Нина Мечковская
Филологический факультет,
Белорусский государственный университет
(Filozofska fakulteta,
Beloruska državna univerza)
nina.mechkovskaya@gmail.com

Slavistična revija 73/3 (2025): 435–450
UDK 81'322:81'374.2
DOI 10.57589/srl.v73i3.4171
Tip 1.01

Корпусная лингвистика, частотная и толковая лексикография: векторы взаимодействия

При общем росте количества корпусов, их объемов и разнообразия происходит специализация корпусов в зависимости от состава их таргетированного контента. Электронные корпусы первого поколения (объемом примерно 100 млн словоупотреблений), называемые или осознаваемые как “национальные” или “государственные”, сохраняют относительную сбалансированность подкорпусов и широкую социально-гуманитарную адресацию. По мере увеличения объемов более поздних корпусов происходит их специализация по двум векторам: 1) содержательно ориентированные мониторные (пополняемые) мегакорпусы газетно-журнальных текстов; в целевые группы корпусного контента данного класса входят социологи и политологи, экономисты, демографы, журналисты и др.; 2) тематически безграничные (неизбирательные) корпусы, аккумулирующие оцифрованные тексты (печатные и электронные), используемые в информатике как сырье для “обработки естествнного языка” (natural language processing): машинного предобучения нейронных сетей и создания статистических алгоритмов самосвязываемости слов в адекватные текстовые реакции искусственного интеллекта.

Названы две наиболее значительные новаторские разработки в корпусной лексикографии: 1) синтез толкового и частотного словарей в словарях Macmillan (2007), позже Collins, Longman; 2) компонентный семантический анализ 100-тысячного словника с использованием в качестве семантических компонентов 2.500 самых частых лексем в Macmillan 2007. Возможности корпусов в скором времени приведут к крупным достижениям в диахронической лингвистике.

Ключевые слова: частотные словари, синтез толкового и частотного словарей в Macmillan (2007), семантический компонентный анализ 100-тысячного словника; высокочастотные слова как семантические множители.

Corpus Linguistics, Frequency and Explanatory Dictionaries: Interaction Vectors

With the overall growth in the number of corpora, their volumes and diversity, there is a specialization of corpora depending on their targeted content. Electronic corpora of the first generation (with a volume of approximately 100-million-word tokens), called or perceived as “national” or “state”, retain a relative balance of subcorpora and a broad social science and humanities audience. As the volumes of later corpora increase, there is a specialization of their purpose along two vectors: 1) content-oriented monitor (replenished) megacorpora of newspaper and magazine texts; the target groups of corpus content of this class include sociologists and political scientists, economists, demographers, journalists, etc. 2) thematically unlimited (non-selective) corpora accumulating digitalized texts (printed and electronic) used in computer science as raw material for “natural language processing” (machine pre-training

of neural networks) and creation of statistical algorithms for self-linking words into adequate verbal responses of artificial intelligence.

Two most significant innovative developments in corpus lexicography are named: 1) synthesis of explanatory and frequency dictionaries in the Macmillan dictionaries (2007) and later Collins, Longman; 2) component semantic analysis of a 100,000-word dictionary using the 2,500 most frequent lexemes in Macmillan (2007) as semantic components. The capabilities of corpora will soon lead to major achievements in diachronic linguistics.

Keywords: frequency dictionaries, synthesis of explanatory and frequency dictionaries in the Macmillan dictionary (2007), semantic component analysis of a 100,000-word dictionary; high-frequency words as semantic multipliers.

1. О значении корпусной лингвистики в науке о языке

Почему из всех информационно-компьютерных технологий, входящих в круг лингвистических задач и интересов, корпусная лингвистика оказалась наиболее востребованной, разработанной и результативной? По ряду причин. За несколько десятилетий корпусная лингвистика стала полнокровной исследовательской отраслью науки о языке, возможности которой ширятся по мере обогащения разметок и внедрения новых исследовательских инструментов. Здесь открываются новые горизонты для лингвистов разных профилей – от акцентологов и исследователей жестикуляции (на основе мультимодальных подкорпусов, включающих фрагменты видео и кинофильмов) до специалистов по синтаксису и стиховедению, а также еще по ряду смежных гуманитарных дисциплин. Благодаря использованию вычислительной техники, в корпусных исследованиях и лексикографии становятся возможны небывалые прежде масштабы и скорости работы. Важно и то, что корпусная лингвистика и тесно связанная с ней частотная лексикография отвечают давним – еще доэлектронным и докомпьютерным, – запросам лексикологии и лингводидактики: потребовались частотные словари и представительные собрания текстов в качестве основы для грамматик и словарей разных жанров.

2. Важная черта в структурировании корпусов: в них проведена внутренняя разветвленная и глубокая жанрово-тематическая диверсификация текстов

Диверсификация текстовой основы частотного словаря проводилась и в докомпьютерное время. Вот факты из близких ареалов. В «Частотном словаре русского языка» под редакцией Л.Н. Засориной (основан на текстах объемом 1 млн с/у; словник около 40 тыс. лексем) различались четыре разновидности текстов: 1) публицистические (на деле – идеологические) и научно-популярные тексты; 2) художественная проза; 3) драматургия; 3) журнальные и газетные тексты (Засорина 1977). Для каждого из 40 тыс. слов было указано, сколько раз оно встретилось в текстах каждой из четырех разновидностей.

