Джерела та література:
1. Кочерган М. П. Загальне мовознавство : підруч. / М. П. Кочерган. – 2-ге вид., розшир. та допов. – К. : Академія, 2006. – 464 с.
2. Вахтин Н. Б. Социолингвистика и социология языка : учеб. пособие / Н. Б. Вахтин, Е. В. Головко. – СПб. : ИЦ "Гуманитарная Академия"; Изд-во Европейского ун-та в Санкт-Петербурге, 2004. – 336 с.
3. Петренко А. Д. Некоторые методические аспекты социофонетики / А. Д. Петренко // Методы экспериментально-фонетического исследования звучащей речи. – К. : КГПИИЯ, 1991. – С. 17-30.
4. Храбскова Д. М. Фонетична еволюція французького мовлення етничних французько-німецьких білінгвів : дис. … канд. філол. наук : 10.02.05 / Д. М. Храбскова. – К., 2007. – 241 с.
5. Гируцкий А. А. Общее языкознание : учеб. пособие для студ. вузов / А. А. Гируцкий. – Минск : ТетраСистемс, 2001. – 304 с.
6. Лабов У. Исследование языка в его социальном контексте / У. Лабов // Новое в лингвистике. – М. : Прогресс, 1975. – Вып. 7 : Социолингвистика. – С. 96-181.
7. Беликов В. И. Социолингвистика / В. И. Беликов, Л. П. Крысин. – М. : Изд. центр РГГУ, 2001. – 439 с.
8. Ахманова О. С. Словарь лингвистических терминов / О. С. Ахманова. – 3-е изд. – М. : КомКнига, 2005. – 576 с.

Перепечкина С.Е. **УДК 81’33**

КОРПУСНАЯ ЛИНГВИСТИКА: ТЕОРЕТИЧЕСКИЕ И ПРАКТИЧЕСКИЕ АСПЕКТЫ

Целью данной статьи является рассмотрение центральных понятий корпусной лингвистики и возможностей практического использования ее достижений. Поиск новых подходов к изучению языка во всех формах его существования, применение современных компьютерных технологий в работе с языковым материалом, необходимость совершенствования методов обучения языку и приемам перевода выводят исследования в области компьютерной, прикладной и корпусной лингвистик на уровень **актуальных** и востребованных.

В целом, под корпусной лингвистикой понимается раздел языкознания, занимающийся разработкой общих принципов построения корпусов текстов, а также использованием их в качестве инструмента лингвистического исследования [1-3]. Описание языка, основанное на корпусном подходе, восходит еще к 13 веку (подробнее о нецифровых корпусах [4, с.13-18; 5, с.1-14]). В своей современной форме корпусная лингвистика существует лишь с начала 1960-х годов [6, с.15] и понятие корпус соотносится сегодня почти исключительно с наличием электронных собраний данных: “machine-readability is a *de facto* attribute of modern corpora” [7, с.6]. Лингвистический корпус – это «совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой» [2]. В англоязычной научной литературе встречается следующее краткое определение корпуса, данное Мак Энери: “A corpus is a collection (1) of machine-readable (2) authentic texts […] which is (3) sampled to be (4) representative of a particular language or language variety” [7, с.5]. Из определения следует, что таким свойствам текстов как **наличие их в электронной форме, аутентичность и репрезентативность** (представительность) придается особое значение. Знания о функциях языка, его структуре и об особенностях использования можно получить только на основе аутентичного языкового материала – таков основополагающий принцип корпусной лингвистики. Корпусы как раз и представляют собой “collection[s] of naturally occurring language text” [8, с.171], т.е. они составляются из речевых высказываний, которые возникли в естественном контексте с определенной коммуникативной целью [9, с.162]. Репрезентативность корпуса заключается в том, что «он должен соответствовать той области функционирования языка, которую будет отражать. Это значит, что корпус должен быть хотя бы минимально достаточным по объёму входящих в него отдельных текстов, чтобы можно было судить обо всей сфере» [1].

