Источники и словари

A.G. – Götze A., Volz H. Frühneuhochdeutsches Lesebuch. 5. Aufl. Göttingen, 1967. 172 S.

F.L. – Frühneuhochdeutsches Lesebuch / Hrsg. von O.Reichmann und K.P.Wegera. Tübingen, 1988. 240 S.

N.W. – von Wyle N. Translationen / Hrsg. von A.von Keller. Stuttgart, 1861. 373 S.

DWB. – Deutsches Wörterbuch / Hrsg. von J. und W.Grimm. Leipzig, 1854 1971. Bd. 1-16.

M.L. – Matthias Lexers mittelhochdeutsches Taschenwörterbuch. 36. Aufl. Leipzig, 1980. 504 S.

С.Е. Вадяев*

Электронная лексикография и корпусная лингвистика

Vadjaew S.

Elektronische Lexikographie und Korpuslinguistik

Elektronische Wörterbücher sind in den letzten Jahren zu komplexen lexikographischen Objekten geworden, die ein neues Verfahren bei ihrer Verfassung und Repräsentation benötigen. Insbesondere betrifft das die zweisprachigen Wörterbücher. Mit dem Wechsel des Mediums sind auch die Schranken des „Papierformats" überwunden. Für die funktionale Erweiterung der Wörterbücher kann ein Parallelkorpus eine bedeutende Rolle spielen, indem es als Sammlung von Belegen für ein Wörterbuch dienen kann. Außerdem kann ein Parallelkorpus in anderen Bereichen benutzt werden, wie z.B. bei der kontrastiven Sprachanalyse oder im Fremdsprachenunterricht.

Термин "электронный словарь" стал уже привычным в современной лексикографии. При этом, атрибут "электронный" так же мало говорит о своем объекте, как и противоположный ему атрибут "бумажный" - о традиционном словаре. Обычно подразумевается, что словарь на компьютере – это введенный в него бумажный словарь, снабженный удобными средствами поиска и отображения лексикографического содержания. То есть, создатели

* С.Е. Вадяев, Самарский государственный университет

электронных словарей делают из бумажных словарей эквивалентный электронный вариант, отображаемый на мониторе компьютера.

Компьютерная лексикография, как область прикладной лингвистики, производящая такие словари, оказывается лишенной собственного языкового предмета. На ее долю оставляется только эффектная демонстрация канонического содержания. Однако, это совсем не так, поскольку компьютерная лексикография является отдельным направлением в практической лексикографии со своими собственными подходами не только к отображению, но и к содержанию словаря. ***Электронный словарь*** - это особый лексикографический объект, в котором могут быть реализованы и введены в обращение многие продуктивные идеи, не востребованные по разным причинам в "бумажных" словарях.

Плоды традиционной практической лексикографии страдают от трех фундаментальных противоречий. Они могут быть сформулированы относительно некоторого безусловного качества, которое наиболее важно для пользователей словаря – ***полноты***. Здесь под этим качеством подразумевается стремление охватить как можно больше лексических значений и при этом дать максимум разнообразной информации об этих значениях: от этимологии до сочетаемости. В итоге:

а) чем полнее словарь, тем сложнее им пользоваться. Это противоречие привело к поляризации рынка бумажных словарей: имеется большая группа массовых изданий, довольно простых, но относительно удобных, которой противостоят единичные профессиональные издания большого объема, непригодные для быстрого получения информации;

б) чем полнее словарь, тем менее он соответствует текущей языковой и культурной ситуации;

в) чем ближе концепция словаря к достижениям современной лингвистики, тем менее он полон.

Действительно, универсальные бумажные словари демонстрируют печальное отсутствие влияния достижений теоретической лексикографии на лексикографическую практику.

Компьютерная реализация бумажного словаря сама по себе позволяет преодолеть часть указанных проблем. К ***новым возможностям электронного словаря*** относятся:

1. Более гибкие средства для представления содержания словарной статьи, включая возможность частичного показа по

разным критериям (различные "проекции" словаря), разнообразные графические средства, которые не используются в обычных словарях. Таким образом, один и тот же электронный словарь может быть легко адаптирован и для школьника, и для профессионального переводчика или лингвиста.

