Л. Н. Беляева

ПОТЕНЦИАЛ АВТОМАТИЗИРОВАННОЙ ЛЕКСИКОГРАФИИ И ПРИКЛАДНАЯ ЛИНГВИСТИКА

Прикладная лексикография является одним из направлений прикладной лингвистики, в ее задачу входит создание и различных проблемно- и предметно-ориентированных автоматизированных и автоматических словарей и баз данных. Полнота и адекватность лексикографических систем в значительной степени определяют уровень и надежность извлечения информации и знаний из текстов различного состава, структуры и назначения.

Ключевые слова: автоматический словарь, терминологическая база данных, лексикографическая система, информационные технологии, создание и ведение словаря.

L. Belyaeva

AUTOMATIC LEXICOGRAPHY SCOPE AND APPLIED LINGUISTICS

Applied lexicography is one of the special branches of applied linguistics, its task is creation and updating different automated and automatic dictionaries and databases, which are problem and subject oriented. Completeness and adequacy of lexicographic systems to a considerable extent determine the level and reliability of information and knowledge extraction from the texts of various composition, structure and assignment.

Keywords: automatic dictionary, terminological database, lexicographic system, information technologies, dictionary management.

Современная прикладная лексикография превращается сегодня в важную отрасль прикладной лингвистики.

Развитие информационных технологий (ИТ), составляющих сегодня часть профессиональной подготовки специалистов-филологов, определяет возможность разработки и использования компьютерных средств и специализированных технологий в области анализа и обработки естественного языка для создания и ведения различных типов лексикографических систем. Полноценность и адекватность спектра лексикографических систем в значительной степени определяют уровень и достоверность извлечения информации и знаний из текстов различного состава, структуры и назначения, а информационные технологии позволяют ускорить процесс сбора и анализа материала.

Автоматизированная лексикография является особым направлением прикладной лексикографии и прикладной лингвистики с особыми подходами не только к способам отображения информации, но и к содержанию словаря. Это связано с тем, что современные задачи словарной науки и практики заключаются:

1) в совершенствовании существующих типов словарей;

2) в создании новых типов словарей на основе новых актуальных принципов и технологий;

3) в соединении уже известных типов словарей в новые словарные комплексы и терминологические системы;

4) в создании компьютерной методологии лексикографической деятельности на основе машинных фондов языков [6; 7].

Автоматизированный словарь — это особый лексикографический объект, обладающий рядом специфических структурных особенностей [15]. Под автоматизированными словарями будем понимать электронные словари, описывающие конкретную терминологию, энциклопедические и толковые словари на машинных носителях, а также автоматические словари систем компьютерной переработки информации, включая автоматические словари систем машинного перевода. Важность создания и ведения таких словарей постоянно возрастает, поскольку сегодня изменились функции терминолога: именно терминолог идентифицирует в тексте термины, проверяет фиксацию их и их переводов в базе данных и ведет всю лексикографическую работу по установлению перевода и описанию нового термина. Только после этого «размеченный» и/или обработанный системой машинного перевода текст передается переводчику [16, с. 19–20].

Автоматизированные словари существенно отличаются от словарей традиционных, во-первых, способом хранения, во-вторых, строгой формализованностью записи информации [1, с. 46]. Фундаментальные «бумажные» словари — неизбежно словари устаревшие или устаревающие, так как они не обладают динамикой [13]. А для программных продуктов, таких как словари систем автоматической переработки текстов, характерны частая смена версий, наличие постоянной обратной связи с пользователями, динамичность существования, актуальность и открытость относительно каждого временного среза. Кроме того, автоматизированные словари, в отличие от традиционных, сочетают большой объем информации с удобством использования благодаря автоматизации механизмов поиска по соответствующему запросу.

Автоматизированные и автоматические словари — словари систем переработки данных ориентированы на использование программами автоматического анализа и переработки текстов на естественном языке. В современном варианте для таких словарей создается специализированная система их создания и ведения, обеспечивающая удобную работу лексикографа с информацией, которая может быть представлена в формальном и/или закодированном виде. Так, принято различать:

- автоматизированные словари, предназначенные для конечного пользователя-человека;
- автоматизированные словари, предназначенные для программной обработки текста — автоматические словари.

