолжительность и состояние субъекта	Дальняя, трудная, обратная
Способ перемещения	Дорога как полоса	Прилагательные, транспонирующие понятие дороги в нестандартную среду	Подземная, канатная, воздушная
Средство перемещения	Дорога как полоса	Прилагательные, определяющий свойства дороги через соотнесение ее с соответствующим транспортом	Автомобильная, военная, санная
Структура покрытия	Дорога как полоса	Прилагательные, характеризующие физические свойства дорожного покрытия	Пыльная, немощеная, грунтовая
Топография	Дорога как вектор	Прилагательные, характеризующие свойства дороги как вектора или кривой на карте, в том числе и относительно других объектов карты	Новая, боковая, окружная

Семантический класс	Тип значения	Основания выделения класса	Примеры самых частотных прилагательных класса
Трафик	Дорога как полоса	Прилагательные, характеризующие заполненность дороги перемещающимися объектами или их отсутствием	Пустая, свободная
Хорошая	Дорога как полоса	Прилагательные положительной оценки, связанные с проходимостью дороги и качеством дорожного покрытия	Хорошая, проезжая, ровная
Цвет	Дорогая как полоса	Прилагательные визуального восприятия, характеризующие дорогу по цвету	Белая, черная, рыжая
Центральная	Дорога как вектор	Прилагательные, характеризующие дорогу как значимую, соединяющие два важных локуса, большую по протяженности и интенсивности движения	Большая, столбовая, почтовая, шоссейная
Эпитет	Дорога как полоса	Прилагательные, характеризующие дорогу как часть пейзажа	Пустынная, снежная, родная

Выделенные семантические классы неравнозначны по количеству элементов и их совокупной частотности. Наиболее широким, как и следовало ожидать, является класс «направление», объединяющий 526 прилагательных, наиболее узким по количеству элементов, кроме «железной дороги», представляющей собой отдельное словосочетание, является класс «центральная», в который вошло 60 прилагательных. В таблице 4 для каждого класса представлены данные по общему количеству совокупных вхождений, по количеству входящих в класс прилагательных и отношение этих двух показателей: среднее число вхождений на одно прилагательное. Медиана среднего вхождений для всех классов равна 6,4. Таким образом, можно оценить более и менее «плотные» классы. Напомним также, что многие прилагательные входят в разные классы, поскольку могут в разных контекстах иметь разное значение.

Таблица 4. Количество элементов в семантических классах

Семантический класс	**Совокупное количество вхождений**	**Количество прилагательных**	**Количество вхождений на одно прилагательное**
эпитет	182	60	3,0
направление	1723	526	3,3
цвет	168	39	4,3
плохая	412	91	4,5
трафик	75	16	4,7
маршрут	568	118	4,8
средство перемещения	214	41	5,2
способ перемещения	220	42	5,2
топография	301	51	5,9
метафора	1442	224	6,4

Семантический класс	Совокупное количество вхождений	Количество прилагательных	Количество вхождений на одно прилагательное
время перемещения	99	14	7,1
путешествие	1188	161	7,4
перцептивные свойства	298	34	8,8
локальная	487	53	9,2
структура покрытия	835	81	10,3
хорошая	571	55	10,4
местность	453	25	18,1
центральная	1443	60	24,1
железная	5154	11	468,5

Далее каждый из семантических классов был представлен совокупностью своих вхождений в 8 выбранных временных периодах, нормализованных относительно объема подкорпусов. Эти данные легли в основу дальнейшего анализа. Они приводятся в таблице 5.

Таблица 5. Распределение частотности семантических классов по периодам в ipm

Названия строк	XVIII век	1801-1850	1851-1900	1901-1920	1921-1940	1941-1960	1961-1980	1981-2000
маршрут	4,48	6,71	2,74	1,33	3,16	3,78	3,45	1,89
хорошая	3,69	3,69	4,94	1,47	3,01	3,36	2,80	1,55
плохая	4,74	4,53	2,64	1,24	1,44	1,58	2,11	2,01
локальная	0,53	4,03	2,31	0,67	2,85	2,63	3,49	2,58
перцептивные свойства	1,32	2,85	1,42	0,71	1,28	1,73	1,66	1,79
топография	1,84	2,35	0,89	0,90	1,16	1,79	1,50	2,55

Названия строк	XVIII век	1801-1850	1851-1900	1901-1920	1921-1940	1941-1960	1961-1980	1981-2000
местность	0,00	1,76	1,52	0,81	2,69	4,47	3,28	2,53
средство перемещения	0,00	1,01	0,86	0,14	1,10	2,26	1,62	1,15
структура покрытия	0,26	3,10	2,81	1,71	5,32	6,67	6,24	4,88
способ перемещения	0,26	0,17	0,38	0,57	1,60	1,68	3,36	0,64
эпитет	0,00	0,76	0,63	0,57	1,66	1,16	0,97	0,91
время перемещения	1,32	1,26	0,51	0,14	0,34	0,58	0,49	0,54
цвет	0,00	0,17	0,20	0,71	1,53	1,52	1,42	0,74
трафик	0,00	0,42	0,15	0,10	0,53	0,47	0,69	0,47
путешествие	2,11	6,80	4,51	2,28	6,48	6,83	9,89	7,17
метафора	6,59	6,96	11,02	3,23	7,86	8,51	8,39	5,18
центральная	15,02	24,92	12,54	3,80	6,42	6,62	3,77	2,21
направление	17,13	25,68	10,19	10,45	8,24	8,41	4,91	4,57
железная	0,00	7,47	32,33	42,14	32,69	30,17	26,27	15,61

Динамика хронологических изменений лучше может быть показана на графике, однако график, отражающий поведение девятнадцати кривых, анализировать затруднительно. Поэтому было принято решение кластеризовать данные, объединив похожие по поведению тренды. Это не просто техническое решение — то, какие классы оказываются в одном кластере, само по себе является фактом, требующим интерпретации. Ниже будут представлены результаты иерархической кластеризации.

2.3. Статистический анализ данных

Кластеризация семантических классов выполнялась следующим образом. Каждый из семантических классов был представлен как вектор значений. Между векторами было посчитано эвклидово расстояние, на основании которого далее была произведена иерархическая кластеризация. Результаты кластеризации представлены на рисунке 1.

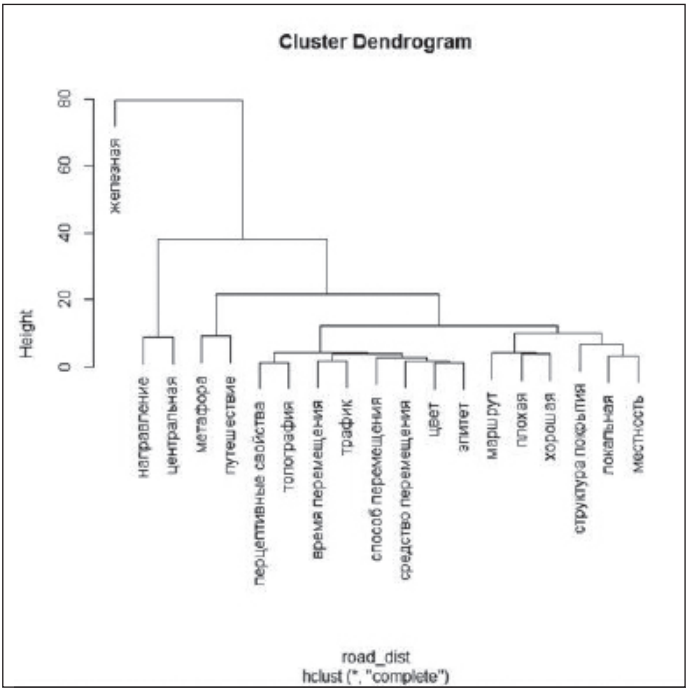
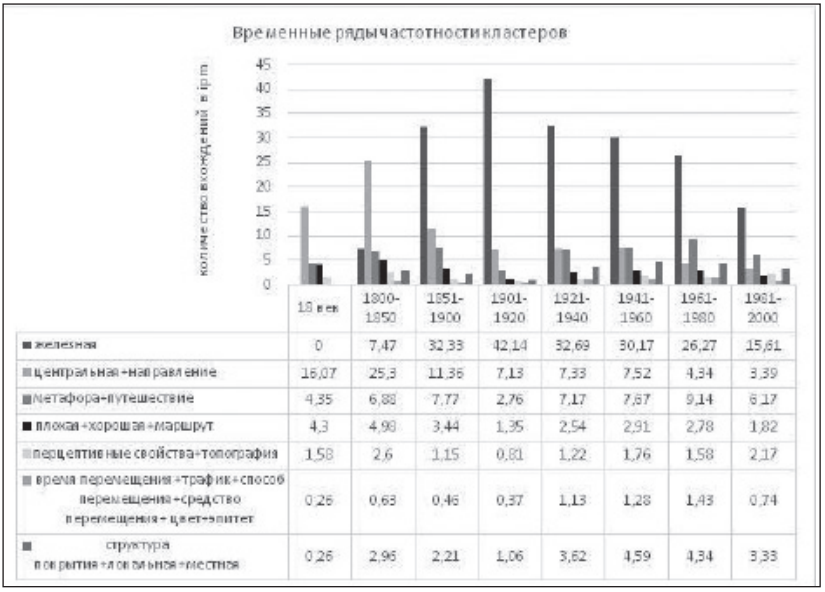


