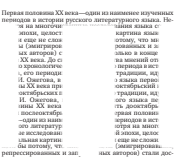
С. О. Савчук

Корпус текстов первой половины XX века:

текущее состояние и перспективы



1.

Первая половина XX века — один из наименее изученных периодов в истории русского литературного языка. Несмотря на многочисленные исследования языка советской эпохи, целостная и детальная картина языковой жизни еще не сложилась, хотя бы потому, что многие тексты (эмигрировавших, репрессированных и запрещенных авторов) стали доступны только в конце 80-х годов XX века. До сих пор нет единства мнений относительно хронологических границ этого периода в истории языка, его периодизации.

Согласно традиции, идущей от С. И. Ожегова, в истории русского языка первой половины XX века принято выделять дооктябрьский и три послеоктябрьских периода. Первый период — до конца 20-х — начала 30-х годов; второй период — 30-е — самое начало 40-х годов; третий период — Великая Отечественная война 1941–1945 годов и первые послевоенные годы[1].

[1] Ожегов С.И. К вопросу об изменениях словарного состава в русском языке в советскую эпоху // Вопросы языкознания. 1953. № 2; Бельчиков Ю.А. Русский язык. XX век. М., 2003; Скворцов Л.И. Сергей Иванович Ожегов – человек и словарь. М., 2001.

Одни исследователи предлагают начинать отсчет дооктябрьского периода с 70-х[2] или 90-х[3] годов XIX века, связывая общий вектор развития языка с процессом демократизации общественной жизни. Октябрьская революция при этом рассматривается как фактор, ускоривший эволюционные процессы[4]. По мнению других исследователей, октябрьский переворот вызвал слом, разрушение старого стандарта и замену его новым стандартом, продержавшимся до конца советского строя, то есть до 90-х годов XX в.[5]

Как представляется, создание современного корпуса текстов первой половины XX века будет способствовать формированию более объективной картины происходивших в языке данного периода процессов и уточнению научных представлений, сложившихся в истории литературного языка.

Этот корпус по своему типу относится к историческим, или диахроническим корпусам. Достижения компьютерной лингвистики в области создания диахронических корпусов значительно уступают успехам в конструировании корпусов современных текстов, что объясняется прежде всего трудоемкостью процесса перевода старых текстов в электронную форму и значительными материальными затратами[6]. В этих условиях описание конкретного опыта разработки исторического корпуса, как кажется, может представлять интерес для специалистов.

---

[2] Грановская Л.М. Русский литературный язык в конце XIX и XX вв. М, 2005.

[3] Мещерский Н. А. История русского литературного языка. Л., 1981.

[4] Поливанов Е.Д. Революция и литературные языки Союза ССР // За марксистское языкознание. М., 1931. С. 73-94; Селищев А.М. Язык революционной эпохи: Из наблюдений над русским языком последних лет. 1917–1926 // Селищев А.М. Труды по русскому языку. Т. 1. М., 2003.

[5] Живов В.М. Язык и революция. Размышления над старой книгой А.М. Селищева // Отечественные записки. 2005. №2.

[6] Corpus Linguistics: Critical Concepts in Linguistics / Ed. By W. Tewbert & R. Krishnamurthy. V.I. L; NY: Routledge, 2006. P. 32-33; C. Onelli, D. Proietti, C. Seidenari, F. Tamburini. The DiaCORIS project: a diachronic corpus of written Italian // Proceedings of the 5th International Conference on Language Resources and Evaluation/ Genoa, 2006; Gau, M. The State of Historical Corpus Linguistics with Special Focus on the Russian Language. M.A. thesis, University of Regensburg, Institute for Slavonic Languages and Literatures, 2005; Xiao R.Z. Diachronic corpora // Xiao R.Z. Well-known and influential corpora: A survey. In Lüdeling A., Kytö M., McEnery A. (eds.) Corpus Linguistics: An International Handbook. Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin, Germany, 2007.

Корпус текстов первой половины XX века создавался в несколько иных условиях и с иными установками, чем основная часть НКРЯ[7]. С самого начала, еще на стадии проектирования, был задан его объем, ограниченный 30 млн словоупотреблений, и стратегия пополнения состава[8], которая включала в себя следующие моменты:

1) Репрезентативность корпуса.
2) Сбалансированность состава текстов.
3) Высокие темпы пополнения корпуса.
4) Отработка технологии подготовки электронных версий текстов разных форматов.

Репрезентативность состава корпуса обеспечивалась тем, что отбор текстов осуществлялся в соответствии с составленной заранее базой данных авторов, куда вошли писатели, принадлежащие к различным литературным направлениям (в том числе все писатели «первого ряда»), видные ученые, религиозные мыслители, деятели культуры, политики — представители различных партий и группировок и т.д. В корпусе представлены основные сферы коммуникации — художественная, публицистическая, научная, официально-деловая, бытовая, церковно-богословская.

Для каждой из этих сфер была установлена доля относящихся к ней текстов в общем объеме корпуса, и подготовка текстовых материалов к размещению на сайте включала обязательный контроль над соблюдением пропорций.

