УДК 81.322

Маник Светлана Андреевна
доктор филологических наук,
директор Института гуманитарных наук
ФГБОУ ВО «Ивановский
государственный университет»
г. Иваново, Россия

Manik Svetlana
Doctor of Philology
Director of Institute
of Humanities Ivanovo State University
Ivanovo, Russia
maniksa@ivanovo.ac.ru

ЦИФРОВАЯ ЛЕКСИКОГРАФИЯ: К ВОПРОСУ О КОНЦЕ ИЛИ ПЕРСПЕКТИВАХ

Статья посвящена рассмотрению перспектив цифровой лексикографии в новом тысячелетии. Описывается само понятие «цифровая лексикография», выделяются основные этапы автоматизации труда лексикографов. В заключении систематизируются ее преимущества и ограничения в эпоху искусственного интеллекта.

К л ю ч е в ы е с л о в а: *цифровая лексикография; корпусная лексикография; автома-тизация; корпус; искусственный интеллект.*

DIGITAL LEXICOGRAPHY: SPEAKING OF THE END OR PERSPECTIVES

The article is devoted to the prospects of digital lexicography in the new millennium. The very concept of "digital lexicography" is described, the main stages of lexicographers' labor automation are highlighted. In conclusion, the advantages and limitations of digital lexicography in artificial intelligence era are systematized.

K e y w o r d s: *digital lexicography; corpus lexicography; automation; corpus; artificial intelligence.*

Автоматизация в современную эпоху информационных технологий и искусственного интеллекта все активнее входит в лексикографию, совершив значительный прорыв с момента зарождения сначала компьютерной лингвистики (Computational Linguistics) в 50-х годах (эксперименты по машинному переводу в Джорджтаунском университете) и затем корпусной лексикографии (Corpus Lexicography) в 60-х годах прошлого столетия [1; 2]. Так, целесообразно отметить разработку и постепенное внедрение усовершенствованных методов и инструментов для автоматического создания списков слов и ключевых терминов, вычленения примеров предложений или сочетаемостных возможностей из больших корпусов. Это в значительной степени облегчает каталогизацию языкового материала, которая ранее отнимала много времени и ресурсов. Автоматизация обработки естественного языка в лексикографии, как правило, применяется для составления словарной статьи и справочника в целом, тем самым упрощая и ускоряя процесс создания или обновления словаря. Сейчас искусственный интеллект довольно быстро справляется со многими задачами по работе с большими данными. Можно констатировать, что некоторые ключевые лексикографические задачи начинают в значительной степени передаваться от человека к машине. В современном научном дискурсе довольно активно звучит следующая тема: автоматизация не только экономит усилия, но и часто приводит к более надежному и систематическому описанию языка, вытесняя тем самым человека. Представляется важным в настоящем исследовании рассмотреть основные этапы цифровизации лексикографии, выделить ее преимущества и ограничения.

Понятие «цифровой» ‘*digital*’ подразумевает несколько составляющих согласно толковым словарям русского и английского языков. Например, в Научно-техническом энциклопедическом словаре под «цифровым» понимают «термин, описывающий информацию, выраженную при помощи чисел. ДАННЫЕ, такие, как слова, изображения, звуки, представляются в виде набора цифр (1 и 0) в ДВОИЧНОЙ СИСТЕМЕ, которая используется в КОМПЬЮТЕРАХ» [3]. В английских дефинициях также речь идет про систему цифровых кодов (*using a system that can be used by a computer and other electronic equipment, in which information is sent and received in electronic form as a series of the numbers 1 and 0*) [4]. Вместе с тем, под «цифровым» часто понимают всё, что связано с компьютером, его использованием, а также с Интернетом (*using or relating to computers and the internet; relating*

to computer technology, especially the internet) [5] или всё доступное в электронном формате (*available in electronic form; readable and manipulable by computer*) [6]. Следовательно, в понятие «цифровая лексикография» возможно вкладывать все электронные и онлайн – лексикографические продукты и инструменты по обработке значительных объемов лингвистических данных в электронном формате.