Еще раньше жанрово-стилистическая диверсификация корпуса была разработана при создании частотного словаря белорусского языка (Минск,

1976–1992), выходившего под редакцией А.Е. Супруна. В пяти отдельных книгах представлена лексика пяти разновидностей белорусской речи: художественная проза, публицистика, устное народное творчество, разговорная речь, поэзия (Мажэйка, Супрун 1976, 1979, 1982, 1989, 1992). Суммарный объем текстов – 1,5 млн с/у; количество разных лексем – 61 тыс.

Когда запросы информационного общества и возможности компьютерных технологий привели к созданию огромных и тематически не ограниченных национальных корпусов текстов, то их разработчики, понимая интересы лингвистики и образования, с самого начала планировали корпус как комплекс подкорпусов с частотными списками слов внутри подкорпусов и их подразделений.

Над созданием первого электронного корпусно-частотного словаря в Ланкастерском университете почти 30 лет работал профессор английского языка и математической лингвистики Джеффри Лич с сотрудниками (Leech 2001). Словарь основан на Британском национальном корпусе (BNC 1994), объем 100 млн с/у; разработан под руководством профессора корпусной лингвистики из университета Бригама Янга (штат Юта) Марка Дэвиса (Davies).

Первое разделение текстов, проводимое в словаре Лича и сотрудников, – это оппозиция письменной и устной (затранскрибированной) речи. Это разделение, вошедшее в название словаря, – *Word Frequencies in Written and Spoken English*, ценно не только в собственно лингвистическом, но и в социолингвистическом и демократизирующем аспектах. В Предисловии к словарю Лич назвал один из важных результатов, увиденный благодаря диверсификации: из первых по частоте 50 слов письменной речи только 33 совпадают с 50 топ-лексемами речи устной.

Вторым по важности разделением словаря Лича стала грамматическая классификация лексики: были получены частотные списки лемм в составах 10 частей речи (различаемых словарем). Дальнейшая диверсификация лексики проведена по таким направлениям: а) образное (*imaginative*), т. е. художественное VS информативное письмо; б) в устном подкорпусе (*Spoken English*) противопоставлены два вида речи: разговорная речь (*conversation Spoken English*) vs речь, ориентированная на решение задачи (*task-oriented Spoken English*); в) 24 «поля интереса», например, «прилагательные, указывающие на отношение к регионам и народам»; цветообозначения, ранжированные по частоте; частоты обозначений лиц мужского и женского пола; ряд других разделений.

В результате диверсификации словарь Leech (2001) содержит более 30 частотных списков, в которых показаны различия в лингвистическом поведении английских слов. Основные частотные списки: 1) общий список лемм по убывающей частоте; он заканчивается на частоте 75 (словоупотреблений); 2) алфавитный список лемм с указанием их частот (доступные в архивированном виде версии различаются по длине); 3) два раздельных (для письменной и устной речи) списка лемм, ранжированных по убыванию их частот; 4) два алфавитных списка лемм

(также раздельно для письменной и устной речи) с указанием их частот; 5) 10 частотных списков лемм в составе частей речи; 6) частотные списки лемм в текстах, противопоставленных как образная (*imaginative*) речь VS информативная речь; еще 16 других частотных списков.

«Новый частотный словарь русской лексики» О.Н. Ляшевской и С.А. Шарова ([первое электронное издание], Ляшевская, Шаров 2009; [печатные издания:] Ляшевская, Шаров (2009, 2015); сокращенная ссылка Ляш/Шар) базируется, как и Leech (2001), на 100 млн с/у Национального корпуса русского языка (НКРЯ).[1] Однако, поскольку хронологические рамки частотного словаря было решено ограничить периодом 1950–2007 гг., то объем корпуса, по оценке составителей, не 100 млн, а в 92 млн. с/у. В НКРЯ 16 подкорпусов с многоступенчатой иерархией. Поскольку разработка новых подкорпусов и пополнение существующих продолжается, то в данных о количестве корпусов в НКРЯ, их разделениях и объемах возможны расхождения.

Первое разделение лексического материала в русском частотном словаре – это его грамматическая (частеречная) классификация, определяющая ранговые (по убыванию частот) списки слов в границах каждой из 7 частей речи. Затем следует длинный ряд функционально-стилистических и тематических разделений лексики, что в совокупности приводит к более сложной, чем в британском словаре, диверсификации материала: тексты разделены на 8 групп (группа художественных текстов и 7 функционально-стилистических разновидностей нехудожественных текстов); это разделение дополняет детализирующий открытый список жанров (или “типов” текстов, в терминологии составителей) – таких, например, как интервью, инструкция, закон, личное письмо (более 100 типов). Тематика текстов представлена списком из 54 категорий, в разной степени дробных: от «экономика» или «политика и общественная жизнь» до «путешествия» и «вооруженные конфликты».

Во Введении (раздел 6 «Структура частотного словаря») перечислено 19 частотных списков), но печатный Ляш/Шар содержит 34 частотных списка. В целом он позволяет анализировать лексику текстов, распределенных примерно по 100 разрядам. Данные о частотности лексем в грамматических и функционально-стилистических разрядах слов позволяют исследовать не только статистическую и грамматическую организацию словаря, но и языка в целом – в его функциональных и жанрово-стилистических вариантах.

