УДК 81’33


С.А. Маник, Н.А. Шамова
Ивановский государственный университет
Санкт-Петербургский государственный университет

КОРПУСНАЯ ЛЕКСИКОГРАФИЯ: ДИАХРОНИЧЕСКИЙ РАКУРС

Статья посвящена описанию становления корпусной лексикографии от предпосылок ее зарождения до настоящего времени. Увеличивающийся репертуар технических средств, перечисляемых в работе, свидетельствует о динамике развития корпусной лексикографии и о ее многообещающих перспективах.

Корпусная лексикография, лингвистический корпус, национальный корпус, узкоспециальный корпус, Веб как корпус, корпусные технологии.

В XXI веке возрастает необходимость обработки большого объема неупорядоченной информации в связи с научно-техническим прогрессом и переходом к информационному обществу. Компьютерные технологии вносят существенный вклад в процесс анализа сведений и активно применяются во многих областях, в том числе в лексикографии, тем самым значительно облегчая и оптимизируя работу лексикографов. Вместе с тем, лексикография прошла длинный путь технологического развития от ручной обработки данных, написания специальных карточек и глоссариев до миллиардных компьютеризированных лингвистических корпусов.

Одним из основных понятий корпусной лексикографии является «конкорданс», под которым понимают список слов к определенному произведению с контекстами их употребления. Исследователи отмечают, что конкорданс, который активно изучается в рамках корпусной лингвистики и корпусной лексикографии, был изобретен давно, в начале XIII века появилась первая «конкорданция» (*Concordantiae morales sacrae scripturae* – «Нравственная конкорданция Священного Писания»), которая представляла собой «предметный конкорданс к текстам Библии» [2, с. 17]. М.И. Солнышкина и Г.М. Гатиятуллина справедливо подчеркивают, что конкордансы доэлектронной эпохи подразумевали «некий указатель места употребления слова или словосочетания», а само понятие «конкорданс» приравнивалось к понятиям «словарь» и «указатель» [5, с. 136]. В цифровую эпоху произошли изменения, которые будут рассмотрены позднее.

Работа с английскими текстами классической литературы как источниками нормативного материала осуществлялась при составлении словарей в XVIII–XIX веках. В.П. Захаров и С.Ю. Богданова указывают на то, что в XVIII веке С. Джонсон во время работы над знаменитым словарем *Dictionary of the English language* (1755) отбирал из книг предложения, которые называл цитатами, чтобы наглядно показать функционирование слов в контексте [2, с. 19]. А в XIX веке грамматисты использовали примеры для иллюстрации своих утверждений «из произведений признанных авторов» [1, с. 9].

Появление машинописи способствовало тому, что на смену рукописному написанию пришел книгопечатный способ, который вывел лексикографический процесс на новый уровень. Однако, несмотря на доступные в то время технические возможности работы с информацией, многое по-прежнему опиралось на интуицию лексикографа и его личные предпочтения относительно необходимости включения той или иной лексической единицы в словарь, опыт, т.е. лексическую интроспекцию (*lexical introspection*).

Долгое время словари были предписывающими (*prescriptive*), они отражали литературную норму и не фиксировали ненормированные варианты, которые оставались за пределами словаря. Потребность в наблюдениях за речью носителей языка явилась предпосылкой к созданию специальных электронных коллекций языкового материала, отражающих реальное словоупотребление.

Первым корпусом текстов, разработанным в 1960-х годах в Брауновском университете в США, является *Brown Corpus*. Корпусные технологии (в основе которых лежит лингвистический корпус) вызывают особый интерес ученых, которые успешно извлекают пользу из них для многих сфер. Отдельной областью применения корпусных технологий является корпусная лексикография (*corpus lexicography*), сформированная благодаря успешному объединению методов традиционной лексикографии и возможностей компьютерных технологий.

