Вестник Новосибирского государственного педагогического университета
5(27)2015 www.vestnik.nspu.ru ISSN 2226-3365

DOI: 10.15293/2226-3365.1505.10

УДК 81’33 + 81’27

К ВОПРОСУ О ПОСТРОЕНИИ БАЗЫ ДАННЫХ ПРАГМАТИЧЕСКИ МАРКИРОВАННОЙ ЛЕКСИКИ*

И. И. Саженин (Новосибирск, Россия)

В последние десятилетия все больший интерес исследователей привлекает прагматический компонент лексического значения. Среди наиболее значимых проблем фундаментальной лингвистики, лежащих на пересечении структурно-системного и коммуникативного описания лексического значения слова, находится, в частности, выявление прагматически маркированных фрагментов словаря и построение типологии микрокомпонентов, входящих в структуру прагматического макрокомпонента, а также исследование их речевого потенциала. В настоящей статье затрагиваются вопросы и принципы построения базы данных прагматически маркированной лексики, корпус которой составляется с учетом способов отражения лингвистически значимой информации: идеологического, национально-культурного, гендерного, социального, эмотивно-оценочного и др. компонентов в значении слова. Принципы построения базы данных как с технических позиций, так и с содержательных, учитывают механизмы построения языковых корпусов (Национальный корпус русского языка, Хельсинский аннотированный корпус, Брауновский корпус и др.), опыт создания электронных лексикографических ресурсов (ABBYY LINGVO, «Русские словари», DICT; Cambridge Online Dictionary; Shorter Oxford English Dictionary Webster’s Dictiomary) и традиции составления бумажных словарей разных типов.

Ключевые слова: *прагматика, языковой корпус, словарь, электронная лексикография лингвистический анализ, лингвистическая база данных.*

Выявление и лексикографическое описание прагматической семантики слова относится к наиболее актуальным задачам современной коммуникативной лексикологии и лексикографии. Целенаправленный семантико-прагматический анализ является необходимым условием разработки схем лексикографирования прагматически маркированной

* Статья подготовлена в рамках реализации проекта «Прагматический потенциал языкового знака: семасиологический и лексикографический аспекты» (грант РГНФ № 15-04-00122) и проекта «Конфликтный текст в зеркале лингвистического экспертирования: оскорбление, унижение чести и достоинства, порочащая информация в публичной коммуникации региона» (грант РГНФ № 15-14-54001)

Саженин Игорь Игоревич – кандидат филологических наук, доцент кафедры современного русского языка, Институт филологии, массовой информации и психологии, Новосибирский государственный педагогический университет.
E-mail: sajana84@mail.ru

лексики, например, в словарях активного типа и/или в специальных компьютерных базах данных. Исследование прагматически маркированного языкового знака и возможных способов его семантизации предполагает разработку концепции лексикографической интерпретации прагматически маркированных фрагментов лексической системы языка и, наконец, создание компьютерной базы данных прагматически маркированной лексики, доступной для широкого круга пользователей: исследователей, преподавателей русского языка и русского языка как иностранного, лингвистов-экспертов, переводчиков.

Некоторым препятствием на пути исследований в данном направлении до последнего времени был сам язык как динамическая система, с одной стороны, а с другой – довольно скудные человеческие возможности в области обработки и хранения больших массивов информации. Однако развитие информационных технологий позволяет взглянуть на проблему несколько более оптимистично. Дело в том, что до недавнего времени традиционно основным способом фиксации, хранения и предоставления доступа к различным типам информации о слове были классические бумажные словари со всем набором проблем традиционной лексикографии (см. «Парадоксы словарной статьи») [9]. С момента начала применения информационных технологий в гуманитарной сфере прошло не одно десятилетие. За это время успели выделиться в самостоятельные направления такие области филологической науки, как компьютерная лингвистика, корпусная лингвистика и компьютерная лексикография. В рамках этих, в первую очередь, прикладных направлений были созданы инструменты, позволившие во многом облегчить труд исследователя. Мы имеем в виду, прежде всего, языковые корпусы и электронные лексикографические ресурсы. Создание и развитее данных инструментов шли отличными друг от друга путями. Если в области корпусной лингвистики главную роль играли, все-таки, специалисты-филологи, то компьютерная лексикография была отдана во многом на откуп профессиональным программистам и бизнесменам, и это не в последнюю очередь заслуга господства в современном обществе рыночных отношений: когда спрос на электронные лексикографические продукты определяется потребностями практического характера. По этой причине наиболее совершенными и многофункциональными лексикографическими продуктами на сегодняшний день являются переводные многоязычные словари [10–16]. Специализированные же лексикографические продукты прогресс если не обошел стороной, то, по крайней мере, не сказался должным образом на их развитии.