Обязательным элементом корпуса является разметка, или лингвистическое аннотирование – описание каждой единицы текста с помощью специальной системы отметок, содержащих дополнительную информацию о том, в каком значении употреблена та или иная форма. Разметка способствует, в частности, разграничению случаев лексической, грамматической или фонетической омонимии, напр.: *руки* – мн.ч., им.п. и *рук**и*** – ж.р., ед.ч., р.п. С помощью корпусного менеджера – специальной программы, подключенной к готовой базе данных – исследователь получает в виде списка все контексты (т.е. фрагменты текста с указанием источника и гиперссылками на исходный текст), в которых встретилось интересующее его слово или группа слов. Результат выдачи представляет собой конкорданс (множество контекстов, в котором встретилось запрашиваемое языковое выражение).

Первым большим корпусом на цифровом носителе считается Брауновский корпус (БК, англ.. *Brown Corpus*, BC [10]), созданный в начале 60-х годов при Университете Брауна (США) для частотного словаря американского варианта английского языка, он содержал 500 фрагментов текстов по 2 тысячи слов в каждом (1 млн. словоупотреблений). В 80-е годы XX века в Великобритании появляются б**о**льшие по объему собрания текстов – Банк Английского (Bank of English, насчитывающий сегодня более 2,5 миллиардов слов [11]) и Британский Национальный Корпус (БНК, British National Corpus, BNC [12]). В

СССР таким проектом был Машинный Фонд русского языка [2]. Создание и расширение универсальных корпусов того или иного языка является одним из приоритетных направлений корпусной лингвистики. Уже разработаны электронные корпусы для русского (Национальный корпус русского языка, ок. 140 млн. слов [13]), немецкого (LIMAS [14], COSMAS [15]), словенского, чешского, финского, новогреческого, китайского, японского и др. языков.

Детальная классификация корпусов может основываться на следующих **критериях**: 1) функциональность (цель, для которой создается или применяется корпус – для составления, к примеру, немецко-финского переводного словаря или для решения лингводидактических задач); 2) выбор языка/языков – одно-, дву- и многоязычные корпусы; многоязычные подразделяются далее на параллельные и сопоставительные; 3) по способу реализации – устные, письменные и мультимедийные корпусы; 4) устойчивость/константность – статичные и динамичные; 5) объем – референтные (объемные) и специальные (на неск. тысяч словоупотреблений).

Рассмотрим далее **типологию** корпусов в связи с возможностями их применения в исследовательских и учебных целях.

Лингвистический корпус может представлять эмпирическую основу для исследований в таких областях как лексикография, педагогика, прагматика, дискурсивный анализ, переводоведение и мн. др. Материал для изучения функционирования языка и его особенностей предлагают как **одноязычные** (к примеру, IdS-Korpora [16], составленные Институтом немецкого языка в Мангейме; British National Corpus; Национальный корпус русского языка), так и дву- либо **многоязычные** корпусы (пример – Oslo Multilingual Corpus [17]). Представительные корпусы, созданные для лексикографических целей, дают возможность наиболее полного исследования языка, поскольку составляются, в первую очередь, по принципу “no data like more data” [18, c.ix]. Отметим, что самым большим корпусом английского языка является на сегодня Google N-Gram Corpus, насчитывавший еще в 2006 г. 1 биллион слов [19]. Такие типичные **референтные** корпусы, как Bank of English, British National Corpus или DWDS-Kernkorpus [20] стремятся охватить наибольшее количество текстов из самых разных дискурсов – письменных и устных, официальных и частных. **Специальные** же корпусы всегда рассматривают лишь один отдельный вариант языка и могут послужить изучению его в какой-либо специальной области или коммуникативной ситуации, например, корпусы официальной речи или общения в чате (Corpus of American and British Office Talk или Das Dortmunder Chat-Korpus). **Статичные** (т.е. неизменные, по завершении их создания состав данных и структура не меняются; дополнения или изменения маркируются как новая версия) корпусы – самые распространенные – подходят для исследования стилей и жанров (корпусы художественной литературы, текстов писателей). Напротив, **мониторные** корпусы постоянно расширяются и предоставляют новые, актуальные «мгновенные фото» языка. С их помощью могут быть идентифицированы, к примеру, неологизмы или новые прочтения слов, исследована динамика языкового материала (корпусы по публицистике – ZEIT-Korpus [21], Korpus des Mannheimer Morgens [22] и др.).