Для обеспечения скорости пополнения корпуса решено было в первую очередь обрабатывать готовые электронные версии текстов (полученные из издательств и открытых электронных библиотек) с тем, чтобы сократить технологический цикл подготовки за счет сканирования и распознавания. И здесь следует отметить положительную тенденцию, характерную для русского сектора интернета: быстрое пополнение отечественных электронных библиотек

---

[7] Работа над корпусом велась в период 2006-2008 гг. в рамках программы ОИФН РАН «Русский язык, литература и фольклор в информационном обществе: формирование электронных научных фондов» и при поддержке РГНФ (грант № 06-04-03817в).

[8] Описание см. в работе Савчук С.О., Пискунова С.В. Опыт создания корпуса текстов первой половины XX века // Труды Международной конференции «Корпусная лингвистика — 2006» 10–14 октября 2006 г. – СПб: Изд-во С.-Петербургского университета, 2006. С. 356-360.

культурно значимыми текстами, значительно опережающее темпы подготовки текстов для аннотированного корпуса. Кроме того, за последние годы было создано немало электронных ресурсов по истории России, содержащих тексты, малодоступные для современного читателя (архивные материалы, газеты, журналы и другие издания начала или 30–40-х годов прошлого века) и представляющие большой интерес для корпуса. Можно упомянуть сайты «СовЛит», «Старые газеты», «Хронос», «Открытая русская электронная библиотека», «Военная литература», ЭБ исторического факультета МГУ и др. Таким образом, в 2006–2008 гг. подготовка текстов первой половины XX века в значительно большей степени, чем это предполагалось заранее, осуществлялась на основе электронных изданий, что позволило превысить первоначально запланированный объем корпуса[9].

Вместе с тем была продолжена работа, начатая еще при подготовке корпуса второй половины XX века, по конвертации текстов из различных форматов в формат XML. При формировании корпуса первой половины XX в. была освоена технология перевода текстов из форматов .pdf, .tif, .jpg, .djvu, поскольку многие тексты (газеты, документы и пр.) представлены в электронных библиотеках в графических форматах. Кроме того, в электронных библиотеках обнаружился дефицит произведений литературы социалистического реализма 30–50-х годов, представляющих интерес для корпуса; электронные версии таких текстов приходится изготавливать путем сканирования и распознавания печатных изданий. Оптимальным же способом подготовки электронных версий рукописных текстов оказался компьютерный набор с последующей корректурой.

Несмотря на то, что хронологическая глубина корпуса первой половины XX века относительно невелика, его разработка потребовала решения тех же задач, что и при формировании корпуса текстов XVIII в. и XIX в. Остановимся на этих задачах подробнее.

---

[9] Основная работа по подготовке электронных версий текстов проведена коллективом разметчиков, среди которых следует отметить Е. Красильщикову, М. Русанову, Е.Н. Морозову, Е.Н. Ловлю. Организационную работу осуществляли С.В. Пискунова и автор этих строк.

1. Состав и структура корпуса

Объем корпуса первой половины XX века в настоящее время составляет более 37 млн словоупотреблений. При отборе текстов для корпуса учитывалась уникальность этого периода в истории русской культуры и русского литературного языка: разнообразие стилей и языковых средств и их стремительная эволюция, раскол русской речевой стихии и параллельное существование двух языковых коллективов — «советского» и «эмигрантского», для которых характерны различные стилевые (отчасти даже собственно языковые) установки.

В корпусе представлены все основные сферы функционирования русского языка, а внутри каждой сферы мы стремились отразить максимальное разнообразие течений общественной мысли и направлений литературного творчества. Прежде всего это касается художественной литературы и публицистики, так как именно в этих сферах общественно-речевой практики происходили самые значительные события, определившие развитие русского литературного языка XX века.

Рубеж XIX и XX столетий (Серебряный век) — эпоха эстетической революции в художественном сознании, период формирования и утверждения художественного сознания модернистского типа, которое наряду с реалистическим стало определять литературный процесс XX столетия[10]. Ведущие эстетические течения русского модернизма (символизм, акмеизм, имажинизм, футуризм) представлены в корпусе прозой и публицистикой А. Белого, А. А. Блока, В. Я. Брюсова, К. Д. Бальмонта, З. Н. Гиппиус, Д.С. Мережковского, Вяч. И. Иванова, И. Ф. Анненского, Ф. К. Сологуба, Н. С. Гумилева, А. А. Ахматовой, Г. В. Иванова, О. Э. Мандельштама, М. А. Кузмина, А. Мариенгофа, Р. Ивнева, В. В. Маяковского, В. Каменского, В. Хлебникова, а также авторов, не причислявших себя к какой-либо группировке (А. М. Ремизова, В. В. Розанова, М. А. Волошина, В. Ф. Ходасевича, М. И. Цветаевой и др.). Реалистическое направление представлено творчеством М. Горького, И. А. Бунина, Л. Н. Андреева, В. Г. Короленко, И. А. Куприна, М. Алданова, Б. К. Зайцева, И. С. Шмелева, А. С. Новикова-Прибоя и др.

---

[10] История русской литературы XX века (20-50-е годы): Литературный процесс. Учебное пособие. М.: МГУ, 2006. С. 7.

