УДК 811.161.1’374 + 81’33 + 81-119

М. Ю. Мухин

БАЗЫ ДАННЫХ ЛЕКСИКОГРАФИЧЕСКОЙ ГРУППЫ «РУССКИЙ ГЛАГОЛ»: СТАТИСТИЧЕСКИЕ ПАРАМЕТРЫ И ВОЗМОЖНОСТИ ИХ ЛИНГВИСТИЧЕСКОЙ ИНТЕРПРЕТАЦИИ*

В статье рассматриваются статистические параметры идеографических баз данных лексикографической группы «Русский глагол» и Уральской семантической школы (руководитель — проф. Л. Г. Бабенко). Обсуждается история создания и развития словарных баз данных. Приводятся примеры лингвистической интерпретации количественного соотношения денотативных сфер и особенностей распределения стилистически маркированной лексики.

К л ю ч е в ы е с л о в а: лексикография; база данных; идеографический словарь; тезаурус; Уральская семантическая школа; статистика; семантический класс.

Лексикографическим базам данных межвузовской проблемной группы «Русский глагол» и Уральской семантической школы недавно исполнилось 20 лет. Первая база на основе СУБД FoxPro была создана на кафедре современного русского языка в 1993 г. для работы с идеографическим «Толковым словарем русских глаголов» [1999] (2-е издание вышло под названием «Большой толковый словарь русских глаголов» [2007]). В среде, работавшей под DOS, были созданы процедуры, обеспечивающие удобный на тот момент ввод данных. Согласно требованиям идеографической организации, данные были размещены в связанных друг с другом таблицах, отражающих уровни семантической иерархии [Мухин]. После заполнения контентом словаря глаголов в базе середины 1990-х гг. было

* Исследование выполнено при финансовой поддержке РГНФ (проект 13-04-00322 «Русская лексика как межчастеречная система: полное идеографическое описание в лексикографических параметрах»), РФФИ (проект № 13-06-00444 «Русские синонимы в системном освещении: структурно-семантический, идеографический, когнитивный и культурологический аспекты») и при поддержке средств, полученных из Программы повышения конкурентноспособности (номер соглашения 02.A03.21.0006).

© Мухин М. Ю., 2014

около 6,5 тыс. записей. Отдельную проблему в то время составлял экспорт информации в текстовые процессоры и издательские системы в силу разницы в форматах, а также технические сложности с рядом внешне совпадающих латинских и кириллических символов.

База данных словаря глаголов в своем развитии прошла ряд этапов в связи с совершенствованием инструментальных возможностей, а также с изменением самого контента: появлением дополнительных типов информации и выделением отдельных записей для членов видовых пар — глаголов совершенного вида, которые привязаны в словаре к вокабулам (глаголам несовершенного вида). Впоследствии по модели глагольного словаря была спланирована архитектура данных для ряда идеографических словарей: «Большого толкового словаря русских существительных» [2005], «Словаря-тезауруса русских прилагательных» [2011], а также для «Словаря-тезауруса синонимов русской речи» [2007] и «Большого толкового словаря синонимов русской речи» [2008]. Эти базы были сделаны уже на основе СУБД MS Access в силу ее доступности для непрофессиональных пользователей и интегрированности в пакет MS Office, обеспечивающей элементарный обмен данных между основными интересующими нас приложениями: Access — Word — Excel.

В 2012–2013 гг. в рамках работы над проектом Универсального идеографического словаря, который в будущем должен содержать описание слов всех частей речи, включая служебные, было проведено слияние данных разных словарей в одну базу и разработка удобной формы для работы с тезаурусом. Таким образом, новая база данных объединила идеографические структуры и словники четырех словарей: 1) «Большого толкового словаря русских глаголов»; 2) «Большого толкового словаря русских существительных»; 3) «Словаря-тезауруса русских прилагательных»; 4) «Словаря-тезауруса синонимов русской речи». Для более точного и полного формирования словника из соответствующих словарных статей были выделены объекты описания, не являющиеся заголовочными словами (вокабулами). Это фонетические, словообразовательные, морфологические варианты слов (*планер — планёр*, *отличье — отличие*, *жирафа — жираф*), глаголы совершенного вида, имеющие пары несовершенного вида, и производные существительные, обозначающие лиц женского пола (*закройщица ← закройщик*, *дошкольница ← дошкольник*). Каждая подобная лексема получила отдельную запись и соотнесена в базе данных с основной вокабулой, организующей словарную статью.