3. Глобальный рост количества корпусов, их объемов и разнообразия

Новейшие интернет-компьютерные национальные и мультинациональные корпусы по объемам обрабатываемой информации конкурируют с крупнейшими поисковиками ранга Google, AOL, Yahoo!, Яндекс. Многомиллиардные корпусы

[1] НКРЯ открыт для общего пользования в 2004 г.; создан на основе Машинного фонда русского языка, созданного в 1983 г., и является его развитием.

являются пополняемыми и мониторными, т. е. регистрируют тексты и лексику текущего времени. Данные по 5-и крупнейшим в мире корпусам представлены в следующей ниже таблице.

Таблица 1: Корпусы.

Год создания	Название корпуса. Страна или город. Язык. Хронологические границы	Объем (в с/у)
1964	Deutsches Referenzkorpus (DeReKo), известен также как Мангеймский корпус. Аффилирован с исследовательским Институтом немецкого языка в Мангейме. Постоянно пополняется; содержит только полные и только лицензионные тексты от начала XIX в.	57,6 млрд
1991	Český národní korpus (ČNK). Разработка велась в специально созданном Институте национального Корпуса чешского языка при Философском факультете Карлова университета в Праге.	9 млрд
2010	News on the Web (акроним NOW). Мониторный мегакорпус “мирового английского”, аккумулирует тексты сетевых газет и журналов на английском в 21 стране. Ежегодно прирастает на 1,5 – 2 млрд. Разработчик Марк Дэвис.	19,6 млрд
2013	Генеральный Интернет-Корпус Русского Языка (акроним ГИКРЯ). Включает нередактированные тексты Рунета (социальной сети ВКонтакте, блогов Живого Журнала), Журнального Зала, а также несетевых СМИ и журналов.	20 млрд в 2024 г.
2018	iWeb. Корпус англоязычных сегментов всемирной паутины. Разработчик Марк Дэвис.	14 млрд

По скорости пополнения, полноте охвата англоязычной интернет-коммуникации и разнообразию аспектов ее структурирования NOW превосходит все поисковики. «NOW Corpus – это единственный ресурс, позволяющий узнать частотность слов, фраз и словосочетаний по году, месяцу и дню – в тысячах газет и журналов со всего мира. В то время как другие ресурсы, такие как *Google Trends*, показывают, что́ ищут люди, *NOW Corpus* – единственный структурированный корпус, который показывает, что́ на самом деле происходит с языком […]. В этом смысле NOW – самый надежный мониторный корпус английского языка.»

В начале XXI в. Марк Дэвис разрабатывает первые специализированные корпусы: в 2001–2012 гг. – *Corpus of American Soap Operas* (100-миллионный корпус мыльной оперы за 10-лет американского телевидения); в 2006 г. – *TIME*

Magazine corpus; создан на основе архива еженедельника *TIME* за 83 года (275 тыс.статей), начиная с первого номера 1923 г. по 2006 г.; объем 100 млн с/у; в 2008–2010 – *Corpus of Historical American English* (акроним *COHA*); охватывает 200 лет в истории языка (1810–2010 гг.); содержит более 475 млн с/у (что в 50 и 100 раз больше объемов других исторических корпусов английского языка); корпус сбалансирован по жанрам десятилетие за десятилетием (В сети); в 2014 г. – *The Wikipedia Corpus*; содержит полный текст англоязычной Википедии: это 1,9 млрд с/у в более чем 4,4 млн статей; в 2020 г. – Корпус короновируса на основе почти 1,9 млн текстов с января 2020 по декабрь 2022 г.; объем 1,5 млрд с/у; это непревзойденная по полноте фиксация социального, культурного и экономического воздействия COVID-19 на англоязычные страны. М. Дэвис, создатель Британского национального корпуса (1994), разработал также специализированные исторические корпусы испанского (2002) и португальского (2011) языков.

4. Корпусная индустрия: продукция тройного назначения

4.1. Как многомиллионные корпусы используются в преподавании и исследовании языков. Опыты адаптации

С распространением корпусов в разных странах все чаще звучали обиженные голоса пользователей, которые не смогли преодолеть технические сложности на пути к самостоятельной работе с данными корпуса. Лингвисты-исследователи обращались к корпусу чаще всего при сборе материала для диссертаций и статей, не всегда обращая внимания даже на даты, а тем более на статистику фактов. Огромные возможности ресурсов и исследовательского инструментария корпусов не только не использовались, но и оставались малоизвестными. Почти не возникало вопросов, которые предполагали бы обращение к большим массивам данных. Нечеловечески огромные корпусы поражали масштабом и останавливали потенциальных пользователей Что касается преподавателей языков в средней школе, то их возможности использовать корпус при подготовке к урокам, а тем более приобщать к поискам в корпусе школьников на практике оказались еще меньше. В начале XXI в. была осознана необходимость адаптации корпусов к запросам и возможностям рядовых гуманитариев — преподавателей языков, лингвистов-исследователей, лексикографов, издателей, журналистов.

