КОМПЬЮТЕРНАЯ ЛЕКСИКОГРАФИЯ КАК ОБЪЕКТ ИЗУЧЕНИЯ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Девдариани Наталья Валерьевна
Курский государственный медицинский университет
г. Курск, Россия

Аннотация. В статье представлены основные концепции и методы компьютерной лексикографии в качестве обзора к другим, более специализированным исследованиям, в которых основное внимание уделяется разработке корпуса лексикона для использования в операционных системах, в частности в системах разговорного языка. Обсуждается представление материала, не затрагивая вопроса получения лексической информации.
Ключевые слова: лексикография, лингвистика, автоматизированные системы, лексика, комьютеризация.

Проблемы современной компьютерной лексикографии подвергаются анализу и разработке, начиная с конца XX века. Существует достаточно много исследований этой области научного знания, с которыми следует ознакомиться для получения дополнительной информации и последующего изучения. Так, Богураев и Бриско (1989) [6], Замполли, Чиннони и Питерс (1990) [11], Уилкс, Слейтор и Гатри (1996) [8] достаточно полно представляют устоявшиеся методы в этой области, и ориентированы на обучение, по конкретным аспектам компьютерной лексикографии и смежным областям. Также к их числу можно отнести Батлера (1992), Гарсайда, Лича и Макинери (1997), Хандке, Майерса и Томаса (1995) [11].

Во многих проектах лексикографических исследований и разработок, результаты которых представлены в исследованиях, были предложены методы автоматического извлечения информации из текстовых массивов и из существующих машиночитаемых словарей. Традиционный взгляд на лексикографию можно найти в работах Ландау (1989), Томащика и Левандовской-Томащик (1990) и других [11].

Существует также множество исследовательских и опытно-конструкторских проектов, связанных как с новыми методами компьютерной лексикографии, так и с созданием лексикографических источников, как на основе корпусов, так и на основе существующих машиночитаемых словарей [1].

Исследователям, заинтересованным в получении практического опыта использования базы текстовых корпусов, рекомендуется извлечь соответствующие лексикографические термины (включая программные средства, ресурсы лексической базы данных, определения технических терминов и т.д.) в числе уже представленных в сети Интернет [5].

Лексикография – это раздел прикладной лингвистики, занимающийся разработкой и конструированием лексики для практического использования. Lexica может варьироваться от бумажной lexica или энциклопедии, предназначенной для использования человеком и хранения на полках, до

электронной lexica, используемой в различных системах обработки человеческого языка, от портативных баз данных Word до текстовых процессоров и программного обеспечения для обратного чтения (путем синтеза речи в системах преобразования текста в речь) и диктовки (с помощью автоматической речи системы распознавания) [3].

С появлением компьютеров лексикографические проекты стали развиваться и систематизироваться более быстрыми темпами. Однако, согласно закону Паркинсона, корпус лексики также увеличился в размерах, а соответственно и разработка и составление достаточно объемного словаря, насчитывающего более десятков или сотен тысяч слов, является серьезной задачей, требующей многих лет работы над спецификацией, дизайном, сбором лексических данных, структурированием информации, ориентированной на пользователя, с последующим форматированием презентации материала.

С другой стороны, лексикология – это раздел дескриптивной лингвистики, занимающийся лингвистической теорией и методологией описания лексической информации, часто фокусирующийся, в частности, на вопросах значения и смысла.

Теория лексикона, в отличии, как от лексикологии, так и от лексикографии, направлена на изучение универсальных, в частности формальных свойств лексики, с точки зрения теоретической лингвистики, языков представления общих знаний в искусственном интеллекте, построения лексикона. Алгоритмы доступа в компьютерной лингвистике или когнитивные условия, влияющие на лексические способности человека в эмпирической психолингвистике, представлены, в частности, в работах Байена, Шредера и Спроута [7].