Кроме того, сопоставление с электронными версиями существующих толковых и частотных словарей выявило частотные лексемы, для которых в Новом частотном словаре русской лексики под ред. О. Н. Ляшевской и С. А. Шарова [Новый частотный словарь...] установлен показатель *ipm* не менее 3 и которые не были ранее семантизированы в идеографических словарях группы «Русский глагол». Эти лексемы включены в новый словник. В него вошли также частотные неоднословные лексические единицы с семантикой наречий, вводных слов, предлогов, союзов и частиц. В итоге стартовый вариант базы данных, рассчитанный на создание Универсального словаря и разработку других проектов, включает 96 966 записей — лексико-семантических вариантов (ЛСВ) слов разных частей речи. Из них от словаря глаголов унаследовано 10 443 записи, от словаря

существительных — 14 898, от словаря прилагательных — 23 048, от тезауруса синонимов — 42 693. К этому количеству в базу добавлено 5 884 записи новых ЛСВ. Специфика разных словарей стала причиной количественной асимметричности слов разных частей речи (особенно в плане ЛСВ имен прилагательных) и совпадения ряда значений в тезаурусе синонимов и других словарях. Эти проблемы должны быть преодолены в процессе окончательной выборки материала для Универсального словаря.

Под руководством проф. Л. Г. Бабенко значительно изменена и дополнена синоптическая схема сводного тезауруса. Выявлены различия между структурами разных словарей, соотнесены рубрикации, произведена переиндексация основной части тезаурусов. Приняты решения по соединению и, наоборот, разделению ряда словарных групп, а также по устранению логико-понятийных нестыковок. На сегодняшний день эта структура объединяет в базе 962 записи — наименования денотативных сфер, подсфер и реальных лексических групп. Схема имеет 6 уровней и проявляет наибольшую дробность в сфере «3. Живая природа» (например, группа «3.2.2.3.2.3. Грызуны»).

Кроме новых лексикографических задач, которые позволяет решить сегодняшняя структура данных, у нее есть базовые статистические возможности, обусловленные самим контентом: семантической классификацией и полями, которые в основном соотносятся со словарными зонами словарей. Приведем из этих возможностей только наиболее очевидные:

- количественное соотношение денотативных сфер и групп: выявление специфики языковой картины мира на лексикографическом материале;
- стилистическое распределение внутри тезауруса (пометы): выявление функционально-стилистических, эмоциональных, оценочных приоритетов в разных денотативных классах, а также у слов, вступающих в синонимические отношения;
- распределение прямых и переносных значений: выявление образно-метафорических приоритетов;
- соотношение описываемых слов и идентификаторов, употребляемых в толкованиях: автоматизированное выявление гипо-гиперонимических отношений в лексике;
- частеречное соотношение и распределение грамматических категорий (например, глагольного вида, рода и числа существительных и т. д.): выявление грамматических особенностей денотативных классов.

Чтобы показать, что отмеченные направления вовсе не декларативны, рассмотрим здесь первые два из них на примерах конкретных сфер, групп и слов.

Общеязыковое количественное соотношение денотативных сфер корректно оценить на суммированном материале записей словарей глаголов, существительных, прилагательных, а также нового материала — значений слов других частей речи, добавленных в базу данных будущего Универсального идеографического словаря. По количеству представленных ЛСВ в базе данных сферы распределяются следующим образом: «11. Общественно-государственная сфера» (27,4 % всего материала), «8. Человек и его внутренний мир» (15,3 %), «12. Конкретная

физическая деятельность» (11,1 %). Замыкает перечень сфера «7. Родственные и семейные отношения» (0,4 %) (см. рис. 1). Всего элементов 1-го уровня классификации в новой структуре 15.