Термин «автоматический словарь» относится именно к словарю, используемому системой автоматической переработки текста и не предназначенному для использования человеком [3, с. 67]. В такой формулировке под автоматическим словарем подразумевается автоматический переводной словарь (АПС), предназначенный в основном для обработки научно-технических текстов. Такие словари, являясь упорядоченным массивом лингвистической информации, регистрируют и хранят лексические единицы (слова и словосочетания) с их морфологическими, синтаксическими и семантическими характеристиками, необходимыми для синтаксического анализа (парсинга) и автоматического перевода текста.

Автоматические словари ориентированы на перевод специальных текстов и задачи конкретного пользователя, а также отражают отдельную предметную область знаний и соответствующую ей узкоспециальную терминологию. В связи с этим бóльшая часть состава автоматического словаря представлена именно терминологической лексикой. Однако АПС включает не только узкоспециальную терминологическую лексику, но и наиболее употребительную лексику смежных областей, общеспециальную терминологическую и нетерминологическую лексику, лексику общеупотребительную.

Автоматизированный словарь рассматривается как база данных, в которой каждая статья представлена как текст, расписанный по зонам словарной статьи, при этом каждая зона имеет свое уникальное имя [10]. Такой словарь должен иметь жесткую организацию и высокий уровень формализации представления данных. Соответственно автоматизированный словарь включает словарные файлы, индексы к ним, описание микро- и макроструктуры словаря и систему программ, обеспечивающую создание этой конструкции, ее поддержание в рабочем состоянии и обращение к ней за справками из других программ [2, с. 124].

Классификация автоматических словарей и в целом лексикографических продуктов, созданных с использованием информационных технологий и используемых с их применением, может проводиться по нескольким параметрам:

1) По способу создания словарей:

- словарь может создаваться методами традиционной лексикографии, но может быть доступен и в электронном формате. В этом случае мы имеем дело со словарем на электронном носителе, полностью воспроизводящем микро- и макроструктуру исходного «бумажного» варианта, но доступным для работы на компьютере и соответственно снабженном средствами поиска информации. Как правило, в таком словаре поиск информации осуществляется по слову в его канонической (словарной) форме или по указанному пользователем словосочетанию. Система поиска дает возможность выбрать все словарные статьи, содержащие слово/словосочетание запроса;
- словарь может создаваться с использованием информационных технологий непосредственно в электронной форме и в соответствии с установленными принципами.

2) По назначению словарей:

- для традиционных филологических исследований (частотные, алфавитные и обратные словники, конкордансы),
- для использования в системах автоматической переработки информации.

Использование ИТ в лексикографии предполагает ряд возможностей:

1. Для поддержания работы лексикографа при создании и ведении словарей:

– решение задач отбора лексических единиц (ЛЕ), их лексикографирования, извлечения информации о лексической единице из проблемно-ориентированного массива текстов;

– создание, редактирование и коррекция оригинал-макета словаря, последующего издания;

– создание и ведение словников на основе отбора ЛЕ из лексикографических баз данных по заданному критерию или набору критериев;

– создание и ведения терминологических баз данных и онтологий.

2. Для поддержания работы специалиста и/или переводчика при обращении к словарям разного типа и на разных носителях:

– извлечение информации из различных лексикографических источников (автоматических, автоматизированных, резидентных словарей);

– исследование лексического состава и динамики лексического спектра конкретного языка/ подъязыка.

Необходимость использовать весь потенциал ИТ на этапах создания и ведения современных лексикографических систем определяется сегодня потребностями научного и технического сообщества, задачами исследования терминологии, развитием множества (более 300) [8, с. 23] языков для профессиональных целей, а также собственно потребностями современной лексикографической работы.

Основные преимущества электронного словаря по сравнению с бумажным состоят:

- в значительном ускорении процесса поиска слов и словосочетаний, а также переводных значений в развернутой словарной статье,
- в различных возможностях просмотра содержания словарной статьи, в максимальной локализации требуемой информации,
- в возможности быстрой подстановки переводных эквивалентов в текстовые файлы (путем копирования в буфер обмена),
- в относительной гарантии того, что благодаря частым обновлениям словарь соответствует современному состоянию языка,
- в возможности оперативного внесения добавлений в словарь.