В послеоктябрьский период прежде единая национальная литература вынужденно разделилась на два потока — литературу метрополии и диаспоры, развитие которых пошло разными путями. В метрополии десятилетие интенсивных идейно-эстетических исканий (1920-е–1932 гг.) сменилось эпохой директивного утверждения в литературе единого господствующего стиля и единого художественного метода — социалистического реализма, что привело к возникновению третьего потока — «потаенной литературы» — произведений, которые публиковались в 20-е годы, но перестали издаваться в 30–50-е годы и нашли путь к читателю только во время оттепели или в конце 80-х годов (М. А. Булгаков, Е. И. Замятин, А. П. Платонов, Л. И. Добычин, Б. Л. Пастернак, И. Э. Бабель, Ю. К. Олеша, Б. А. Пильняк и др.).

В корпусе нашли отражение и многостилье прозы 20-х годов (А. Веселый, Вс. В. Иванов, В. М. Зензинов, А. С. Неверов, Л. Н. Сейфулина, М. М. Зощенко, В. Каверин, Б. А. Лавренев, Д. И. Хармс, В. Шкловский), и творчество писателей, продолжающих традиции реализма (М. Горький, М. М. Пришвин, К. А. Федин, Л. М. Леонов, Б. К. Паустовский, А. Н. Толстой, М. А. Шолохов), и советская литература (Ф. В. Гладков, Б. А. Лавренев, Б. Л. Горбатов, А. Гайдар, Н. Н. Ляшко, В. П. Катаев, А. С. Макаренко, Н. А. Островский, А. С. Серафимович, А. А. Фадеев, Д.А Фурманов, М. С. Шагинян, И. Эренбург), и «потаенная литература» 30–50-х годов (М. А. Булгаков, Е. И. Замятин, А. П. Платонов, Л. И. Добычин, Б. Л. Пастернак, И. Э. Бабель, Ю. К. Олеша, Б. А. Пильняк, К. К. Вагинов, М. М. Зощенко, С. Н. Клычков, С. Д. Кржижановский), и литература зарубежья — как творчество писателей старшего поколения (Д. С. Мережковский, З. Н. Гиппиус, И. А. Бунин, Р. Б. Гуль, В. Ф. Ходасевич, И. С. Шмелев, Б. К. Зайцев, М. А. Осоргин, Г. В. Иванов, Вяч. И. Иванов и др.), так и произведения молодых авторов, пришедших в литературу уже в эмиграции (В. В. Набоков, Г. А. Газданов, Н. Н. Берберова). Учтено жанровое разнообразие художественной литературы: в корпус включены детская литература (В. А. Каверин, А. С. Некрасов, В. Губарев, К. И. Чуковский, Л. И. Лагин, А. М. Волков, И. С. Соколов-Микитов, П. П. Бажов, Б. В. Шергин, Р. Штильмарк), фантастика (А. Р. Беляев, И. А. Ефремов, В. А. Обручев, Я. Ларри), историческая и историко-биографическая проза

(С. Д. Мстиславский, П. П. Муратов, Б. А. Садовской, Ю. Н. Тынянов, О. Д. Форш, Г. И. Чулков, В. Ян), сатирическая и юмористическая проза (А. Т. Аверченко, И. Ильф и Е. Петров, П. С. Романов, Н. А. Тэффи, С. Черный, Д. И. Хармс).

Публицистические тексты составляют в корпусе около 30%. Значимость этой сферы в структуре литературного языка на протяжении XIX века неуклонно росла, что, по мнению В. В. Виноградова, было следствием процесса демократизации русского литературного языка, выражавшегося в продвижении разговорной стихии в книжные стили. К середине XIX века «изящная словесность», художественная речь перестает быть образцом литературной нормы, и «доминирующее положение постепенно занимают стили журнально-публицистической, газетной и научно-популярной речи»[11].

Сферу публицистики в корпусе формируют газетно-журнальные тексты (около 13%) и мемуарно-биографическая литература (около 17%). Общественно-политические тексты отбирались таким образом, чтобы дать представление об острой партийной борьбе начала века и периода революций (Н. И. Бухарин, В. И. Ленин, А. В. Луначарский, Г. В. Плеханов, И. В. Сталин, Л. Д. Троцкий, П. Н. Милюков, П. А. Новгородцев, Б. В. Савинков, П. А. Сорокин, И. Л. Солоневич, Н. С. Трубецкой, Н. В. Устрялов и др.). Газетные тексты («Правда», «Известия», «Звезда», «Борьба», «Гудок», «Пионерская правда», «Культурная жизнь», «Ленинградский университет» и др.) отражают изменения в стиле советской агитации и пропаганды в период 1922–1950 гг.

Что касается мемуарно-биографических текстов, то они разнообразны с точки зрения социальной, политической и профессиональной принадлежности их авторов. Больше всего в корпусе дневников и мемуаров писателей и журналистов (М. А. Алданов, И. Э. Бабель, П. П. Бажов, П. Д. Боборыкин, В. В. Вишневский, В. А. Гиляровский, Л. Я. Гинзбург, Б. К. Зайцев, Б. К. Лившиц, Ю. К. Олеша, М. М. Пришвин, М. И. Цветаева, В. Ф. Ходасевич, Л. К. Чуковская и мн. др.). Значительное место занимают воспоминания политиче-

[11] Виноградов В.В. Очерки по истории русского литературного языка XVII–XIX веков. М., 1982. С. 423.

