Гадомский А. К.

О НЕКОТОРЫХ ПРОБЛЕМАХ СОСТАВЛЕНИЯ ЛИНГВИСТИЧЕСКИХ СЛОВАРЕЙ И СПОСОБАХ ИХ РЕШЕНИЯ

Наиболее подвижной, активной, реагирующей на все изменения, происходящие в обществе, частью языка является лексика. До недавнего времени в лексикографии существовало два основных способа фиксации лексики в словарях: алфавитный (по сходству означающего) и идеографический (по сходству означаемого). В последнее десятилетие очень активно заявляет о себе еще один способ фиксирования словарного материала – электронный. Данной проблемой сегодня занимаются лингвисты всего мира.

Настоящая статья отражает достижения в этой области, полученные английскими, российскими, чешскими, польскими лексикографами. Особое внимание уделяется работам таких польских лингвистов, как: M. Bańko, M. Łazinski, K. Wengrzynek и др[1-9]. Пристальное внимание автора статьи к польской лексикографии объясняется следующими причинами.

Во-первых, географическое положение Польши, занимающей промежуточное место между Востоком и Западом, определило развитие польской лингвистики, впитавшей в себя и развившей идеи западных и восточных исследователей. Эти достижения могут и должны быть использованы в восточнославянской лексикографии.

Во-вторых, настоящая статья отражает результаты подготовительного этапа работы по составлению русско-польского словаря религиозной лексики и знакомит читателя с возможными источниками выбора лексикографического материала и возможной методикой его отбора, систематизации, хранения и описания.

С этой целью написания статьи нами были также изучены основные этапы эволюции лексикографии.

Первые словари, известные современным исследователям, были идеографическими, поскольку одним из первых видов письменности является идеография (письмо понятиями). Информация, отраженная в этих словарях, могла располагаться только с учетом близости понятий. Слова были объединены в классы, группы и располагались в определенной логической, смысловой, тематической последовательности. Для того, чтобы убедиться в этом, достаточно обратиться к таким словарям, как словарь Поллукса, словарь «Амаракоша», словарь К. Касареса. Очень подробно о них пишет в своей книге Ж. П. Соколовская [4]. Выполнены такие словари были на папирусе, пергаменте, позднее на бумаге.

Пришедшая на смену идеографии, фонография упростила способ фиксирования языкового материала и подарила лексикографии совершенно иные возможности. Стало возможным располагать слова в словаре не с учетом степени близости понятий (означаемого), а по формальному признаку, в алфавитном порядке, с учетом близости формы (означающего). Все народы, использующие фонографическое письмо, начали создавать алфавитные словари. Они достаточно просты в обращении, но имеют принципиально важный недостаток: за формой теряется система понятий. Об этой проблеме говорят многие лингвисты. Достаточно познакомиться с работами В. В. Морковкина, Ю. Н. Караулова [1, 2].

Неоднократно предпринимались попытки, используя фонографическое письмо, создавать идеографические словари, стремясь тем самым максимально приблизить способ хранения информации на бумаге к способу ее хранения в человеческом мозгу, где, естественно, «словарный материал» фиксируется не в алфавитном порядке.

Подобные попытки предпринимались и в русском языке. Немало в этом направлении трудились Ю. Н. Караулов [1], В. В. Морковкин [11], М. В. Баранов [10], Н. Ю. Шведова [12] и другие лингвисты.

Были ли эти попытки удачны? Можно утверждать, что попытки были удачны, но далеко не всегда были удачны сами словари. Они, как правило, сложны в обращении и рассчитаны прежде всего на специалистов. Возникает вопрос: нужен ли вообще носителю языка такой идеографический словарь? Ответы могут быть самыми разными. В настоящей статье мы предлагаем свои варианты ответов на поставленный вопрос.

Во-первых, эта проблема соприкасается с семиотикой, ментальностью того или иного народа, типом письма и алфавита, которым пользуется тот или иной народ. Невзирая на то, что до 70% населения земного шара использует латиницу, 30% все-таки остается за другими видами письменности. Характерно, что нелатиницей пользуются не закрытые этнические группы, а успешно развивающиеся народы. Очевидно, все-таки есть причины, по которым отдельным носителям языка удобнее писать не слева направо, а наоборот; не латиницей, а кириллицей или какими-нибудь другими знаками; использовать не фонографическое письмо, а идеографическое.