Интересно отметить, что один из самых больших онлайн-корпусов английского языка, ежедневно обновляемых, NOW Corpus (News on the Web) регистрирует следующие наиболее распространенные коллокации с прилагательным «digital»: *digital articles* ‘цифровые статьи’, *transformation* ‘трансформации’, *content* ‘содержание’, *subscription* ‘подписка’, *access* ‘доступ’, *media* ‘медиа’, *economy* ‘экономика’, *marketing* ‘маркетинг’, *news* ‘новости’, *advertising* ‘реклама’, *assets* ‘доходы/активы’, *currency* ‘валюта’, *platforms* ‘платформы’, *versions* ‘версии’, *world* ‘мир’, *age* ‘эпоха’, *technology* ‘технология’, *services* ‘сервисы’, *editions* ‘издания’, *payments* ‘платежи’ и т. п. [7]. К сожалению, в корпусе нет примеров *digital lexicography* ‘цифровая лексикография’, однако есть *digital dictionary* ‘цифровой словарь’, что также интересно, поскольку лексикография понимается как наука об изучении и составлении словарей. Безусловно, словарь прошел определенный путь от рукописного к печатному изданию, а затем от оцифрованного к индексированному изданию с возможностями поиска, обновления и дополнения информации и интеракции с читателями. Далее важно кратко описать наиболее значимые шаги к цифровизации в лексикографической науке.

Лоуренс Урданг (Laurence Urdang), редактор словаря английского языка Random House Dictionary of the English Language [8], был одним из первых, кто увидел потенциал компьютеров для облегчения и рационализации процесса сбора, хранения и обработки аутентичных примеров для словаря, следуя традициям С. Джонсона. Именно с этого момента идея словаря как некой базы данных, в которой каждый из компонентов словарной статьи имеет свою собственную ячейку (контейнер), прочно укоренилась. Одним из важных преимуществ такого подхода было то, что перекрестные ссылки можно было проверять более систематически: компьютер генерировал отчет о несовпадении перекрестных ссылок, но ошибки затем устранялись вручную. Таким образом, рутинная задача была частично передана компьютерам. Затем в электронный словарь добавили опцию по ограничению (фильтрации) вокабуляра: как известно, в учебных словарях в дефинициях важно использовать только те слова, которые включены в корпус издания согласно регистрируемому уровню сложности. Поэтому следующим шагом в автоматизации труда лексикографа стал автоматический контроль за языком определений словарной статьи, используя ограниченный лексикон, аналогичные методы можно было также применять для того, чтобы не допустить попадания запрещенных (табуированных) слов. В дальнейшем первое издание словаря современного английского языка Longman Dictionary of Contemporary English [9] включало некоторые категории данных (в частности, сложную систему

семантического кодирования), которые никогда не должны были появиться в самом словаре. В подобных проектах процесс первоначального составления текста оставался практически неизменным, но последующее редактирование, как правило, осуществлялось на страницах, созданных линейными принтерами, а изменения вносились в базу данных техническими специалистами.

Как отмечалось ранее, появление компьютерной лингвистики связывают с первыми экспериментами в области машинного перевода. В 1979 году в Японии был разработан один из первых цифровых словарей – The Pocket Electric Translating Machine. Физически он был очень похож на современные цифровые словари и получил высокую оценку за скорость и точность перевода. В 1980-е и 1990-е годы рынок цифровых словарей расширялся, что оказало значительное влияние на рынок бумажных словарей.

