

© 2009

А.В. Венцов, Е.В. Грудева

**АКЦЕНТНО РАЗМЕЧЕННЫЙ КОРПУС РУССКОГО  
ЛИТЕРАТУРНОГО ЯЗЫКА КАК ИСТОЧНИК НОВЫХ СЛОВАРЕЙ  
(«СЛОВАРЬ ОМОГРАФОВ РУССКОГО ЯЗЫКА» И «ЧАСТОТНЫЙ  
СЛОВАРЬ СЛОВОФОРМ РУССКОГО ЯЗЫКА»)**

В настоящее время активно развивается такое относительно новое прикладное направление в языкознании, как корпусная лингвистика. Особенно большие надежды возлагаются на создание больших (национальных) аннотированных корпусов того или иного языка. Уже более 10 лет существует Британский национальный корпус, который во многом является ориентиром для разработчиков корпусов других языков. Существует несколько аннотированных корпусов русского языка, которые отличаются друг от друга как объемом материала, так и качеством его обработки. Как отмечают создатели русскоязычных корпусов, организация Национального корпуса русского языка — абсолютно необходимая предпосылка для создания новой академической грамматики и академического словаря русского языка, которые послужили бы базой для разработки семейства грамматик и словарей разной ориентации, в том числе школьных, а также самых разных пособий и справочников ([Вербицкая, Казанский, Касевич 2003], [Плунгян 2005] и др.).

Мы остановимся на некоторых особенностях Корпуса русского литературного языка ([www.narusco.ru](http://www.narusco.ru)), создаваемого в лаборатории моделирования речевой деятельности Санкт-Петербургского государственного университета (научный руководитель лаборатории — доктор филологических наук, профессор В.Б. Касевич), и в связи с этим на двух словарях, которые изданы благодаря имеющемуся в нашем распоряжении корпусу.

Корпус русского литературного языка (далее — КРЛЯ) объемом 1 млн. словоупотреблений создан силами сотрудников лаборатории в период с 2002 по 2007 гг. КРЛЯ представляет собой совокупность четырёх относительно самостоятельных подкорпусов: художественной литературы (30 % от общего объема корпуса), публицистики (30 %), научно-популярной литературы (20 %), а также драмы как некоторого приближения к зафиксированной на письме разговорной речи (20 %). Все тексты относятся к периоду с 50-х гг. XX в. до нашего времени.

Особенностью корпуса является то, что все тексты в нём акцентуированы (каждой словоформе приписан символ ударения), последовательно восстановлена буква «ё», а также специальной разметке подверглись так называемые «составные слова» (единицы типа *в обнимку*, *в головах*, которые по всем лексико-грамматическим признакам являются словами, но пишутся — в силу традиции — отдельно).

Поскольку в типологическом отношении русский язык относится не только к языкам с развитой морфологией (флективным), но ещё и к акцентным языкам, причём, как известно, ударение в русском языке разноместно и подвижно, то тем самым каждая единица текста (словоформа) должна быть акцентно оха-

рактирована. Долгое время исследователи русского языка недооценивали роль омографии в русских текстах, в учебных пособиях обычно приводится 2-3 примера (типа *крУжки — кружкИ, зАмок — замОк, мУка — мука*), и в результате складывалось впечатление, что явление омографии в русском языке носит периферийный характер.

В ходе акцентной разметки всех текстов, вошедших в корпус, обнаружилось, что омографы в современном русском языке исчисляются *тысячами*. При этом «сравнительно мало лексем, которые различаются только ударением в своих словарных формах — именно это и объясняет тот факт, что данное явление до сих пор мало привлекало внимание исследователей. Ударение, как своего рода подсобное средство, «работает» в основном «внутри» парадигмы, а также в сфере различения форм, принадлежащих разным парадигмам» [Венцов, Касевич, Сведенцова 2004: 187]. Речь в данном случае идет о том, что в классификации полученных омографических пар (и реже — троек) наиболее наполненными оказались два класса: первый связан с различением за счет ударения двух форм одного и того же глагола — 2-е л. мн.ч. наст. или буд. времени в противоположность 2-му л. мн.ч. императива (типа *лЮбите — любИте, ввАлитесь — ввалИтесь*), а второй представляет собой различение за счет ударения двух словоформ разных слов (типа *бЕлкам — белкАм*), тогда как класс омографов, различающихся ударением в своих словарных формах, по наполняемости занимает одно из последних мест в классификации.

На основании полученного материала и был составлен «Словарь омографов русского языка» [Венцов, Грудева, Касевич, Корешкова и др. 2004]. Словарь содержит более четырех тысяч омографических пар, т.е. слов и форм, которые пишутся одинаково, но читаются (произносятся) по-разному. Материал в словаре упорядочен двояко: в первой части все омографические пары представлены в алфавитном порядке, во второй части те же омографы разбиты по грамматическим классам. В обзорной статье, помещенной в словарь, дается анализ связи между типом омографии и семантикой омографов.