ских и военных деятелей (С. Ю. Витте, Л. М. Каганович, Н. И. Махно, С. П. Мельгунов, Н. Н. Суханов, Л. Д. Троцкий, В. М. Чернов, П. Н. Врангель, А. И. Деникин, А. А. Игнатьев, П. Г. Курлов), деятелей искусства и культуры (Н. Ф. Балиев, С. М. Волконский, И. М. Дьяконов, В. И. Мухина, И. Е. Репин, К. С. Станиславский, П. Н. Филонов, Ф. И. Шаляпин и др.), науки и техники (П. К. Козлов, А. Н. Крылов, Е. М. Мелетинский, И. И. Сикорский, А. С. Яковлев).

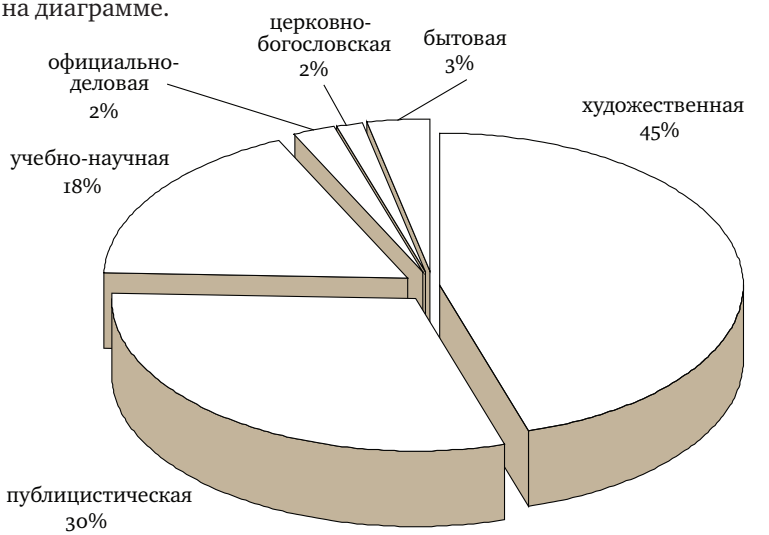
Учебно-научные тексты составляет 18% в общем объеме корпуса. Репрезентативность этой сферы достигается тем, что, с одной стороны, в корпусе собраны тексты, относящиеся к разным областям знания. С другой стороны, тексты научной сферы неоднородны по жанровой принадлежности: наряду с собственно научными статьями и монографиями в корпус включены учебные тексты (учебники и учебные пособия) и научно-популярная литература, возникновение и массовое распространение которой приходится как раз на первую половину XX века (популярные брошюры, художественно-познавательная литература и под.).

Широко представлены гуманитарные науки, в несколько меньшей степени — естественные и прикладные науки: филология (М. М. Бахтин, П. М. Бицилли, В. В. Виноградов, М. О. Гершензон, Н. К. Гудзий, А. К. Дживелегов, В. Я. Пропп, Ю. Н. Тынянов, О. М. Фрейденберг, К. И. Чуковский, Л. В. Щерба, Л. П. Якубинский), история (А. В. Арциховский, Ф. Ф. Зелинский, В. В. Зеньковский, Н. Г. Порфиридов, Е. В. Тарле), философия и культурология (Л. С. Аксельрод, Н. А. Бердяев, С. Н. Булгаков, И. А. Ильин, Л. П. Карсавин, Н. О. Лосский, Н. К. Рерих, В. В. Розанов, Г. П. Федотов, П. А. Флоренский, Г. Флоровский, С. Л. Франк, Л. И. Шестов, Н. С. Трубецкой), социология и право (П. А. Сорокин, Н. В. Устрялов, П. А. Новгородцев, А. Ф. Кони), психология (С. Л. Рубинштейн); математика (А. Н. Крылов), физика (П. Л. Капица), химия (А. Е. Арбузов, Н. Д. Зелинский), геология, география (Д. Н. Анучин, А. Е. Ферсман), биология, медицина (В. М. Бехтерев, Н. И. Вавилов, П. Б. Ганнушкин, В. Х. Кандинский, В. А. Гиляровский, Ю. В. Каннабих, И. И. Мечников, И. П. Павлов, П. П. Семенов Тян-Шанский, И. В. Мичурин), техника (И. И. Сикорский, Н. А. Рынин). Наблюдающийся в текущем составе корпуса перекос в сторону текстов гуманитарных наук имеет временный характер и будет устранен по мере пополнения корпуса новыми текстами.

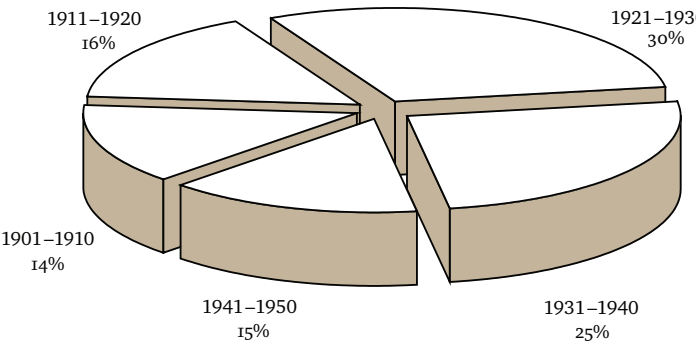
Официально-деловая сфера (около 2%) представлена текстами периода революций начала века, Великой Отечественной войны 1941–1945 гг. Наряду с партийными, правительственными, законодательными, дипломатическими документами в корпусе присутствует административная и канцелярская документация (протоколы, отчеты, приказы, донесения, докладные записки и т.д.). В сфере официально-деловой письменности после Октябрьской революции происходят значительные изменения, связанные с перестройкой государственно-административного аппарата. Витиеватость и архаика, характерная для официального стиля конца XIX века, сменяется аффектированностью и избыточной метафоричностью документов первых лет советской власти (принятой в агитационно-пропагандистской литературе), а затем, в 30–50-е годы, на смену ей приходит нейтрализация и стандартизация официальной речи. Все эти особенности можно изучать на материале документов, собранных в корпусе.