Компьютеризация лексикографии и условия нового информационного пространства обусловливают необходимость разработки принципов построения автоматизированных словарей и терминологических баз данных (ТБД), к которым относятся:

- модульность [2; 12; 3] — словарные статьи формируются как относительно независимые объекты, но организуются в системы, «которые в дальнейшем могут рассматриваться как подсистемы некоей конкретной системы обработки данных» [2, с. 107]
- динамичность [4] — этот принцип подразумевает возможность оперативно, и без существенных трудностей корректировать номенклатуру словаря, состав и объем информации в словарной статье, а также включать новые программы и файлы в уже созданную систему [2];
- гибкость — заключается в способности уже созданных баз данных и словарей удовлетворять новым требованиям без глобальной перестройки их организации, что подразумевает возможность информационного поиска и выборки данных, а также обратимость словаря [4];
- сбалансированность — определяет системность комплектования баз данных и словарей, исключает произвольность и стихийность в отборе информации и подразумевает тщательный отбор материала, структурность базы, полноту представляемой информации [5, с. 13];
- дружественность — обеспечивает наиболее эффективный характер взаимодействия между системой и пользователем; она выражается в использовании графических изображений, разного вида меню, в звуковом сопровождении, т. е. в средствах, делающих работу пользователя максимально удобной. Такой принцип взаимодействия получил название «дружественного интерфейса», а также «интуитивно-понятного интерфейса» [14, с. 16].

Выбор конкретной макро- и микроструктуры словаря определяет оценку соответствующих информационных ресурсов, которые способны поддерживать выбранный формат представления и хранения данных. Это могут быть стандартные базы типа Microsoft Access или специально создаваемые средства типа систем ABBY Lingvo Content, Polyglossum и им подобные.

Первоначально электронные словари создавались в виде электронных версий традиционных и проверенных бумажных словарей, однако при таком подходе они обновлялись не чаще, чем переиздавался печатный первоисточник.

Такие словари до сих пор существуют в электронном формате и доступны через сеть Интернет (например, Большой англо-русский словарь под редакцией Ю. Д. Апресяна (http://www.rambler-co.ru/dict/enru/) и на компакт-дисках. Однако гораздо более полно современное состояние языка отражают такие электронные словари, которые являются самостоятельной (независимой от бумажных версий) единицей и постоянно обновляются и пополняются.

В некоторых случаях бумажный словарь используется в качестве отправной точки для создания электронного словаря (например, так составлялись словарь общей лексики Контекст и специализированные словари Lingvo), но в дальнейшем словарь пополняется по мере изменения лексического состава языка. Поскольку выпуск новой версии электронного словаря или же просто обновление предыдущей версии занимает гораздо меньше времени, чем переиздание словаря бумажного, а также благодаря установленной электронной обратной связи разработчиков с пользователями словарей (например, для Lingvo), электронные словари чаще модернизируются и соответственно не устаревают.

При создании электронных словников или электронных версий словников на базе «бумажных» словарей должны быть решены следующие задачи:

1) *Перевод материала из печатной формы в электронную*. Как правило, такое преобразование осуществляется путем сканирования информации с помощью современных средств распознавания. Будем называть результат сканирования словарной информации электронным словарным текстом.

Электронный словарный текст может быть конечным результатом преобразования словаря в электронную форму. В этом случае мы имеем дело со словарем на электронном носителе, который, как правило, оформляется как сопровождающий диск и может использоваться параллельно с «бумажным» словарем. Формат работы с таким словарем может быть различным: как с резидентным словарем, копируемым в память компьютера и формирующим часть автоматизированного рабочего места специалиста, или как с текущим рабочим словарем, устанавливаемым в соответствующий дисковод и доступным только в этом режиме. В последнем случае словарь хранится на диске и устанавливается при выполнении конкретной работы (см., например, электронную версию «Нового англо-русского медицинского словаря», которая не копируется в память компьютера и может быть использована только с диска);

2) *Объединение словников*, представленных в виде электронных словарных текстов, для введения в словарную базу или формирования нового словаря;

3) *Перевод информации* из формата электронного словарного текста во внутренний формат (микроструктуру) автоматизированного или автоматического словаря.