Еще в начале 1980-х гг. в нашей стране велись работы в области специализированной компьютерной лексикографии. В данной области были заняты такие ученые как А. П. Ершов, Ю. Н. Караулов, В. М. Андрющенко, А. Я. Шайкевич и др. Ряд идей и разработок Машинного фонда русского языка впоследствии лег в основу создания Национального корпуса русского языка. Однако, по словам В. М. Андрющенко, информатизация русистики в тех организационных и финансовых условиях как направление оказалось нежизнеспособным. Возможно, по этой причине до сих пор для компьютерной лексикографии не сформирован собственный предмет изучения, а лексикографическая теория отстает от компьютерной лексикографической практики.

Технический инструментарий, используемый при разработке электронных лексикографических ресурсов, изначально не ориентирован на обеспечение работы со словарным

содержанием, а наиболее перспективные методы, разработанные в области, например, корпусной лингвистики, не столь активно применяются в практике создания электронных лексикографических ресурсов [4; 6; 8].

Основной проблемой при создании электронных лексикографических ресурсов является то, что машина не способна в полной мере работать с текстом на естественном языке для репрезентации пользователю информации, соответствующей возможному спектру его запросов [4; 6]. Данная проблема была решена специалистами, разрабатывающими корпусы текстов посредством использования такого инструмента, как разметка. Как отмечают М. В. Копотев и А. Мустайоки «…современная корпусная лингвистика, несмотря на относительно короткую историю существования, является хорошо разработанным направлением языкознания, тесно связанным с компьютерной и когнитивной лингвистикой. С первой она связана технологией и инструментами обработки языкового материала, со второй совпадает в базовой предпосылке: как когнитивная, так и корпусная лингвистика интересуется речевой деятельностью, представленной в бесконечном числе текстов. <…> Каждый новый этап в развитии машинной обработки языкового материала открывает новые возможности сначала для создателей корпусов, а затем и для лингвистов, осуществляющих исследования на основе существующей разметки» [13]. Как целостный инструмент лингвистических исследования корпус, имеет две основных составляющих, а именно: непосредственно массив данных (текстов); корпусный менеджер (специализированная поисковая система), которая и позволяет производить отбор необходимых исследователю единиц из всего массива данных, на основании разметки (аннотации) [5].

Создание любого словаря, после определения его концепции, конечно, начинается со сбора и фиксации материала – единиц словарного писания. В случае с созданием электронного информационного ресурса – оптимальным способом накопления и хранения языкового материала является электронная база данных. Существует множество определений понятия «база данных» применительно к сфере информационных технологий. Приведем некоторые из них. Самое обще определение базы данных дает толковый словарь английского языка *Concise Oxford English Dictionary*: *database – a structured set of data held in a computer*.

По большому счету, данное определение не является исчерпывающим. Действительно, под это определение может подойти, например, список расположенных по алфавиту файлов, именованных определенным образом, например, с расширением *doc*, или *pdf*, однако будет ли такой набор данных являть собой базу данных – остается вопросом.

Иное определение дает американский специалист по базам данных, автор учебника «Введение в системы баз данных» К. Дж. Дейт: саму же базу данных можно рассматривать как подобие электронной картотеки, т. е. хранилище или контейнер для некоторого набора занесенных в компьютер файлов данных. Пользователям этой системы предоставляется возможность выполнения над такими файлами различных операций [2].

Как видим, данное определение представляет базу данных уже не как структурированный набор неких данных, а как хранилище файлов этих данных, с возможностью для пользователя осуществлять операции с этими данными.

Рассмотрим еще два определения понятия базы данных. Так, авторы книги «Базы данных. Проектирование, реализация и сопровождение. Теория и практика» вносят еще ряд

элементов в определение, не встреченных нами в предыдущих двух: база данных – совместно используемый набор логически связанных данных (и описание этих данных), предназначенный для удовлетворения информационных потребностей [3]. Во-первых, в данном определении появляется такой семантический компонент, как «совместность использования», во-вторых, согласно данному определению, база данных должна содержать описание находящихся в ней данных, в-третьих, в определении появляется информация о целевом назначение подобной структуры – удовлетворении информационных потребностей.