Первые крупные и успешные шаги для сближения корпусной лингвистики и образовательной практики были сделаны исследовательской компанией *Lexical Computing*, основанной в 2003 г. Адамом Килгарриффом (1960–2015) в Великобритании, с филиалом в университете Масарика в Брно (руководитель исследований Павел Рыхлы (Rychlý)). В настоящее время работа фирмы локализована в Брно. *Lexical Computing* поставляет базы данных слов, базы данных n-грамм для использования их в других программах или для лексикографических онлайн-проектов. С 2003 г. *Lexical Computing* становится ведущим поставщиком решений в NLP (*natural language processing*), компьютерной лингвистики и электронной лексикографии.

Флагманская разработка А. Килгарриффа *Sketch Engine for Language Learning* (SkELL) включает облачное хранилище корпуса текстов живого языка (английского или другого изучаемого идиома) и пакет программ для лингвистических запросов к корпусу, управления им и анализа корпусных данных (в том числе для визуализации ряда данных). Килгаррифф назвал свою разработку *Sketch Engine*, потому что ее ключевая функция состоит в том, чтобы по данным корпуса автоматически составить по запросу для любого слова (для английского – из первых по частоте 60 тыс. слов BNC) его *sketch* ('эскиз, набросок') – одностраничную картину языкового поведения каждого слова. В настоящее время *Sketch Engine* разработан для более 800 текстовых корпусов на более чем 100 языках, в том числе и для обучения русскому языку.

Одно из направлений адаптации заключалось в сокращении объемов языкового материала, включаемого в адаптированные версии корпусных разработок. Если адаптация корпусов для нужд лингводидактики требует «уменьшения объема корпуса в соответствии с учебными целями ресурса», то встает вопрос, зачем создавать миллиардные корпусы. В ответах на подобные вопросы корпусы предстали как продукция фактически "тройного назначения".

Первое из "назначений" – гуманитарные дисциплины, и лингводидактика в первую очередь. В СССР/ СНГ в преподавании западноевропейских языков в высшей школе лексический минимум определялся в 3–4,5 тыс слов (с различиями в зависимости от специальности студентов). Это немало: как известно, в любом тексте 2 тыс самых частых слов покрывают примерно 70–80% словоупотреблений. Первостепенная задача в преподавании иностранных языков виделась в том, чтобы студент мог легко и правильно использовать эти 2–3 тысячи слов в порождении и восприятии устной и письменной речи. Понятно, что лексический запас преподавателя иностранного языка или составителя учебника должен в разы превышать студенческий минимум, но все же это не 50–60 тыс слов неродного языка (примерная длина используемых частотных списков лемм в *Sketch Engine*).

Корпусные выдачи, доступные благодаря *Sketch Engine* (в первую очередь собственно "Скетчи", затем ранжированные списки коллокаций, конкордансы, хронологические графики динамики *ipm* лемм, диахронические и диаграмматические по сути облака "похожих слов") всё шире входят в исследовательскую лингвистическую практику и преподавание информатики для гуманитариев. Исследовательский инструментарий корпусов становится все более разнообразным и результативным, но лингвистические задачи не выходят за пределы корпусов объемом в 100 млн с/у. И пока неизвестно, чтó и в какой мере изменится в вероятностях и ранговых списках, если всё будет пересчитано для корпусов в 1 млрд или 10 млрд с/у. Тем более неизвестно, скажутся ли эти изменения на лексическом составе и рангах 1-й тысячи самых частых слов, на полноте представления языков в преподавании.

Чем больше объем корпуса, тем больше в нем малоинтересных "хвостовых" данных, завершающих списки по убыванию частот, – многих тысяч лемм

с одинаково низкой частотой (ниже 10). Судя по словарю Засорина (1977), 40-тысячный словник которого показывает все разные леммы, в нем лексемы с частотой 1, 2, 3 в сумме составляют 22.489 слов, т. е. 57,8 % от всего словника (Засорина 1977: 915). Это объясняет, почему ни корпусы, ни частотные словари не приводят полные списки своих лемм. Так, в Leech (2001) алфавитный список лемм (50 или 60 тыс) в Сети не доступен, а частотный список заканчивается на частоте 75; в его частеречных словниках списки лемм доведены до частоты 10 включительно. В Ляш/Шар (2015) в алфавитном списке 50 тыс; в частотном, доведенном до частоты 33, – 20 тыс лемм (сс. XIV–XVI).

Иная картина в частотных словарях, основанных на корпусах "всего" в 1 млн с/у: "порог вхождения" слова в ранговый список преодолевается значительно легче. Так, в словаре Засорина (1977) пороговая частота составляет 10 с/у, а ранговый список насчитывает 9.044 лемм. Русский частотный словарь Леннарта Лённгрена, созданный в 1993 г. на основе корпуса в 1 млн с/у Упсальского университета и широко известный по ряду переизданий, в ранговом списке насчитывает 10 тыс. слов при пороге вхождения 9 с/у включительно (Lonngren 1993).

Пока в преподавании языков миллиардные корпусы не используются, а данные 100-миллионных корпусов (как BNC или НКРЯ) нуждаются в основательной редукции, что и делается в адаптирующих разработках типа *Sketch Engine for Language Learning*.