Несмотря на небольшой по сегодняшним меркам объем информации (1 млн слов) *Brown Corpus* и его ограниченные технические возможности, в 1969 году был скомпилирован словарь *American Heritage Dictionary* на основе его корпусных данных [16, p. 428]. В 1970-е годы был создан *COBUILD project* [10, p. 334]. Целью проекта являлась подготовка и публикация материалов для изучения и обучения английскому языку, основанных исключительно на анализе компьютеризированных корпусов, содержащих уже тексты в устной и письменной форме [8].

С развитием науки и техники в XIX веке перед лингвистами также встал вопрос о стандартизации терминологического пласта, который активно ворвал-

ся в язык. Лингвисты систематизировали терминологию разнообразных сфер при помощи автоматизированных терминологических банков данных (ТБД), ориентированных на конкретные группы пользователей и отдаленно напоминающих корпусы. В западной практике активно применяются термины *terminological databanks* и *terminological databases* (TDBs).

М.И. Солнышкина и Г.М. Гатиятуллина упоминают о том, что к середине 1970-х годов были сформированы специальные базы для хранения электронных корпусов: *Oxford Text Archive (OTA)* и *International Computer Archive of Modern English (ICAME)* [5, с. 137]. Подобные базы были необходимы в связи с увеличением количества корпусов, которое произошло во многом благодаря пониманию алгоритмов их составления в результате сформированных ценных теоретических навыков. Постепенно в создаваемые корпусы начинает добавляться разметка. В 1968 году был впервые использован термин *метаразметка* (*metadata*), применяемый для обозначения информации о текстах [13, p. 195].

П. Хэнкс указывает на то, что в 1990-е годы британские издательства словарей, особенно издательства, связанные с производством учебных словарей, переиздавали свои словари на основе корпусных данных [9, p. 80]. Словари и учебные пособия XXI века часто сопровождаются пометой *corpus-built*, которая свидетельствует об обращении к корпусным данным. На основе корпусных данных составлены следующие авторитетные словари, фиксирующие большее количество языковых особенностей, чем словари XX века, например, издательством *Collins* выпущен *English Learner's Dictionary* (2011 г.), издательством *Longman – Dictionary of Contemporary English* (2009 г.), издательством *Cambridge University Press – Business English Dictionary* (2011 г.) и т.д. Подчеркнем, что в XXI веке практически все крупные издательства, выпускающие словари, используют корпусные данные, которые успешно зарекомендовали себя как источник надежной информации.

Особую ценность для лингвистов и лексикографов имеют репрезентативные корпусы национальных языков. Если в докомпьютерную эпоху на создание словарей национального языка уходили целые декады, то с применением автоматизированных корпусов национальных языков этот процесс начал занимать значительно меньшее количество времени. Представляется важным акцентировать внимание на том, что ранее корпус воспринимался в основном как источник для фундаментальных исследований, поскольку доступ к нему, как правило, имели только ученые. В настоящее время создание корпусов, в частности узких корпусов, становится все более доступным практически для любого пользователя.

Появление Интернета не могло не отразиться на принципах формирования и форме существования лексикографических справочников. В XXI веке в эпоху Интернета появляется термин «киберлексикография» (*cyberlexicography*). О.М. Карпова среди новинок киберлексикографии выделяет такие «словарные ресурсы, как *Sketch Engine*, *Google Ngram Viewer*, *WorldNet*, *the Pattern Dictionary of English Verbs*, *VerbNet*, *FrameNet*, *Google Image Search*» [4, с. 32].

В связи с увеличением технических возможностей происходит расширение рамок понятия «лингвистический корпус». На современном этапе Интернет все активнее начинает рассматриваться в качестве корпуса *Web as Corpus*, *Corpus as Web.* В начале 2000-х годов было создано сообщество *The Web-As-Corpus Kool Yinitiative (WaCky)* [2, с. 80]. Его участники разрабатывали «набор инструментов (и интерфейсов для существующих инструментов), позволяющих лингвистам искать в Интернете тексты (веб-страницы) для наполнения корпусов, обрабатывать их, индексировать и в итоге создавать из них корпусы, которые могут насчитывать миллиарды токенов» (Там же). Составление корпусов на основе сведений из Интернета указывает на значительный технический прорыв.