Нужно отметить, что способ фиксирования материала, вид письменности определяют форму его восприятия. Поэтому, на наш взгляд, идеографический словарь хорош для тех, кто использует идеографическое письмо, а словарь, в котором слова располагаются в алфавитном порядке, более приемлем для народа, использующего фонографическое письмо. И нет необходимости менять эту схему: создавать идеографический словарь, на базе фонографического письма. Ведь никто же не пытается составить алфавитный словарь с применением идеографического письма.

Помимо названной проблемы современным составителям словарей приходится решать и другие проблемы.

1. Проблему способа фиксирования материала. Эта проблема может быть решена только в том случае, если будет использоваться какая-то новая форма письменности.

2. Проблему источника выбора материала. Источником материала могут служить:

а) ранее написанные словари;

б) тексты, выбранные из различных источников;

в) компетенция составителей словарей.

3. Проблему концепции словаря, учитывающую особенности потенциального пользователя словаря. Попытки творить для массового пользователя, как правило, терпят неудачу: словарь, написанный для всех, не служит никому.

4. Проблему нормативности словаря. Поскольку любой словарь служит для того, чтобы отразить, передать особенности языка, то возникает вопрос: какой язык следует отражать и что принять за эталон. Здесь возможны две крайности:

а) можно пропустить в язык все, оставив за автором право передаточного звена, сведя его активность до минимума;

б) может иметь место так называемый «синдром Пушкина». То есть, автор может занять позицию контролера и «пропускать» в словарь только те слова, которые употреблялись авторитетными носителями языка. Что отражено в авторитетных источниках.

5. Проблему функциональности словарей: каждый лексикограф стремится к тому, чтобы составленный им словарь не только отражал реальное положение вещей, давал возможность человеку как понимать тексты, так и творить их на данном языке.

Где же выход?

Одной из многочисленных, но наиболее удачных попыток решения перечисленных проблем, на наш взгляд, является развитие корпусной (электронной) лексикографии. «Под корпусом понимают собрание разнообразных текстов современного языка, осуществленного при помощи компьютера» [7].

Первые корпусы текстов начали создаваться в 60-х годах XX века. Но ограниченные возможности компьютеров, которыми в то время были оснащены лаборатории, не позволили выйти за рамки 1 млн. слов. К их числу относятся американский корпус Brown Corpus, который был собран учеными в Brown University USA, и европейский корпус LOB Corpus, который был собран учеными в Lancaster-Oslo-Bergen.

«Польским корпусом того времени является собрание в 500 тысяч слов, составленное из текстов пяти разновидностей: научно-популярных текстов, мелких газетных статей, публицистики, художественной прозы и драмы. Из каждого стиля было выбрано по 100 тысяч слов в контекстах, протяженностью по 50 слов каждый. Опубликовано это было в 1974–1977 гг. под общим названием «Słownictwo współczesnego języka polskiego» и объединено в томе «Słownik frekwencyjny polszczyzy współczesnej» [17].

В 80-х годах был создан корпус английского языка Collins Birmingham University International Language Database (COBUILD). В самом начале работы он включал 7,3 млн слов, на завершающем этапе число примеров составило порядка 20 млн. На его базе в 1987 году был создан словарь «Collins Cobuild English Language Dictionary» (CCELD) – первый корпусный словарь. Это издание 1 млн слов (500 текстов по 2 тысячи слов каждый).

В 90-х годах заявил о себе British National Corpus (BNC), который включал в себя 100 млн. слов. Разговорным текстам в нем отведено 10%. Этот корпус, если говорить о его структуре, может служить примером для лингвистов, занимающихся корпусной лексикографией (http://www.hcu.ox.ac.uk/BNC).

В 1995 году вышло в свет новое издание оригинального словаря под названием «Collins Cobuild English Dictionary», опиравшегося на корпус в 200 млн. слов.

В октябре 2000 г. этот корпус, известный как «Bank of English», насчитывал 415 млн. слов. Он также доступен в Интернете: http://titania.cobuild.collins.co.uk.

В 90-х годах были созданы корпусы и других языков:

1) корпус чешского языка (http://ucnk.ff.cuni.cz);
2) корпус немецкого языка (http://corpora.ids-mannheim.de);
3) корпус русского языка (htpp://www.sfb441.uni-tuebingen.de/bl/korpora.html) [7].