4.2. "Второе назначение". Целевые группы контента мегакорпусов: социологи, политологи, экономисты, демографы

Сопоставление объемов (под)корпусов в НКРЯ показывает, что самым крупным является не "Основной" корпус (389 млн с/у), но корпус, названный *Газетным*: совокупный объем его 2-х подкорпусов (*Центральные СМИ* и *Региональные СМИ*) в начале 2025 г. составлял 850 млн с/у, что более чем в 2 раза больше *Основного* корпуса. Встает вопрос: зачем так много газетно-журнальных текстов? Почему Марк Дэвис создавал корпусы не романов Драйзера или Фолкнера, но полный корпус журнала *Time* за 80 лет его истории? Ведь, как известно, язык журналистики беднее языка художественной литературы. Значит, и разработчиков НКРЯ и М. Дэвиса интересовал НЕ язык, а то **содержание** повседневной жизни и текущей политики, о котором рассказано на газетно-журнальном языке. Таков мировой тренд: в корпусной индустрии настоящего времени приоритетны корпусы, аккумулирующие информацию новостных лент. Особенно это характерно для международных мегакорпусов и близких к ним проектов – таких, как *NOW*, *iWeb*, *Linguistic Data Consortium* (*LDC*, Консорциум лингвистических данных) и др. Дело не в языке (не в словарях и учебниках языка, не в лингводидактике и даже не в лингвострановедении), но в востребованности гораздо более полной информации о повседневной жизни общества – о настроениях, конфликтах, о том, чего люди ждут от будущего, чего боятся. Как известно из истории разведок, 4/5 "разведданных" извлекаются из отрытых источников. Вот почему объемы

корпусов периодики превышают все другие корпуса. Идет глобальный мониторинг текущей повседневности, чреватой будущим. Детальная информация о жизни народов, стран, парламентов и правительств нужна экономистам, политикам, социальным психологам, историкам, философам, демографам, футурологам. Все хотят держать руку на пульсе сегодняшней жизни, чтобы понимать или предвидеть, что будет со страной и человечеством завтра. Существенны также ретроспективные возможности мегакорпусов — при историческом анализе и художественном осмыслении недавнего прошлого, когда важны все подробности и хронометраж происходившего по дням и часам, иногда по минутам, и имена всех участников.

4.3. “Третье назначение”. Мегакорпусы как текстовое сырье для статистических алгоритмов машинной обработки языка

Относительно недавно, в 2022 г., после гигантского взлета в разработках искусственного интеллекта (далее ИИ), оказалось, что основной пользователь и, скорее всего, главный заказчик миллиардных корпусов – это корпорации, создающие нейронные сети и модели *GPT* –трансформеры, генеративные и предобученные (*Generative Pre-trained Transformer*). ChatGPT-модели генерируют текст в ответ на заданный вопрос, принимая во внимание как вопрос, так и прежде введенные в него тексты. Восприятие (понимание) и генерирование информации осуществляется в сознании человека и в ИИ-устройстве на основе функционирования нейронных сетей головного мозга (или их компьютерных имитаций) в соединении с алгоритмической обработкой накопленных и пополняемых данных на основе статистики связей между внутренними (внутри устройства) и внешними элементами информации.

Революционный прогресс в ИИ-разработках, произошедший в 2010-х гг., был революцией НЕ в разработке нового поколения нейронных сетей, но в создании, во-первых, гигантских вычислительных и энергетических ресурсов, которые сделали возможным это ускорение, и, во-вторых, в открытии возможности вводить в трансформер текст на естественном языке, т. е. без трудоемкой предварительной разметки. То, что называют *Языковой моделью* и *Большой языковой моделью*, – это всепроникающий статистический анализ миллиардных корпусов текстов на основе алгоритмов **самосвязываемости**. Модель обрабатывает последовательные текстовые данные, генерируя по двум-трем словам запроса осмысленное высказывание. По первому высказыванию генерируется следующее, по первым двум высказываниям — третье и так далее. Статистические предсказания текстовых последовательностей тем более вероятны и, значит, тем более надежны, чем бóльшие массивы текстов в них введены. В речи разработчиков предобученных трансформеров часто звучит слово *скармливать*: речь идет о том, какого объема текстовые массивы им приходилось вводить в трансформер для его предобучения, а еще о том, что трансформеры постоянно *голодны*. Некоторые специалисты (напр., Иван П. Ямщиков) считают, что к настоящему времени все оцифрованные тексты планеты уже введены в память генеративных устройств; однако более существенным (но преодолимым в перспективе) тормозом в развитии ИИ является нехватка энергоресурсов.

Примером исследовательского центра, в котором на основе мониторных мегакорпусов осуществляются разнопрофильные исследования, может служить *Консорциум лингвистических данных* (*Linguistic Data Consortium*, *LDC*) – открытое объединение университетов, компаний и государственных исследовательских лабораторий, занимающееся созданием, сбором и распространением речевых и текстовых баз данных, лексиконов и других ресурсов для исследований и разработок в области лингвистики, корпусной лингвистики и информатики. Штаб-квартира LDC находится в Филадельфии, его хост-институтом является Пенсильванский университет. LDC был основан в 1992 г. при поддержке *Управления перспективных исследовательских проектов* Министерства обороны США (*Defense Advanced Research Projects Agency*, сокращённо: *DARPA*, существует с 1958 г.). Консорциум способствует развитию технологий обработки естественного языка, машинного перевода, машинного обучения и др. Здесь разрабатываются алгоритмы и статистические модели, используемые в системах искусственного интеллекта. Ресурсы, предоставляемые LDC, широко используются в академических и промышленных исследованиях, а также в образовательных целях. Легко, однако, видеть, что психология и лингводидактика, приоритетные для корпусной лингвистики в начале ее истории (характерен, напр., частотный 3-томный словарь Эдварда Торндайка *Teacher's Word Books* 1922–1944 гг.), сегодня занимают гораздо более скромное место.