2. Использование для доступа к содержанию различных лингвистических технологий, таких как морфологический и синтаксический анализ, полнотекстовый поиск, распознавание и синтез звука и т.п. Наиболее разработанными на данный момент являются технологии поиска, в том числе и с применением морфологического и синтаксического анализа. Это значит, что в электронном словаре возможен поиск не только по леммам, но и по текстам словарных статей (что часто бывает очень необходимым). При этом возможно ограничить поиск на одном из тематических полей словарной статьи, например, толкование слова, примеры употребления, этимология, стилистические пометы и т.п. Технология морфологического анализа позволяет найти исходную форму слова по любой его грамматической форме, а также отобразить все однокоренные или семантически родственные слова. Особенно полезно использование этой функции в двуязычных словарях для пользователей, которые начинают изучение иностранного языка, в частности, немецкого.

С точки зрения пользователя смысл реализации в электронном словаре всех этих технологий состоит в том, что становится возможным быстро получить информацию, которая содержится где-то в недрах словаря и непосредственно отвечает тому запросу, который сформулирован пользователем в удобной для него форме. При традиционном подходе минимальной единицей доступа является лексема (имя словарной статьи): необходимо прочитать всю статью, чтобы определить, содержится ли в ней ответ на запрос пользователя. Для многотомных словарей это представляет серьезную проблему, поскольку некоторые слова (особенно глаголы) могут иметь до нескольких десятков, а то и сотен основных значений. Например, в словаре *DUDEN - Das große Wörterbuch der deutschen Sprache* глагол *kommen* имеет 22 основных значения, многие из которых имеет несколько подзначений; в Оксфордском словаре английского языка глагол *set* имеет 400 только основных значений. При таком объеме словарной статьи довольно затруднительно быстро найти в ней необходимую информацию.

Пользователь хотел бы, чтобы словарь максимально локализовал релевантную информацию. При этом речь все же не

идет об автоматическом переводе (если мы рассматриваем переводной словарь). Специфика словарного ответа состоит в том, что он дает весьма разнообразную информацию о слове или словосочетании, а не просто переводное соответствие, предполагает активный выбор пользователя из нескольких возможных хорошо обоснованных альтернатив.

Электронная реализация словаря позволяет также решить проблему "монофункциональности" бумажных словарей. К примеру, особенностью большинства бумажных переводных словарей является ориентация описания структуры лексического значения в исходном языке на лексическую систему языка перевода и на реализацию ровно одной функции – собственно перевода с языка А на язык Б в предположении, что язык А является иностранным, а язык Б – родным. Такое ограничение делает словарь исключительно неудобным при необходимости перехода от пользовательской модели "Читатель" к модели "Писатель". Фактически сегодня такие модели реализуются разными типами словарей, что достаточно неудобно для читателя. Как уже указывалось, фундаментальные бумажные словари – неизбежно словари устаревшие. Особенно это характерно для разговорной лексики. Функции фиксации текущего состояния языка берут на себя словари малого объема, обычно не полные и поверхностные. Новые значения в них оторваны от своих языковых корней, плохо, поверхностно или произвольно объяснены.

Для массовых программных продуктов, каковыми являются электронные словари, характерны частая смена версий и наличие постоянной обратной связи с тысячами пользователей. Поэтому компьютерная лексикография – это неизбежно актуальная лексикография. Эту тенденцию можно проследить на примере, пожалуй, самого удачного на сегодняшний день отечественного электронного словаря *"Lingvo"* (www.lingvo.ru). Данный программный продукт обновляется с интервалом один-два года. При этом меняется не только программная оболочка словаря, но и дополняется его лексикографическое содержание. На форуме компании производителя в интернете пользователи активно обсуждают достоинства и недостатки словаря, а также указывают на неточности и ошибки, которые обычно исправляются уже в следующей версии словаря.