Интернет предоставляет разнообразные инструменты поиска, поэтому не удивительно, что, по мнению исследователей, *Google* и другие поисковые системы могут приравниваться к корпусным инструментам [11, p. 34–35]. Таким образом, в широком смысле корпусные инструменты – это средства, позволяющие упорядочивать информацию.

Отмечается успешное функционирование в режиме онлайн гибридных форм справочных ресурсов, направленных на объединение доступных технических достижений. Так, онлайн словарь *Pattern Dictionary of English Verbs (PDEV)* основан на корпусных данных [14, p. 463]. В качестве перспектив ученый видит создание платформ, объединяющих корпус и словарь (*dictionary-cum-corpus platform*), используя которые пользователи могут загружать свои собственные корпусы с целью получения информации об особенностях функционирования слова в контексте (Там же, p. 476–477).

Особое внимание следует уделить антропоцентричности корпусных технологий XXI века, которые (в отличие от разработок XX века) стараются удовлетворить запросы конкретного пользователя. В настоящее время многие лексикографические проекты применяют корпусные инструменты, представленные в режиме онлайн, поэтому пользователям не нужно устанавливать какое-либо программное обеспечение на свой компьютер [11, p. 35–36].

В сравнении с первыми корпусными технологиями, которые были лишены каких-либо подсказок, помогающих пользователю эффективно работать с ними, многие современные корпусы, напротив, имеют информативный «справочный блок» разного объема. Например, программа *AntConc* во вкладке *Help (View Help File)* содержит документ в формате *pdf*, в котором подробно рассматривается функционал программы и принципы работы с ней [6]. Безусловно, быстрый доступ к технологиям, удобство при их использовании и понятный интерфейс делают их более привлекательными для лексикографов и лингвистов.

Если на этапе зарождения корпусных технологий в арсенале у исследователей был преимущественно конкорданс, позволяющий работать с контекстом, то корпусы нового столетия предлагают существенно больше технических возможностей, что способствует расширению задач, которые ставятся перед корпусной лексикографией. Функциональные возможности современных корпусных программ довольно разнооб-

разны, и даже привычный поиск информации может осуществляться по разным параметрам, что было недоступно в доцифровую эпоху.

Большой список задач, с которыми успешно справляются современные корпусные технологии, вдохновляет исследователей на создание новых технических возможностей. В рамках прикладной и корпусной лингвистики активно разрабатываются более совершенные функции и инструменты, позволяющие лексикографам автоматически извлекать необходимую информацию, обрабатывать ее, получать статистические данные в соответствии с поисковыми запросами. Программное обеспечение, делающее информацию из корпуса доступной для лексикографов, получило название *corpus query system (CQS)* 16, p. 431].

Высоким функционалом обладает технология *TickBox Lexicography (TBL)*, благодаря которой лексикограф выбирает необходимые элементы автоматического анализа слова для включения в словарь [12]. Опция *Good Dictionary Example (GDEX)* автоматически отсортировывает примеры из корпуса и помещает неподходящие варианты, например, те, которые имеют длинные предложения, редкие слова, варианты с более чем одной заглавной буквой и т.д. в конец списка [7].

Возможности для осуществления разметки (аннотирования) в XX веке значительно отличались от возможностей, доступных в XXI веке. Доступный на современном этапе анализ структуры материала помогает сформировать наиболее полное представление об особенностях употребления лексической единицы. П. Хэнкс акцентирует внимание на том, что при разработке лексикографического издания на основе корпусных данных необходимо обращать внимание не только на частотность лексической единицы, но также и на принципы распределения ее в тексте [9, p. 79]. Так, например, сделанные наблюдения об использовании определенных слов конкретным автором или группой авторов могут внести вклад в авторскую или писательскую лексикографию.