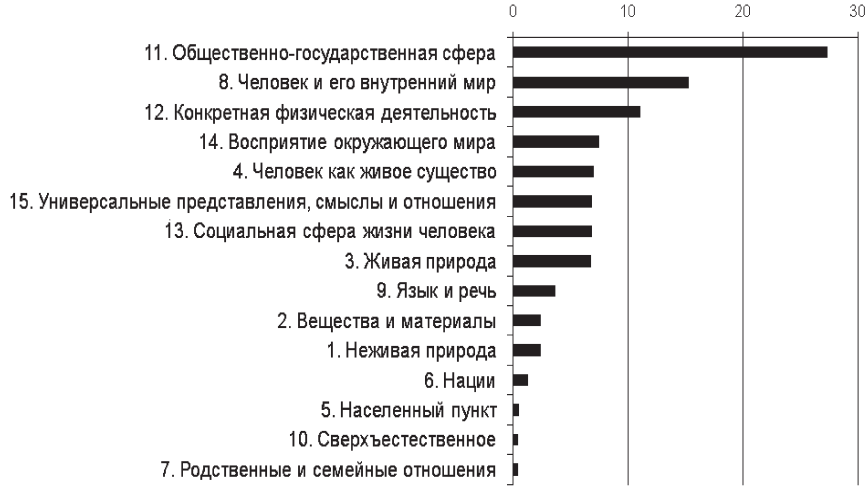


Рис. 1. Количественное соотношение денотативных сфер

Посмотрим, как внутри основных денотативных сфер соотносятся классы 2-го уровня. В лидирующей по количеству записей «Общественно-государственной сфере» выделено 20 подсфер. Среди них количественно наиболее выражены: «Искусство» (11 %), «Экономика» (7,9 %), «Медицина» (7,5 %), «Техника» (7,3 %), «Военная служба» (6,9 %) (см. рис. 2).



Рис. 2. Количественное соотношение
внутри «Общественно-государственной сферы»

В денотативной сфере «Человек и его внутренний мир» преобладает класс «Эмоции» (58,5 %) (см. рис. 3). Таким образом, лексика эмоций составляет почти 9 % от анализируемого объема. Еще более показательные сведения дает анализ синонимического материала. Среди слов, вступающих в синонимические отношения (по данным тезауруса русских синонимов), доля эмотивов составляет около 20 %, т. е. 1/5 часть русского синонимического лексикона.



Рис. 3. Количественное соотношение внутри сферы «Человек и его внутренний мир»

В структуре денотативной сферы «Конкретная физическая деятельность» очевидно преобладание слов со значением физического воздействия на объект (30,7 %). Этот класс формирует лексика нанесения удара, давления, прикосновения, повреждения объекта и др. (всего 14 денотативных групп).

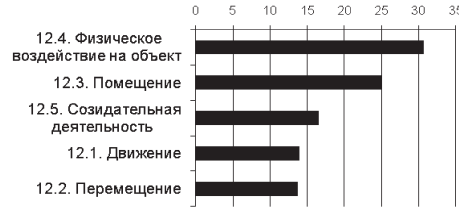


Рис. 4. Количественное соотношение внутри сферы «Конкретная физическая деятельность»

Очевидно, что, несмотря на лексико-грамматические различия разных словарей, формирующих базу данных, отмеченные соотношения классов отражают представления о мире, закрепленные в языке. Об этом говорит само покрытие лексической номинацией разных фрагментов действительности. К слову, статистику реального употребления лексики различных денотативных и семантических классов можно было бы проследить по текстовым корпусам, но на сегодняшний день о качестве семантической разметки даже в Национальном корпусе русского языка [НКРЯ] говорить пока рано.

Обсудим теперь некоторые особенности распределения по конкретным денотативным группам стилистически маркированной лексики. Для этого показательнее привлечь данные одного словаря — тезауруса русских синонимов: в нем изначально представлены слова разных частей речи, а употребление помет выдержано в одном ключе.

