С.О. Савчук

КОРПУС ТЕКСТОВ ПЕРВОЙ ПОЛОВИНЫ XX ВЕКА: ИЗ ОПЫТА РАБОТЫ[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 90-е годы 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[2].

Одни исследователи предлагают начинать отсчет дооктябрьского периода с 70-х[3] или 90-х[4] годов XIX века, связывая общий вектор развития языка с процессом демократизации общественной жизни. Октябрьская революция при этом рассматривается

[1] Работа выполнена при поддержке РГНФ, грант № 06-04-03817в.

[2] *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.

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

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

как фактор, ускоривший эволюционные процессы[1]. 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.[2]

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[3]. В этих условиях описание конкретного опыта разработки исторического корпуса, как кажется, может представлять интерес для специалистов.

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 них подробнее.

---

[1] *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.

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

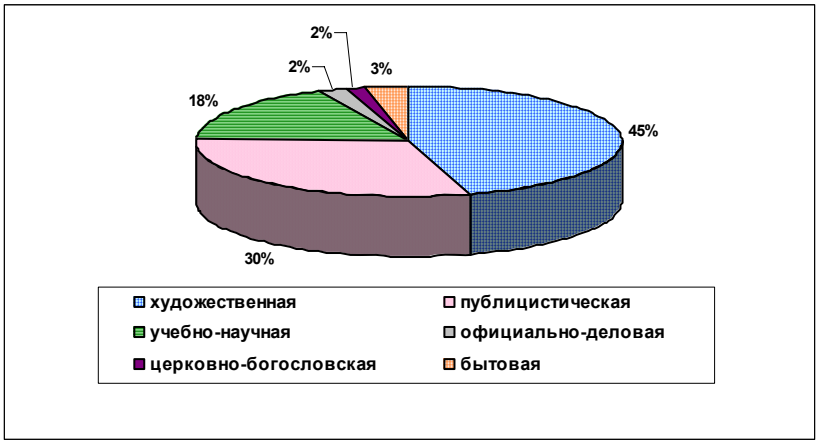
[3] *Corpus Linguistics*: Critical Concepts in Linguistics / Ed. by W. Tewbert & R. Krishnamurthy. V.I. L; NY: Routledge, 2006. P. 32–33; *Onelli C., Proietti D., Seidenari C., Tamburini F.* The DiaCORIS project: a diachronic corpus of written Italian // Proceedings of the 5th International Conference on Language Resources and Evaluation / Genoa, 2006.

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 существования двух языковых коллективов – «советского» и «эмигрантского», для которых характерны различные стилевые (отчасти даже собственно языковые) установки. В состав авторов включены писатели, представляющие основной спектр литературных направлений (Л.Н. Андреев, М.А. Алданов, П.П. Бажов, А.А. Бек, М. Горький, В.А. Гиляровский, А.Н. Толстой, А.И. Куприн, Ю.К. Олеша, М.М Пришвин, М.А. Булгаков, А.С. Грин, А. Белый, В.Я Брюсов, Н.С Гумилев, Г. Газданов, В. Жаботинский, Б.К. Зайцев, Е.И. Замятин, В.П. Катаев, С.Д. Кржижановский, В.В. Маяковский, В.В. Набоков, А.П. Платонов, Н.А. Тэффи, М.И. Цветаева, В.Ф. Ходасевич и мн. др.), религиозные мыслители и ученые (Н.А. Бердяев, С.Н. Булгаков, П.А. Флоренский, С.Л. Франк, Г. Флоровский, М.М. Бахтин, В.В. Виноградов, В.Я. Пропп, О.М. Фрейденберг, Ю.Н. Тынянов, В.М. Бехтерев, Е.В. Тарле, В.И. Вернадский, Н.И. Вавилов, Л.С. Выготский, П.Б. Ганнушкин, Н.Д. Зелинский, П.Л. Капица, А.Ф. Кони, П.К Козлов, А.Н. Крылов, И.И. Мечников, Н.А. Рынин, А.Е. Ферсман и др.), политические деятели (С.Ю. Витте, П.Н. Врангель, В.И. Ленин, Г.В. Плеханов, И.В Сталин, П.А. Столыпин, Л.Д. Троцкий и др.) – всего более 300 имен.

В корпусе представлены тексты, относящиеся к различным функциональным сферам: художественная литература различных жанров и направлений, критика, публицистика, в том числе тексты журналов и газет, агитационные тексты, мемуары, научные и философские работы, официальные документы, а также тексты, изначально не предназначенные для публикации: частные дневники, личная переписка.

Распределение текстов по сферам функционирования показано на диаграмме.



По периодам создания тексты распределяются следующим образом: 1901–1910 – 5,2 млн, 1911–1920 – 5,8 млн, 1921–1930 – 11,1 млн, 1931–1940 – 9,1 млн, 1941–1950 – 5,6 млн.

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, связанного 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, издан-

ных 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 т.д.).

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

варе 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}[1]. Предполагается, что по мере пополнения состава таких единиц ручная обработка 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 неправиль-

[1] 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 принято разработчиками других диахронических корпусов, см.: *Kučera K.* Hyperlemma: A Concept Emerging from Lemmatizing Diachronic Corpora // Computer Treatment of Slavic and East European Languages / Editoři: J. Levická, R. Garabík. Bratislava: Tribun, 2007; S. 121–125; *Meyer R.* The Regensburg Diachronic Corpus of Russian // Труды международной конференции «Корпусная лингвистика–2006». СПб., 2006. С. 244.

ные 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 позволит снизить количество ошибочных разборов. Другой способ снижения шума, который в настоящее время опробуется программистами, – это обучение программы-парсера на подкорпусах однородных текстов (например, разговорных, 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 публицистка первой половины XX века дают многочисленные примеры образования разных категорий слов по продуктивным моделям: *взаимоприспособление*, *благовоззрение, главноначальствующий*, *главноуправляющий,* в *противность* последующим уверениям, *невыборка* номерного знака, при обнаружении *нерегистрации* и *несообщении* в Горсовет, идея *приправ-*

нения, с целью подыскания, неродимость северной почвы; комчванство, химопыты, *спецгазометы, регсбор* (регистрационный сбор), *завдомы, партаппарат, комвуз, крайКК* РКИ, *наркомзем, колхозцентр, райколхозсоюз, трудкнижка, техучеба* по *техпропаганде, партполитработа,* полевые *культстаны, культбригада*, *агитпропгруппа, агитмашина; полуукоризненно,* к *полуцирковому* «Горячему сердцу», *полусовдеповское* временное правительство, *полуброненосный* фрегат, *полуспособный*, *полуобщественный*.

Изучение этого материала позволит выявить активные способы пополнения словаря языка в изучаемую эпоху. Часть этих единиц, а именно тех, которые преодолели определенный порог частотности, целесообразно включить в словарь корпуса.

На ближайшее будущее разработчики корпуса ставят перед собой следующие задачи. Во-первых, планируется пополнение корпуса новыми текстами, пока недостаточно в нем представленными (в частности, относящимися к периоду 1900–1920-х гг.), прошедшими процесс соответствующей орфографической обработки.

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.