Для сопоставительного переводоведческого анализа и обучения приемам перевода используются корпусы **параллельных** текстов – идентичные тексты на разных языках либо несколько переводов одного и того же оригинала. Примером таковых являются корпусы OPUS [http://opus.lingfil.uu.se/] и Europarl-Korpus [23], который содержит протоколы Европейского Парламента в переводах на разные языки; JRS-Aquis Multilingual Parallel Corpus [24] (тексты из области Европейского права на 22 языках). Электронные корпусы параллельных текстов – это ресурсы, из которых будущий переводчик может почерпнуть образцы профессионального перевода и которые могут послужить ему эталоном в процессе овладения этим мастерством. Время на обращение к реферативной информации (переводным или толковым и пр. словарям) при этом значительно сокращается. Весьма актуальным в современном индустриальном мире становится обучение переводу специальных текстов и освоение терминологии, что делает разработку корпусов научно-технических, публицистических и деловых текстов насущной необходимостью.

Материалы для сопоставительного дискурсивного анализа предлагает т.н. **сравнительный многоязычный** корпус, включающий различные тексты по дискурсам, напр., Википедия (статьи на множество тем на более 200 языках). Сравнительные корпусы могут представлять различные национальные варианты одного языка, как, напр., International Corpus of English [25].

Поскольку корпусный подход ориентирован на прикладное изучение языка, т.е. изучение его функционирования в естественной среде, то он чрезвычайно полезен в преподавании иностранного языка и лингвистической педагогике в целом. Так, лексикографический анализ на базе корпусов позволяет выявить контексты употребления той или иной синонимической единицы (*schmal – eng; lieber – besser; unentgeltlich – umsonst – kostenlos – kostenfrei – frei – gratis*), полнее раскрыть ее семантику, показать сочетаемость с другими словами и частотность ее употребления. Немаловажен и факт доступности ряда **письменных и устных** иноязычных корпусов, в том числе и **учебных**: Gutenberg Texts, British National Corpus Sampler, The Longman Corpus, International Corpus of Learner English (ICLE), LIMAS, COSMAS, корпусы новостей Рейтер, электронные архивы крупных газет и журналов (The Times, Zeit, Spiegel). Учебные корпусы отличаются от остальных тем, что содержат тексты лиц, изучающих иностранных язык, например, эссе студентов продвинутого языкового уровня (ICLE [26]). Анализ такого рода корпусов позволяет выработать эффективные методы освоения изучаемого языка, так как дает возможность выявить типичные лексические и синтактико-грамматические ошибки учащихся, установить их частотность, провести статистический анализ их вокабуляра. Кроме того, корпусы устных текстов дают представительную базу для проведения социофонетических исследований (например, речи детей, программа CHILD), поскольку включают транскрипции устных разговоров вместе с фонетической и просодической аннотацией.

Всестороннему исследованию коммуникации людей, которая в основе своей мультимодальна, способствуют **мультимедийные** корпусы, содержащие текст, изображение, аудио -и видеоматериалы, например, MUSA Multilingual Multimodal Corpus [http://sinfos.ilsp.gr/musa] или Das Archiv für Gesprochenes Deutsch [27] (архив немецкой разговорной речи).

Все вышесказанное позволяет сделать следующие **выводы**:

1. Достижения и наработки корпусной лингвистики открывают новые возможности для широкого спектра исследований теоретического и прикладного характера.

2. Метод анализа на базе корпусов текстов дает не только лингвистическую, но и квантитативную характеристику описываемой единицы и является надежным и достоверным, поскольку основывается на эмпирическом подходе, использует репрезентативный объем аутентичных текстов и современные компьютерные технологии для анализа и статистической обработки данных.

3. Электронные корпусы и корпусный подход могут эффективно использоваться как в учебном процессе (методике преподавания иностранного и родного языков, обучение переводу), так и в научно-исследовательской работе в области филологии.