Стилистическая дифференциация классов показывает следующее. Книжная лексика (помета *книжн.*) преобладает в группах сфер «Язык и речь» —«Речевое

сообщение», «Особенности изложения...», «Речевое воздействие» (*вещать*, *изрекать*, *инсинуация*, *порицать*, *менторский* и многие другие слова) и «Интеллект» (*диагностировать*, *интерпретация*, *дефиниция*). Слова с пометой *высок.* в первую очередь обнаруживаем в группах «Награждение, признание заслуг или лишение чего-либо, наказание» (*лепта*, *стяжать*, *кара*), «Управление, власть, принуждение» (*властелин*, *повелевать*), «Создание объекта» (*вершить*, *творить*, *созидать*).

По разговорной лексике далеко впереди всех прочих группа «Звуки, их разновидности и источник»: *ахнуть*, *бабахнуть*, *бахнуть*, *бацнуть*, *бу́хнуть*, *громыхнуть*, *грохнуть*, *тарарахнуть*, *ухнуть* (примеры только из одного синонимического ряда). Разговорно-сниженную и жаргонную лексику, что совсем не удивительно, встречаем в первую очередь в группах «Физиологические действия, состояния и процессы, их признаки», «Внутренние и внешние части тела», «Повреждение тела», «Процесс питья и приведения в состояние опьянения», «Характеристика человека по отношению к еде и питью», «Общая оценка».

Распределение эмоциональных помет не менее любопытно. Например, по помете *груб.* лидирует физиологическая лексика. Ирония (пометы *ирон.*, *шутл.*) чаще всего сопровождает обозначение частей тела (*брюшко*, *телеса*, *кумпол*), состояния опьянения (*вздрогнуть*, *лечиться* и др. — о принятии спиртного), особенностей изложения, текста (*сладкоречивый*, *вития*, *златоуст*).

Неодобрение, пренебрежительное и презрительное отношение преимущественно связано с употреблением слов, относящихся к группам «Серьезное и несерьезное поведение человека» (*вертихвостка*, *кривляка*), «Человек определенного возраста» (*молокосос*, *сопляк*), «Речевое сообщение» (*трепь*, *брехня*), «Характеристика человека по отношению к еде и питью» (*пьянчуга*, *нажраться*), «Человек по его отношению к труду» (*лоботряс*, *тунеядец*). А вот помету *ласк.* (*уменьш.-ласк.*) встречаем чаще всего в группах «Животные» (*киса*, *зверюшка*, *заюшка*), «Человек определенного возраста» (*бабуля*, *парнишка*, *девчушка*), «Предмет одежды» (*платочек*, *свитерок*, *штанишки*), «Семья и ее члены» (*мамочка*, *папочка*).

Эти и многие другие подобные наблюдения еще ждут своей системной лингвистической интерпретации. Разработка тезаурусов русского языка — как традиционных бумажных, так и электронных — продолжается. Надежных, проверенных временем данных о количественных аспектах семантических классов русских слов в словаре и дискурсе сегодня еще нет. Как видно, разработка даже чисто статистических аспектов баз данных Уральской семантической школы имеет сегодня большой научный потенциал, что, конечно, отразится в новых исследованиях на лексикографическом материале.

Большой толковый словарь русских глаголов: Идеографическое описание. Синонимы. Антонимы. Английские эквиваленты / под ред. проф. Л. Г. Бабенко. М., 2007. 576 с. (Сер. «Фундаментальные словари») [Bol'shoj tolkovyj slovar' russkih glagolov: Ideograficheskoe opisanie. Sinonimy. Antonimy. Anglijskie jekvivalenty / pod red. prof. L. G. Babenko. M., 2007. 576 s. (Ser. «Fundamental'nye slovari»)]

Большой толковый словарь русских существительных: Идеографическое описание. Синонимы. Антонимы / под ред. проф. Л. Г. Бабенко. М., 2005. 864 с. [Bol'shoj tolkovyj slovar' russkih sushhestvitel'nyh: Ideograficheskoe opisanie. Sinonimy. Antonimy / pod red. prof. L. G. Babenko. M., 2005. 864 s.]