Выделяя важные в истории автоматизации лексикографии вехи, целесообразно отметить проект COBUILD, в котором компьютеры с самого начала занимали центральное место. В 1981 г. словарь был составлен на основе более обширной базы данных, лексикографы создавали словарные статьи, используя массив цветных листков для записи информации разных типов [10]. Каждый лингвистический факт, выявленный лексикографами, подкреплялся эмпирическими данными в виде корпусных отрывков. Корпусный проект COBUILD 1980–85-х гг. был разработан для целей учебной лексикографии и, таким образом, исключал некоторые виды текстов (исторические, технические, региональные, поэзию, детский язык и т. п.), но делал акцент на наиболее востребованные и популярные на данный отрезок времени тексты художественной литературы, произведения, рекомендованные при изучении английского языка как иностранного, контексты устной коммуникации. Впервые, как пишет Кришнамурти, с нуля было создано масштабное описание английского языка, отражающее реальное употребление, проиллюстрированное в большом для того времени и разнообразном корпусе текстов [10]. Систематическое применение этой корпусной методологии представляет собой смену парадигмы в лексикографии. То, что было революционным в 1981 году, теперь, спустя поколение, является нормой для любого серьезного лексикографического труда. Однако с точки зрения баланса между человеком и машиной достижения COBUILD были относительно скромными. Создание корпусов по-прежнему было трудоемким делом. Поскольку использование сканеров дополняло работу с клавиатурой, сбор данных был несколько менее трудоемким, чем методы, доступные Г. Кучере двумя десятилетиями ранее, когда он с помощью перфокарт превратил миллион слов в Брауновский корпус в 60-е гг. (Brown Corpus) [11].

Брауновский корпус в качестве нового системного и репрезентативного на тот момент источника аутентичных данных по американскому варианту английского языка мог быть использован для автоматического составления компьютером списков контекстов с ключевым словом (*key-word-in-context list*), что значительно облегчало задачу лексикографу по поиску примеров употребления. Также в этом корпусе было еще одно преимущество: часте-

речная разметка (*part-of-speech tagging*), что давало возможность исследователю после некоторых компьютерных доработок искать не только слова, но и синтаксические паттерны. Важно напомнить, что почти одновременно в Великобритании появился аналогичный по объему и принципам отбора корпус британского варианта английского The Lancaster-Oslo-Bergen Corpus.

Таким образом, к концу 1990-х годов использование компьютеров для анализа данных и составления словарей стало стандартной практикой (по крайней мере, для английского языка). Вместе с тем, создание корпусов оставалось ресурсоемким делом. Анализ корпусов стал проще и быстрее благодаря аннотированию (токенизации, лемматизации и тегированию частей речи) и добавлению новых опций, например, усовершенствованных систем корпусных запросов. Однако размеры корпусов увеличивались с миллиона до миллиардов, следовательно, лексикографы начали работать с гораздо большим количеством данных.

Корпусная революция трансформировала традицию описания языкового материала. Если сначала лексикографы, опираясь на классические тексты художественной литературы или периодических изданий, предписывали нормативное употребление, регистрировали образец письменных канонов, то начиная с конца XX века словарь начал восприниматься как зеркало языка во всем его многообразии. Для пользователя подобные перемены были мало уловимы, поскольку постепенное добавление в корпус толкового справочника жаргонизмов, диалектизмов, просторечных выражений, ряда терминов и т. п. не меняло образа лексикографического справочника как модели правильного словоупотребления. Однако для лексикографов технический прорыв значительно улучшил основу для описания языка, технологию и инструменты автоматизированной обработки текстов, позволил создавать более качественные словари за меньший временной отрезок. До этого исследователи вручную изучали тысячи строк конкорданса с описываемым словом, вычленяли оттенки значения, наиболее продуктивные шаблоны и сочетаемостные возможности лексической единицы. В настоящее время данную работу выполняет искусственный интеллект, а роль редактора заключается в проверке качества автоматической выборки, в интерпретации полученных данных, анализе и обобщении результатов.

Вместе с тем, в 1990-е для лексикографов стало нормой работать на собственных компьютерах, а не зависеть от технического персонала для ввода данных, и было создано первое поколение специализированных систем составления словарей (dictionary-writing system – DWS) с целью облегчить работу по составлению словарных статей [1]. Помимо поддержки лексикографов в их работе они обеспечивают надежное хранение данных и эффективную систему управления всего проекта, который может охватывать все этапы – от разработки концепции до конечного продукта. Данное программное обеспечение может включать редактор, базу данных, веб-интерфейс и различные инструменты управления. Как правило, оно работает на основе словарной грамматики, которая определяет структуру словаря. В прошлом

был разработан ряд программных пакетов, наиболее известными из которых являются такие коммерческие продукты, как IDM DPS, TLex, ABBY Lingvo Content и iLEX, предлагающие готовые решения для производства большого количества словарных продуктов.