В Польше корпус польского языка начали создавать в Институте польского языка отделения Польской Академии Наук в Кракове. Он доступен только в Интернете Ягеллонского университета в Кракове (Польша). Включает тексты, начиная с 1956 г. Особое внимание уделено текстам 90-х годов. Эти даты считаются переломными в развитии польского языка. Условно, по характеру собранного материала, он может быть разделен на три равные части: литературные тексты, половина из которых – это примеры из художественной литературы; газетные тексты; тексты из учебников для студентов, обучающихся в вузах по программе для бакалавров [10]. По причине отсутствия финансирования в настоящее время работа над ним приостановлена. Самым большим корпусом польского языка считается Лодзьский корпус. Собран на кафедре английского языка Лодзьского университета (г. Лодзь, Польша). Доступен в Интернете: htpp://www.uni.lodz.pl. В его основу положены: газетные тексты – 45%; книжные тексты – 40%; разговорные тексты – 7%; другие – 8%. Этот корпус в основном опирается на газетные материалы (данная информация получена нами в ходе участия в Пленарном заседании Комитета языкознания Польской Академии Наук, которое проходило 25.11.02 в Варшаве).

С некоторых пор свой корпус создает PWN – Государственное научное издательство в Варшаве. В конце 2000 г. он состоял из 50 млн. слов, в начале 2003 г. – 62 млн. слов (информацию об этом корпусе да-

ет М. Банько). «Размеры отдельных частей корпуса поданы в мегабайтах (1 Мв – это около 1,05 млн. знаков). Приблизительное число слов можно определить, если разделить число Мв на 7: столько знаков в среднем приходится в польских письменных текстах на одно слово и отступ, в устном – короче. Перечень текстов, источников, на базе которых был создан корпус PWN, дает М. Лазински во вступлении к «Innemu słowniku języka polskiego» [9].

Далее приводятся таблицы дающие представление о структуре корпуса PWN.

Таблица 1. Структура корпуса PWN: виды текстов

Виды текстов	Кол-во текстов	Объем текстов (Мв)	Объем текстов (%)
Художественная Литература	292	94,1	26,9
Небеллетристическая Литература	307	146,2	41,9
Периодика	1387	81,9	23,4
Разговорные тексты	1003	22,7	6,5
Рекламные тексты (печатная реклама)	141	4,4	1,3
ИТОГО	3130	349,3	100,0

[7, с. 36]

Таблица 2. Структура корпуса PWN: хронология (за исключением устных текстов).

Период	Кол-во текстов	Объем текстов (Мв)	Объем текстов (%)
1918–1945	103	22,6	6,9
1946–1970	161	53,2	16,3
1971–1989	167	47,5	14,5
1990–2000	1636	189,2	57,9
Тексты различных периодов(н-р, антологии)	60	14,1	4,3
ИТОГО	2127	326,6	100,0

[7, с. 36]

Таблица 3. Структура корпуса PWN: тексты.

Разновидность языка	Количество текстов	Объем текстов (Мв)	Объем текстов (%)
Официальный	719	17,3	76,2
Просторечный	284	5,4	23,8
ИТОГО	1003	22,7	100,0

[7, с. 36]

Этот корпус можно найти в Интернете: http://slowniki.pwn.pl/korpus.

Информация в корпусе PWN располагается следующим образом: в центре строки помещено слово, выделенное жирным шрифтом, слева и справа – контекст, продолжительностью до 50 знаков (вместе с отступами). При этом контекст может начинаться не с начала предложения («левый контекст») и заканчиваться не в конце предложения («правый контекст»). Если пользователя интересует какое-то слово, то достаточно вписать его в соответствующее окошко на дисплее: машина выдаст общее количество употреблений этого слова в корпусе и предложит все имеющиеся примеры. Если контекста недостаточно, то следует подвести курсор к каждому из примеров, выделенных жирным шрифтом: машина выдаст точные выходные данные контекста, в котором находится пример. Если же пользователя интересует сам текст, ему придется поработать в библиотеке.

Настоящий корпус привлекает своей доступностью в Интернете и простотой в эксплуатации. Корпус имеет важное значение как для развития лексикографии, так и для решения важных социальных вопросов. Достаточно познакомиться с тем, что пишут по этому поводу авторы корпуса PWN.