Другой словарь, который также явился результатом работы с акцентуированным корпусом, — частотный словарь словоформ русского языка [Венцов, Грудева 2008]. С одной стороны, выборка текстов в 1 млн. словоупотреблений в настоящее время считается слишком маленькой для создания частотного словаря. Считается также, что гораздо интереснее и ценнее получить частотный словарь на материале, скажем, 100-миллионного корпуса. С другой стороны, оказывается, что создать 100-миллионный корпус лингвистически аннотированных текстов вручную — трудновыполнимая задача. При создании же большого корпуса *автоматически* аннотированных текстов появляется, с нашей точки зрения, недопустимое число ошибок, искажающих представления о языке. Разбор ошибок при составлении частотных словарей на многомиллионных автоматически размеченных корпусах русского языка представлен в следующей публикации [Венцов, Грудева 2007].

В то же время для решения целого ряда задач вполне достаточным оказывается корпус и меньшего объема. Ср. замечание редактора uppsalского частотного словаря Л. Лённгрена: «В наш корпус входит 1 миллион словоупотреблений. На вопрос, достаточно ли этого, однозначного ответа нет: всё зависит от

того, для каких исследований будет употребляться материал корпуса. Например, для изучения относительно высокочастотных явлений в языке достаточно и меньшего объёма выборки. С другой стороны, даже корпус, во много раз превышающий 1 миллион словоупотреблений, не может гарантировать “правильное” ранжирование низкочастотных лексем, составляющих бульшую часть словарной сокровищницы» [Лённгрен 1993: 13-14].

Как известно, частотный словарь языка с развитой морфологией может создаваться как минимум двумя путями, в зависимости от выбора основной единицы словаря, — либо как словарь словоформ, либо как словарь лексем. Практически все известные частотные словари русского языка — это словари лексем, а не словоформ. Для решения многих задач (например, для отбора лексического минимума при обучении иностранному языку), действительно, гораздо важнее иметь представление о частотных рангах именно лексем. Однако для решения многих других проблем, напр., для моделирования процессов восприятия речи, крайне необходим частотный словарь именно словоформ.

Как известно, в русском языке словоформа не всегда равна графическому слову (ср. уже упоминавшуюся выше проблему «составных слов»). Наконец, хорошо известно, что реальной частотностью в языке (особенно в таком морфологически богатом языке, как русский) обладают словоформы, а не лексемы. Это было хорошо показано уже в одном из первых частотных словарей русского языка — словаре Э.А. Штейнфельдт [Штейнфельдт 1963].

Интересно также отметить, что составители частотных словарей отмечают *привычность* основной словарной единицы — лексем, а также тот факт, что при сведении словоформ в лексем лингвисты могут использовать разные принципы, что приводит к более высокой доле субъективности в количественных показателях по лексемам по сравнению с количественными данными по словоформам. Ср. замечание Л. Лённгрена: «Лемма (исходная форма) для каждой словоформы должна указываться вручную. Это означает, что, прежде чем приступить к лемматизации, нужно установить принципы, по которым она будет проводиться. Эти последние могут отличаться от применявшихся в других работах принципов и дать результаты, которые невозможно будет полностью сравнить с уже существующими. С этой точки зрения *количественные языковые факты, опирающиеся только на уровень словоформ, являются более объективными и надёжными*» [Лённгрен 1993: 28-29; курсив наш. — А.В., Е.Г.].

Таким образом, в качестве единицы описания словаря выступает акцентно размеченная словоформа (всего в словарь вошло 133 267 словоформ, включая имена собственные и «составные слова»). Следует отметить, что по многим показателям данный словарь создан впервые, поскольку аналогов акцентно размеченного КРЛЯ, в котором последовательно восстановлена буква Ё и маркированы «составные слова», в настоящее время, насколько нам известно, не существует.