Рисунок 1. Иерархическая классификация семантических классов

Из рисунка 1 видно, что «железная дорога» очень сильно отличается от всех остальных семантических классов: деление самого верхнего уровня противопоставляет класс «железная дорога» всем остальным семантическим классам. На следующем уровне отделяется пара классов «направление» и «центральная». Еще ниже таким классом, противопоставленным всем остальным, является пара «метафора» и «путешествие». Следует отметить, что кластеризацию семантических классов с невысоким уровнем частотности нельзя считать очень надежной. Тем не менее, и на этом уровне мы видим несколько объединений, которые связаны не только цифрами частотности, но и семантически, т. е. такие, в которых используются близкие по значению прилагательные, или точнее прилагательные актуализируют близкие свойства значения существительного *дорога.* Это такие классы как «перцептивные свойства» и «топография», «цвет» и «эпитет», оценочные классы «плохая» и «хорошая»,

классы, с помощью которых описываются небольшие дороги — «локальная» и «местность». Вообще говоря, объединение названных классов в общие кластеры не является тривиальным фактом. Так, например, вполне могло быть так, что со временем негативная оценка дорог падает, а положительная возрастает (или наоборот). Однако оказывается, что ругают дороги примерно так же часто, как и хвалят. Распределение частотностей объединенных кластеров по временным периодам представлены на диаграмме 1.



	18 век	1800-1850	1851-1900	1901-1920	1921-1940	1941-1960	1961-1980	1961-2000
железная	0	7,47	32,33	42,14	32,69	30,17	26,27	15,61
центральная+направление	16,07	25,3	11,36	7,13	7,33	7,52	4,34	3,39
метафора+путешествие	4,35	6,88	7,77	2,76	7,17	7,67	9,14	6,17
плохая+хорошая+маршрут	4,3	4,98	3,44	1,35	2,54	2,91	2,78	1,82
перцептивные свойства+топография	1,58	2,6	1,15	0,81	1,22	1,76	1,58	2,17
время перемещения+трафик+способ перемещения+средство перемещения+цвет+эпитет	0,26	0,63	0,46	0,37	1,13	1,28	1,43	0,74
структура покрытия+плохая+местная	0,26	2,95	2,21	1,06	3,62	4,59	4,34	3,33

Диаграмма 1. Распределение частотностей кластеров

Представленная диаграмма является продуктом последовательной подготовки medium data для анализа в рамках задач культуромики, это и есть те данные, которые имеет смысл проецировать на наши знания об изменениях в обществе в определенную эпоху.

3. Культуромика российских дорог: анализ данных.

Стартовая гипотеза о том, как соотнесены изменения частотности словосочетаний, состоящих из имени прилагательного и существительного *дорога*, с изменениями, происходящими в обществе, состо-

ит в том, что общественное развитие, прежде всего экономическое, связывается с интенсификацией мобильности населения, а стало быть, мы можем ожидать и увеличения частотности использования лексемы *дорога*. В очень огрубленном виде это можно представить так: чем больше люди путешествуют и перемещаются, тем чаще они об этом пишут в текстах соответствующей эпохи. Однако при более тщательном рассмотрении графиков оказывается, что, во-первых, ни один из семантических классов не показывает линейного роста, а во-вторых, что семантические классы (или их кластеры) в целом очень различаются по своему поведению. Есть ряд кластеров, главным трендом которых является падение, например, кластер «центральная» + «направление», или же кластер оценочных прилагательных «плохая» + «хорошая». Есть, наоборот, кластеры, которые растут, например, кластер «метафора» + «путешествие». Далее мы рассмотрим более подробно наиболее значимые кластеры.

3.1. «Железная дорога»

Семантический класс «железная дорога», как уже было сказано, противопоставлен всем остальным кластером. Более того, этот класс составляет почти треть от всех вхождений выборки. Безусловно, и график изменения частотности этого класса является очень показательным, см. диаграмму 2.

Диаграмма 2. График изменения частотности класса «железная дорога»

Во-первых, обращает на себя внимание феноменальный взлет частотности семантического класса «железная дорога» во вторую половину XIX века, мы видим, что количество вхождений увеличилось более, чем в 4 раза. Это при том, что в этот период корпус обширно представлен художественной литературой (39% текстов). Действительно, по-видимому, для носителя русского языка второй половины XIX века железная дорога сразу стала чрезвычайно мощным символом искусства, отражающим современную эпоху. Тут можно вспомнить и стихотворение Некрасова «Железная дорога» (1864 год), и начало романа Достоевского «Идиот» (1868 год), и, конечно же, «Анну Каренину» Толстого (1878 год)[1]. Характерно, что наиболее высоких показателей этот семантический класс достигает в период 1901–1920 годов, ср. знаменитый 1913 год, который рассматривался как эталонный год экономического развития дореволюционной России. Однако не менее интересным представляется последующее падение графика. К последнему временному периоду (1981–2000) число вхождений семантического класса «железная дорога» в два раза меньше аналогичных показателей второй половины XIX века. Иначе говоря, можно предположить, что к концу двадцатого века железная дорога, во-первых, утратила свое символическое значение, а во-вторых, перестала быть ключевым фактором транспортного развития. Заметим, что разница предпоследнего периода (1961–1980) и последнего составляет 40%. Замечательным образом изменения частотности вхождений словосочетания «железная дорога» (и близких, но существенно более редких синонимических выражений) прекрасно соотносится с абсолютно экономическим графиком, отражающим объем гражданских авиаперевозок в СССР, который начался существенно расти именно в 70-е годы XX-го столетия, см. рисунок 2[2].

[1] Железная дорога стала символом не только в литературе, ее можно обнаружить на полотнах импрессионистов, ср. также знаменитое «прибытие поезда» братьев Люмьер.

[2] Источник рисунка: аналитическая записка рейтингового агентства Эксперт http://raexpert.ru/docbank/376/cf4/340/a2b12706903f30f88c522e4.pdf

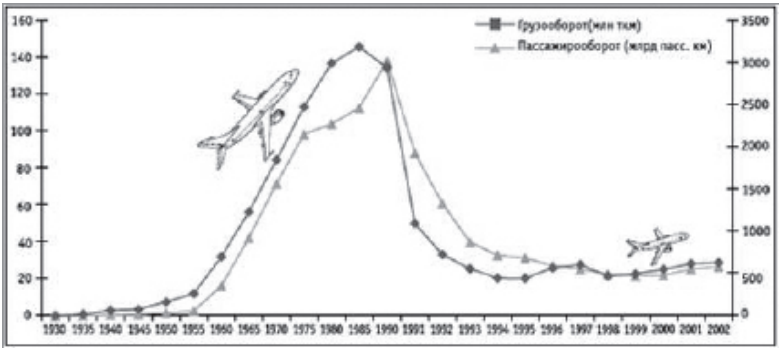


Рисунок 2. График роста пассажирооборота авиаперевозок

При этом железная дорога так или иначе соотносится с дорогой как физическим объектом, она чрезвычайно материальна, железнодорожные пути — это распространенный мотив живописи XX века. Рост воздушного перемещения вытесняет «железную дорогу», не как собственно дорогу, но как опыт путешествия. Таким образом, уже только по графику изменения семантического класса «железная дорога» можно увидеть, как меняется концепт мобильности в общественном сознании XIX и XX-го веков. В XIX веке он чрезвычайно актуализирован, связан прежде всего со строительством железных дорог, связывающих разные точки и локации друг с другом. В XX-м веке меняется сам характер мобильности, перемещение в гораздо меньшей степени зависит от конкретной дороги, оно может быть сложным, можно ехать разным транспортом, можно сидеть сутки в аэропорту, и это тоже будет считаться «изнурительной дорогой». Сеть железных дорог уже не играет такой ключевой роли, однако, как будет показано ниже, показатели частотности семантического класса «путешествия» имеют совсем иную динамику.

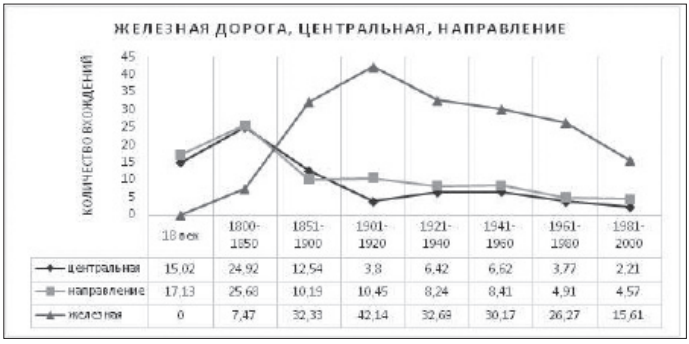
3.2. «Центральная» и «направление»

Два класса, тесно связанных друг с другом, «центральная» и «направление», — это классы, характеризующие дороги, так сказать, государственного значения. С точки зрения семантики, эти классы представлены совершенно разными прилагательными: са-

мое частотное прилагательное в классе «центральная» это *большая*, а класс «направление» представлен относительными прилагательными, связанными с определенными топонимами (*Московская, Смоленская* и т. д.), однако динамика изменения частотностей этих классов очень и очень близка. Ниже на диаграмме 3 представлен график изменения частотности этих двух классов. На графике также отражена и кривая класса «железная дорога». Соотношение этих трех классов чрезвычайно важно для того, чтобы увидеть то, как различалось представление о социальной мобильности в разные эпохи. Прежде чем анализировать графики, остановимся чуть более подробно на семантическом классе «центральной дороги». В словаре Даля в статье *дорога* указано следующее противопоставление:

Дороги бывают: большие, т. е. общие почтовые и торговые; малые, уездные, от городка до городка, в стороне от больших; проселочные, от селенья к селенью, в стороне от малых.