Выделяя вехи в автоматизации лексикографической работы, следует также отметить другое ПО. В то время как система составления словарей, будучи специализированной системой, а не универсальным редактором, в основном поддерживает составление словарей и редактирование словарных статей, существует другой тип программного обеспечения, которое помогает лексикографам в составлении словарей, а именно системы корпусных запросов (corpus-query systems). Они часто используются для анализа и отбора данных. В последние годы системы корпусных запросов стали стандартным инструментом в лексикографической работе [12]. Корпусно-запросная система может использоваться в дополнение к системе составления словарей или быть ее неотъемлемой частью.

Начало нового тысячелетия ознаменовалось бурным развитием интернета. Следовательно, одним из наиболее ярких событий XXI века в цифровой лексикографии стал «веб-корпус» (web corpus). Теперь корпуса в большей степени создаются на основе текстов из Интернета. Эти тексты очень разнообразны. С появлением новых технологий стало возможно создавать разножанровые по регистрам корпуса, насчитывающие миллиарды слов. Такие программные инструменты, как WebBootCat, обеспечивают единую операцию, в ходе которой тексты отбираются в соответствии с заданными пользователем параметрами, «очищаются» и лингвистически аннотируются. Время создания большого лексикографического корпуса сократилось с нескольких лет до нескольких недель, а небольшого корпуса в специализированной области – с нескольких месяцев до нескольких минут. Тексты в Интернете, по определению, уже находятся в цифровой форме. Как основной результат: резко сокращаются как человеческие усилия, затрачиваемые на создание корпусов, так и расходы на данную операцию.

2020-е годы в лексикографии детерминируются превалирующим влиянием нейронных сетей и искусственного интеллекта при составлении словаря. Так, Г. М. де Шривер своим выступлением с презентацией возможностей ChatGPT при составлении словаря вызвал бурную дискуссию о будущем профессии лексикографа и о полном вытеснении ИИ человека [13]. Известны работы, посвященные экспериментам с нейронными сетями по моделированию словаря, словарной статьи и других разделов словаря. Все они доказывают, что несмотря на преимущества внедрения технологий полное вытеснение человека невозможно. Возможно констатировать революционную трансформацию лексикографии XXI века от доперсональных компьютеров с данными на карточках и бумажных лентах до сегодняшнего инновационного программного обеспечения, размещенного не на жестком диске компьютера, а в облаке, которое самостоятельно готовит проекты словарных статей, основанных на массивном корпусе данных, для рассмотрения лекси-

кографами и/или для пользователей, которые могут внести свой вклад через краудсорсинг. Развитие идет в сторону персонализации представления запрашиваемой информации в цифровом формате на основе обработки и анализа ИИ внушительного объема лингвистического материала.

Итак, для многих современных пользователей цифровой словарь – это электронный онлайн-словарь или машиночитаемый словарь (Machine Readable Dictionary), который используется для проверки произношения, орфографии и грамматики, значения слова, получения сочетаемостной информации или примера употребления. В настоящее время довольно популярны разные лексикографические справочные ресурсы онлайн: словарные генераторы, регистрирующие дефиниции из разных словарей (например, Dictionary.com; Wordreference.com; Vocabulary.com; the Free Dictionary; Wordnik и т. п.); Lexipedia, которая в основном работает как лексическая энциклопедия; цифровой лексический профиль, который основан на обменном интерфейсе между словарем, тезаурусом и энциклопедией (например, Text Inspector); WordNet, так называемая лексическая база данных, онтология, которая строится на семантических отношениях между словами и определяет смысловые вариации (к данной группе также можно отнести VerbNet, FrameNet, система ЛЕКСИКОГРАФ; цифровая лексическая база данных, хранящая отсортированный по алфавиту список слов и лексических единиц; банк терминов, содержащий только научные и технические термины одного языка с указанием области применения (например, Term Bank, UNTERM, TERMIUM Plus и т. п.); цифровой словник или глоссарий, в котором собраны только слова определенной предметной области и их семантические классификации. Целесообразно подчеркнуть, что для некоторых людей цифровой словарь может означать встроенные справочники, которые используются в компьютерных программах обработки текстов или программах-переводчиках.