Вместе с тем, какими бы высокими ни были цели в области описания компьютерной базы данных, и какой бы обоснованной ни была теория, лежащая в основе программного обеспечения, лексикография в первую очередь ориентирована на решение задач, а lexica в первую очередь предназначена для конкретных целей. Следовательно, основным вопросом, лежащим в основе лексикографического проекта, как и проекта по разработке программного обеспечения, является спецификация требований, т.е. формулировка практических целей, которые впоследствии будут использоваться для оценки результатов проекта.

Лексикон для использования в программе автоматического поиска и требования к нему, будут отличаться от требований к бумажному словарю – ему не требуется никакой информации о произношении, тем более статистики, и, по сути, ему может потребоваться не более чем список словоформ в их стандартной орфографии, в то время как для других типов лексики требуется множество различных типов лексической информации.

Разработчику системы разговорного языка, например, потребуется лексикон как часть программного обеспечения (lingware, т.е. машиночитаемые

лингвистические данные и модели, а также инструменты для создания лексикона и доступа к нему), содержащего информацию в основном о следующем:

– статистические данные о связи между словами и акустическими признаками,

– воплощенные в скрытой марковской модели (разновидности вероятностного конечного автомата), в сочетании со статистической информацией о количестве слов в последовательности,

– в сочетании с другими видами лексической информации, в зависимости от приложения, например, в качестве интерфейса для программы редактирования текста или система запроса к базе данных.

Основанный на спецификации требований, дизайн автоматизированного словаря охватывает следующий важный набор условий для его построения:

– структуру лексикона с точки зрения отношений между лексическими элементами (макроструктура), проявляющаяся, например, в различии между семасиологической, ориентированной на форму, часто основанной на словах лексикой с орфографической альфа-бетической сортировкой;

– лексику произношения или рифмованной лексикой, и ономасиологическими иерархическими таксономиями, которые характерны, с семантической точки зрения, для связанного семейства слов;

– тип лексической информации для каждой записи (микроструктуры), от поверхностных деталей произношения и орфографии до деталей структуры слова в целом;

– простые или сложные слова, синтаксические свойства лексических единиц, их значение и прагматика их использования в реальных контекстах и т.д.

Методы, используемые в лексикографии, неизменно основаны на теоретико-знаковых соображениях. Так, традиционная лексикографическая знаковая модель Соссюра, проводит различие между формой слова, с одной стороны, и его значением, интерпретируемым как понятие, с другой. Эта простая знаковая модель обеспечивает основу для фундаментального различия между типами словарей, основанного на процедурном критерии вывода стратегии поиска:

– семасиологический словарь: традиционный тип словаря, в котором ключом поиска является словоформа (как правило, орфографическая), а требуемая информация – семантическая.

– ономасиологический словарь: словарь типа тезауруса, в котором ключом к поиску является понятие (фактически слово или семейство слов, представляющее понятие или область понятий), а требуемой информацией является форма слова / «название», указанное в техническом термине.

По мере развития области генеративной лингвистики лексика приобретает все более важную роль в описании как специфических, так и регулярных свойств языка. Лексикону, который всегда рассматривался как естественная среда, в первые годы трансформационной грамматики уделялось относительно мало

внимания. Затем в 1970 году Хомский предложил (Chomsky, 1970), что сходство в структуре девербальных именных фраз и предложений может быть выражено в терминах лексической связи между глаголом и его производными [12].

Акендофф (1975) описал дальнейшие лексические закономерности как в морфологии, так и в семантике, а Бреснан (1976, 1982) стал пионером в разработке синтаксической структуры (лексической функциональной грамматики), в рамках которой основные грамматические явления, такие как пассивизация, могут быть объяснены в рамках лексикона. Параллельное направление работы Газдара (1981) под названием «Обобщенная грамматика структуры» было направлено на создание нетрансформационной синтаксической структуры путем использования мета-форм вместо контекстно-свободной грамматики [9].