Последнее определение мы взяли из образовательного Интернет-ресурса «Информатика в школе»: базы данных (БД) – это организованный набор фактов в определенной предметной области. БД – это информация, упорядоченная в виде набора элементов, записей одинаковой структуры. Для обработки записей используются специальные программы, позволяющие их упорядочить, делать выборки по указанному правилу. Базы данных относятся к компьютерной технологии хранения, поиска и сортировки информации. Данное определение дает нам следующую дополнительную информацию: база данных, помимо самих данных и их описания, содержит также набор компьютерных программ, которые, собственно, и позволяют проводить необходимые операции с данными. Кроме того, данное определение в каком-то смысле раскрывает формулировку «удовлетворение информационных потребностей». То есть, такими потребностями являются хранение, поиск и сортировка информации.

Таким образом, основываясь на рассмотренных признаках базы данных, мы можем представить внутреннее устройство базы прагматически маркированной лексики, как организованный набор элементов, которыми будут являться в нашем случае лексические и фразеологические единицы.

Следующим этапом по развитию ресурса станет его разметка в соответствии с теми прагматическими компонентами, что присущи той или иной отдельно взятой языковой единице. Так, в своей работе «Введение в прикладную лингвистику» А. Н. Баранов пишет: «Размеченные в соответствии с описанными параметрами тексты представляют собой лишь сырой материал. Отметим, что в традиционной технологии это и есть окончательный результат. В технологии динамического корпуса текстов размеченный исходный массив является источником для формирования конкретных корпусов, более точно отражающих информационную потребность пользователя. Массив хранится в виде базы данных, а каждый отдельный текст – в одной записи (параметры – в текстовых и числовых полях, сама статья – в поле MEMO). Перевод размеченных текстов в формат базы данных осуществляется с помощью специальной служебной утилиты» [1].

Основываясь на приведенном высказывании, мы, во-первых, спроецируем описанное свойство корпусной базы данных на наш объект, а, во-вторых, поясним ряд моментов относительно некоторых использованных понятий. Обычно база данных являет собой некоторым образом заполненную таблицу, а чаще – набор таких таблиц, атрибутами которых являются строки (записи), столбцы (поля). В табличной структуре адрес данных определяется пересечением строк и столбцов. Поля формируют структуру базы данных, а записи составляют информацию, содержащуюся в базе данных. Текстовое поле – символьные или числовые данные, которые не требуют вычислений. Такое поле может содержать до 255 символов.

Поле **МЕМО** предназначено для ввода текстовой информации, объем которой превышает 255 символов. Такое поле может содержать до 65 535 символов. Этот тип данных отличается от типа **Текстовый** тем, что в таблице хранятся не сами данные, а ссылки на блоки данных, хранящиеся отдельно. За счет этого ускоряется обработка таблиц (сортировка, поиск и т. п.).

Числовой тип используется для размещения числовых данных, необходимых для математических расчетов.

Простейшая база данных, как правило, представляет собой одну такую таблицу. Если бы информация размещалась и хранилась в таких достаточно простых структурах, то для взаимодействия пользователя с табличными данными не требовались бы специальные системы управления базами данных (СУБД). Однако в действительности типы информации, размещаемые в базах данных, имеют большое количество качественных и количественных отличий и находятся между собой в отношениях связанности. Поэтому для работы с более сложными структурами принято использовать несколько таблиц связанных друг с другом. Базы данных, имеющие связанные таблицы, именуются реляционными базами данных.

Применительно к объекту нашей работы – корпусу прагматически маркированной лексики структура базы данных будет зависеть, во-первых, от формальной структуры словарной статьи, поскольку языковая единица без соответствующего описания ее разноплановых свойств ничего не дает исследователю, во-вторых, будет зависеть и от набора речевых информационных характеристик, тех прагматических потенций, что заложены в единице в силу тех или иных причин.

