ФИЛОЛОГИЯ

УДК 81

DOI 10.21779/2542-0313-2022-37-4-81–86

И.В. Балканов

Корпусная лингвистика: новые возможности лексикографии

МГИМО(У) МИД России; Россия, 119454, г. Москва, пр. Вернадского, 76; i-balkanov@mail.ru;

Военный университет им. князя Александра Невского; Россия, 123001, г. Москва, ул. Б. Садовая, 14; i-balkanov@mail.ru

В статье рассматриваются перспективы развития теоретической и практической электронной лексикографии в контексте достижений прикладной (корпусной) лингвистики, в диахронии исследуются возможности корпусов текстов при создании толковых словарей английского языка, анализируется влияние объема корпуса текстов на работу лексикографа.

Развитие электронных ресурсов корпусной лингвистики привело к изменению роли лексикографа в процессе составления словаря. Сегодня вместо отбора данных в корпусе текстов для последующего их включения в словарь составитель сосредоточен на просмотре и редактировании подборки упорядоченной лексикографической информации, сформированной компьютером из корпуса текстов с помощью специальных инструментов.

В качестве перспективного направления дальнейших исследований видится поиск приемов и способов взаимодействия лексикографа с инструментарием корпуса текстов и разработка стратегий составления толкового и переводного словарей в цифровую эпоху.

Ключевые слова: *электронная лексикография, корпус текстов, корпусная лингвистика, лемматизация, семантизация*.

Постановка проблемы

В конце XX в. в практической лексикографии наметилась тенденция перехода от традиционных, или бумажных, словарей к словарям электронным. Так, в 2010 г. стало известно, что очередное издание Oxford English Dictionary (OED) появится только в электронном виде. В то же время в Оксфорде задумались о судьбе другого своего словаря – Advanced Learner's Dictionary, целевую аудиторию которого составляют пользователи 17–24 лет, так называемые «цифровые аборигены», которые родились в эпоху информационных технологий и с детства привыкли получать и обрабатывать данные в Интернете [1].

Издатели и составители словарей столкнулись с проблемой. С одной стороны, онлайн-словарь способен объединить на своей платформе словари различных типов (толковые, переводные, лексической сочетаемости), предложить читателю корпус, не ограниченный объемом печатной книги, и удовлетворить запросы различных категорий пользователей. Составление такого словаря частично или полностью выполняется с помощью специального программного обеспечения на базе корпусов текстов, что способствует увеличению объема словника, поиску связей между леммами путем построения предметно-ориентированных онтологий и ассоциативных карт, интеграции онлайн-словаря с другими электронными ресурсами переводчика. С другой стороны, «цифровые аборигены» не привыкли платить за доступ к информации, а значит издателям словарей предстоит выработать новую эффективную бизнес-модель.

В условиях рыночной экономики цель любого издательства – создать словарь, способный удовлетворить запросы целевой аудитории и принести прибыль. Для этого необходимо понимать возможности информационных технологий в лингвистике в целом, и в электронной лексикографии в частности. Такие технологии обеспечивают переход от статичной печатной книги к динамично обновляющемуся веб-ресурсу, что вынуждает ученых переосмысливать базовые положения теоретической лексикографии, меняет роль пользователя и лексикографа в процессе создания словаря.

Одна из основных задач составителя – поиск и извлечение из языкового материала лексических единиц для их последующего описания (создание толкового словаря) и/или подбора переводных эквивалентов (создание двуязычного словаря). До 1970-х гг. эти операции делались вручную, затем с помощью аналоговых вычислительных машин, а сегодня – через корпуса текстов и программы работы с ними. Иллюстрирование полученных словарных статей примерами исходного языка и/или языка перевода также выполняется с помощью корпуса текстов и является важным этапом в создании словаря.

Таким образом, цель нашего исследования – изучение истории взаимоотношений корпусной лингвистики и электронной лексикографии, корпуса текстов и словаря в диахронии.

Материал исследования – корпусы текстов и толковые словари английского языка, доминирующая роль которого в научной и деловой средах в эпоху глобализации и создания цифрового общества не вызывает сомнения. Большинство технологий электронной лексикографии создаются для английского языка и уже потом попадают в словари других языков.

Влияние объема корпуса текстов на процесс составления словаря

Сбор данных для корпуса текста изначально требовал значительных временных затрат. Лексикографы не верили в перспективы корпусной лингвистики и возможности применения ее достижений в теории и практике составления словарей: в начале 1980-х гг. лишь некоторые печатные тексты были переведены в цифровую форму, не существовало единого способа обработки, кодирования и декодирования лексической информации [2]. Программисты могли поместить в корпусы текстов не более пяти миллионов слов, а высокая цена создания корпуса делала данный инструмент лексикографа доступным только для крупных издательств: первые корпусы – Birmingham Collection of English Texts (1987) и The Longman-Lancaster Corpus (1993) – создавались для получения конкурентных преимуществ на рынке востребованных у читателей печатных учебных словарей английского языка.