Противопоставление больших и малых дорог в XVIII–XIX веках является ключевым: большие дороги — это дороги, обеспечивающие товарный обмен, мобильность населения. Они же назывались также почтовыми, столбовыми. Это дороги путешествий, длинных перемещений. Кроме собственно *большой, почтовой* и *столбовой*, в этот класс входят прилагательные *главная, караванная, шоссейная, основная, стратегическая* и т. д. Иначе говоря, основные дороги в стране могут называться либо функционально, с точки зрения их предназначения (*почтовая*), либо привязываться к названиям ключевых местностей и населенных пунктов на карте (класс «направления»), либо просто выделяться с помощью оценочного прилагательного (*большая, главная, основная*). Важно то, что все эти характеристики выделяют некоторый класс объектов (ключевые транспортные артерии), который, как и железная дорога, обладает собственной динамикой развития.



	18 век	1800-1850	1851-1900	1901-1920	1921-1940	1941-1960	1961-1980	1981-2000
центральная	15,02	24,92	12,54	3,8	6,42	6,62	3,77	2,21
направление	17,13	25,68	10,19	10,45	8,24	8,41	4,91	4,57
железная	0	7,47	32,33	42,14	32,69	30,17	26,27	15,61

Диаграмма 3. Распределение семантических классов «центральная», «направление» и «железная дорога»

Из диаграммы видно, что значимость больших дорог в общественном сознании в XVIII веке сопоставима со значимостью железных дорог в середине XIX-го (и, кстати, со значимостью железных дорог в конце XX-го). Мы видим, как до середины XIX-го века показатели частотности центральных дорог увеличиваются, их рост опять же сравним с ростом частотности железной дороги на следующем этапе. Вторая половина XIX-го века показывает столь же резкое падение частотности: связь между отдаленными точками теперь осуществляется с помощью поездов. Значимость падения частотности семантического класса «центральных» дорог в период с 1901 по 1920 год сложно оценить, выше уже говорилось о несбалансированности корпуса в этот период. Однако интересно, что со второй половины XX века тренд частотности вхождений железных дорог и центральных дорог опять идет в общем направлении. Как уже было сказано выше, маршруты становятся гораздо сложнее и в меньшей степени связанными с собственно дорожным полотном или железнодорожными путями. Весьма примечателен поэтому тот факт, что семантический класс со значением путешествия, напротив, повышает свою частотность, и, что еще более интересно, так же ведет себя и класс «метафоры», который попадает с «путешествием» в один кластер.

3.3. «Метафора», «путешествие» и «способ перемещения»

Семантический класс «путешествие» объединяет употребления существительного *дорога* в значении перемещения субъекта

(наблюдателя) из точки начала его движения в точку его окончания. Фактически это метонимический перенос с собственно физического объекта — дороги как полосы земли — на обозначение действия перемещения по дороге или на состояние человека, перемещающегося по дороге. Это значение переноса актуализируется в том числе и с помощью употребляемых в этом контексте прилагательных, например, прилагательных времени (*продолжительная дорога, шестидневная дорога*) или прилагательных оценки (*приятная дорога, утомительная дорога, легкая дорога*), которые, на самом деле, определяют состояние путешественника, а не собственно физический объект. Весьма интересны в этом контексте такие прилагательные, которые так или иначе характеризуют опыт путешественника — совместная дорога, обратная дорога. Можно было бы предположить, что частотность класса «путешествия» должна была бы оставаться более-менее стабильной, этот класс сложнее напрямую связать с экстралингвистическими факторами, чем класс «железной дороги» и даже класс «центральной дороги». Однако мы наблюдаем интересный эффект во второй половине XX-го века (см. диаграмму 4):

Диаграмма 4. Распределение семантических классов «железная дорога», «путешествие», «центральная», «направление»

Частотность семантического класса «путешествие» растет в период 1961–1980, т. е. именно тогда, когда и «железные дороги», и кластер «центральных» дорог начинают снижаться. Иначе гово-

ря, перемещения становятся более сложными и менее связанными с физическим передвижением по дороге, и важность обозначения совокупного опыта такого перемещения возрастает. Любопытным образом именно в этот период вырастают показатели семантического класса «способ перемещения» — другого метонимического класса значений, в которых *дорогой* обозначается не столько физический объект — полоса земли, сколько тем или иным образом обеспеченная возможность для перемещения — *канатная дорога, морская дорога, речная дорога, воздушная дорога*, см. диаграмму 5.

Диаграмма 5. Распределение семантических классов «путешествие» и «способ перемещения»

Мы видим, что динамика и пропорции изменений кривых частотностей классов «путешествие» и «способ перемещения» практически одинаковы. Снижение частотностей в последнее двадцатилетие XX-го века также требует интерпретации. Предварительная и весьма спекулятивная гипотеза состоит в том, что по мере снижения значимости дороги как физического объекта для мобильности в целом, метонимические отношения между источником и целью переноса утрачивают свою близость. Сама идея путешествия, перемещения все меньше связывается с некоей (пусть абстрактной) непрерывной протяженной линией, связывающей пункт отбытия с пунктом назначения. Очень характерны в этом смысле изменения лексического состава класса «способ передвижения». В подкорпусе периода с 1961 по 1980 год вдоль воображаемой абстрактной линии

передвигаются все виды транспорта (и не только транспорта, как в примере (6)):

4) *Тут речная дорога была, а вот там, — говорил он, — можно было, не зная этих боров, легко заблудиться и к медведям в гости попасть, и ребят в старину пугали рысями и медведями* [Л. Ф. Зуров. Иван-да-марья (1956–1969)].

5) *Спрашивается, возможно ли проложить бесперебойную и безаварийную* ***воздушную дорогу*** *через грозы в данный район?* [Даниил Гранин. Иду на грозу (1962)].

6) *Разве не шедевр этот крик журавлей и их величавый перелет по неизменным в течение многих тысячелетий* ***воздушным дорогам****?* [К. Г. Паустовский. Наедине с осенью (1963)].

7) *За ужином принесли эту радиограмму, когда мы обсуждали, поедая блинчики с мясом, варианты встречи Нового года — семидесятого года; решили как раз отойти в сторонку от главных* ***морских дорог*** *в океане, лечь там в дрейф и встречать Новый год без лишней нервотрепки, и вдруг — Касабланка…* [Виктор Конецкий. Вчерашние заботы (1979)].

В подкорпусе 1981–2000 годов из всего многообразия линий перемещения, подобных дороге, остается только *канатная дорога*, которая, с одной стороны, превратилась в устойчивое словосочетание аналогичное словосочетанию *железная дорога*, а с другой стороны, в буквальном смысле реализует идею линии, связывающей две точки. Похожий эффект наблюдается и в изменениях лексического состава семантического класса «путешествия» — уменьшается частотность прилагательных, объединяющих семантику протяженности во времени с протяженностью в пространстве — *длинная, дальняя* и *долгая*.

Семантический класс «метафора», объединяющий употребление существительного *дорога* в переносных значениях, по количеству вхождений и их дистрибуции во временных периодах входит в один кластер с классом «путешествие», эти два класса ведут себя очень похоже. Нельзя не заметить то, что между этими классами есть и семантическая близость: фактически метафорический перенос от «перемещения, изменения точки нахождения в пространстве» к «развитию, изменению состояния во времени» происходит непосредственно через значение *дороги* как «путешествия». Опыт

путешественника транспонируется в экзистенциальный опыт. Тот факт, что вхождения *дороги* как «путешествия» и *дороги* в переносном смысле составляют примерно близкие по объему доли от совокупности вхождений *дороги* во всех значениях, представляется весьма примечательным. Однако заметим, что у семантического класса «метафора» есть своя собственная специфика диахронической дистрибуции. Более того, как видно из диаграммы 6, кривым двух классов соответствуют разные линейные тренды: тренды класса «путешествия» — это тренд повышающий, тогда как тренд «метафоры» — это тренд стабильный с небольшим снижением.

Диаграмма 6. Распределение классов «метафора» и «путешествие»

Очевидно, что разница трендов происходит из разницы стартовых позиций двух семантических классов: количество вхождений *дороги* в переносных значениях в XVIII веке не сильно отличается от конца XX-го века, в то время как распространенность использования *дороги* в значении «путешествия» увеличивается почти в 4 раза. Интерпретация этого факта пока не вполне понятна, однако стоит отметить, что он несколько противоречит намеченной выше идее того, что метонимический перенос служит промежуточным звеном метафорического переноса (полоса земли > путешествие как перемещение по полосе > жизнь как перемещение во времени). Однако более всего обращает на себя внимание взлет метафорических употреблений во второй половине XIX-го века. Анализ лексического состава семантического класса «метафора» этого периода времени

показывает, что есть целый ряд прилагательных, которые именно в этот период встречаются десятки раз, а в остальные, в том числе и последующие периоды — единицы: это *новая, ложная, настоящая*, т. е. можно предположить, что относительная высокая частотность вхождений этого класса во второй половине XIX века происходит именно благодаря названным прилагательным. Семантически эти три прилагательных объединены идеей выбора, вернее, оценкой сделанного или предстоящего выбора, которая противопоставляет один из путей/направлений развития другому с точки зрения его соответствия конечной цели.