Кроме того, в корпусе представлены также тексты, изначально не предназначенные для публикации: частные дневники, личная переписка, составляющие бытовую сферу (около 3%).

Распределение текстов по сферам функционирования показано на диаграмме.



По периодам создания тексты распределяются следующим образом:



2. Проблема орфографической вариативности

Поскольку корпус первой половины XX-го века является частью Национального корпуса, тексты, включенные в него, должны быть переданы только средствами современной орфографии. Это влечет за собой проблему редактирования оригинала, связанную с орфографической модернизацией текстов дореволюционных изданий. Редактирование текстов в НКРЯ осуществляется в соответствии с эдиционными принципами, принятыми для изданий академического типа или близких к ним (в том числе для филологических электронных библиотек), согласно которым электронная версия приводится в соответствие печатной. Таким образом, если воспроизводится современное издание текстов первой половины XX века, то орфография в нем соответствует правилам 1956 года; при воспроизведении текстов, изданных до 1956 года, а также дореволюционных и эмигрантских изданий в них сохраняются все особенности орфографических норм соответствующего периода, за исключением тех изменений в графике, которые были внесены реформой 1918 года (исправляются только такие написания, которые могут быть восстановлены автоматически, например, ъ после твердого согласного в конце слова, і перед гласным и й и т. д.).

Возникающая при этом множественность орфографических вариантов передачи одного и того же слова или формы может представлять интерес для специалистов, изучающих историю и совре-

менное состояние орфографических норм, однако создает проблемы при лингвистической аннотации текстов и поиске в корпусе. Решить эту проблему предлагается путем нормализации орфографии и расширения словаря за счет внесения в него вариантов, в том числе орфографических.

Нормализация орфографии не означает ее унификацию в текстах в соответствии с современными правилами. Ее назначение состоит не в том, чтобы исправить в тексте все отклонения от современных норм, а в том, чтобы снабдить все вариативные написания соответствующим нормативным вариантом. В процессе морфологической разметки разбирается нормативная форма, а набор грамматических признаков приписывается всему комплексу, так что на поисковый запрос выдаются контексты, содержащие запрашиваемое слово во всех вариантах написания, при этом оно отображено на экране в том реальном виде, в котором представлено в тексте.

Хотя эта операция требует дополнительных затрат труда лингвиста-эксперта, они оправданы тем, что во-первых, на выходе сохраняется оригинальная орфография текста, во-вторых, обеспечивается поиск всех орфографических вариантов слова по морфологическим признакам (без этой операции найти в корпусе устаревший вариант написания можно только при точном поиске), в-третьих, происходит пополнение словаря корпуса. В словаре формируются единицы (леммы), объединяющие словоформы не только в современных, но и в вариативных написаниях, соответствующих нормам предшествующих периодов. Например, *инфлюэнца* = f,inan,nom,norm {инфлюэнца| инфлуэнца| инфлуенца| инфлюэнция| инфлуэнция| инфлюенция}[12]. Предполагается, что по мере пополнения состава таких единиц ручная обработка

[12] Для наименования таких единиц предложен термин орфографическая лемма, или—шире—гиперлемма, если учесть, что такая единица может объединять не только орфографические, но и морфологические варианты. Аналогичное решение предложено разработчиками Чешского национального корпуса, см.: Kučera, K. Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora // Levická, J.; Garabík, R. (ed.). Computer Treatment of Slavic and East European Languages. Slovanské a východoeurópské jazyky v počítačovom spracovaní. Bratislava: Slovak National Corpus, Slovak Academy of Sciences, 2007, s. 121–125; ср. также Meyer, R. The Regensburg Diachronic Corpus of Russian // Труды международной конференции «Корпусная лингвистика-2006». СПБ, 2006. С. 244.

текстов будет уменьшаться, и варианты будут опознаваться автоматически.

3. Проблема грамматической вариативности

Помимо орфографических вариантов корпус текстов первой половины XX века отличается повышенной степенью вариативности на других уровнях — морфологии, словообразования, синтаксиса. Морфологические варианты, которые в словаре корпуса, отражающем современную литературную норму, не опознаются как формы соответствующих слов и недоступны при поиске, предполагается включить в состав леммы, с тем чтобы они получали морфологическую аннотацию наряду со стандартными формами (как это сделано для вариантных форм тв. п. сущ. жен. р. на -ой/-ою, -ей/-ею). Это касается прежде всего таких частотных случаев, как варианты слов с основами на –j-: *сомненье/сомнение, уменье/умение, питанье/питание* и др. (такие формы, как на *распутьи*, в *поместьи*, в *нетерпеньи* и вовсе получают неправильные разборы, например распутьи), форм род. п. сущ. (*грузинов, турков, сапогов, яблоков, грабель* при нормативных формах *грузин, турок, сапог, яблок, граблей*) и т.д.