Перевод словаря из бумажной формы в электронную требует проведения тщательного сканирования с обязательной последующей «ручной» проверкой. Такая проверка должна основываться на сопоставлении исходного материала словаря и распечатки его отсканированного варианта. Необходимость подобной трудоемкой и требующей тщательного выполнения и скрупулезности работы связана с тем, что в словарях любой сложности используются разные алфавиты и надстрочные знаки, распознавание которых при сканировании может вызывать ошибки.

Следует отметить, что при создании таких ответственных продуктов, как лексикографические системы, ручная проверка распечатанных материалов является обязательным этапом работы: опыт показывает, что при визуальной проверке материала на дисплее часть ошибок и/или опечаток остаются не выявленными.

После того, как текст традиционного «бумажного» словаря переведен в электронную форму и выверен, должны быть реализованы алгоритмы распознавания его макро- и микроструктуры.

Распознавание макроструктуры словаря определяет способ представления словарных статей — алфавитный, гнездовой, алфавитно-гнездовой, тематический и т. д. Это распознавание может производиться алгоритмически или задаваться лексикографом в режиме диалога или при «ручной» обработке.

Микроструктура словаря и параметры описания лексических единиц (ЛЕ) должны устанавливаться алгоритмически. При этом выделяются границы заглавия словарной статьи, параметры описания и все компоненты описания ЛЕ. Особую сложность представляет собственно установление границ компонентов, поскольку в «бумажных» словарях они не всегда выделяются и/или маркируются специальными символами, и соотнесение выделенных компонентов с параметрами описания.

Решение этой задачи может облегчаться и упрощаться при условии обучения алгоритма в диалоге с лексикографом. В процессе такого диалога на первом этапе должны предъявляться выявленные компоненты описания, которые диагностируются человеком-лексикографом в режиме ограниченного выбора с использованием разворачивающихся меню. Так могут диагностироваться параметры описания и соответствие выделенного компонента определенному параметру. Спектр возможных вариантов в меню определяется микроструктурой создаваемой словарной базы.

Так, например, если микроструктура описания ЛЕ в создаваемой базе данных включает следующие параметры: заглавие, морфологическая характеристика, стилистическая помета, толкование, перевод, то предлагаемый набор параметров должен быть иерархически организован и после установления границ заглавия информация должна представляться в формате «признак–значение».

Например, в словаре терминов электроэнергетики [9] словарная статья слова *switch* имеет следующий вид:

switch 1. переключатель; коммутационный аппарат; коммутатор || переключать; коммутировать **2.** выключатель; прерыватель; разъединитель; рубильник || выключать; прерывать; разъединять **3.** *ж-д* стрелка

При переводе этого формата в формат базы данных на основании символов-разделителей, используемых в словарной статье, должны быть выполнены преобразования, позволяющие заполнить часть информации в базе данных:

заглавие	— *switch*
морфологическая характеристика	— *существительное*
стилистическая помета	—
толкование	—
перевод	*1. переключатель* *коммутационный аппарат* *коммутатор* *2. выключатель* *прерыватель* *разъединитель* *рубильник* *3. стрелка*
морфологическая характеристика	— *глагол*
стилистическая помета	—
толкование	—
перевод	*1. переключать* *коммутировать* *2. выключать* *прерывать* *разъединять*

Структура формируемой статьи зависит от выбранного формата и системы поддержки базы данных, приведенный пример демонстрирует возможный вариант анализа.

После установления модели соответствия «электронный словарный текст – микроструктура» целесообразен переход к автоматическому решению задачи с обращением к диалогу только в случаях с возможными множественными решениями.

Следует особо подчеркнуть, что результат преобразования из электронного словарного текста в базу данных также должен тщательно проверяться «вручную».

Создание словарей на основе использования информационных технологий и имеющихся автоматизированных ресурсов может производиться различными способами:

1) напрямую средствами систем управления базами данных — использование специальных программных средств позволяет создавать новые словари, вносить в них описания, добавлять новую информацию и новые словарные статьи; таким же образом можно удалять или модифицировать словарные статьи в уже существующих словарях;

2) путем компиляции словника из исходного текста на специализированном формальном языке — как правило, в современных системах предусмотрен специальный формальный язык, на котором можно описать любую систему атрибутов и произвольное множество слов с установленными атрибутами, которые могут храниться и обрабатываться системой; в пакет часто входит компилятор, который на основе исходного текста словаря создает готовую базу данных; при обнаружении ошибок в исходном тексте компилятор выдает соответствующий отчет [11, с. 2003];