8) *Сколько раз думал я, глядя на Веру: да, прав Гете: «Добрый человек в неясном своем стремлении всегда чувствует, где* ***настоящая дорога****».* [И. С. Тургенев. Фауст: рассказ в девяти письмах (1856)].

9) *По упомянутому объяснению, Белинский возродил русскую литературу, в особенности критику, и поставил ее на* ***настоящую дорогу****, так что если бы она и дальше пошла по этой дороге, то мы имели бы и в настоящее время литературу и критику как следует — не в упадке, а на надлежащей высоте.* [М. А. Антонович. Причины неудовлетворительного состояния нашей литературы (1876)].

10) *Поднялся шум, и грохот, и ярые возгласы. — Россия на* ***ложной дороге****! — кричал какой-то политик.* [Н. Г. Помяловский. Молотов (1861)].

11) *Он иногда восклицал наедине со мной и почти с отчаянием про себя, что он — «так необразован, что он на такой* ***ложной дороге****!* [Ф. М. Достоевский. Подросток (1875)].

12) *Я невольно увлекался его страстями; но был слишком горд, чтобы идти по его следам, и слишком молод и несамостоятелен, чтобы избрать* ***новую дорогу****.* [Л. Н. Толстой. Отрочество (1854)].

13) *Разумеется, нет людей совершенных и непогрешимых, и потому, если мы сами не чувствуем себя в силах проложить* ***новую дорогу*** *и вести по ней других, то нам, чтобы не стоять бесполезно на месте, нужно идти за кем-нибудь и для этого выбрать себе руководителя.* [Н. А. Добролюбов. Всероссийские иллюзии, разрушаемые розгами (1860)].

Интересно, что в XX-м веке метафора дороги как жизненного пути уже не связывается с выбором, остаются лишь те прилагательные (имеющие высокие показатели и во второй половине XIX-го века), которые связаны с центральностью и значимостью:

14) *Перед каждым человеком теперь открывается* ***широкая дорога*** *для его стремлений, для его счастья и благополучия.* [А. С. Макаренко. Книга для родителей (1937)].

15) *Борис Федорович — человек очень талантливый, окончил в свое время Академию художеств, перед ним была открыта* ***широкая дорога*** *к славе, известности, к тому, что, по мнению Китова, только и достойно уважения.* [Анатолий Рыбаков. Кортик (1946–1948)].

16) *Нет, не здесь проходит вдаль* ***большая дорога*** *жизни — не в бедной любви, не в кишках и не в усердном разумении точных мелочей, как делает Сарториус*. [А. П. Платонов. Счастливая Москва (1936)].

Замечательным примером здесь является употребление прилагательного *столбовая*. Так, в выборке второй половины XIX-го века это прилагательное встречается 28 раз в значении «центральной дороги» и лишь 2 раза в метафорическом значении. В сопоставимой с ней по объему выборке 1921–1940 годов (39 млн словоупотреблений во второй половине XIX-го века, 31 млн в втором двадцатилетии XX-го) прилагательное *столбовая* используется для определения *дороги* в центральном значении лишь 4 раза, и 12 раз в метафорическом значении центрального направления развития:

17) *По* ***столбовой дороге*** *реализма и гуманизма пришел лучший наш театр к нынешнему расцвету, и этим он в большой степени обязан Горькому, тому направлению, которое дал он молодому театру в решающий момент его жизни.* [М. С. Гус. Великий мастер русской сцены. 75-летие Константина Сергеевича Станиславского. В борьбе за реализм (1938.01.18)].

В следующих периодах это соотношение примерно сохраняется. Иначе говоря, по мере того, как семантический класс «центральной дороги» снижает свою значимость и стало быть частотность, прилагательное этого класса «мигрирует» в класс «метафоры», сохраняя при этом семантику центральности.

Можно было бы предположить, что для советской риторики чужда идея оценки выбора жизненного пути или направления развития: правильный выбор уже заранее предопределен партией и правительством. Однако, скорее, наблюдаемый сдвиг связан все с теми же изменениями концептуализации дороги, которые мы наблюдали, обсуждая снижение частотности словосочетаний со значением способа перемещения. Метафора настоящей/ложной дороги как правильной/неправильной судьбы базируется на представлении о дороге, как о единственной прямой, соединяющей пункт назначения с пунктом отбытия. Именно это представление постепенно и размывается в XX-м веке, что выражается и в снижении частотности семантического класса «направление», и в последующем, запаздывающем на один временной интервал, снижении частотности метонимического обозначения маршрута перемещения — воздушная дорога, морская дорога. В этом смысле очень интересна динамика употребления относительных прилагательных со значением профессии или вида деятельности и существительного дорога, употребленном в переносном значении — артистическая дорога, педагогическая дорога, научная дорога, отраженная на диаграмме 7.

Диаграмма 7. Распределение частотности прилагательных со значением профессии или статуса

Из диаграммы 7 видно, что во второй половине XX-го века употребление *дороги* в переносном значении заданного направления

резко снижается. Это явление лежит опять же в русле общей тенденции, связанной с ослаблением определенных семантических компонентов значения лексемы в целом, прежде всего идеи заданного направления, которое обеспечивается дорогой. Обобщая наблюдения, сделанные выше, можно сказать, что в XX-м веке утрачивает свою значимость представление о дороге, как об основном средстве связанности между населенными пунктами. Возможно, этому способствовало не только изменение характера мобильности, но и с то, что связанность стала обеспечиваться другими способами, например, при помощи телефонной связи. В целом, мы видим снижение значимости дорог (во всех смыслах) для общественного сознания в XX-м веке. И это снижение находит свое отражение и в языке. Мало заментные для лексикографа и едва ли достаточные для того, чтобы даже пытаться отразить их в словаре, основания для выбора тех или иных языковых выражений тем не менее оказываются чрезвычайно чувствительными к социальным изменениям в стране и обществе. Можно сказать, что глобальные общественные сдвиги влекут за собой сдвиги языковые, которые проявляются прежде всего на уровне изменения лексической сочетаемости и распространенности тех или иных выражений, а также изменения лексических отношений между значениями, прежде всего устойчивости метонимических и метафорических переносов. Предложенный выше анализ ставил своей целью показать, как с помощью выборок формата medium data можно увидеть за изменением сочетаемости слов изменения смыслов, а изменения смыслов в языке связать с изменениями смыслов в общественном сознании.

Литература

Арутюнова Н. Д. Путь по дороге и бездорожью // Логический анализ языка. Языки динамического мира. / Отв. ред. Н.Д. Арутюнова, И.Б. Шатуновский. Дубна, 1999. С. 3–17.

Замятин Д. Н. Геокультура и процессы межцивилизационной адаптации: стратегии репрезентации и интерпретации ключевых культурно-географических образов // Цивилизация. Восхождение и слом. Структурообразующие факторы и субъекты цивилизационного процесса. М.: Наука, 2003. С. 213–256.

Щепанская Т. Б. Культура дороги в русской мифоритуальной традиции XIX–XX вв. М.: Индрик, 2003.

Падучева Е. В. Наблюдатель: типология и возможные трактовки // Труды международной конференции Диалог. М.: РГГУ, 2006. С. 403–413.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. // Science, 331(6014), 176–182.

A. A. Bonch-Osmolovskaya
National Research University Higher School of Economics
(Russia, Moscow)
abonch@gmail.com

CULTUROMICS IN THE RUSSIAN NATIONAL CORPUS: THREE CENTURIES OF RUSSIAN ROADS

Culturomics, known as an innovative method of cultural research, studies cultural trends through quantitative analysis of the immense data of Google digitized books. This paper suggests the concept of medium data methodology, which uses much smaller but cleaner data sets. Medium data allows statistical analysis as well as more complicated linguistic preprocessing and use of complicated mark-up schemes. The study, proposed as an example of medium data research, deals with the concept of road, which is reflected in lexical data of 18th–20th centuries, received from the Russian National Corpus. The data comprises more than 15800 examples with the noun *road* and an adjective which serves as a diagnostic context for polysemy disambiguation. Each example is linked with one of the eight temporal periods. All the examples are divided in 19 semantic classes, and the method of hierarchical clusterization is used to reveal the most important trends and similarities in the behavior of different classes. The changes in charts are connected to social processes and changes in cultural perception.

Key words: the Russian National Corpus, semantics, road, social development, cultural development, Russian language

References

Arutiunova N. D. [The way on the road and without the road]. *Logicheskii analiz yazyka. Yazyki dinamicheskogo mira* [Logical analysis of language. The Languages of the Dynamic World]. Arutiunova N. D., Shatunovskii I. B. (gen. ed.), Dubna, 1999, pp 3–17. (In Russ.)