Лингвисты, лексикографы, преподаватели языков — это хронологически первая группа создателей корпусов, частотных словарей и исследователей статистики речи. Эту деятельность можно назвать собственно корпусной лингвистикой. Ее основной и оптимистический результат состоит в разработке и реализации концепции корпуса как виртуального собрания репрезентативного массива текстов и как комплекса эффективных инструментов исследования языка. Результаты корпусных исследований существенны для приращения знаний о языке и для университетского преподавания языков и языкознания.

Хронологически вторая группа корпусных исследователей – это те, кто стремится "держать руку на пульсе" общества: социологи, политологи, журналисты, историки, демографы, литературоведы. Это именно их профессиональные интересы способствовали тому, что в мире резко выросли объемы и число корпусов газетно-журнальных текстов, включая сетевые. К названному классу корпусов примыкает корпусная лексикография языка писателей или отдельных произведений.

Третью группу корпусных исследований составляют специалисты по искусственному интеллекту. Если для гуманитариев в корпусах интересно содержание текстов, а для лингвистов язык, то для информатиков тексты – это прежде всего сырьё для "обработки языка", предобучения нейронных сетей и создания статистических алгоритмов "самосвязываемости" слов в адекватные словесные реакции трансформера в разговоре с человеком.

5. Как корпусы используются в лингвистических исследованиях

В литературе по корпусной лингвистике пока резко преобладают публикации о поисках и решениях, связанных с созданием корпусов и перспективами внедрения корпусных инструментов в исследовательскую практику и лингводидактику. Информации о результатах корпусных исследований языка значительно меньше. Между тем это небывалые достижения, поражающие прежде всего своим не кабинетно-лабораторным, но промышленным масштабом. Новый масштаб исследований, использующих корпусы, стал возможным благодаря гигантской скорости компьютерных вычислений, гигантским массивам текстов, противопоставленных по ряду релевантных признаков (по времени создания, типам речи, сферам бытования, жанрово-стилистическому своеобразию и др.), гигантской информации о каждом слове – информации, аккумулированной в корпусных аннотациях и представленной в эскалирующей (ступенчатой и расширяющейся по мере продвижения пользователя вглубь словарной статьи) подаче лингвистической информации в онлайн словарях (таких, как Macmillan, Collins и близких к ним по идеологии и архитектуре). Вот примеры таких достижений.

5.1. Самое поразительное новаторство корпусной лексикографии: в словарях Macmillan (2007) и Collins (2011) осуществлен синтез толкового и частотного словаря

Будущий электронный словарь Macmillan вначале готовился на бумаге и в 2002 г. в бумажном (*hard*) виде был издан как *Macmillan English Dictionary for Advanced Learner*s (Macmillan 2002; далее MED)[2]. В нем было 100.000 словарных статей и 80.000 примеров. Второй источник будущего словаря – это частотный словарь Leech (2001).

Создатели MED соединили информацию толкового словаря с информацией о частотах слов наглядно и выразительно, как в букваре. Технологию назвали *Красные слова и звезды*. В 100-тысячном словнике MED красным цветом были набраны первые по частоте 7.500 слов. Внутри этих 7.500 самых частых и потому "красных" слов были маркированы красными звездочками (в разном их количестве) три группы слов с разной частотой. Красный цвет слова и три звезды рядом с ним указывали на максимальную важность слова и объясняли пользователю, почему о таком слове информации больше, чем о слове с одной звездой или о слове, напечатанном обычным черным цветом и без звезды. Так MED визуально показывал читателю разную нужность в коммуникации разных по частоте слов. Новаторское включение частотных данных в MED было использовано в последующих брендовых английских словарях – Collins, Longman.

[2] В 2002 г. словарь получил высшую награду по литературе для изучающих английский язык – Приз Герцога Эдинбургского. В 2004 г. словари издательства Macmillan – *Macmillan English Dictionary for Advanced Learner*s и *Macmillan Essential Dictionary* получли новую престижную награду – приз Британского Совета за инновации в создании справочных материалов по английскому языку.

5.2. Компонентный лексический анализ в “промышленных масштабах”

Согласно принципу, принятому в MED, лексика, используемая в исходных (первых) словарных дефинициях (в онлайн-версиях MED и Collins такие дефиниции даются на первой ступени в эскалированной подаче информации о слове), не должна выходить за пределы первых по частоте 2.500 слов. Эти слова (2,5% от 100-тысячного словника), образуя ядро MED, выступают в функции, невозможной в MED для его остальных 97.500 лексем, — в функции семантических множителей, используемых для описания семантики всех 100.000 лексем словаря. Так составители MED, реализуя давнее правило здравого смысла и логики — не определять неизвестное через неизвестное, — фактически провели компонентный анализ лексики в небывалом масштабе: определили 100.000 слов с помощью 2.500 семантических множителей, выработав более мягкую версию компонентного анализа и попутно показав его практическую полезность. Список первых по частоте 2.500 слов (т.е. семантических компонентов описаия) приводится в конце бумажной версии словаря (см. Macmillan 2007).