Фундаментальной концепцией лингвистической репрезентации в HPSG (Грамматика структуры фраз, управляемая головой), является высоко лексикализованная грамматика, основанную на знаках. Знак в HPSG – это набор различных видов свойств или информации, включая фонологические, синтаксические, семантические и контекстуальные ограничения, представленные в виде типизированной матрицы значений атрибутов (AVM), где каждому значению атрибута также присваивается тип, возможно, с дополнительными ограничениями. Поскольку такие ограничения на тип могут вводить допустимые или подходящие объекты для этого типа называются условиями соответствия, или, в более общем смысле, объявлениями объектов. Слова представлены в виде знаков, и фразы также являются знаками, при этом многие, но не все из них имеют одинаковые атрибуты или функции, общие для слов и фраз.

Таким образом, область компьютерной лексикографии весьма обширна, например, интеллектуальный анализ текста для построения лексики на основе корпусов, создание lexica для систем обработки естественного языка (NLP), автоматическое получение синтаксической или семантической информации из текстов, повторное использование машиночитаемых словарей для новой лексики, машиночитаемых словарей (MRD) в целом или компьютерного создания lexica для использования человеком и т.д., с чем ученым как лингвистам, так и программистам, предстоит работать в тесном взаимодействии для создания массивных корпусов многочисленных языков, в том числе и русского языка.

Литература:

1. Андрющенко В. М. Вычислительная лексикография и автоматические словари // Вопр. языкознания. 1986. № 3. С. 42–53.
2. Беляева Л. Н. Лингвистические автоматы в современных информационных технологиях. СПб.: Изд-во РГПУ им. А. И. Герцена, 2001.
3. 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.

4. 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.
5. Рубцова, Е. В. Возможности использования компьютерных технологий в процессе изучения иностранного языка в условиях самоизоляции / Е. В. Рубцова, Н. В. Девдариани // Балтийский гуманитарный журнал. – 2020. – Т. 9, № 4(33). – С. 134-137. – DOI 10.26140/bgz3-2020-0904-0033. – EDN MHOBCE.
6. Boguraev, Bran & Ted Briscoe, eds. (1989). Computational Lexicography for Natural Language Processing. London: Longman
7. Butler, Christopher S., ed. (1992). Computers and Written Texts. Oxford: Blackwell.
8. Coward, David F. & Charles E. Grimes (1995). Making Dictionaries: A guide to lexicography and the Multi-Dictionary Formatter. Waxhaw, North Carolina: Summer Institute of Linguistics.
9. Gazdar, G.: 1981, 'Unbounded Dependencies and Coordinate Structure'. Linguistic Inquiry 12, 155-84.
10. Pollard, Carl & Ivan A. Sag (1987). Information-based Syntax and Semantics Volume 1: Fundamentals. Stanford: CSLI.
10. Wilks, Yorick, S1ator, Brian & Guthrie, Louise (1996). Electric Words-Dictionaries, Computers, and Meanings. Cambridge, Mass.: MIT Press.
11. Zampolli, Antonio, L. Cignoni & C. Peters, eds. (1990). Computational Lexicology and Lexicography: Special issue of Linguistica Computazionale dedicated to Bernard Quemada. 2 Vols. Pisa, Giardine.
12. Britannica, The Editors of Encyclopaedia. «Syntactic Structures». Encyclopedia Britannica, 21 Apr. 2023, https://www.britannica.com/topic/Syntactic-Structures. Accessed 19 September 2024.

COMPUTER LEXICOGRAPHY AS AN OBJECT OF STUDY OF APPLIED LINGUISTICS

Devdariani Natalia Valeryevna
Kursk State Medical University
Kursk, Russia

Abstract. The article presents the basic concepts and methods of computer lexicography as a practical guide to other, more specialized studies, which focus on the development of a lexicon for use in operating systems, in particular in spoken language systems. The presentation of the material is discussed, without touching on the issue of obtaining lexical information.
Keywords: lexicography, linguistics, automated systems, vocabulary, computerization.