Paducheva E. V. [The Observer: Typology and Possible Interpretations]. *Trudy mezhdynarodnoy konferentsii "Dialog-2006"* [Proceedings of International Conference "Dialog–2006"]. Moscow, RSUH Publ., 2006, pp. 403–413. (In Russ.)

Shchepanskaya T. B. *Kultura dorogi v russkoi miforitual'noi traditsii XIX-XX vekov* [Road culture in Russian mythic and ritual tradition of the 19th-20th centuries]. Moscow, Indrik Publ., 2003.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 176–182.

Zamiatin D. N. [Geoculture and processes of intercivilization adaptation: strategies of representation and interpretation of key cultural and geographic symbols]. *Tsivilizatsiia. Voskhozhdeniie i Slom. Strukturoobrazuyuschie Factory i Sub'iekty Tsivilizatsionnogo Protsessa* [Civilization. Ascending and fracture. Structure-forming factors and actors in the civilization process]. Moscow, Nauka Publ., 2003, pp. 213–256. (In Russ.)

А. И. Левинзон
Национальный исследовательский университет
«Высшая школа экономики»
(Россия, Москва)
anna.j.levinzon@gmail.com

КОРПУСНОЕ ПРЕПОДАВАНИЕ В РОССИЙСКОЙ ШКОЛЕ

До сих пор на уроках русского языка в российской школе практически не используются электронные корпуса. Цель статьи — продемонстрировать возможности НКРЯ как инструмента эффективной работы с детьми. Мы анализируем как основные достоинства различных методов корпусной педагогики, так и сложности, которые предстоит преодолеть учителю, выбравшему, например, метод обучения на основе анализа данных.

Ключевые слова: корпусная педагогика, Национальный корпус русского языка, обучение на основе анализа данных, русский язык в школе

Если корпусная лингвистика в России в последнее время переживает очевидный подъем, корпусное преподавание русского языка остается до сих пор занятием отдельных энтузиастов. Цель статьи — представить возможности использования НКРЯ на уроках школьной словесности. Предложенные методы, с одной стороны, базируются на теоретических принципах корпусной педагогики, предложенных еще Дж. Личем, Р. Квирком, С. Гринбаумом и Я. Свартвиком[1], с другой — разработаны на основании практического опыта занятий с детьми.

Начнем с перечисления учебных курсов, в ходе которых учителю было бы удобно применять задания, связанные с НКРЯ:

[1] См., например, [Quirk, Greenbaum, Leech, Svartvik, 1997]

1. **РКИ для детей**: это направление преподавания совсем недавно было осознано как значимое для российской системы образования[1]. При огромном спросе на качественные учебные материалы, способные облегчить подготовку к уроку учителю, не имеющему специального образования в области обучения русскому языку иностранцев, количество таких материалов до сих пор крайне мало.

2. **Русский язык как второй родной**: еще одно направление, активно развивающееся сегодня; высокие темпы эмиграции очевидно увеличат в ближайшем будущем количество учеников русских школ за рубежом. Заметим, что здесь, как и в случае с РКИ, максимально эффективных результатов удается добиться, только если преподаватель подстраивает задания под конкретную группу: разнообразие речевого опыта детей, равно как и разница в структурах их первых и второго языков не позволяет ограничиться стандартными, созданными «для всех», упражнениями и заставляет ответственного учителя разрабатывать собственные пособия для урока.

3. **Занятия по подготовке к ГИА/ЕГЭ**: сборники упражнений, созданных с помощью НКРЯ, разумеется, могли бы пополнить и без того богатую коллекцию учебников, призванных готовить школьников к аттестации. Между тем, на наш взгляд, использовать Национальный корпус стоит не столько при подготовке к сдаче экзамена, сколько при составлении самих экзаменационных тестов. Ниже мы рассмотрим несколько основных преимуществ экзаменационных заданий, составленных с помощью НКРЯ.

4. **Уроки развития речи**: область, которая до сих пор считалась в российской школе второстепенной, несмотря на то, что внятное выражение мысли — умение, тренировка которого в теории составляет суть занятий родным языком. Характерно, что учителя пренебрегают уроками развития речи, заменяя их занятиями по орфографии и пунктуации, в частности, из-за отсутствия простых упражнений, представляющих собой цепочки коротких примеров и позволяющих привычным образом отрабатывать материал: задания, связанные с сочинением и анализом текстов часто требуют слишком большого подготовительного труда и нуждаются в долгой проверке. НКРЯ с его разными видами поиска — ресурс, способный снабдить

[1] Так, кафедра ЮНЕСКО, задача которой – координировать программу «Международное образование и интеграция детей мигрантов в школе», открылась в МИОО только в 2011 году.

учителя наборами недлинных контекстов, удобных для тренировки многих типов навыков.

5. **Уроки литературы**: при существующем количестве учебников по школьной литературе преподавателю, по-видимому, не имеет смысла осваивать новые способы подготовки к уроку; именно поэтому мы рекомендовали бы работать с НКРЯ в первую очередь в междисциплинарном пространстве, скажем, при обращении на уроках литературы к истории или к географии. В этой ситуации Национальный корпус становится инструментом быстрого поиска и организации информации, для сбора которой в других источниках учителю и ученику пришлось бы потратить немало сил.

6. **Олимпиады по словесности**: олимпиадное движение, в частности, составление олимпиад по лингвистике и литературе, как известно, является давней российской традицией[1]. Участие в олимпиадах, до недавнего времени бывшее занятием избранных, все больше становится массовым увлечением. Тренировка олимпиадников оказывается, таким образом, с одной стороны, необходимым, с другой — чрезвычайно энергоемким для школьного учителя делом. НКРЯ, как нам кажется, можно рассматривать как справочник, с помощью которого удобно тиражировать заданный образец творческих заданий.

7. **Школьные исследовательские проекты**: необходимый, согласно новым образовательным стандартам, вид учебной деятельности. Неудивительно, что большинство детских научных работ, организованных школами, отличает общий недостаток — отсутствие оригинальности: ребенок в силу возраста, во-первых, мыслит несамостоятельно, во-вторых, не обладает достаточной настойчивостью для сбора интересного материала. НКРЯ — пространство, в котором даже небольшие усилия юного исследователя могут быть вознаграждены весьма любопытными находками, минимально осмыслив которые, он в состоянии получить красивый результат.

Перейдем теперь к демонстрации связанных с преподаванием словесности трудностей и методов их преодоления с помощью НКРЯ.

1) Несомненно, главную сложность для любого составителя

[1] Пример тому — Международная олимпиада по лингвистике, организованная по примеру московских лингвистических олимпиад.

школьных упражнений и хрестоматий представляет выбор материала. Автору пособий, предназначенных для детей, бывает важно таким образом подобрать тексты, чтобы не только их тематика, но и язык соответствовали уровню интеллектуального и эмоционального развития предполагаемого читателя. Именно ориентация на условную «детскую речь», на наш взгляд, становится причиной того, что учебники для младшей и средней школы во многих случаях становятся для ребенка объектом насмешек и отторжения, а не приятия или солидарности. Педагоги, занятые разработкой заданий по русскому языку, по большей части не имеют возможности строить свои тексты исходя из результатов, полученных социолингвистами, описывающими реальные особенности живой письменной и устной речи детей. В то же время авторы пособий могли бы, как нам кажется, сделать входящие в них тексты менее искусственными и, соответственно, более приемлемыми для читателей, если бы хотя бы иногда использовали в качестве проверочного образца НКРЯ.

Приведем характерный пример: стремясь подражать детской манере общения, составители учебников обычно активно используют существительные с уменьшительно-ласкательными суффиксами. Так, в многочисленных методических пособиях по развитию речи у детей[1] предлагается использовать в качестве материала для выучивания, выразительного чтения и анализа отрывок из стихотворения С. Погореловского:

Вот он — хлебушко душистый
Вот он — теплый, золотистый!
В каждый дом, на каждый стол
Он пожаловал-пришел.

Как известно, учебники, предназначенные для детей младшего школьного возраста стараются оперировать ограниченным набором лексических единиц. Ниже мы обсудим возможность вычленения такого базового списка лексики с помощью НКРЯ, пока же сформулируем следующий вопрос: насколько уместно использовать слово «хлебушко» в качестве элемента развития речи семи — восьмилетнего ребенка? Национальный корпус послужит нам инструментом, позволяющим быстро установить как частотность употребления лексемы, так и обычные контексты ее бытования. Поиск в корпусе показывает следующий результат: на общий объем

[1] См., например, [Каландарова, 2009: 112].

в 192 840 904 слова интересующее нас существительное дает всего 65 вхождений. Недолгий просмотр примеров позволяет обнаружить следующий любопытный контекст: *С этой войны слова новые появились, которые в большом ходу: барыга, ксивы (документы), шансы (деньги), крохобор, прохаря (сапоги),* ***хлебушко*** *и многое другое. Такие выражения: "будем посмотреть", "еще тот", "война спишет"; про что-нибудь хорошее, прочное говорят: "Миколаевское". Про здорового, удалого: "С Волги" и т. д. (...) 6.10.45 г.* [Владимир Чивилихин. «Моя мечта — стать писателем», из дневников 1941–1974 гг. (2002) // «Наш современник», 2002.06.15]. Как видим, слово, которое авторами учебников для детей воспринимается, по-видимому, как принятое ласковое и уважительное обозначение хлеба, в действительности, является крайне малоупотребительным и, если не жаргонным, то, во всяком случае, ярко просторечным. Включение его в учебники по развитию речи для детей представляется, таким образом, неоправданным.