Словообразовательные, фонетические, лексические варианты (*импровизованный, патентирование, гиероглифы, конфекты, шкап, двухкратный* и под.) могут пополнить словарь в статусе самостоятельных единиц.

Однако эта гипотеза требует дальнейшей проверки на материале корпуса, которая позволит выяснить, насколько такое пополнение словаря будет способствовать уменьшению количества ошибочных разборов. Другой способ снижения шума, который в настоящее время тестируется программистами, — это обучение программы-парсера на подкорпусах однородных текстов (например, устных, диалектных или XVIII–XIX вв.) и настройка таких программ на морфологическую разметку текстов определенного типа. По мнению специалистов, такая настройка позволит программе приписывать словоформе наиболее вероятные разборы.

4. Расширение словаря

Исторические корпуса содержат большое количество несловарных слов — единиц, не отраженных современными словарями и потому не вошедших в словарь корпуса. Это архаизмы, историзмы, окказионализмы и специфические для текстов первой половины XX века советизмы, не удержавшиеся в языке и перешедшие в разряд устаревших слов. В частности, официальные документы и публицистика первой половины XX века дают многочисленные примеры образования разных категорий слов по продуктивным моделям: *взаимоприспособление*, *благовоззрение, главноначальствующий*, *главноуправляющий,* в *противность* последующим увериениям, *невыборка* номерного знака, *доразборка* частей, *неприсылка* снарядов; при обнаружении *нерегистрации* и *несообщении* в Горсовет, идея *приравнения, с целью подыскания, факт получения и скрытия, неродимость северной почвы, незакономерность действий, неблагомысленность, неблагоидейность; предварилка (тюрьма предварительного заключения), Учредилка (учредительное собрание), потребилка (потребительская кооперация), генералка (генеральная репетиция), обогатилка (обогатительная фабрика), реалка (реальное училище),* взрыв *бензинки; полуукоризненно,* к *полуцирковому* «Горячему сердцу», *полусовдеповское* временное правительство, *полуброненосный* фрегат, *полуспособный*, *полуобщественный; архиправославная* семья, *архикабинетный* человек, *архиполицейский* и т.д.

Большой интерес представляют окказионализмы: ее *быстроговорение* все-таки не поспевает за *богатствомыслием* [Л. К. Чуковская. Памяти Тамары Григорьевны Габбе]. Можно смело сказать, что опыт этой работы положил начало новой отрасли инженерии и искусства — *статуестроению* [В. И. Мухина. Автобиография]. Ничего одиноче его вечной *обступленности*, *обсмотренности*, *обслушанности* я не знала [М. И. Цветаева. Пленный дух (Моя встреча с Андреем Белым) (1934)]. …как новые биологи утверждают *всюдность* жизни, так и я убежден, что близкие мне люди находятся почти в равном числе во всякой среде [М. М. Пришвин. Дневники (1929)], Его *Высокотоварищество* Господин Пролетарий вышел откуда-то из трущобы и занял место Его Высокопревосходительства [М. М. Пришвин. Дневники (1917)]. Не помогло *рапполепство*. За упокой РАППа божия [Л. Гинзбург. Записные книжки. Воспоминания. Эссе (1920–1943)].

Изучение этого материала позволит выявить активные способы пополнения словаря языка в изучаемую эпоху и, возможно, уточнить сложившиеся представления и разрушить стереотипы. В частности, на основании работ 20-х годов (С. И. Карцевский, А. М. Селищев, Е. Д. Поливанов) сложилось мнение, что обилие аббревиатур — исключительная особенность языка советской России. Однако расширение анализируемого материала показывает, что, во-первых, модель становится продуктивной еще до революции, в начале XX века, и, во-вторых, активно используется в 20-е годы не только в советской России, но и в речи эмиграции (Грановская 2005: 212–216, 252–256). Материалы корпуса подтверждают это наблюдение. Так, в дореволюционной служебной переписке встречаем: *Гос. Дума*, *Мориском* (Морская историческая комиссия), *шифртелеграмма, комфлота, старлейт, каперанг, кавторанг, главарт, штафлот, наштафлот, натрадив, намобор, наоперу, главкомев* (М. К. Бахирев, Отчет о действиях Морских сил Рижского залива 29 сентября — 7 октября 1917 г.). Многочисленные аббревиатуры из советской прессы 1920–1930-х годов (*комчванство, химопыты*, *спецгазометы, регсбор* (регистрационный сбор), *завдомы, партаппарат, комвуз, крайКК* РКИ, *наркомзем, колхозцентр, райколхозсоюз, трудкнижка, техучеба* по *техпропаганде, партполитработа,* полевые *культстаны, культбригада*, *агитпропгруппа, агитмашина* и т. д.) соседствуют с аналогичными примерами из текстов, созданных за пределами России: *главковерх, Главком, Командарм, командармдобр* (Командующий Добровольческой армией), *Донармия*, *Добрармия*, *ВСЮР* (Вооруженные силы Юга России), *ревком, эс-эры*, *совдеп*, *совдепия* и т. д.