Со временем объем корпусов текстов увеличивался: первый корпус издательства HarperCollins, составленный в середине 1980-х гг., насчитывал около 20 млн слов, что не позволяло лексикографам делать объективных выводов о составе английского языка [3]. Британский национальный корпус на 100 миллионов слов, собранный в начале 1990-х гг., стал настоящим прорывом. Однако только в начале XXI в. специалисты в области прикладной лингвистики научились при относительно небольших затратах собирать, кодировать, аннотировать и хранить корпусы, насчитывающие миллиарды слов. Этому есть несколько объяснений. Во-первых, вычислительная мощность обычных компьютеров достигла уровня, позволяющего анализировать массивные базы данных. Во-вторых, программное обеспечение для автоматического кодирования лингвистических данных (например, программы-лемматизаторы, приводящие словоформы к лемме) стало быстрым и надежным. В-третьих, повсеместное распространение Интернета сделало объем текстов, существующих в цифровой форме, практически бесконечным.

В англоязычных странах лексикографы регулярно работают с корпусами, состоящими из двух и более миллиардов слов, а современный Collins Corpus насчитывает более четырех с половиной миллиардов слов [3]. Другие языки не отстают: например, в базе данных Sketch Engine, предназначенной для создания, анализа и управления корпусами текстов, хранятся корпусы итальянского и немецкого языков объемом в два миллиарда слов [4], а национальный корпус русского языка сегодня насчитывает около полутора миллиардов слов.

Что же дает объем корпуса лексикографу? В исследовании по корпусной лингвистике английского языка утверждается, что среднечастотные слова английского языка в первых корпусах текстов, рассчитанных на несколько миллионов слов, получали 10–20 примеров употребления [5]. В качестве примера приводится слово *seal*: как глагол оно находится примерно на 4300 месте по распространенности в Британском национальном корпусе, а как имя существительное занимает 4400 место. В корпусе на 7,3 млн слов, который составил основу первого словаря COBUILD, всего 23 совпадения для всех форм этого слова: *seal, seals, sealed, sealing*. В упомянутом выше первом корпусе HarperCollins менее 15 совпадений. В корпусе на 1,7 млрд слов, созданном в 2008–2010 гг. на базе программного обеспечения Sketch Engine для лексикографического проекта DANTE (Database of ANalysed Texts of English), исследователь отмечает уже 42 138 совпадений для слова *seal* (примерно половина из них – имена существительные), а в Британском национальном корпусе мы находим уже более 100 000 совпадений, в том числе многочисленные модели употребления этой лексемы в речи [6]. Например, глагол *seal* часто сочетается с частицами *off, up, on* и *in*, образуя с ними устойчивые фразы и фразовые глаголы. Как правило, он сочетается с именами существительными *envelope, victory, meat* и *leak*, наречиями *tightly, hermetically, securely*, образует предложные фразы в сочетании, среди прочих, с предлогами *into, against* и *with*. Каждое из этих сочетаний дает лексикографу новые возможности семантизации и филиации леммы *seal*, позволяет отобрать наиболее частотные примеры ее употребления. В результате использования корпуса онлайн-словарь Oxford Dictionary может предложить пользователю пять значений глагола *seal* и четыре значения имени существительного *seal*, два из которых включают три дополнительных значения. Запрос *seal* в Британском национальном корпусе предлагает пользователю 17 отдельных значений, 6 идиоматических фраз и 3 фразовых глагола (каждый из которых имеет как минимум 2 значения). Однако при изучении моделей употребления не основных значений слова *seal* мы находим в корпусе в среднем не более 10–20 примеров для каждого из значений. Объем корпуса в 6 млрд слов с трудом удовлетворяет запрос лексикографа и позволяет ему отобрать языковые примеры для иллюстрирования не частотных значений слова [6].

Второе преимущество корпусов текстов большого объема: закономерности по количеству значительно превосходят исключения. Вместе с текстами стилей художественной литературы и публицистического в корпус попадают контекстуальные авторские значения слов, не характерные для языка в целом [7]. Наш опыт работы с небольшими по объему корпусами текстов свидетельствует, что лексикограф может ошибиться и принять исключение за правило. Мы полагаем, что увеличение объема корпуса в сочетании с инструментами, которые автоматически скрывают единичные ситуации употребления слов, позволяет лексикографу сосредоточиться на главном – поиске языковых закономерностей.