2) Другая непростая задача, которую приходится решать учителю русского языка — борьба со скукой на уроке. Рамки статьи не позволяют нам останавливаться на критериях, представляющихся значимыми при подборе примеров для упражнений и текстов для анализа; заметим только, что одним из основных свойств удачных, с точки зрения возбуждения у ребенка желания работать с языковым материалом, отрывков является присутствие конфликта. Тексты о природе, в изобилии представленные в учебниках русского языка, как правило, лишены именно этой составляющей. *На поляну вышел огромный однорогий лось. Зверь постоял с минуту, как бы к чему-то прислушиваясь, и зашагал дальше* [Никитина, 2009:154] — вот всего лишь один из многочисленных примеров таких лишенных внутреннего движения текстов. Работая с НКРЯ, автор учебника мог бы подобрать куда более динамичные короткие отрывки, затратив при этом совсем немного усилий. Предположим, учителю хочется найти текст именно про лося. В этом случае, задав подкорпус детской литературы, получаем, например, такие контексты: *Дело было ясное. Кто-то стрелял в лосёнка и ранил его. Стрелял браконьер, дурак. Он знал: лосей бить запрещено. Выстрелив, он напугался того, что сделал, убежал домой.* [Юрий Коваль. У Кривой сосны (1979)] или *Лось наклонил голову и стал шумно, большими глотками пить воду.— Эй, рогатый! — закричал лосю*

маленький лягушонок. — Это не твой ручей! [Сергей Георгиев. Маленький зеленый лягушонок // «Мурзилка», 2001] или *Заяц сидел под кустом и мыл себе уши. — Возьми еловую шишку! — крикнул Ёжик. — Еловой шишкой лучше отмывается! — А чем вы посоветуете почистить мне рога? — спросил, выйдя на опушку перед Ежикиным домиком, Лось* [Сергей Козлов. Правда, мы будем всегда? (1969–1981)]. Интрига, присутствующая во всех найденных нами примерах, делает их в глазах ребенка привлекательными объектами для исследования.

Еще один способ поддержать интерес детей к упражнению — выстроить его вокруг персонажа, которому ребенок заранее симпатизирует. Упражнение на отработку постановки знаков препинания, в котором 10–15 сложных предложений посвящены рыцарским битвам, образу жизни принцесс или быту пиратов, почти невозможно придумать, но относительно просто скомпоновать с помощью НКРЯ.[1]

Наконец, сама возможность взаимодействовать со сложной и в то же время удобно организованной электронной системой делает процесс обучения на уроках и выполнения домашних заданий делом вполне приятным. Обучающий корпус был выделен в составе НКРЯ именно с этой целью: дать возможность ученику и преподавателю работать с текстами, подобранными и размеченными в соответствии со стандартами школьного образования. Для самостоятельной работы ребенка в корпусе потребуется, разумеется, наличие в школе достаточного числа компьютеров, однако, даже в отсутствие возможности работать всем классом одновременно, учитель может предлагать детям домашние задания в НКРЯ. Виды таких заданий могут варьироваться от совсем простых (подобрать пример на изученное явление) до действительно сложных (провести собственное мини-исследование на заданную тему).

3) Требование постоянного повторения, а значит, необходимость использовать большое число однотипных упражнений — трудность, которая также может быть преодолена с помощью НКРЯ. Учитель часто оказывается в ситуации, когда существующих пособий ему оказывается недостаточно: все задания на конкретную тему сделаны, а ученики все еще не достаточно хорошо освоили материал.

[1] Примеры таких тематически подобранных упражнений можно найти в сборнике [Левинзон, 2014].

Придумывание примеров в этом случае невозможно: во-первых, качество придуманного в соответствии с собственной языковой интуицией автора материала окажется невысоким; во-вторых, фантазия даже самого творческого сочинителя при выдумывании многочисленных фраз одного типа быстро истощается. В этой ситуации мы рекомендовали бы использовать метод стилистического эксперимента[1] при составлении Тренажера по академическому письму[2]: в корпусе подбирается пример на изучаемое языковое явление, далее этот пример «портится» — в нем появляется пропуск, который следует заполнить или характерная ошибка, которую надо найти.

Представим в качестве примера создание упражнения на отработку правильного употребления двойных союзов. «Ухудшение» хорошего примера проводится в несколько этапов:

Шаг 1: задаем двойной союз и подбираем ряд примеров в НКРЯ. Возьмем союз *как… так и* и рассмотрим всего одну из множества полученных фраз: *Галль полагал, что как тело состоит из отдельных органов со своими функциями, так и мозг состоит из многих «ментальных органов», занимающихся теми или иными мыслительными процессами* [Игорь Лалаянц. Детектор лжи на молекулярном уровне? Завтра, завтра… послезавтра! // «Знание — сила», 2003].

Шаг 2: сокращаем пример, устраняем тавтологию. *Галль полагал, что как тело состоит из отдельных органов со своими функциями, так и мозг составлен из многих «ментальных органов».*

Шаг 3: заменяем неизвестное школьнику имя собственное. *По мнению ученых восемнадцатого века, как тело состоит из отдельных органов со своими функциями, так и мозг составлен из многих «ментальных органов».*

Шаг 4: заменяем правильный двойной союз на неверный. *По мнению ученых восемнадцатого века, как тело состоит из отдельных органов со своими функциями, но и мозг составлен из многих «ментальных органов».*

Действуя по этому алгоритму, учитель может без труда подобрать нужное число предложений для исправления/заполнения про-

[1] Метод стилистического эксперимента, суть которого состоит в подстановке неудачного варианта в образцовый текст и последующем анализе причин стилистической неудачи, был предложен А.М. Пешковским [Пешковский, 1930].

[2] Адрес тренажера http://web-corpora.net/acad/accounts/register/

пусков — в случае необходимости тексты могут быть ограничены жанром, тематикой или характером адресата.

4) Выше мы отмечали уже принципиальное несоответствие между временем, имеющемся в распоряжении учителя на подготовку к уроку, и затратами, необходимыми для составления нестандартных заданий, без которых невозможна тренировка потенциальных участников олимпиад. Как нам кажется, НКРЯ может выполнять функцию образовательной среды, в рамках которой создание упражнений, требующих творческого решения, оказывается делом, посильным даже занятому повседневной работой школьному учителю.

Предложим здесь несколько заданий, придуманных нами с использованием контекстов из НКРЯ и книги К. Чуковского «Живой как жизнь». Подчеркнем, что в данном случае речь идет именно о тренировке работы с непривычным языковым материалом — наши вопросы требуют долгого размышления, в некоторых случаях привлечения дополнительных сведений и не являются собственно «олимпиадными», то есть решаемыми в отсутствие не представленного в задаче материала.

*— Что означала фраза **художник полюбил водолазку** в 1950-х годах и что она значит сейчас?*

*— В текстах XIX века можно встретить выражения **запросто приехать на обед**. Что означало в этих текстах слово запросто?*

*— В историко-политических сочинениях XVIII века можно встретить выражения **члены верхней камеры** и **члены нижней камеры**. О каких камерах идет речь?*

*— В научных трудах XVIII века можно встретить выражения **противное трение**, **противная сила**. Что означало в этих текстах слово противный?*

*— Почему в начале XX века заявление **он обязательно тебя побьет** могло показаться смешным?*

*— Фраза **в церковь не пустили никого, кроме поезда** была совершенно нормальной в начале XIX века. Как вы ее объясните?*

Другой тип работы, подготавливающий ребенка к участию в олимпиаде и в то же время развивающий у него вкус к научной деятельности, можно предлагать ученикам, уже обученным поиску в НКРЯ. Мы советовали бы начинать с простых и в то же время любопытных для ребенка вопросов, например:

— Кто из русских писателей, согласно НКРЯ, первым написал о ***ерунде****?*

— Какой русский писатель, согласно НКРЯ, одним из первых заставил своих героев есть ***чипсы****?*

— В современном школьном и студенческом жаргоне слово ***ботать*** *означает учиться, работать. Что оно значило в XIX веке?*

Более сложная задача связана с поиском лексических несоответствий в художественных текстах. Удобным материалом здесь служат исторические романы, в частности, хорошо известные старшеклассникам детективы Б. Акунина[1]. Рассмотрим одно из таких заданий и ход его выполнения.

Б. Акунин пишет, что действие второго тома романа «Алмазная колесница происходит в 1878 году. С помощью НКРЯ докажите, что речь героя отличается от той, что могла звучать в указанный период.

Лошадей в Японии мало, и они дороги, а людей много, и они дешевы. Рикша — профессия из новых, лет десять назад про нее здесь не слыхивали. Колесный транспорт считается тут европейским новшеством. Этакий бедолага пробегает за день верст шестьдесят. Зато плата по местным понятиям очень хорошая. Если повезет, можно пол-иены заработать, это по-нашему рублишко. Правда, долго рикши не живут — надрываются. Годика три-четыре, и к Будде в гости.