Принцип ограничения лексики словарных дефиниций словами из предварительно выделенного списка высокочастотных слов проводится также в бумажном и онлайн-словаре Collins (2011): здесь состав металексики определений ограничен 2.000 первых по частот слов.

5.3. Новые горизонты в историческом языкознании

Библиография выдающегося создателя корпусов Марка Дэвиса говорит о том, что корпусных лингвистов влечет к истории языка. Романист по первой университетской специальности, М. Дэвис не начинал как историк языка, но стал им. Создатель 11 корпусов английского языка (включая Википедию, Google, Web, самый объемный исторический корпус американского английского), составитель синхронных и исторических корпусов испанского и португальского языков и их частотных словарей, Дэвис называет историческое языкознания своим основным научным интересом (после компьютерной лингвистики) и основным (после компьютерной лингвистики) университетским курсом, который он последние полтора десятилетия читал в исследовательском университете Бригама Янга. Дэвис издал в Японии книгу о использовании корпусов в исследовании языковых изменений; в избранной библиографии Дэвиса из 80 статей 10 относятся к историческому языкознанию. Дэвис показал, в частности, что американский английский лексически существенно дистанцировался от британского не в XIX в., но в XX.

Обращаясь к современному языку (или, шире, к лингвистической синхронии на разных этапах истории), корпусные лингвисты исследуют в первую очередь вариантность языка. Ведь сегодняшние синхронические колебания и варианты, их конкурентное сосуществование прогнозируют завтрашнее изменение баланса, и именно в ту сторону, которую предсказывает статистическая картина варьирования.

Представляется не случайным, что в русистике едва ли не первой книгой, в которой показаны результаты корпусных исследований, стали очерки по истории русской лексики: «Два века в двадцати словах» (Добрушина, Даниэль (ред.) 2016).

5.4. О перспективах корпусной проверки гипотез И.А. Мельчука о предмете фразеологии и количестве фразем в языке

Еще в 1960 г. в статье *О терминах "устойчивость" и "идиоматичность"* И. А. Мельчук (1960) сформулировал тезисы, которые изменили фразеологию как лингвистическую дисциплину. Он показал, что, свойства "устойчивость" и "идиоматичность" фразем взаимно независимы: сочетание лексем может быть устойчивым и при этом неидиоматичным (как *високосный год* или *кромешный ад*) или, напротив, идиоматичны, но неустойчивы, как *он собаку съел* (*в этом деле*), *синий чулок*, *на голубом глазу* и др. Было показано также, что оба свойства градуальны, поэтому присутствуют в конкретных словосочетаниях в разной степени (от нуля до 100 %) и допускают измерение. В результате объект фразеологии стал пониматься существенно шире: это не только идиомы и клише, но всё необозримое множество по-разному несвободных словосочетаний, в разной степени обладающих свойствами устойчивости и идиоматичности.

Мысли о более широком, чем думалось прежде, присутствии в речи свойств и признаков фразеологичности получили распространение в лингвистике и дальнейшее развитие. В книге Иорданская, Мельчук (2007) количественное соотношение слов и фразем охарактеризовано так: «Люди говорят не словами, а фраземами. Количественно фраземы превосходят слова в словарях примерно в соотношении 10 к 1» (С. 218). Понятно, что это пока не строгая формулировка (поскольку не сказано, о каких словарях речь и как определяется фразема). Однако, благодаря корпусам, у лингвистов появились инструменты для исследования механизмов фразеологизации на несравненно более обширном материале.

Корпус и его менеджеры-алгоритмы по запросу пользователя выстраивают тысячи таблиц с ранжированными по частоте коллокациями и данными о частотах их компонентов; конкордансы "в одно касание" открывают сотни высказываний по любому запрошенному слову; на помощь приходят таблицы, графики, облака "близких слов", диаграммы по годам, по темам, по жанрам, авторам… Всё для того, чтобы увидеть самые ранние семантические процессы, создающие идиоматичность, и количественную весомость этих процессов — всё для того, чтобы всё это удалось понять.

6. Заключение. Корпус как новый способ существования языка

Электронный корпус текстов представляет собой новый, хронологически третий способ существования языка (наряду с устным и письменно-печатным) – виртуальное существование. Благодаря диверсификации подкорпусов и их внутренним разделениям, а также благодаря лингвистической и метатекстовой

аннотации (разметке) корпусов, они представляют собой аналитические виртуальные модели языка. В отличие от письменно-печатного языкового существования, реализуемого визуально, корпусы существуют в электронной кодировке, которая визуализируется средствами разных семиотик (естественный язык, цифры, пиктограммы, диаграммы, графики, облака, а также фото- и киноизображения), но становится доступной пользователю через ограниченное экранное посредство компьютера, однако с возможностью последующей печати и тиражирования.