Жизнь электронного словаря должна быть похожа на жизнь других программных продуктов: она обусловлена с одной стороны, стремлением пользователей обнаружить очередную ошибку или

лакуну, и, с другой стороны, возможностью и необходимостью внести необходимые исправления сейчас, а не через десятилетия. Такой подход всего лишь фиксирует естественное положение дел: коллективное авторство на словарное содержание принадлежит всем носителям языка, задача лексикографа – фиксация языковых фактов и их методически правильное описание.

Если взглянуть на те пути решения лексикографических проблем, то становится понятным, что они едины для всех направлений компьютерной лингвистики. Однако словарные технологии являются в некотором смысле более базовыми. Они не связаны непосредственно с задачами распознавания сложных синтаксических структур или задачами логического анализа содержания, которые возникают в системах понимания и обработки естественного языка.

Для любого словаря наиболее важным аспектом является то, насколько полно он отражает структуру языка (двух или более языков для переводных словарей) и в каком объеме позволяет находить средства для адекватного выражения мыслей на родном или иностранном языке.

Понятие "лексической функции", позволяющее систематически описывать *несвободную сочетаемость* слов (например, то, что "войну ведут", а "экзамен – держат", что "теории выдвигают", а "мысли подают" и т.п.) играет при этом очень важную роль. Каждый носитель языка интуитивно находит подходящие сочетания для выражения той или иной мысли, человек, изучающий иностранный язык, должен искать такие в словаре.

Несвободная сочетаемость слов, результатом которой являются устойчивые сочетания слов, очень важны для овладения языком. Словосочетанием принято считать два или более синтаксически связанных слова. В этом смысле синтаксически связанными словами являются не только *хорошая возможность, возложить обязанности, встретиться позже*, но и *у реки, не имела, войти в*. Не будут являться словосочетаниями встречающиеся вместе смежные слова *чтобы скрытно, пугать всякими, чтобы они* и т.п., поскольку непосредственной синтаксической сочетаемости между составляющими эти фрагменты словами здесь нет.

И.Мельчук [Mel'chuk 1989] выделяет следующие классы словосочетаний:

- ***полные фраземы***, т.е. устойчивые словосочетания, смысл которых не выводится из смысла входящих полнозначных слов: *синий чулок, клевать носом, положить глаз (на кого-то)*;

- ***полуфраземы***, т.е. устойчивые словосочетания, смысл которых включает смысл одного из сочетаемых слов, в то время как второе слово, подбираемое к первому, берется не в основном своем смысле. Известно и другое название полуфразем – ***коллокации*** или словосочетания с ***лексическими функциями***: *уделять внимание, войти в доверие, точный инструмент, крепкий чай, неверное решение, потворствовать агрессии.*

Все прочие словосочетания считаются ***свободными***, т.е. их смысл прямо выводится из смысла сочетаемых слов: *предвещать грозу, видеть лес, интересоваться физикой* и т.п.

Многочисленны случаи, когда данное словосочетание может трактоваться либо как свободное, либо как фразема (полуфразема). Например, *желтый билет* может означать не только билет желтого цвета (свободное словосочетание), но и документ (полная фразема).

При использовании любого современного лингвистического приложения (например, программы для редактирования текста, обучения иностранному языку и др.) часто возникает необходимость использования дополнительной лексикографической информации, к которой в первую очередь относится информация о сочетаемости данной лексемы.

Человеку, готовящему текст или обучающемуся иностранному языку, информация о сочетаемости слов необходима по двум основным причинам:

1. Нужно знать, как различные ограничения, налагаемые языком на сочетаемость слов, влияют на тип конкретного словосочетания.

2. Опираясь на слова, сочетающиеся с данным омонимичным словом, можно пытаться разрешить лексическую омонимию. Только опираясь на синтаксический контекст слова, можно с высокой вероятностью разрешить его омонимию.

В обычном словаре источником информации о сочетаемости слов являются примеры их употребления. Как правило, объем таких примеров строго ограничивается рамками "бумажного словаря". Помимо этого они представляют собой лишь самые релевантные случаи, выбор которых имеет, тем не менее, довольно субъективный характер. Отбор таких примеров для словарных статей осуществлялся с помощью картотеки, составляемой авторами словарей.