Влияние инструментов корпуса текстов на процесс составления словаря

Первые корпусы текстов представляли собой простую поисковую систему: они не могли определять части речи или приводить словоформы к начальному значению, не

предлагали пользователю различные фильтры (функциональный стиль, историческая эпоха, регион употребления и т. п.). Лексикограф вручную просматривал каждую строку, изучал примеры употребления и отбирал нужную информацию для включения в словарь. Это требовало больших временных затрат, особенно при анализе высокочастотных значений слов, когда приходилось просматривать сотни или даже тысячи примеров [2]. Последний толковый словарь английского языка без опоры на корпусы текстов Longman Dictionary of Contemporary English был создан небольшой командой лексикографов менее чем за три года и появился в продаже в 1978 г., а первый проект по созданию словаря на основе корпуса текстов, длившийся семь лет, потребовал усилий большого числа специалистов и был завершен только в 1987 году.

Поскольку в конце 80-х – начале 90-х гг. корпусы текстов увеличились в объеме, ручное взаимодействие с ними стало занимать гораздо больше времени. Это потребовало от специалистов в области прикладной лингвистики создания специальных программ автоматической разметки загружаемых текстов, что привело к повышению скорости обработки данных и сделало работу с корпусом текстов более эффективной. Так, при создании словаря Macmillan English Dictionary в 2001 г. впервые были использованы программы-лемматизаторы, приводящие все словоформы к их начальной форме, лемме, и программные инструменты, способные осуществлять поиск заданных лексикографом N-грамм – последовательностей, состоящих из N элементов (звуков, морфем, слогов, слов или словосочетаний) [2; 9]. Данный инструмент позволил найти ряды устойчивых словосочетаний, в результате чего составители словарей Macmillan получили экономически эффективный способ автоматического поиска и определения коллокаций для 7 500 наиболее частотных слов английского языка, отобранных для включения в первое издание Macmillan English Dictionary. В результате предоставления пользователю исчерпывающей информации о коллокациях словарь получил конкурентное преимущество [8].

Следующим шагом в развитии корпусной лингвистики стало создание программного обеспечения для поиска так называемых «эскизов слов» (*word sketch*) – «наиболее частотных словосочетаний (*word sketches*), которые распределены по лексико-синтаксическим шаблонам» [9, с. 106]. Под лексико-синтаксическим шаблоном мы понимаем «модель грамматического и синтаксического согласования лексем в языковых выражениях, основанную на определенных грамматических характеристиках и синтаксических правилах употребления слов» [10, с. 322]. Эскиз предоставлял лексикографу удобное одностраничное описание значимых моделей словоупотребления, что повысило эффективность процесса составления словарей: появилась возможность в автоматическом режиме обрабатывать большие объемы данных. В дальнейшем эффективность корпусов текстов в лексикографии повышалась за счет адаптации эскизов к конкретным словарным проектам.

Дальнейшее развитие корпуса текстов привело к созданию инструментов, основная задача которых – поиск релевантных иллюстративных примеров в корпусе и их последующий автоматический перенос в программу для составления словаря [2; 7]. Толковые словари содержат десятки тысяч примеров словоупотреблений, а корпуса текстов предлагают миллионы их вариантов. Возможность одним щелчком мыши скопировать целое предложение позволяет лексикографом экономить до пяти секунд времени при поиске каждого примера и значительно сокращает общее время работы над составлением словаря. Как правило, программное обеспечение корпуса текстов находит и отображает все возможные примеры для включения в словарную статью, а лексикограф отмечает только те из них, которые кажутся ему релевантными, после чего данные автоматически переносятся из корпуса текстов в словарь.

Перечисленные выше инструменты корпусной лингвистики дополняют друг друга в процессе создания словаря. Сначала по заданной N-грамме корпус текстов выполняет поиск возможных коллокаций. Затем лексикограф вручную отбирает те из них, которые будут включены в словарь. И, наконец, корпус формирует список примеров употребления, а лексикограф лишь отмечает те из них, которые будут использованы в словарной статье. Мы полагаем, что со временем специалисты в области прикладной лингвистики научат корпуса текстов самостоятельно формировать первоначальный словник словаря по заданным параметрам сортировки и отбора данных, что значительно снизит финансовые и временные затраты на создание онлайн-словарей. Таким образом, изменится роль лексикографа – вместо отбора данных в корпусе текстов для последующего их включения в словарь составитель сможет сосредоточиться на просмотре и редактировании сформированной компьютером подборки упорядоченной лексикографической информации.