Благодаря корпусно-компьютерным технологиям возможности познания языка выходят далеко за пределы тех вопросов о языке, ответы на которые лингвисты находили в текстах, словарях и грамматиках и вносили ответы в новые словари и грамматики. Знание, которое открывается благодаря корпусам и тестируется на искусственном интеллекте, – это ответы на вопросы, которые лингвисты еще не формулировали. Остается открытым ключевой вопрос: достанет ли у лингвистики сил увидеть человеческие смыслы в умении общаться с ИИ.

Источники и литература

Collins Online English Dictionary. Glasgow: HarperCollins Publishers. В сети.

Collins Concise Dictionary, 2011: Glasgow: HarperCollins Publisher.

Нина Р. Добрушина, Михаил А. Даниэль (ред.), 2016: *Два века в двадцати словах* Москва: Издательский дом Высшей школы экономики.

[Nina R. Dobrušina, Mihail A. Danièl, (red.), 2016: *Dva veka v dvadcati slovah*. Moskva: Izdatel'skij dom Vysšej školy èkonomiki.]

Любовь Н. Засорина (ред.), 1977: *Частотный словарь русского языка. Около 40 тысяч слов*. Москва: Русский язык.

[Ljubov' N. Zasorina (red.), 1977: *Častotnyj slovar' russkogo jazyka. Okolo 40 tysjač slov*. Moskva: Russkij jazyk.]

Лидия Н. Иорданская, Игорь А. Мельчук, 2007: *Смысл и сочетаемость в словаре*. Москва: Языки славянских культур.

[Lidija N. Iordanskaja, Igor' A. Mel'čuk, 2007: *Smysl i sočetaemost' v slovare*. Moskva: Jazyki slavjanskih kul'tur.]

Geoffrey Leech, Paul Rayson, Andrew Wilson, 2001: *Word Frequencies in Written and Spoken English, based on the British National Corpus*. В сети.

Lonngren Lennart et. al., 1993: *Частотный словарь современного русского языка*. Uppsala: Acta Univ. Ups. (Studia Slavica Upsaliensia).

Ольга Н. Ляшевская, Сергей А. Шаров, 2009: *Новый частотный словарь русской лексики*. Москва: Азбуковник. В сети.

[Ol'ga N. Ljaševskaja, Sergej A. Šarov, 2009: *Novyj častotnyj slovar' russkoj leksiki*. Moskva: Azbukovnik. В сети.

Ольга Н. Ляшевская, Сергей А. Шаров, 2015: *Частотный словарь современного русского языка на материалах Национального корпуса русского языка*. Москва: Словари.ру.

Ol'ga N. Ljaševskaja, Sergej A. Šarov, 2015: *Častotnyj slovar' sovremennogo russkogo jazyka na materialah Nacional'nogo korpusa russkogo jazyka*. Moskva: Slovari.ru].
Macmillan English Dictionary for Advanced Learners, 2002. Oxford: Macmillan ELT
Macmillan English Dictionary for Advanced Learners, 2007. Oxford: Macmillan ELT.
Macmillan English Dictionary. Oxford: Macmillan ELT.
Надзея С. Мажэйка, Адам Я. Супрун, 1976–1992: *Частотны слоўнік беларускай мовы*. Мінск: Выдавецтва БДУ.
[Nadzeja S. Mažèjka, Adam Ja. Suprun, 1976–1992: *Častotny sloŷnik belaruskaj movy*. Minsk: Vydavectva BDU.]
Игорь А. Мельчук 1960: О терминах "устойчивость" и "идиоматичность". *Вопросы языкознания* 4. 73–80.
[Igor' A. Mel'čuk 1960: O terminah "ustojčivost'" i "idiomatičnost'". *Voprosy jazykoznanija* 4. 73–80.]

Povzetek

S splošnim naraščanjem števila korpusov, njihovega obsega in raznolikosti prihaja do specializacije korpusov glede na sestavo njihove ciljne vsebine. Prva generacija elektronskih korpusov (z obsegom približno 100 milijonov besednih oblik), ki se imenujejo ali dojemajo kot »nacionalni« ali »državni«, ohranjajo relativno ravnotežje podkorpusov ter široko socialno in humanitarno usmerjenost. Z večanjem obsega kasnejših korpusov se specializirajo na dveh vektorjih: 1) vsebinsko usmerjeni spremljajoči (dopolnjevani) megakorpusi časopisnih in revijalnih besedil; ciljne skupine vsebin korpusa tega razreda so sociologi in politologi, ekonomisti, demografi, novinarji itd.; 2) tematsko neomejeni (neselektivni) korpusi, ki akumulirajo digitalizirana besedila (tiskana in elektronska), ki se v računalništvu uporabljajo kot surovina za »procesiranje naravnega jezika«: za strojno predusposabljanje nevronskih mrež in izdelava statističnih algoritmov za samopovezovanje besed v ustrezne besedilne reakcije umetne inteligence.

Med najpomembnejšimi inovativnimi dosežki v korpusni leksikografiji sta navedena dva: 1) sinteza razlagalnih in frekvenčnih slovarjev v slovarjih *Macmillan* (2007), pozneje *Collins*, *Longman*; 2) komponentna pomenska analiza 100.000-besednega slovarja z uporabo 2500 najpogostejših leksemov v Macmillanu (2007) kot pomenskih komponent. Možnosti korpusov bodo kmalu pripeljale do velikega napredka v diahronem jezikoslovju.