В качестве основы для выборки примеров служат оригинальные литературные произведения, периодические издания, а также

специализированные публикации, относящиеся к различным областям науки и техники. Так, например, в списке лексикографических источников словаря *DUDEN - Das große Wörterbuch der deutschen Sprache* [Duden 1999] приводится свыше тысячи наименований. Среди них помимо художественных произведений немецких авторов и периодических изданий можно найти издания по электронике и электротехнике, машиностроению, медицине и юриспруденции. В двуязычных словарях выборка примеров употребления во многом опирается на материал одноязычных или специализированных словарей. Так, к примеру, объем лексикографических источников *Немецко-русского (основного) словаря* под редакцией К.Лейна составляет чуть более пятидесяти наименований, подавляющая часть которых также является словарями.

В последние годы некоторые издательства, специализирующиеся на издании словарей (например, немецкое издательство "DUDEN" и британское "Collins"), перешли с использования традиционной "бумажной" картотеки на электронные корпусы текстов, которые помимо всего прочего позволяют статистически оценивать релевантность слова или словосочетания, а также выявлять случаи несвободной сочетаемости. Автоматический анализ корпуса текстов не способен полностью заменить работу редакторов, однако они позволяют существенно повысить уровень объективности, а также оперативно изымать из словаря устаревшие примеры и словосочетания и добавлять неологизмы.

Использование машинных корпусов для составления словарей отражает общие тенденции развития в современной лингвистике. Поскольку оно опирается на сбор и анализ объемного языкового материала, новые информационные технологии играют для составителей словарей все большую роль. Они позволяют существенно сократить время обработки и поиска языковых данных, а также повышают достоверность выводов.

Под **корпусом данных** в широком смысле понимается "сформированная по определенным правилам выборка данных из проблемной области" [Баранов 2001, С. 115]. **Корпус текстов** представляет собой "вид корпуса данных, единицами которого являются тексты или их достаточно значительные фрагменты [...]" [Баранов 2001, С. 115]. В зависимости от структуры и назначения корпусов выделяются различные их типы.

Особенно полезными для улучшения функциональности и полноты двуязычных словарей могут быть **корпусы параллельных текстов**. Корпус параллельных текстов представляет собой множество текстов на языке-источнике и множество текстов на языке-цели, которые являются переводами текстов языка-источника. Корпус параллельных текстов может служить для научных (контрастивные исследования двух и более языков) и практических целей (преподавание иностранных языков). В зависимости от назначения и сферы применения информация в параллельном корпусе может храниться в виде **неструктурированного текста** или **структурированного формата хранения** (текст со специальной разметкой). При структурированном формате хранения информации оба текста разбиваются на параллельные сегменты (фрагменты текста, абзацы, предложения и даже слова), содержание которых соответствует друг другу. Помимо этого, корпус может содержать морфологическую, синтаксическую и другую релевантную информацию об отдельном слове и о тексте в целом.

В основе корпуса параллельных текстов могут лежать тексты различных типов (литературные произведения, газетные статьи и т.д.), для которых имеется перевод на иностранный язык. Тексты переводятся в электронную форму, сегментируются и подвергаются параллелизации.

Для структурирования, а также для кодирования дополнительной лингвистической информации, как правило, используются специальные языки разметки текстов. Автоматический анализ и обработка корпуса текстов предполагает точное описание всех элементов текста, а также отношений между ними. SGML (Standard Generalized Markup Language) и XML (Extensible Markup Language) являются языками разметки, позволяющими описывать структуру текста и его элементы на метаязыковом уровне. Такая разметка позволяет осуществлять более гибкую обработку текстов корпуса и снижает вероятность ошибочной интерпретации при автоматическом анализе.