Заключение

В ходе исследования мы выяснили, что электронные ресурсы корпусной лингвистики постепенно берут на себя многие задачи практической лексикографии. Это вынуждает теоретическую лексикографию переосмысливать роль лексикографа в процессе составления словаря – перейти от создания корпуса словаря к проверке решений, принятых программным обеспечением. Автоматизации подвергаются не только процессы создания корпуса текстов и последующего анализа представленных в нем данных, но и начальный этап процесса создания словаря, когда электронные инструменты лексикографа по заданным параметрам автоматически извлекают из корпуса текста наполнение макро- и микроструктуры словаря, формируют словарные статьи, в которых леммы уже разбиты на значения, а каждое из значений дополнено релевантными примерами употребления.

Развитие инструментов построения корпусов текстов и автоматического составления словников электронных толковых и переводных словарей играет важную роль в реализации языковой политики, поскольку «в эпоху глобализации изменениям подвергаются все области человеческой жизнедеятельности, в том числе и язык» [11, с. 93]. Изменения, произошедшие за последние двадцать лет в области корпусной лингвистики и электронной лексикографии, освободили лексикографов от выполнения монотонных задач (поиск устойчивых словосочетаний и примеров употребления), решение которых требовало больших временных затрат, минимизировали финансовые расходы издательств на составление словарей. Мы полагаем, что перспективным направлением дальнейших исследований могут стать поиск и выработка приемов и способов взаимодействия лексикографа с инструментарием корпуса текстов, разработка стратегий составления толкового и переводного словарей в цифровую эпоху.

Литература

1. *Никифорова Н.В.* Существуют ли "цифровые аборигены"? Информационное поведение и визуальные практики в современной цифровой культуре // Россия в глобальном мире. 2015. № 7 (30). – С. 209–218.

2. *Dash N., Ramamoorthy L.* Processing Texts in a Corpus // Utility and Application of Language Corpora. 2019. – Pp. 73–90.

3. *Pollach I.* Taming Textual Data: The Contribution of Corpus Linguistics to Computer-Aided Text Analysis // Organizational Research Methods. 2012. Vol. 15, Issue 2. – Pp. 236–287.

4. *Kilgarriff A.* The Sketch Engine: Ten Years On // Lexicography. 2014. Issue 1 (1). – Pp. 7–36.

5. *Rundell M.* The road to automated lexicography // Electronic Lexicography. – Oxford, 2012. – Pp. 15–30.

6. British National Corpus (BNC). – URL: https://www.english-corpora.org/bnc/

7. *Conrad S.* Corpus Linguistics Texts // TESOL Quarterly. 2003. no. 37. – Pp. 559–561.

8. *Black K.* Evaluating Lemmatization Models for Machine-Assisted Corpus-Dictionary Linkage // Language Resources Evaluation Conference. 2014. – Pp. 3798–3805.

9. *Костина И.А.* Исследование в системе Sketch Engine на примере глаголов физического восприятия в английском языке // Новая наука: Проблемы и перспективы. 2016. № 53 (79). – С. 106–112.

10. *Mitrofanova O.A., Zaharov V.P.* Automatic Analysis of Terminology in a Russian Corpus of Texts on Corpus Linguistics // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference. – Bekasovo, 2019. – Pp. 321–328.

11. *Юдаева О.В.* Инструменты реализации языковой политики в Российской Федерации // Вестник Дагестанского государственного университета. Сер. 2: Гуманитарные науки. 2021. Т. 36, № 1. – С. 85–93.

Поступила в редакцию 7 июля 2022 г.

UDC 81

DOI: 10.21779/2542-0313-2022-37-4-81–86

Corpus Linguistics: New Opportunities of Lexicography

I.V. Balkanov

MGIMO University; Russia, 119454, Moscow, Vernadskiy av., 76; i-balkanov@mail.ru;

Prince Alexander Nevsky Military University; Russia, 123001, Moscow, B. Sadovaya st., 14; i-balkanov@mail.ru

The article examines the prospects of theoretical and practical electronic lexicography in the focus of applied (corpus) linguistics. It explores in diachrony the role and possibilities of text corpus in the creation of explanatory dictionaries of English and analyzes the impact of the volume of the text corpus on lexicographers.

Electronic resources of corpus linguistics have changed the role of lexicographers in the process of dictionary compilation. The author proves that instead of selecting data in a text corpus for their subsequent inclusion in the dictionary, the compiler is focused on reviewing and editing a selection of lexicographic data generated by special text corpus tools.

The author sees as a promising direction of further research the search for methods and ways of interaction between the lexicographer and the corpus tools, as well as the development of strategies for compiling explanatory and translational dictionaries in the digital era.

Keywords: *electronic lexicography, text corpus, corpus linguistics, lemmatization, semantization*.

Received 7 July 2022