Ученику, взявшемуся за выполнение задания, придется догадаться, что «доказательство» в данном случае сводится к поиску слов и выражений, не употреблявшихся в конце XIX-го века и пришедших в роман из словаря самого писателя — человека века XX-го. От школьника в этом случае потребуется определенная языковая интуиция, поскольку выяснять время вхождения в язык всех лексических единиц в отрывке, разумеется, окажется слишком сложно. Развитие такой интуиции, впрочем, и является одной из целей упражнения. В нашем случае в тексте присутствуют три главные точки несоответствия, которые предстоит выделить ученику, работая с НКРЯ: это выражения *колесный транспорт*, появившееся во время Второй мировой войны и относившееся первоначально исключительно к военной технике, и *по местным понятиям*, дати-

[1] Идея задания на материале текста Б. Акунина принадлежит Н.Р. Добрушиной — см., например, ее задание на www.studiorum.ruscorpora.ru

руемое корпусом 1970 годом; лексема бедолага, по данным НКРЯ, также не могла встретиться раньше 20-х годов XX века. Заметим, что предложенный здесь анализ отрывка является весьма поверхностным и ни в коем случае не претендует на полноту: вполне вероятно, что вдумчивый школьник, сопоставив материалы Национального корпуса и акунинского текста, сделает еще много любопытных наблюдений.

Вкратце представив задачи, которые школьный учитель в состоянии решить, пользуясь НКРЯ, обобщим описание методов работы с Национальным корпусом на уроке словесности.

Корпусное преподавание может быть условно разделено на два типа: «легкий» и «трудный». Пользуясь «легким» методом, преподаватель предлагает ученикам заранее подготовленный материал, представленный либо в бумажном виде, либо в виде загодя сконструированного файла или презентации. В ходе подготовки к занятию учитель

— выделяет внутри корпуса подкорпус текстов, наиболее актуальный для решения поставленных учебных задач;

— получив искомые контексты, фильтрует данные, отсеивая примеры, не укладывающиеся в представляемую теоретическую модель;

— упрощает выбранные для демонстрации примеры таким образом, чтобы их содержание, во-первых, являлось понятным и, во-вторых, не отвлекало учеников от собственно лингвистического анализа;

— предлагает полученные в НКРЯ данные в виде списков контекстов, учебных упражнений или иллюстраций высказанных теоретических положений.

Достоинства метода очевидны: во-первых, на подготовку учебных заданий тратится минимум времени; во-вторых, список примеров является выборкой реальных словоупотреблений — НКРЯ в этом случае становится инструментом экспертизы, дополняющим интуицию; в-третьих, используя статистику, преподаватель может выбрать из нескольких возможных примеров те, что наиболее характерны для изучаемого языкового регистра.

Недостатков у предложенного способа работы немного, однако, они достаточно существенны для того, чтобы в целом ряде ситуа-

ций предпочесть менее предсказуемый, однако куда более творческий «трудный» способ. Речь идет, во-первых, о получении преподавателем возможности невольно или сознательно манипулировать полученным материалом, «подгоняя» его под готовые концепции. Во-вторых, что не менее важно, работа ребенка носит в этом случае механистический характер: ученик всего лишь воспроизводит ответы, полученные до него учителем.

«Трудный» метод предполагает самостоятельную работу ученика с НКРЯ — в идеале такая работа проводится в классе под контролем учителя; в случае недостаточной оснащенности школы корпусное задание может быть дано на дом. Этот способ также имеет два варианта. Вариант первый строится следующим образом:

— готовясь к занятиям, преподаватель собирает и анализирует данные Обучающего корпуса НКРЯ. Обратим внимание на то, что здесь, в отличие от «легкого» способа, все задания строятся на материале именно Обучающего корпуса: предлагать ребенку самому проводить поиск во всем многообразии текстов, представленных в НКРЯ, по-видимому, нецелесообразно;

— из нескольких возможных типов поиска в качестве учебного задания выбирается тот, что дает наименее противоречивую подборку примеров;

— ученику предлагается от начала до конца пройти проделанный преподавателем путь: получив учебное задание, ребенок входит в Обучающий корпус, подбирает контексты и анализирует их.

— результат работы ученика оценивается как положительный, если он с возможной точностью повторяет результат, полученный учителем.

В этом случае ребенок, во-первых, получает удовольствие от работы с современным электронным ресурсом; во-вторых, исследует реальный языковой материал; в-третьих, самостоятельно замечая языковые явления, он лучше усваивает информацию о них. Результат работы может быть представлен в форме короткого текста. Несмотря на то, что само исследование в данном случае будет носить фиктивный характер, отчет о нем окажется полезным упражнением в области развития письменной речи.

Организуя «трудное» взаимодействие с НКРЯ, учителю стоит помнить о том, что метод окажется неподходящим для двух групп учеников:

1) тех, кто не склонен к самостоятельности, не верит в свои силы или теряется перед обилием материала;

2) тех, кто не в состоянии увлечься работой, требующей значительных затрат времени и внимания, но лишенной серьезного творческого потенциала.

Если для первой группы более предпочтительным окажется «легкий» метод, для последней мы предлагаем второй вариант метода «трудного». Вариант этот базируется на концепции DDL (Data Driving Learning — обучение на основе анализа данных), разработанной Тимом Джонсом и представленной, в частности, в его работе «Компьютеры в изучении языка»[1]. К основным принципам DDL обычно относят:

1. отказ от дедуктивного обучения в пользу индуктивного;

2. формулировку полученной учеником информации не в виде набора правил –исключений, но в виде распознанных языковых тенденций;

3. внимание не к «правильному», но к редкому и неожиданному в языке.

Ход работы учителя, решившегося на этот вариант, выстраивается так.

— Преподаватель, исходя из своих представлений о возможностях и желаниях ученика, предлагает ему для решения несколько возможных задач.

— Ребенок выбирает задачу, входит в Обучающий корпус НКРЯ и приступает к сбору материала.

— Полученные данные ученик обрабатывает и формулирует на их базе собственные выводы, которые могут расходиться с предположениями преподавателя или учебника.

— Результат представляется в форме мини-исследования. Такое исследование имеет очень узкую направленность и небольшой объем, однако обладает реальной научной ценностью и может быть в дальнейшем развито.

Важнейшее достоинство метода DDL — возможность в полной мере реализовать принципы автономного обучения. Ребенок чувствует ответственность за результат занятия, как следствие, повышается его мотивация. Другой фактор повышения мотивации —

[1] См., например, Johns, T. Whence and whither classroom concordancing? in [Bongaerts, De Haan, Lobbe, Wekker, 1988].

стремление удовлетворить любопытство, пробуждаемое неоднозначными и противоречивыми примерами, найденными в корпусе. Знание, полученное на занятии, является новым и не повторяет информацию, которую ребенок мог бы прочитать в учебнике или услышать от учителя. Наконец, в классе создается естественная рабочая обстановка: сотрудничая с преподавателем и коллегами по учебе, ребенок усваивает правила поведения в профессиональном сообществе. Преподаватель, пользующийся DDL, перестает быть истиной в последней инстанции: его задачей становится организация условий работы, создание четких инструкций и содержательное обсуждение результатов. Учитель, таким образом, передает ученику самую ценную из имеющейся у него информации: информацию о стратегиях анализа данных.

Принятое в российском школьном филологическом образовании уважение к правилам и четкости формулировок не позволяют выстраивать на базе DDL существенную часть занятий. Использованию метода, кроме того, мешают технофобия учителей и нежелание учеников выполнять задания, требующие исследовательских навыков. Неумение пользоваться представленным в НКРЯ терминологическим аппаратом, равно как и небольшой объем нормативного Обучающего корпуса также могут служить помехой для эффективного использования метода на уроках. Между тем, предложенный способ работы с НКРЯ представляется нам чрезвычайно удачным при выполнении проектных заданий: напомним, что сколь бы ограниченную область ни рассматривал в этом случае школьник, его результаты, тем не менее, с большой вероятностью будут отличаться аккуратностью и оригинальностью — свойствами, наиболее ценными для детских исследовательских работ.

Завершим наш анализ использования НКРЯ на уроках русского языка и литературы рассмотрением нескольких конкретных учебных задач, решаемых с помощью НКРЯ.

1. Составление списков слов для начинающих читателей.