Использование **структурированного корпуса параллельных текстов** в электронной лексикографии поможет восполнить пробелы традиционных словарей:

· В корпусе параллельных текстов возможен поиск не только эквивалентов для отдельных слов, но и для словосочетаний, что необходимо при переходе от пользовательской модели "Читатель" к модели "Писатель". Это очень важно для людей, изучающих

иностранный язык, поскольку в отличие от носителя языка, который интуитивно находит подходящие сочетания для выражения той или иной мысли, человек, изучающий иностранный язык, должен искать такие словосочетания в словаре.

· Пользователь видит не отдельно взятое слово или словосочетание, оторванное от контекста и снабженное переводом. Корпус дает возможность показать реализацию слова в контексте, что играет важную роль для понимания его смысла и закономерностей его употребления.

· Использование корпуса параллельных текстов может быть полезным и для составителей словарей. Особенно интересно использование параллельного корпуса для составления и обновления специализированных словарей.

· Параллельный корпус может использоваться при изучении иностранного языка.

Корпус параллельных текстов может применяться не только для прикладных лексикографических задач, но и для контрастивных лингвистических исследований в области синтаксиса и лексикологии, а также для преподавания иностранного языка. Использование параллельного корпуса в контрастивных лингвистических исследованиях, а также в прикладных задачах открывает новые возможности как для исследователей, так и для простых пользователей электронных словарей.

Библиографический список

1. Mel'chuk, Igor. Fraseologia y diccionario en la linguistica moderna // I. Uzcanga Vivar et al. (Eds.) Presencia y renovacion de la linguisticafrancesa. Salamanca: Ediciones Universidad, 1989. P. 267-310.

2. Mel'chuk I., Zholkovsky A. The explanatory combinatorial dictionary // M. Evens (ed.) Relational models of lexikon. Cambrige, Eng land: Cambridge University Press, 1988. P.41-74.

3. Wanner, L. (Ed.). Lexical Fuctions in Lexicography and Natural Language Processing. Studies in Language Companion Series, ser. 31. Amsterdam / Philadelphia: John Benjamin, 1996.

4. Bergenholtz H., Mugdan J. Formen und Probleme der Datenerhebung II: Gegenwartsbezogene synchronische Wörterbücher // Wörterbücher: ein internationales Handbuch zur Lexikographie. Berlin; New York: de Gruyter, 1990, Bd. 5, Teilbd. 2.

5. Баранов А.Н. Введение в прикладную лингвистику. Москва, 2001.

6. DUDEN - Das große Wörterbuch der deutschen Sprache, die 3. völlig neu bearbeitete und erweiterte Auflage in 10 Bänden. Mannheim, 1999.

7. Немецко-русский словарь (основной) / Под ред. К.Лейн. М.: Русский язык, 1992.

8. Lenders W. (Hg.): Computereinsatz in der angewandten Linguistik – Konstruktion und Weiterverarbeitung sprachlicher Korpora. Frankfurt a.M.: Lang, 1993.

9. Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher / Hrsg. von I.Lemberg. Tübingen: Niemeyer, 2001.

10. Schneider R. Rechtschreibung auf Knopfdruck: Elektronische Wörterbücher // Lexicographica 9, 1993, S. 220-229.

11. Storrer A.: Hypermedia-Wörterbücher: Perspektiven für eine neue Generation elektronischer Wörterbücher // Wörterbücher in der Diskussion III. Vorträge aus dem Heidelberger Lexikographischen Kolloquium. Hrsg. von Herbert Ernst Wiegand. Tübingen. 1998. S.106-131.

12. Johansson, Stig: Towards a multilingual corpus for contrastive analysis and translation studies. SPRIKreports: Reports from the project Languages in Contrast (Språk i kontrast), Department of British and American Studies, University of Oslo 2000; http://www.hf.uio.no/german/sprik.

13. English-Norwegian Parallel Corpus (ENPC), более подробная информация о ENPC, включая публикации находится на сайте: http://www.hf.uio.no/iba/prosjekt/

14. Johansson, Stig: On the role of corpora in cross-linguistic research. // S. Johansson and S. Oksefjell (eds.), Corpora and cross-linguistic research: Theory, method, and case studies, 3-24. Amsterdam and Atlanta, GA: Rodopi, 1998.