«Сегодня, как никогда, словари народного языка, так еще именуются корпусы, становятся символом

единства народа (этот аспект, очень важен в Германии, где работа над корпусом финансируется правительством). Идея народного корпуса находит благодатную почву в объединяющейся Европе, где распространяется убеждение, что создание корпусов служит европейской интеграции. Даже созданы организации, занимающиеся рекламой корпусов разных языков и их доступностью для пользователей – в разных странах (в Европе такая организация ELRA – European Language Resources Association)» [7].

Что же дает нам корпус в сравнении с картотекой? Во-первых, корпус обеспечивает большую скорость и селективность выискивания информации. Во-вторых, существует возможность своевременного использования найденной информации в других компьютерных программах, хотя бы возможность переноса их в другие тексты. В-третьих, администратор постоянно расширяющегося корпуса может менять количество, размер и раздел составляющих текстов, чтобы обеспечить ему соответствующую структуру и наибольшую репрезентативность по сравнению с языком. В-четвертых, с учетом того, что сказано ранее, корпус является более удобным источником информации, чем обычная картотека. В-пятых, создание, развитие и обслуживание корпуса дешевле картотеки [7, с. 36–38].

Корпусная лексикография позволяет постоянно увеличивать количество лексических единиц, не ожидая очередного переиздания словаря. Кроме того существует возможность постоянно фильтровать и упорядочивать словарный материал. Усовершенствование компьютерных программ, используемых для составления компьютерных картотек, а также использование веб-страниц в роли компьютерной картотеки упрощает доступ к нему массового пользователя.

Электронная (нефонографическая, неидеографическая) запись материала снимает проблему систематизации лексического материала. В данном случае не так неважно, в алфавитном или тематическом порядке располагаются лексические единицы: каждую из них очень легко найти, используя различные критерии.

Работа с электронными картотеками позволяет решать по-разному проблему нормативности языка. Если до начала 90-х годов в языке господствовал принцип: норма – запрет, то изменение в жизни общества привели к тому, что норма позволяет выбирать [3].

Пропасть, «вырытая» Ф. де Соссюром между языком и речью, противоречие между фактом и метафактом, между нормой и узусом может быть решена благодаря корпусной лексикографии.

На сегодняшний день, если между языковым фактом и метафактом возникают противоречия, возможны следующие решения.

Во-первых, можно проигнорировать факт. Так часть делают в словарях, называемых нормативными. Во-вторых, можно зарегистрировать факт и проигнорировать метафакт. Так часто делается в словарях, традиционно называемых описательными. В практике обе тенденции имеют право голоса. Самым функциональным, на наш взгляд, является третье решение, которое включает совмещение факта и метафакта. «Вместо того, чтобы говорить о каком-то слове, что его нет в словарях, следовало бы лучше спросить, почему его там нет. Следовало бы сравнить компетентные отзывы о нем с фактической сферой употребления» [7].

В качестве подтверждения справедливости приведенных выше рассуждений следует сказать о том, что в польской лексикографии уже начала складываться традиция описательного нормативизма, в соответствии с которой в словарях помещаются не только лексические единицы, формы, опирающиеся на авторитеты, но и те, которые опираются на языковые привычки. Примером такого нормативизма является «Jnny słownik języka polskiego», «Słownik wyrazów kłopotliwych», «Słownik poprawnej polszczyzny», Słownik Dunaj, «Nowy słownik poprawnej polszczyzny» [14, 15, 16, 19, 20].

Активная компьютеризация позволила приступить к созданию электронных словарей, которыми изобилует сейчас рынок. Электронные словари имеют достаточно много преимуществ. Многотомный словарь может поместиться на одном CD-диске. Очень часто предусмотрена поисковая система, что позволяет без проблем находить нужное слово.

Однако все электронные словари страдают одним недостатком: они являются копией графических словарей, изданных ранее, и повторяют всех их проблемы. Эта проблема может быть решена, на наш взгляд, только после того, как новое поколение электронных словарей будет создано на базе корпуса текстов.

С целью репрезентации тех преимуществ и перспектив, которые открывает корпусная лексикография, предлагаем познакомиться с выводами к настоящей статье, помещенными в следующей таблице.