Часть несловарных единиц, а именно тех, которые преодолели определенный порог частотности, целесообразно включить в словарь корпуса.

5. Перспективы развития корпуса текстов первой половины XX века

На ближайшее будущее разработчики корпуса ставят перед собой следующие задачи. Во-первых, планируется пополнение корпуса новыми текстами, пока недостаточно в нем представленными и прошедшими процесс соответствующей орфографической обработки. Прежде всего,

это касается текстов, относящихся к периоду 1900–1920-х гг. В корпусе пока слабо отражена бурная философская, научная, литературная полемика начала века и 1920–1930-х годов (например, дискуссии о формализме, о реформе орфографии, о евразийстве, о фрейдизме, о «физическом идеализме», манифесты литературных школ и группировок и под.); ораторская практика эпохи революции и гражданской войны (вспомним, какое внимание уделяла советская власть агитации и пропаганде). Планируется расширить состав газетных и журнальных текстов, существенно пополнить естественнонаучными текстами научный раздел. Не следует забывать также о еще одной составной части корпуса первой половины XX века, которая формально является принадлежностью корпуса устной речи, транскриптов фильмов 30–40-х годов (около 150 тыс. словоупотреблений).

Во-вторых, предполагается проанализировать состав несловарных форм, выделенных в текстах первой половины XX в., произвести ручную лемматизацию орфографических вариантов и отобрать возможных кандидатов для пополнения словаря корпуса.

В настоящее время проанализирован список орфографических вариантов, подготовленный на основе списков, составляемых разметчиками в процессе редактирования текстов. В нем около 600 слов. Большая часть вариантов (около 17%) связана с написанием иноязычных корней. Колебания отмечены в следующих типах орфограмм: написания удвоенных согласных (*аггрегат, аллюминиевый, пуддинг, диффракция, баттарея, веррсия, галлерея, корридор*, которым соответствуют современные написания с одиночными согласными, и *афект, амиак, брилиант, пресованный, геена, гутаперчевый, диференциация,* которым по современным нормам соответствуют написания с удвоенными согласными); написания *э* и *е*: *кафэ, канапэ, купэ, кабарэ, кашнэ, декольтэ, пенснэ, проэкт, траэктория* (ср. совр. *кафе, канапе, купе, кабаре, кашне, декольте, пенсне, проект, траектория*) и *елоквенция, ерудиция, ефиопка* (ср. *элоквенция, эрудиция, эфиопка*); дефисные написания (*порт-плэд, виц-мундир, деми-сезоны, колд-крем* ср. *портплед, вицмундир, демисезоны, кольдкрем*); отдельные написания (*ягдаш*, *эксплоатация*, *кибаб*, *конверзия* ср. совр. *ягдташ*, *эксплуатация*, *кебаб*, *конверсия*). Немало колебаний в написании иностранных имен собственных: *Ботичелли, Савонаролла, Верлэн, Мадлэн, Фихтэ, Уот/Уольт Уитман, Массачуветс* (ср. совр.

Боттичелли, Савонарола, Верлен, Мадлен, Фихте, Уолт Уитмен, Массачусетс и т.д.).

Вторая по величине группа орфографических вариантов — написание наречий: дефисное, которому соответствует современное раздельное (*как-раз, бок-о-бок, друг-дружку, на-бегу, на-днях, заграницу, на-нет* ср. *как раз, бок о бок, друг дружку, на бегу, на днях, за границу, на нет*) и современное слитное написание (*во-время, на-вылет, на-готове, на-долго, по-долгу*, ср. совр. *вовремя, навылет, наготове, надолго, подолгу*); раздельное, которому соответствует нормативное слитное (*в повалку, в роде, за панибрата, на ряду, на веки* ср. совр. *вповалку, вроде, запанибрата, наряду, навеки*); слитное, которому соответствует современное дефисное (*повидимому* ср. *по-видимому*).

Многочисленны колебания в написании сложных слов — существительных и прилагательных: *анти-национализм, архи-глупость, кино-театр, контр-разведка, контр-революция, пионер-отряд, радио-волна, пол-дорога, пол-победы, пол-фунта, пол часа, Вышний-Волочек, Нижний-Новгород; агро-технический, гидро-авиационный, древне-греческий, западно-европейский, мелко-буржуазный, сельскохозяйственный, светлорусый, темнобурый, яркозеленый* и др.; ср. совр.: *антинационализм, архиглупость, кинотеатр, контрразведка, контрреволюция, пионеротряд, радиоволна, полдорога, полпобеды, полфунта, полчаса, Вышний Волочок, Нижний Новгород; агротехнический, гидроавиационный, древнегреческий, западноевропейский, мелкобуржуазный, сельскохозяйственный, светло-русый, темно-бурый, ярко-зеленый*).

Как видно из этих примеров, в ходе реформы 1956 года было значительно сокращено количество дефисных написаний в сложениях, что также коснулось и написания частиц *бы(б), будто, ли(ль), же, то*: *следовало-бы, как-будто, однако-ж, опять-же, приведет-ли, то-есть*, ср. *следовало бы, как будто, однако ж, опять же, приведет ли, то есть* и т.д.