О.М. Карпова выделяет особую современную техническую разработку окулографию (*eye tracking*), которая фиксирует поисковую стратегию пользователя при обращении к справочнику, что позволяет устанавливать перспективу пользователя *user's perspective* [3, с. 26]. Подобное наблюдение за процессом работы пользователя со справочником, выявление его интереса к определенным блокам информации может помочь лексикографу при выпуске новой литературы или переизданию уже существующей в соответствии с потребностями пользователей.

В XXI веке, несмотря на различные технические возможности, связанные с автоматизацией многих процессов, необходимость личного участия лексикографа/лингвиста/переводчика (которое заключается в ручной проверке данных, планируемых для включения в словарь) до сих пор все еще сохраняется. Справедливо отмечается, что компьютеры никогда не заменят человека, но в практическую лексикографию они могут внести значительный вклад [15]. Безусловно, мнение человека, имеющего профессиональные знания, является важным и должно быть учтено на этапе «посткомпьютерной проверки данных».

Подводя итог описанию современного состояния корпусной лексикографии, необходимо акцентировать внимание на практической ценности использования корпусов в качестве источника данных для лексикографии. К списку достоинств корпусов в соответствии с современным состоянием корпусной лексикографии и доступными на данный момент возможностями необходимо отнести следующие достижения:

- корпус предоставляет доступ к неограниченному количеству информации в соответствии с тенденцией рассматривания Интернета в качестве корпуса (*Web as Corpus*, *Corpus as Web*);

- функциональные возможности корпусных технологий становятся все более антропоцентричными, они все чаще используются в лексикографическом процессе;

- репертуар узкоспециальных корпусов существенно расширяется, поскольку возрастает роль пользователя (необязательно ученого), который может (в том числе на бесплатной основе) создать свой собственный корпус, определив его объем и выбрав для него наполнение;

- гармоничное сосуществование корпуса и словаря позволяет им дополнять возможности друг друга, в результате чего появляется новый уникальный гибридный вид справочного ресурса.

Перспективы развития корпусной лексикографии, по нашему мнению, заключаются в разработке большего количества находящейся в свободном доступе электронной справочной литературы со встроенными корпусами, позволяющими пользователям, имеющим минимальный опыт работы с такими технологиями, благодаря интуитивно понятному интерфейсу, самостоятельно анализировать корпусные данные.

Таким образом, корпусные технологии, имеющие высокий потенциал, внесли значительный вклад в выпуск разнообразных авторитетных лексикографических изданий в рамках корпусной лексикографии, которая прошла долгий путь становления. Предпосылкой к созданию электронных коллекций материала о языке явились наблюдения за речью носителей языка, которая включает не только нормированные варианты словоупотребления, но и ненормированные, важность отражения которых в справочниках становилась все более очевидной. Создание электронных корпусов, включающих сначала не более 1 млн слов ознаменовало значительный прорыв в корпусной лексикографии и явилось толчком к разработке в дальнейшем более функциональных технологий. Несовершенные в техническом плане и ограниченные по объему корпусы 1960–1980-х годов заменились на более совершенные, имеющие богатый функционал и ценные метаданные. На их смену в XXI веке приходит новое понимание корпуса как неограниченного по объему самообновляемого ресурса, способного встраиваться в различные формы справочников, тем самым расширяя их репертуар и предоставляя больше возможностей для работы с материалом.

Литература

1. Захаров, В. П. Корпусная лингвистика : учебник для студентов направления «Лингвистика» / В. П. Захаров, С. Ю. Богданова. – 2-е изд., перераб. и доп. – Санкт-Петербург : СПбГУ. РИО. Филологический факультет, 2013. – 148 с.

2. Захаров, В. П. Корпусная лингвистика : учебник / В. П. Захаров, С. Ю. Богданова. – 3-е изд., перераб. – Санкт-Петербург : Издательство Санкт-Петербургского университета, 2020. – 234 с.