Несмотря на то, что в последние годы в англо-американской системе образования все больше укрепляются позиции сторонников звуко-буквенного обучения чтению и письму, по-прежнему обяза-

тельными для составления разнообразных методик остаются так называемые списки Дольча и Фрая[1]. Мы имеем в виду списки слов, которые ребенок должен узнавать зрительно и, соответственно, написание которых должно быть заученно подобно заучиванию иероглифов. Проблема заучивания и зрительного узнавания не слишком актуальна для ребенка, обучающегося чтению по-русски[2], однако, другое направление использования списков Дольча и Фрая, а именно, составление учебных текстов для чтения на их основе, представляется актуальным и для российской школы. Действительно, современные российские буквари, как кажется, включают лексику соответствующую в первую очередь индивидуальным вкусам их авторов. *Мышата Тиша и Глаша — шалунишки. Они шалили, шарили по углам, шуршали и нашли сыр, шпроты и лапшу. Мышата устроили пир. Пригласили ли они на пир кошку Мурку?* — читаем мы в одном из популярных в российской начальной школе пособий[3]. Здесь, как и в описанном выше случае с *хлебушком* неясно, насколько оправданным является выбор конкретных лексем для составления текста. Еще более странным выглядит набор слов, представленных авторами в упражнении, обучающем первоначальным навыкам чтения и письма: в задании ребенку предлагается вписать первую букву (в нашем случае букву *д*) в слова *диво, дума, дрова, дракон, диван* и *друг*.[4] Между тем, обратись авторы азбуки к НКРЯ, простой анализ частотных лемм, начинающихся с *д* в подкорпусе детской литературы, показал бы, что, не считая коротких *дом* и *день*, наиболее актуальными для детского чтения являются не *дума* или *дрова*, а *деревня, двор, доктор, деньги* и *дождик*. Список Фрая, как известно, был составлен на основе The American Heritage Word Frequency Book, чьи авторы[5] еще в 1971 году с помощью компьютерной программы проанализировали рекомендованные для школьников США тексты и, распределив произведения по возрастам и научной тематике, выбрали наиболее часто встречающиеся, а значит, наиболее актуальные для использования детьми, лексические единицы. Создание аналогичного списка с использованием данных НКРЯ несо-

[1] Подробнее см, например, в [Farrell, Osenga, Hunter].

[2] Впрочем, именно на зрительном узнавании частей слов основана, в частности, известная система «кубиков Зайцева».

[3] [Бунеев, Бунеева, Пронина, 2002: 84]

[4] [Бунеев, Бунеева, Пронина, 2002: 33]

[5] [Caroll, Davies, Richmond, 1971]

мненно позволило бы сделать учебники детского чтения более эффективными.

2. Создание экзаменационных заданий.

Парадоксальное сочетание высокого процента школьников, получающих высшее образование, и крайне низкой общей письменной грамотности населения вызвано, в частности, неспособностью школьной программы по русскому языку и литературе подготовить ребенка к осмыслению и составлению текстов, с которыми он встретится в реальной взрослой жизни. Художественная литература долгие годы остается основным источником примеров для учебных заданий; публицистика, используемая в ЕГЭ и ГИА также тяготеет прежде всего к канонам классических литературных образцов. Между тем, такие области письменной речи как, например, научная или деловая, почти полностью исключены из обихода школьной словесности. Обращение к НКРЯ позволило бы быстро накопить базу тренировочных и экзаменационных заданий на основе этих функциональных стилей, умелое пользование которыми чрезвычайно важно для успешной послешкольной жизни ученика.

3. Подготовка междисциплинарных проектов.

Появление новых отраслей знания на стыке не взаимодействовавших прежде наук обусловливает необходимость проведения интердисциплинарных занятий и в школе. Сложность подготовки такого рода проектов связана не только с чисто организационными проблемами сотрудничества учителей разных предметов, но и с отсутствием материала, на котором школьник мог бы выстроить свои относительно самостоятельные выводы. Художественный текст и здесь до последнего времени оказывалась единственным пространством пересечения, скажем, литературы и истории. При этом оригинальность выводов ребенка, работающего с хорошей книгой (а плохие книги детям, очевидно, сложно и едва ли полезно читать и изучать), обычно невелика: большинство сколько-нибудь значимых произведений литературы, в том числе современной, уже проанализированы и прокомментированы взрослыми исследователями. НКРЯ представляется в этой связи чрезвычайно удобным источником информации, на основании которой могут строиться легко вы-

полнимые и вполне любопытные историко-лингвистические изыскания. Так, требующий тщательности, но достаточно несложный анализ частотности и графиков распределения по десятилетиям существительных с суффиксом *ист позволяет нарисовать яркую картинку, которую потом юный автор проекта мог бы расширять в любом интересующем его направлении. Мы видим, например, что 1900-е годы — это время *гимназистов, анархистов, телеграфистов, фельетонистов, аферистов, эгоистов, пессимистов* (единственный период XX-го в., когда их больше, чем *оптимистов*) и *скандалистов*. В 1910-е популярны *металлисты* (появляется специальное слово для обозначения рабочих, льющих металл), *кавалеристы, романисты, индивидуалисты, синтетисты*, объединяющие театр, музыку, живопись и поэзию, а также *идеалисты* — скоро революция. В 1920-е годы обсуждают *коммунистов, марксистов, материалистов* и *интернационалистов*, в моду входят *куплетисты*, вдруг в центре внимания оказываются *букинисты* и *кокаинисты*, а по улицам по-прежнему разгуливают *бакенбардисты*. Остановимся на этом, предоставив нашему читателю-учителю вместе с учениками самому пройти проделанный нами путь, совершая при этом множество мелких, но забавных открытий.

Воспользовавшись НКРЯ, подыщем, наконец, образ для описания учителя, пользующегося на уроках корпусными технологиями: поставив в поле поиска *имя существительное + прокладывать + путь* получим *ледокол* — этой метафорой и закончим наш краткий обзор достоинств работы с НКРЯ на уроках школьной словесности.

Литература

Бунеев Р. Н., Бунеева Е. В., Пронина О. В. Моя любимая азбука М.: Баласс, 2002

Каландарова Н. Н. Уроки речевого творчества: 1, 2, 3, 4 класс. М.: ВАКО,2009

Левинзон А. И. Русский язык. Развитие письменной речи 5 – 6 классы. — М. Вентана-Граф, 2014

Никитина Е. И. Русский язык. Русская речь 5 класс. М., Дрофа, 2009

Пешковский А. М. Вопросы методики родного языка лингвистики и стилистики. М, 1930

Bongaerts P., De Haan P., Lobbe, S.& Wekker H. (Eds.), Computer applications in language learning. Dordrecht: Foris, 1988

Caroll J., Davies P., Richmond B. The American Heritage Word Frequency Book. Boston, Houghton Mifflin,1971

Farrell L., Osenga T., Hunter M. Comparing the Dolch and Fry High Frequency Word Lists [Электронный ресурс] URL: http://www.readsters.com/ (дата обращения 15.10.2014)

Knowles G., Mcenery T., Fligelstone S., Wichman A. (Eds.) Teaching and language corpora . Longman, 1997

Quirk R., Greenbaum S., Leech G. and Svartvik J. A Comprehensive Grammar of the English Language. Harlow: Longman, 1985

A. I. Levinson
National Research University "Higher School of Economics"
(Russia, Moscow)
anna.j.levinzon@gmail.com

USING CORPUS DATA IN RUSSIAN LANGUAGE CLASSROOM

Corpora have not yet become an everyday tool of a Russian language teacher in Russia. My goal is to demonstrate a wide range of usage-based approaches which can be applied in the teaching of L1 to children. I present the benefits of exploiting Russian National Corpus as a valuable pedagogical resource, as well as obstacles which can be caused by using data-based teaching methods.

Key words: Usage-based approaches, data-based teaching, Russian National Corpus, Russian language teaching

References

Bongaerts P., De Haan P., Lobbe S.& Wekker H. (Eds.). *Computer applications in language learning*. Dordrecht, Foris, 1988.

Buneev R. N., Buneeva E. V., Pronina O. V. *Moya lyubimaya azbuka* [My favorite alphabet]. Moscow, Balass Publ., 2002, p. 84.

Caroll J., Davies P., Richmond B. *The American Heritage Word Frequency Book*. Boston, Houghton Mifflin, 1971.

Farrell L., Osenga T., Hunter M. *Comparing the Dolch and Fry High Frequency Word Lists*. Available at URL: http://www.readsters.com/ (accessed 15.10.2014)

Kalandrova N. N. *Uroki rechevogo tvorchestva* [Lessons of language creativity]. Moscow, VAKO Publ., 2009.

Knowles G., Mcenery T., Fligelstone S., Wichman A. (Eds.) Teaching and language corpora. Longman, 1997

Levinzon A. *Russkii yazyk. Razvitie pis'mennoi rechi 5 – 6 klassy* [Russian language. Creative writing]. Moscow, Ventana-Graf Publ., 2014.

Nikitina E. I. *Russkii yazyk. Russkaya rech'* [Russian language. Speaking and writing in Russian]. Moscow, Drofa Publ., 2009 (In Russ.)

Peshkovsky A. M. *Voprosy metodiki rodnogo yazyka lingvistiki i stilistiki* [Questions of methodic, linguistics and stylistic of a first language]. Moscow, 1930.

Quirk R., Greenbaum S., Leech G. and Svartvik J. *A Comprehensive Grammar of the English Language*. Harlow, Longman, 1985.

Для заметок

ТРУДЫ
ИНСТИТУТА РУССКОГО ЯЗЫКА
им. В. В. ВИНОГРАДОВА

Выпуск 6, 2015 г.

Редактор — *Евгений Степанов*
Компьютерная верстка — *Марина Кива*
Дизайн обложки — *Дарья Павлова*

Издательство «Вест-Консалтинг»
109378, г. Москва, Есенинский бульвар,
д. 1/26, корп. 1, офис 34. Тел. (495) 978-62-75

Подписано в печать 10.11.2015 г.
Формат 60X84/16. Гарнитура Times
Бумага офсетная
Тираж 250 экземпляров

Типография ИПК «Квадрат»,
Белгородская обл., г. Старый Оскол,
проспект Комсомольский, 73