Довольно многочисленны варианты написания орфограмм в русских корнях, среди них: *о* или *е* после шипящих (*жолудь, чорт, шопот, шолк, решотка* вм. *желудь, черт, шепот, шелк, решетка*), *и* или *ы* после Ц (*цыфра, цынга, панцырь* вм. *цифра, цинга, панцирь*), удвоенные согласные (*белоруссы, черкесска* вм. *белорусы, черкеска*), чередо-

вание гласных в корне (*возрасла, наростать, срощение, выравнять, пловучий, зорница* вм. *возросла, нарастать, сращение, выровнять, плавучий, зарница*), глаголы *итти*, *притти* (совр. *идти*, *придти*).

Вариативны написания суффиксов существительных (*зрачёк, крючёк, толчёк, волчёнок, ручёнки, семячек, масляница* и др., ср. совр. *зрачок, крючок, толчок, волчонок, ручонки, семечек, масленица*), прилагательных (*большевицких, дешовый, парчевый, серебрянный, смышленный* ср. совр. *большевистских, дешевый, парчовый, серебряный, смышленый*), глаголов (*заведывать, проповедывать, гарцовать, танцовать* ср. совр. *заведовать, проповедовать, гарцевать, танцевать*).

Проведенный анализ позволит продолжить отбор орфографических вариантов по всей диахронической части корпуса, которые затем будут внесены в состав соответствующих лемм, с тем чтобы обеспечить грамматический поиск по всем возможным способам орфографической передачи словоформ.

Помимо теоретической значимости корпус текстов первой половины XX века имеет большое прикладное значение, прежде всего для лексикографии. Материалы корпуса активно используются при работе над новым изданием Большого академического словаря, дополняя материалы Большой словарной картотеки ИЛИ РАН. Корпус рассматривается как основной источник при создании Словаря русского языка первой половины XX века, проект которого готовится к обсуждению в ИЛИ РАН (Гердт 2008, 144–147). Все это свидетельствует о своевременности создания этого лингвистического ресурса и его востребованности, открывает перспективы и вместе с тем уточняет направления его развития.

Литература

Бельчиков Ю. А. Русский язык. XX век. М., 2003.

Виноградов В. В. Очерки по истории русского литературного языка XVII–XIX веков. М., 1982. С. 423.

Гердт А. С. Национальный корпус русского языка — Словарная картотека — Академический словарь // Труды Международной конференции «Корпусная лингвистика — 2008». 6–10 октября

2008 г. — СПб: Изд-во С.-Петербургского университета, 2008. С. 143–147.

Грановская Л. М. Русский литературный язык в конце XIX и XX вв. М., 2005.

Живов В. М. Язык и революция. Размышления над старой книгой А. М. Селищева // Отечественные записки. 2005. №2.

История русской литературы XX века (20–50-е годы): Литературный процесс. Учебное пособие. М.: МГУ, 2006.

Карцевский С. И. Язык, война и революция // Карцевский С. И. Из лингвистического наследия. Т. 1. М., 2000.

Мещерский Н. А. История русского литературного языка. Л., 1981.

Ожегов С. И. К вопросу об изменениях словарного состава в русском языке в советскую эпоху // Вопросы языкознания. 1953. № 2.

Поливанов Е. Д. Революция и литературные языки Союза ССР // За марксистское языкознание. М., 1931. С. 73–94.

Савчук С. О., Пискунова С. В. Опыт создания корпуса текстов первой половины XX века // Труды Международной конференции «Корпусная лингвистика — 2006». 10–14 октября 2006 г. — СПб: Изд-во С.-Петербургского университета, 2006. С. 356–360.

Селищев А. М. Язык революционной эпохи: Из наблюдений над русским языком последних лет. 1917–1926 // Селищев А. М. Труды по русскому языку. Т. 1. М., 2003.

Скворцов Л. И. Сергей Иванович Ожегов — человек и словарь. М., 2001.

Corpus Linguistics: Critical Concepts in Linguistics. Ed. By W. Tewbert & R. Krishnamurthy. V.I. L; NY: Routledge, 2006. P. 32–33.

Gau, M. The State of Historical Corpus Linguistics with Special Focus on the Russian Language. M. A. thesis, University of Regensburg, Institute for Slavonic Languages and Literatures, 2005. *http://www.uni-r.de/Fakultaeten/phil_Fak_IV/Korpuslinguistik/meyer/PDF/melanie.pdf.*

Kučera, K. Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora // Levická, J.; Garabík, R. (ed.). Computer Treatment of Slavic and East European Languages. Slovanské a východoeurópské jazyky v počítačovom spracovaní. Bratislava: Slovak National Corpus,

Slovak Academy of Sciences, 2007, pp. 121–125.
Meyer, R. The Regensburg Diachronic Corpus of Russian // Труды международной конференции «Корпусная лингвистика–2006». СПБ, 2006. С. 244.
Onelli, C., Proietti, D., Seidenari, C., Tamburini, F. The DiaCORIS project: a diachronic corpus of written Italian // Proceedings of the 5th International Conference on Language Resources and Evaluation. Genoa, 2006.
Xiao R. Z. Diachronic corpora // Xiao R.Z. Well-known and influential corpora: A survey. In Lüdeling A., Kytö M., McEnery A. (eds.) Corpus Linguistics: An International Handbook. Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin, Germany, 2007. *http://postgrad/xiaoz/papers/corpus%20survey.htm*.