№	Вид лексикографии / Критерии оценки словаря	Неэлектронная Лексикография	Электронная лексикография (корпусная)
1	Порядок расположения материала.	1. Алфавитный. 2.Идеографический.	Алфавитный с системой поиска.
2	Способ фиксирования материала.	1.Графический. 2. Постоянный.	1. Электронный. 2. Переменный.
3	Источники выбора материала.	1. Словари, написанные ранее.	1.Тексты различных жанров с соблюдением ко-

		2. Тексты. 3. Компетенция автора словаря на первом плане.	лич. пропорции. 2.Компетенция автора словаря не на первом плане.
4	Концепция словаря с учетом адресата.	1. Словарь написан «для всех».	1.Словарь написан «для каждого в отдельности».
5	Нормативность.	1.Нормативный словарь – игнорирование фактов языка. 2.Описательный словарь – игнорирование метафактов языка.	1.Совмещение фактов и метафактов.
6	Функциональность.	Словарь «мертвого языка».	Словарь «живого языка».

Сам факт создания корпусов текстов является, по нашему мнению, серьезным шагом на пути к созданию словарей нового поколения и решению многих проблем современной лексикографии.

Литература

1. Караулов Ю. Н. Общая и русская идеография. – М.: Наука, 1976. – 355 с.
2. Морковкин В. В. Идеографические словари. – М.: Изд-во МГУ, 1970. – 72 с.
3. Русский язык конца XX столетия. (1985–1995). – М.: Языки русской культуры, 1996. – 480 с.
4. Соколовская Ж. П. «Картина мира» в значениях слов. «Семантические фантазии» или «катехизис семантики?». – Симферополь: Таврия, 1993. – 232 с.
5. Соссюр Ф. де. Труды по языкознанию. – М.: Прогресс, 1977. – 695 с.
6. Щерба Л. В. Языковая система и речевая деятельность. – Л.: Наука, 1974. – 428 с.
7. Bańko M. Z pogranicza leksykografii i językoznawstwa. – Warszawa, UW, 2001. – 336 s.
8. Bańko M. Zawartość słownika (W:) Jnny słownik języka polskiego PWN, red. nacz. M. Bańko, t. 1–2. – Warszawa: PWN. – S. XVI–LV.
9. Łaziński M. Korpus PWN (W:) Inny słownik języka polskiego, red. nacz. M. Bańko. – Warszawa: PWN, 2000. – S. LVI–LXI.
10. Wengrzynek K. Projekt komputerowego korpusu współczesnych tekstów polskich. Język Polski LXXV, Z. 4–5, 1995. – S. 332–341.

Словари

11. Баранов О. С. Идеографический словарь русского языка. – М.: Изд-во «Прометей» МГПИ им. В. И. Ленина, 1990.
12. Морковкин В. В. Лексическая основа русского языка: Комплексный учебный словарь. – М.: Русский язык, 1984.
13. Толковый словарь, систематизированный по классам слов и значений / РАН. Ин-т русск. яз. Под общ. ред. Н. Ю. Шведовой. – М.: Азбуковник, 1998.
14. Inny słownik języka polskiego, red. nacz. Bańko, t. 1–2. – Warszawa: PWN, 2000.
15. Nowy słownik poprawnej polszczyzny PWN, red. nauk. A. Markowski. – Warszawa: PWN, 1999.
16. Popularny słownik języka polskiego, red. nauk. B. Dunaj. – Warszawa: Wilga, 1999.
17. Słownik frekwencyjny polszczyzny współczsnej, I. Kurcz, A. Lewicki,
J. Sambor, K. Szafran, I. Woronczak, t. 1–2. – Kraków: PAN, 1990.
18. Słownik języka polskiego, red. nacz. W. Doroszewski. – Warszawa: PWN, 1973.
19. Słownik współczesnego języka polskiego, red. nauk. B. Dunaj. – Warszawa: Wilga, 1996. Wyd 2, poprawione i uzupełnione, t. 1–2. – Warszawa: Reader's Digest Przegląd, 1998.
20. Słownik wyrazów kłopotliwych, M. Bańko, M. Krajewska. – Warszawa: PWN, 1994.

удк – 81': 81'374

Аннотация

В настоящей статье идет речь о способах построения словарей и тех перспективах, которые открывают формирование корпусов текстов и изменение подходов к созданию словарей нового типа.

Ключевые слова

Идеография, лексикография, языковая норма, корпус текстов, электронные словари.