Словарь состоит из следующих частей:

1. Алфавитно-частотный список словоформ;
2. Частотный список словоформ;
3. Алфавитно-частотный список словоформ подкорпуса «драма»;
4. Частотный список словоформ подкорпуса «драма»;

5. *Алфавитно-частотный список словоформ подкорпуса «художественная литература»;*
6. *Частотный список словоформ подкорпуса «художественная литература»;*
7. *Алфавитно-частотный список словоформ подкорпуса «публицистика»;*
8. *Частотный список словоформ подкорпуса «публицистика»;*
9. *Алфавитно-частотный список словоформ подкорпуса «наука»;*
10. *Частотный список словоформ подкорпуса «наука»;*
11. *Алфавитно-частотный словарь омографов;*
12. *Алфавитно-частотный список «составных слов»;*
13. *Частотный список «составных слов»;*
14. *Алфавитно-частотный список имен собственных;*
15. *Частотный список имен собственных;*
16. *Алфавитно-частотный словарь омонимов (имя собственное — имя нарицательное);*
17. *Алфавитно-частотный список ритмических структур;*
18. *Частотный список ритмических структур;*
19. *Некоторые статистические сведения.*

В силу большого объема словаря в публикации [Венцов, Грудева 2008] представлено подробное описание словаря и даны его основные разделы в виде ограниченных по тем или иным принципам выборок. Полную версию словаря планируется разместить на сайте [www.naguso.ru](http://www.naguso.ru).

Как видим, корпус даже небольшого объема, но содержащий важную для русского языка информацию об ударении, может послужить источником создания новых словарей русского языка. Авторы надеются, что представленные словари найдут своего пользователя.

#### ЛИТЕРАТУРА

Венцов А.В., Грудева Е.В. К вопросу о создании частотного словаря словоформ русского языка // *Русская языковая личность: материалы шестой школы-семинара (Череповец, 25-27 октября 2007 г.)* / отв. ред. Е.В. Грудева, Р.Л. Смулаковская. — Череповец: Изд-во Череповец. гос. ун-та, 2007. — С.70—80.

Венцов А.В., Грудева Е.В. Частотный словарь словоформ русского языка: Проект. — Череповец: Изд-во Череповецкого гос. ун-та, 2008.

Венцов А.В., Грудева Е.В., Касевич В.Б., Корешкова Е.И., Сведенцова Е.А., Ягунова Е.В. Словарь омографов русского языка. — СПб.: Изд-во Санкт-Петербург. гос. ун-та, 2004.

Венцов А.В., Касевич В.Б., Сведенцова Е.А. Омография, омофония и восприятие речи // *Человек пишущий и читающий: проблемы и наблюдения: материалы междунар. конфер. 14-16 марта 2002 г.*, — СПб.: Изд-во Санкт-Петербург. гос. ун-та, 2004. — С.182—189.

Вербицкая Л.А., Казанский Н.Н., Касевич В.Б. Некоторые проблемы создания национального корпуса русского языка // *НТИ. Сер. 2. Информационные процессы и системы.* — М.: ВИНТИ, 2003. — № 6. — С.2—8.

Частотный словарь современного русского языка / под ред. Л. Лёнгрена. — Uppsala, 1993.

Плунгян В.А. Зачем мы делаем национальный корпус русского языка? // *Отечественные записки.* — М. — 2005. — № 2(23). — С.

Штейнфельдт Э.А. Частотный словарь современного русского литературного языка. — Таллин: Изд-во НИИ педагогики Эстон. ССР, 1963.

ACCENTUATED TEXT CORPUS OF LITERARY RUSSIAN AS A SOURCE OF NEW  
DICTIONARIES (THE DICTIONARY OF RUSSIAN HOMOGRAPHS  
AND THE FREQUENCY WORD BOOK OF RUSSIAN WORD FORMS)

*A.V. Ventsov, E.V. Grudeva*

The paper is devoted to two new dictionaries of Russian. The dictionary of homographs contains more than four thousand homographic pairs that refute the traditional assumption about a peripheral role of homographs in Russian. The frequency word form book is also unique in its own way as it is the first to employ accentuated word forms as units of word description.

© 2009

К.Р. Галиуллин

ИНТЕРНЕТ-ЛИНГВОГРАФИЯ:  
РУССКИЕ ТЕКСТООПИСЫВАЮЩИЕ СЛОВАРИ

Словари останутся навсегда  
насушной потребностью нашей  
науки.

*И.А. Бодуэн де Куртене*

Анализ тенденций развития словарного дела и информационного потенциала языковых справочников показывает, что наиболее перспективной формой существования словаря является интернет-версия, среди «плюсов» которой:

1) широкий круг пользователей; благодаря обогащению компьютерных технологий интернет-технологиями, что представляет собой очередную информационную революцию, достигается глобальная обобщественность языковых справочников;

2) удобство эксплуатации;

3) многовходовость, возможность многопризнакового поиска;

4) возможности поддержки словаря в актуальном состоянии, постоянного развития, совершенствования (оперативная корректировка текста словаря, внесение необходимых дополнений и т.д.);

5) снятие многих ограничений на объем включаемого в словарь материала;

6) широкие возможности установления связи со сходными сетевыми справочниками и формирования лингвографических интернет-комплексов на основе ресурсов, размещенных как на одном, так и на разных порталах (сайтах).