3) «вручную» на основе специального программного приложения, в котором

предусмотрены операции создания словарей, внесения описаний, добавления, удаления или модификации слов — это наиболее трудоемкий способ создания и редактирования словарей, однако для мелких правок он может быть наиболее удобен;

4) посредством использования имеющихся словарных ресурсов путем включения в лексикографическую систему или автоматизированный словарь новых языков или подъязыков — это позволяет достичь существенной экономии трудозатрат на разработку словарных баз при сохранении поисковых возможностей системы, повысить точность переводов для третьей пары языков, если в распоряжении исследователя имеются словари для первых двух пар языков, можно с учетом информации о тематике, хранящейся в каждом из словарей.

Следует учитывать, что лексикографическая работа даже с использованием возможностей ИТ остается работой творческой и не может быть полностью автоматизирована. В то же время существуют возможности подготовки массивов текстов для автоматизации лексикографического анализа. Идеальным источником материала являются корпусы параллельных текстов, построенные на основе материалов узкой предметной области (статей, монографий, материалов конференций и их переводов на другой язык).

Такой корпус должен быть выровнен по предложениям, что позволяет выявлять и анализировать термины и их переводы, оценивать стандартизованность и единство переводов, распространенность конкретных вариантов. Однако получение такого параллельного корпуса не всегда возможно. Одним из вариантов создания материала для последующего лексикографического анализа является формирование особых корпусов текстов, включающих параллельное представление исходных текстов, их машинных переводов и отредактированных переводов, согласованных с экспертами в конкретной области знаний. Важно отметить, что качество и потенциал такого корпуса в большой степени зависит от сотрудничества с экспертами при отборе исходного материала и редактировании переводов.

СПИСОК ЛИТЕРАТУРЫ

1. *Андрющенко В. М.* Вычислительная лексикография и автоматические словари // Вопр. языкознания. 1986. № 3. С. 42–53.

2. *Андрющенко В. М.* Концепция и архитектура машинного фонда русского языка. М.: Наука, 1989.

3. *Беляева Л. Н.* Лингвистические автоматы в современных информационных технологиях. СПб.: Изд-во РГПУ им. А. И. Герцена, 2001.

4. *Беляева Л. Н., Герд А. С., Убин И. И.* Автоматизация и лексикография // Прикладное языкознание: Учеб. / Отв. ред. А. С. Герд. СПб.: Изд-во С.-Петерб. ун-та, 1996. С. 318–333.

5. *Вигурский К. В., Пильщиков И. А.* Филология и современные информационные технологии // Изв. АН. Сер. лит. и яз. 2003. Т. 62. № 2. С. 9–16.

6. *Дубичинский В. В.* Искусство создания словарей: Конспекты по лексикографии. Харьков: Харьк. гос. политехн. ун-т, 1994.

7. *Дубичинский В. В.* Основные аспекты переводной лексикографии // Актуальные проблемы теоретической и прикладной лексикографии: Межвуз. сб. науч. тр. / Отв. ред. О. М. Карпова. Иваново: Юнона, 1997. С. 112–115.

8. *Кияк Т. Р.* Лингвистика профессиональных языков и терминоведение // Терминология и знание: Материалы I Международного симпозиума. М.: Институт русского языка им. В. В. Виноградова РАН, 2009. С. 21–27

9. *Лугинский Я. Н., Фези–Желинская М. С., Кабиров Ю. С.* Англо–русский словарь по электронике и электроэнергетике. М.: РУССО, 1995.

10. *Марчук Ю. Н.* Основы терминографии: Метод. пособие. М.: Изд-во Моск. ун-та, 1992.

11. Описание системы RU.DICT: Версия 2. 2003 (Электронный ресурс) <http://rudict.noolab.ru/about.phtml?2>

12. *Пиотровский Р. Г.* Лингвистический автомат (в исследовании и непрерывном обучении). СПб.: Изд-во РГПУ им. А. И. Герцена, 1999.

13. *Поминов А. В.* Некоторые вопросы построения многоязычных автоматических словарей (Электронный ресурс) // Труды Междунар. конф. «Диалог' 2003»: (сайт) / Ассоц. компьютер. лингвистики и интеллектуал. технологий. М., 2003 <http://www.dialog-21.ru/archive_article.asp?param=7023&y=2001&vol=6078>

14. *Послед Б. С.* Access 2002: Приложения баз данных: Лекции и упражнения. М.: ДиаСофт, 2002.