Как отмечает Н. С. Дэш, цифровой словарь представляет собой «управляемое компьютером и функционально автоматизированное лингвистическое справочное устройство, специально разработанное для удовлетворения лексикографических потребностей целевой аудитории в веб-интерфейсе для изучения языка» [14].

Таким образом, справедливо отметить, что в области лексикографии под цифровым словарем в настоящее время понимают издание, с одной стороны, разработанное и созданное в цифровом формате с использованием лексического фонда языка, доступного в корпусах, а, с другой стороны, имеющее множество прикладных преимуществ перед своим печатным аналогом для обслуживания речевого сообщества, в частности, в осуществлении поиска и представлении вербальной, аудиовизуальной, анимационной информации. По сути, цифровой словарь является неким инструментом, способным выполнять лексикографические, тезаурусные и энциклопедические функции вместе через расширенный интерфейс, позволяющий использовать, объяснять и демонстрировать понятия, заключенные в словах и других лексиче-

ских единицах, употребляемых в языке. С развитием мобильных приложений словарь будет постепенно становиться все более мощным, расширять свои функции, такие, как аудио-видео и анимация, чтобы служить людям не только как отдельное цифровое устройство, помеченное академическими портфелями, но и как вездесущее устройство, связанное с порталом электронного правительства и встроенное в смартфоны в качестве лексикографического приложения.

ЛИТЕРАТУРА

1. Rundell M., Kilgarrif A. Automating the creation of dictionaries: where will it all end // *A Taste for Corpora. In Honour of Sylviane Granger. Amsterdam*, 2011. P. 257–282.
2. Rundell M., Jakubíček M., Kovář V. Technology and English Dictionaries // *The Cambridge Companion to English Dictionaries*. Cambridge, 2020. P. 18–30.
3. Научно-технический энциклопедический словарь [Электронный ресурс]. URL: https://rus-scientific-technical.slovaronline.com/ (дата обращения 30.06.2024).
4. Cambridge English Dictionary [Electronic resource]. URL: https://dictionary.cambridge.org/dictionary/english/digital (accessed: 30.06.2024).
5. Dictionary.com [Electronic resource]. URL: https://www.dictionary.com/browse/digital (accessed: 30.06.2024).
6. Oxford English Dictionary [Electronic resource]. URL: https://www.oed.com/dictionary/digital_n?tl=true (accessed: 30.06.2024).
7. News on the Web Corpus (NOW) [Electronic resource]. URL: https://www.english-corpora.org/now/ (accessed: 01.07.2024).
8. Stein J., Urdang L. Random House Dictionary of the English Language. New York : Random House, 1966. 2059 p.
9. Procter P. Longman Dictionary of Contemporary English. Harlow : Longman. 1978. 1303 p.
10. Krishnamurthy R. The Process of Compilation // Looking Up: An Account of the COBUILD Project in Lexical Computing. London, 1987. P. 62–85.
11. Brown Corpus [Electronic resource]. URL: https://www.sketchengine.eu/brown-corpus/ (accessed: 30.06.2024).
12. Atkins S., Rundell M. The Oxford Guide to Practical Lexicography. Oxford : Oxford University Press, 2008. 552 p.
13. An overview of digital lexicography and directions for its future: an interview with Gilles-Maurice de Schryver [Electronic resource]. URL: https://euralex.org/wp-content/uploads/2019/12/de-Schryver-et-al.-2019-An-overview-of-Digital-Lexicography.pdf (accessed: 30.06.2024).
14. Dash N. S. Digital Dictionary: A Phoenix in Lexicographic Metamorphosis [Electronic resource]. URL: https://www.researchgate.net/publication/323445151_Digital_Dictionary_A_Phoenix_in_Lexicographic_Metamorphosis (accessed: 08.07.2024).