Большой толковый словарь синонимов русской речи: идеографическое описание, антонимы, фразеологизмы / под ред. проф. Л. Г. Бабенко. М., 2008. 784 с. [Bol'shoj tolkovyj slovar' sinonimov russkoj rechi: ideograficheskoe opisanie, antonimy, frazeologizmy / pod red. prof. L. G. Babenko. M., 2008. 784 s.]

Мухин М. Ю. Структура базы данных Толково-идеографического словаря // Слово и словарь = Vocabulum et Vocabularium : сб. науч. тр. по лексикографии. Гродно, 2002. С. 177–179. [Muhin M. Ju. Struktura bazy dannyh Tolkovo-ideograficheskogo slovarja // Slovo i slovar' = Vocabulum et Vocabularium : sb. nauch. tr. po leksikografii. Grodno, 2002. S. 177–179.]

НКРЯ — Национальный корпус русского языка [Электронный ресурс]. URL: http://www.ruscorpora.ru [Nacional'nyj korpus russkogo jazyka [Electronic resource]. URL: http://www.ruscorpora.ru]

Новый частотный словарь русской лексики / под ред. О. Н. Ляшевской и С. А. Шарова [Электронный ресурс]. URL: http://dict.ruslang.ru/freq.php [Novyj chastotnyj slovar' russkoj leksiki / pod red. O. N. Ljashevskoj i S. A. Sharova [Electronic resource]. URL: http://dict.ruslang.ru/freq.php]

Словарь-тезаурус русских прилагательных, распределенных по тематическим группам / под ред. проф. Л. Г. Бабенко. М., 2011. 232 с. [Slovar'-tezaurus russkih prilagatel'nyh, raspredelennyh po tematicheskim gruppam / pod red. prof. L. G. Babenko. M., 2011. 232 s.]

Словарь-тезаурус синонимов русской речи. М., 2007. 512 с. [Slovar'-tezaurus sinonimov russkoj rechi. M., 2007. 512 s.]

Толковый словарь русских глаголов: Идеографическое описание; Английские эквиваленты; Синонимы; Антонимы / под ред. проф. Л. Г. Бабенко. М., 1999. 704 с. [Tolkovyj slovar' russkih glagolov: Ideograficheskoe opisanie; Anglijskie jekvivalenty; Sinonimy; Antonimy / pod red. prof. L. G. Babenko. M., 1999. 704 s.]

Статья поступила в редакцию 29.09.2014 г.

УДК 81-119 + 811.161.1:378.4 + 811.161.1'374

Ю. В. Казарин

ИДЕОГРАФИЯ — ЗЕРКАЛО ЯЗЫКА И МИРА
К 40-летию кафедры современного русского языка и прикладной лингвистики

Исследуются и подводятся итоги деятельности Уральской семантической школы (основатели — Э. В. Кузнецова и Л. Г. Бабенко), научным центром которой является кафедра современного русского языка и прикладной лингвистики УрФУ, а также словарная группа (зав. кафедрой и руководитель группы — Л. Г. Бабенко). Анализируется и прослеживается поэтапное становление новой теории и методологии лексико-семантических исследований, теории и практики идеографической лексикографии и производства словарей нового типа — словарей XXI века. Комментируются достижения кафедры и научной школы и определяются перспективы дальнейших научных исследований и лексикографической работы.

К л ю ч е в ы е с л о в а: Уральская семантическая школа; кафедра современного русского языка и прикладной лингвистики УрФУ; Л. Г. Бабенко; идеография; толково-идеографические словари глаголов, существительных, прилагательных и синонимов; словарь концептов; Универсальный идеографический словарь.

Традиция изучения языковой, речевой и текстовой семантики на Урале объективно и коренным образом связана с научной деятельностью Уральской

© Казарин Ю. В., 2014