3. Карпова, О. М. Новые вызовы современной английской лексикографии / О. М. Карпова // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. – Воронеж : ВГУ, 2018. – № 3. – С. 24–28.

4. Карпова, О. М. Современная лексикографическая картина Великобритании / О. М. Карпова // Вестник Московского государственного областного университета. Серия: Лингвистика. – 2018. – № 6. – С. 28–36.

5. Солнышкина, М. И. История развития корпусной лингвистики (на примере англоязычных корпусов) / М. И. Солнышкина, Г. М. Гатиятуллина // Вестник Томского государственного университета. Филология. № 63. – Томск : ТГУ, 2020. – С. 132–160.

6. AntConc. – URL: https://laurenceanthony.net/software/antconc/ (дата обращения: 13.07.2021). – Текст : электронный.

7. Frankenberg-Garcia, A. Slipping Through the Cracks in E-Lexicography / A. Frankenberg-Garcia, G. P. Rees, R. Lew // International Journal of Lexicography. – 2021. – Vol. 34. – № 2. – P. 206–234.

8. Groom, N. COBUILD Project / N. Groom // The Encyclopedia of Applied Linguistics / Ed. by C. A. Chapelle. – Blackwell Publishing Ltd, 2013. – P. 650–653.

9. Hanks, P. Lexicography / P. Hanks // The Oxford Handbook of Computational Linguistics / Ed. by R. Mitkov. – Oxford: Oxford University Press, 2003. – P. 71–87.

10. Kilgarriff, A. Introduction to the Special Issue on the Web as Corpus / A. Kilgarriff, G. Grefenstette // Computational Linguistics. – 2003. – Vol. 29, № 3. – P. 333–347.

11. Kilgarriff, A. Corpus Tools for Lexicographers / A. Kilgarriff, I. Kosem // Electronic Lexicography / Ed. by S. Granger, M. Paquot. – Oxford : Oxford University Press, 2012. – P. 31–55.

12. Kilgarriff, A. Tickbox Lexicography / A. Kilgarriff, V. Kovář, P. Rychlý // eLexicography in the 21st Century : New Challenges, New Applications. 2010. – P. 411–418. – URL: https://www.sketchengine.eu/wp-content/uploads/tickbox_lexicography_2010.pdf (дата обращения 13.07.2021). – Текст : электронный.

13. Nguyen, T. H. Big Data Metadata Management in Small Grids / T. H. Nguyen, V. Nunavath, A. Prinz // Big Data and Internet of Things: A Roadmap for Smart Environments. Studies in Computational Intelligence / Ed. by N. Bessis, C. Dobre. Springer, 2014. – Vol. 546. – P. 189–214.

14. Paquot, M. Lexicography and Phraseology / M. Paquot // The Cambridge Handbook of English Corpus Linguistics / Ed. By D. Biber, R. Reppen. – Cambridge : Cambridge University Press, 2015. – P. 460–477.

15. Rundell, M. Good Old-fashioned Lexicography: Human Judgement and the Limits of Automation / M. Rundell // Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S.Atkins / Ed. by M. H. Corréard. – Grenoble, France : EURALEX, 2002. – P. 138–155.

16. Walter, E. Using Corpora to Write Dictionaries / E. Walter // The Routledge Handbook of Corpus Linguistics / Ed. by A. O'Keeffe, M. McCarthy. – London ; New York : Routledge, 2010. – P. 428–443.

S.A. Manik, N.A. Shamova

CORPUS LEXICOGRAPHY: DIACHRONIC PERSPECTIVE

The article describes the formation of corpus lexicography from the prerequisites of its origin to the present state. The increasing repertoire of technical means which are enumerated in this work is an indication of the dynamics of the development of corpus lexicography and its promising perspectives.

Corpus lexicography, linguistic corpus, national corpus, highly specialized corpus, Web as corpus, corpus technologies.