

Н. Б. Мечковская, д. филол. н.

Белорусский государственный университет, Минск, Республика Беларусь

N. B. Mechkovskaya, Doctor of Sc. (Philology)

Belarusian State University, Minsk, Republic of Belarus

**КОРПУСНО-ЧАСТОТНАЯ ЛЕКСИКОГРАФИЯ,
МОДЕЛИ GPT И ТОЛКОВЫЕ СЛОВАРИ:
ВЕКТОРЫ ВЗАИМОДЕЙСТВИЯ**

**CORPUS-FREQUENCY LEXICOGRAPHY
GPT MODELS AND EXPLANATORY DICTIONARIES:
VECTORS OF INTERACTION**

Масштабы корпусных исследований ставят задачи по адаптации многоцелевых корпусов к потребностям и возможностям лингвистики. Показан ряд важных направлений во взаимодействии корпусов и толковых словарей: предварительная внутренняя диверсификация корпусов, лингвистически корректная селекция корпусных данных, их сужение, обобщение или концентрация.

Ключевые слова: предварительная внутренняя диверсификация корпусов; программные продукты SkELL, RuSkELL; конкордансы; коллокации; ChatGPT как анализатор полисемии.

The scale of corpus research poses the challenge of adapting multipurpose corpora to the needs and capabilities of linguistics. A number of important directions in the interaction of corpora and explanatory dictionaries are shown: preliminary diversification of corpora, linguistically correct selection of corpus data, their narrowing, generalization or concentration.

Key words: preliminary internal diversification of corpora; software products SkELL, RuSkELL; concordances; collocations; ChatGPT as a polysemy analyzer.

1. Один из базовых принципов создания электронных корпусов текстов и основанных на них частотных словарей состоит в проведении разнонаправленной диверсификации текстового и лексического материала. Благодаря диверсификации, корпусно-частотные словари (такие, как «Word Frequencies in Written and Spoken English, based on the British National Corpus» Джеффри Лича и сотрудников (2001); толковые онлайн-словари под брендами издательских домов *Macmillan*, *Collins* и др.; «Новый частотный словарь русской лексики» О. Н. Ляшевской и С. А. Шарова (2009) и др.) предстают как частотные списки слов, соотносимые с частеречной классификацией лексики и жанрово-стилистическим разнообразием речи (текстов). Это позволяет

исследовать статистическую организацию не только словаря, но и языка в целом – в его функциональных и жанрово-стилистических вариантах.

2. В словаре *Macmillan English Dictionary* (MED; 2007 г. в бумажном издании, 2009 – онлайн) представлены два выдающихся новаторских результата: 1) впервые в лексикографии осуществлен синтез толкового и частотного словарей; 2) впервые компонентный анализ лексической семантики был проведен на материале 100-тысячного словаря. В MED были выделены красным цветом первые по частоте 7500 лексем. Дальнейшие различия слов по частоте (их отнесенность к одному из четырех частотных диапазонов) сигнализировалось в начале словарной статьи о конкретном слове количеством звезд (или отсутствием звезды). В результате информация толкового словаря соединяется с информацией о частотах слов.

Словарные толкования значений 100 тыс. лексем MED – это, по сути дела, компонентный анализ лексической семантики (в его «мягкой» версии), проведенный в небывалых прежде «промышленных» масштабах. Важно, что был выдержан главный принцип компонентного анализа: имелся предварительно определенный список семантических компонентов. В MED его составили первые по частоте 2500 слов, и лексика дефиниций не выходила за границы этого списка.

Яркая особенность толковых онлайн-словарей (MED, Collins) состоит в эскалирующей подаче лексикографической информации; она размещается по ярусам, уходящим вглубь экрана, и на каждом ярусе информация расширяется – и парадигматически и синтагматически.

3. Качественно новый этап сближения корпусов и образовательной практики связан с активностью исследовательской компании *Lexical Computing*, основанной в Великобритании в 2003 г. Адамом Килгарриффом. Сервисы *Lexical Computing* используют набор из более чем 700 текстовых корпусов размером до 60 миллиардов слов и охватом более 100 языков. Широкую известность получила флагманская разработка Килгарриффа *Sketch Engine for Language Learning* (SkELL), которая включает облачное хранилище текстов живого языка и пакет программ для запросов к корпусу, управления им и анализа корпусных данных. Килгаррифф назвал свою разработку *Sketch Engine* («эскизный движок», или «устройство для скетчей (набросков, эскизов, зарисовок)»; программа может автоматически составить для любого слова его *word sketches* («эскизы») – одностраничную сводку грамматического и коллокационного поведения. В настоящее время *Sketch Engine* облегчает доступ к корпусам на более чем 90 языках и в 100 странах, в том числе

и в РФ – в модификации RuSkELL, созданной в Высшей школе экономики в Москве.

4. Далее на конкретных примерах обсуждается, в чем корпусные конкордансы и списки коллокаций могут помочь академическому словарю. Заслуживают внимания также и встречные вопросы: как справляться с избыточностью корпусных данных? В чем корпус не превзойдет толковый академический словарь (учитывая, что корпус, как и частотный словарь, не занят толкованием слов)? В какой мере (за пределами корпусных данных) бот *ChatGPT* может анализировать и редактировать словарные дефиниции?

5. В интернет-опосредованной лексикографии укрепляются обратные связи между создателями и потребителями онлайн-словарей (сайты «вокруг» брендовых словарей и корпусов; активизация средств рекламы и Public Relations), расширяющие общественное внимание к языку и коммуникации.