15. *Селегей В.* Электронные словари и компьютерная лексикография (Электронный ресурс) // Ассоциация переводчиков Lingvo: (сайт) <http://www.lingvoda.ru/transforum/articles/pdf/selegey_a1.pdf> 2005.

16. Towards Consolidation of European Terminology Resources. Experience Recommendations from Euro TermBank Project / Edited by: Signe Rirdance, Andrejs Vasiljevs. Riga: Tilde, 2006.

REFERENCES

1. *Andrjuschenko V. M.* Vychislitel'naja leksikografija i avtomaticheskie slovari // Vopr. jazykoznanija. 1986. № 3. S. 42–53.

2. *Andrjuschenko V. M.* Koncepcija i arhitektura mashinnogo fonda russkogo jazyka. M.: Nauka, 1989.

3. *Beljaeva L. N.* Lingvisticheskie avtomaty v sovremennyh informacionnyh tehnologijah. SPb.: Izd-vo RGPU im. A. I. Gercena, 2001.

4. *Beljaeva L. N., Gerd A. S., Ubin I. I.* Avtomatizacija i leksikografija // Prikladnoe jazykoznanie: Ucheb. / Otv. red. A. S. Gerd. SPb.: Izd-vo S.-Peterb. un-ta, 1996. S. 318–333.

5. *Vigurskij K. V., Pil'schikov I. A.* Filologija i sovremennye informacionnye tehnologii // Izv. AN. Ser. lit. i jaz. 2003. T. 62. № 2. S. 9–16.

6. *Dubichinskij V. V.* Iskusstvo sozdanija slovarej: Konspekty po leksikografii. Har'kov: Har'k. gos. politehn. un-t, 1994.

7. Dubichinskij V. V. Osnovnye aspekty perevodnoj leksikografii // Aktual'nye problemy teoreticheskoj i prikladnoj leksikografii: Mezhvuz. sb. nauch. tr. / Otv. red. O. M. Karpova. Ivanovo: Junona, 1997. S. 112–115.

8. *Kijak T. R.* Lingvistika professional'nyh jazykov i terminovedenie // Terminologija i znanie. Materialy I Mezhdunarodnogo simpoziuma. M.: Institut russkogo jazyka im. V. V. Vinogradova RAN, 2009. S. 2̃127.

9. *Luginskij Ja. N., Fezi̇-Zhelinskaja M. S., Kabirov Ju. S.* Anglo-russkij slovar' po elektronike i elektroenergetike. M.: RUSSO, 1995.

10. *Marchuk Ju. N.* Osnovy terminografii: Metod. posobie. M.: Izd-vo Mosk. un-ta, 1992.

11. Opisanie sistemy RU.DICT: Versija 2. 2003 (Jelektronnyj resurs) <http://rudict.noolab.ru/about.phtml?2>

12. *Piotrovskij R. G.* Lingvisticheskij avtomat (v issledovanii i nepreryvnom obuchenii). SPb.: Izd-vo RGPU im. A. I. Gercena, 1999.

13. *Pominov A. V.* Nekotorye voprosy postroenija mnogojazychnyh avtomaticheskih slovarej (Jelektronnyj resurs) // Trudy Mezhdunar. konf. «Dialog' 2003»: (sajt) / Assoc. komp'juter. lingvistiki i intellektual. tehnologij. M., 2003 <http://www.dialog-21.ru/archive_article.asp?param=7023&y=2001&vol=6078>

14. *Posled B. S.* Access 2002: Prilozhenija baz dannyh: Lekcii i uprazhnenija. M.: DiaSoft, 2002.

15. Selegej V. Jelektronnye slovari i komp'juternaja leksikografija (Jelektronnyj resurs) // Associacija perevodchikov Lingvo: (sajt) <http://www.lingvoda.ru/transforum/articles/pdf/selegey_a1.pdf> 2005.

16. Towards Consolidation of European Terminology Resources. Experience Recommendations from Euro TermBank Project / Edited by: Signe Rirdance, Andrejs Vasiljevs. Riga: Tilde, 2006.