В первую очередь, структуру словарной статьи необходимо формализовать следующим образом: как известно, базовая структура словарной статьи, например, статьи толкового словаря, имеет вид **лемма – словарная статья.** Проецируя ее на табличную структуру базы данных, получаем две таблицы: в первой таблице (№ 1) хранятся заголовки словарных статей, а во второй таблице (№ 2) – собственно тексты словарного описания. При этом поля обеих таблиц имеют соответствующую корреляцию: одно из полей таблицы № 1, содержащее заголовок словарной статьи связано с полем таблицы № 2, содержащем текст словарного описания, соответствующего данной единице. Составляющими базы данных корпуса прагматически маркированной лексики должны стать также элементы, являющиеся для пользователя параметрами поисковой системы – например, способы семантизации прагматически маркированного языкового знака, прагматические макрокомпоненты [7]; типы и способы реализации прагматической семантики в разных типах дискурсов, способы лексикографического представления прагматической семантики в словарях разных типов.

СПИСОК ЛИТЕРАТУРЫ

1. **Баранов А. Н.** Введение в прикладную лингвистику: учебное пособие. – М.: Едиториал УРСС, 2003. КОЛ_ВО СТР
2. **Дейт К. Дж.** Введение в системы баз данных. – М.: Вильямс, 2006. – 1328 с.
3. **Коннолли Т., Бегг К.** Базы данных. Проектирование, реализация и сопровождение. Теория и практика. – М.: Вильямс, 2003. КОЛ_ВО СТР
4. **Перванов Я.** Языковой резонанс и компьютерная лексикография [Электронный ресурс]. Полное ОПИСАНИЕ по ГОСТ URL: http://www.sedword.com/sed/content/yazykovoi-rezonans-i-

kompyuternaya-leksikografiya (дата обращения: 05.06.15) 25.08.2013 ПОЧЕМУ ДАТЫ ОБРАЩЕНИЯ ЗА 2013???? Автор обязан проверить существование ресурса на сегодняшний день

5. **Саженин И. И.** Словарный корпус как элемент оптимизации исследовательского процесса // Вестник Новосибирского государственного педагогического университета. – 2013. – № 2. – С. 120–127.
6. **Селегей В. П.** Электронные словари и компьютерная лексикография [Электронный ресурс]. URL: http://www.lingvoda.ru/transforum/articles/selegeya1.asp (дата обращения: 05.06.15) 04.09.2013 ПОЧЕМУ ДАТЫ ОБРАЩЕНИЯ ЗА 2013???? Автор обязан проверить существование ресурса на сегодняшний день
7. **Трипольская Т. А., Булыгина Е. Ю.** Идеологическая семантика как объект лексикографирования разновременными словарями // Вестник Новосибирского государственного педагогического университета. – 2015. – № 2. – С. 28–40. DOI: http://dx.doi.org/10.15293/2226-3365.1502.02
8. **Трипольская Т. А., Гончарова Е. А.** Динамические процессы в лексиконе языковой личности // Вестник Новосибирского государственного педагогического университета. – 2014. – № 3. – С. 57–67. DOI: http://dx.doi.org/10.15293/2226-3365.1403.06
9. **Шведова Н. Ю**. Парадоксы словарной статьи // Национальная специфика языка и её отражение в нормативном словаре. – М., 1988. – С. 6-11 ИЗДАТЕЛЬСТВО
10. **ABBYY Lingvo** [Электронный ресурс]. – Изд-во Abby, 2006. – 1 электронн. опт. диск (CD-ROM). Полное ОПИСАНИЕ по ГОСТ
11. **Cambridge Online Dictionary** [Электронный ресурс]. – Режим доступа: http://dictionary.cambridge.org/ (**дата обращения**: 01.08.2013) ПОЧЕМУ ДАТЫ ОБРАЩЕНИЯ ЗА 2013???? Автор обязан проверить существование ресурса на сегодняшний день
12. **DICT** [Электронный ресурс]. – Режим доступа: http://www.dict.org/bin/Dict Полное ОПИСАНИЕ по ГОСТ (**дата обращения**: 01.08.2013) ПОЧЕМУ ДАТЫ ОБРАЩЕНИЯ ЗА 2013???? Автор обязан проверить существование ресурса на сегодняшний день
13. **Kyröläinen A. J.** Low-Frequency Constructions and Salience: a Case Studyon Russian Verbs of Motion of Dative Impersonal Construction Type // Инструментарий русистики: корпусные подходы. – Slavica Helsingiensia, 34. Helsinki University Press, 2008. – С. 176–197.
14. **Random** House Webster's Dictionary Italiana [Электронный ресурс]. – Изд-во Random House Reference, 2005. – 1 электронн. опт. диск (CD-ROM). Полное ОПИСАНИЕ по ГОСТ
15. **Shorter** Oxford English Dictionary [Электронный ресурс]. – Изд-во Sopheon UK, 1993. – 1 электронн. опт. диск (CD-ROM). Полное ОПИСАНИЕ по ГОСТ
16. **Русские** словари [Электронный ресурс]. – Режим доступа: http://www.slovari.ru/ (дата обращения: 01.08.2013) Полное ОПИСАНИЕ по ГОСТ