Источники и литература:

1. Яскевич А. А. Корпусная лингвистика / А. А. Яскевич // Энциклопедия для школьников и студентов : в 12 т. / под общ. ред. В. И. Стражева. – Минск : Белорусская энциклопедия, 2009. – Т. 1 : Информационное общество. XXI век. – С. 167-169.
2. Корпусная лингвистика : [Электронный ресурс] // Википедия. – Режим доступа : http://ru.wikipedia.org/wiki/корпусная_лингвистика
3. Толдова С. Ю. Корпусная лингвистика : [Электронный ресурс] / С. Ю. Толдова. – Режим доступа : http://www.lomonosov-fund.ru/enc/ru/encyclopedia
4. Kennedy Graeme. An Introduction to Corpus Linguistics / Kennedy Graeme. – London, N. Y. : Longman, 1998. – P. 13-18.
5. Meyer Charles. “Pre-electronic Corpora” / Meyer Charles // Corpus Linguistics. An International Handbook. – Berlin : Mouton de Gruyter. – 2008. – № I. – P. 1-14.
6. Tognini Bonelli Elena. “Theoretical Oveview of the Evolution of Corpus Linguistics” / Tognini Bonelli Elena // Routledge Handbook of Corpus Linguistics. – London, N. Y. : Routledge, 2010. – P. 14-27.
7. McEnery, Richard Xiao. Corpus-based Language Studies: an Advanced Resource Book / McEnery, Richard Xiao, Yukio Tono. – N. Y. : Routledge, 2006. – P. 3-12.
8. Sinklair John. Corpus, Concordance, Collocation / Sinklair John. – Oxford : Oxford University Press, 1991.
9. Bergenholtz Henning. Zusammensetzung von Textkorpora für die Fachlexikographie / Bergenholtz Henning, Jette Pedersen. – Tübingen : Narr, 1994. – S. 161-176.
10. Brown Corpus Manual : [Электронный ресурс]. – Режим доступа : http://icame.uib.no/brown/bcm.html
11. Titania : [Электронный ресурс]. – Режим доступа : http://www.titania.bham.ac.uk/
12. British National Corpus : [Электронный ресурс]. – Режим доступа : http://www.natcorp.ox.ac.uk/
13. Национальный корпус русского языка : [Электронный ресурс]. – Режим доступа : http://ruscorpora.ru/
14. LIMAS : [Электронный ресурс]. – Режим доступа : http://www.korpora.org/Limas/
15. COSMAS : [Электронный ресурс]. – Режим доступа : http://korpora.ids-mannheim.de/~cosmas/
16. Das Deutsche Referenzkorpus – DeReKo : [Электронный ресурс]. – Режим доступа : http://www.ids-mannheim.de/kl/projekte/korpora
17. Oslo Multilingual Corpus : [Электронный ресурс]. – Режим доступа : http://www.hf.uio.no/ilos/forskning/prosjekter/spik/english/corpus/index.html
18. Sinklair John. Preface. Small Corpus Studies and ELT: Theory and Practice / Sinklair John. – Amsterdam : John Benjamins, 2001. – vii-xv.
19. Google N-Gram Corpus : [Электронный ресурс]. – Режим доступа : http://books.google.com/ngrams
20. DWDS-Kernkorpus : [Электронный ресурс]. – Режим доступа : http://www.dwds.de/resource/kerncorpus/
21. ZEIT-Korpus : [Электронный ресурс]. – Режим доступа : http://www.dwds.de/resource/zeitungskorpora/#part_3
22. Korpus des Mannheimer Morgens : [Электронный ресурс]. – Режим доступа : http://www.ids-mannheim.de/kl/projekte/korpora/archive/mm.html
23. European Parliament Proceedings Parallel Corpus 1996-2011 : [Электронный ресурс]. – Режим доступа : http://www.statmt.org/europarl/
24. The JRC-Acquis Multilingual Parallel Corpus : [Электронный ресурс]. – Режим доступа : http://optima.jrc.it/Acquis/
25. International Corpus of English : [Электронный ресурс]. – Режим доступа : http://ice-corpora.net/ice/
26. International Corpus of Learner English : [Электронный ресурс]. – Режим доступа : http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html
27. Das Archiv für Gesprochenes Deutsch : [Электронный ресурс]. – Режим доступа : http://agd.ids-mannheim.de/html/index.shtml