2015, Vol. 5, No. 5 http://en.vestnik.nspu.ru ISSN 2226-3365

DOI: 10.15293/2226-3365.1505.10

Sazhenin Igor Igorevich, Candidate of Philological Science, Associated Professor, Modern Russian Language Department, Novosibirsk State Pedagogical University, Novosibirsk, Russian Federation.
E-mail: sajana84@mail.ru

ABOUT CONSTRUCTION OF A DATABASE OF PRAGMATICALLY MARKED VOCABULARY

Abstract

In recent years, an interest of researchers attracts pragmatic component of lexical meaning. Among the most significant problems of fundamental linguistics, lying at the intersection of structural-systemic and communicative description of lexical meanings of the word are, in particular, identifying pragmatically marked fragments of vocabulary and construiction typology of micro-components, which are part of a pragmatic macro-component, as well as the study of their speech development. This article addresses the issues and principles of building of database of pragmatically marked vocabulary, the corpus of which must be constructed taking into account the linguistically meaningful information: the ideological, national-cultural, gender, social, emotive-evaluative and others components of the meaning of the word. Principles of construction of a database with the technical position and with the content position take into account the mechanisms of constructing of linguistic corpuses (Russian National Corpus, Helsinki Annotated Corpus, the Brown Corpus and others.), the experience of creating the electronic lexicographical resources (ABBYY LINGVO, «Russian dictionaries», DICT; Cambridge Online Dictionary; Shorter Oxford English Dictionary Webster's Dictiomary) and the tradition of compiling paper dictionaries of different types .

Keywords

Pragmatics, linguistic corpus, dictionary, electronic lexicography, linguistic analysis, linguistic database

REFERENCES

1. Baranov A. N. *Introduction to applied linguistic.* Moscow, Editorial URSS Publ., 2003. (In Russian).
2. Date C. J. *Introduction to Database Systems.* Moscow, Williams Publ., 2006, 1328 p. (In Russian).
3. Connolly T., Begg K. *Databases. Design, implementation and maintenance. Theory and practice.* Moscow, Williams Publ., 2003. (In Russian).
4. Pervanov Y. *Language resonance and computer lexicography*. Available at: http://www.sed-word.com/sed/content/yazykovoi-rezonans-i-kompyuternaya-leksikografiya (Accessed 05.06.15) (In Russian)
5. Sazhenin I. I. Vocabulary corpus as element of optimization of research. *Novosibirsk State Pedagogical University Bulletin*. 2013, no. 2, pp. 120–127. (In Russian)
6. Selegey V. P. *Electronic Dictionaries and computer lexicography*. Available at: http://www.lingvoda.ru/transforum/articles/selegeya1.asp (Accessed 05.06.15) (In Russian)
7. Tripolskaya T. A., Bulygina E. Y. Ideological semantics as the object of lexicographically different-temporal dictionaries. *Novosibirsk State Pedagogical University Bulletin.* 2015, no. 2, pp. 28–40. DOI: http://dx.doi.org/10.15293/2226-3365.1502.02 (In Russian)

2015, Vol. 5, No. 5 http://en.vestnik.nspu.ru ISSN 2226-3365

8. Tripolskaya T. A., Goncharova E. A. Dynamic processes in the lexicon of a language personality. *Novosibirsk State Pedagogical University Bulletin.* 2014, no. 3, pp. 57–67. DOI: http://dx.doi.org/10.15293/2226-3365.1403.06 (In Russian).
9. Kyröläinen A. J. Low-Frequency Constructions and Salience: a Case Studyon Russian Verbs of Motion of Dative Impersonal Construction Type. *Russian Studies Instrumentation: corps approaches.* Slavica Helsingiensia, 34. Helsinki University Press Publ., 2008, pp. **176–197. (**In Russian).