Г.В. КОЛПАКОВА
(Казань)

КОРПУСНАЯ ЛИНГВИСТИКА И ЛЕКСИКОГРАФИЯ

Охарактеризовано новое филологическое направление, активно развивающееся в Германии – корпусная лингвистика.

Ключевые слова: *корпусная лингвистика, дефиниция, лексикография, текст, интерпретация, метод.*

Корпусная лингвистика сформировалась в самостоятельное научное направление, достижения которого знаменуют новый этап в развитии научной мысли. Вышли в свет работы ведущих представителей компьютерной и корпусной лингвистики, освещающие кардинальные проблемы организации корпусов текстов и применения результатов их анализа в лингвистических исследованиях: K. Scherer «Korpuslinguistik» (2006), «Sprachkorpora – Datenmengen und Erkenntnisfortschritt» (2007), «Lexikalische Semantik und Korpuslinguistik» (2006), U. Langancke «Das Hypermedia-Online-Wörterbuch an der Schnittstelle zwischen Philologie, Kognition und Informatik» (2004). Поворотным моментом в исследовании корпусов текстов явилось требование изучать закономерности организации больших массивов текстов и особенности естественного речевого употребления, опираясь на статистические методы (методы количественного анализа), минуя этап предварительно сформулированного предположения-гипотезы и осуществляя качественный анализ (интерпретацию корпусных данных) лишь на последующем (конечном) этапе.

Для лексикологии проведение исследований при поддержке электронных корпусов стало общепринятым стандартом. На этом фоне перспективными оказываются лексикографические исследования, опирающиеся на данные корпусного анализа и позволяющие обнаружить влияние последнего на структуру словарных статей и выбор иллюстративных примеров. Методы корпусной лингвистики применяются для анализа парафраз значения словарных единиц с целью оптимизации дефиниций значения в электронных словарях. Вызывает дискуссию тема вариантов в корпусе. Ученых интересует вопрос о том, в какой степени в многофункциональном стандартизированном объемном корпусе необходимо учитывать варианты исследуемого феномена и в какой мере архитектура и дизайн словаря должны определяться факторами вариантности. Имеющийся корпус может оказаться недостаточно репрезентативным для языкового употребления определенного лингвосообщества. Однако как отдельный соразмерно организованный пропорциональный корпус он представляет собой срез многообразия коммуникативных жанров. В таком корпусе вариантность в любом случае будет присутствовать на уровне необработанных языковых фактов. При работе с языковым материалом иного культурного ареала возникает проблема соответствующей документации необходимого для анализа культурного фонового знания, особенно важного для изучения региональных разновидностей языков.

В представлении большинства лингвистов текстовый корпус представляет собой множество текстов, используемых в научном исследовании. К. Леманн отмечает, что понятие корпуса – основополагающего понятия корпусной лингвистики – за последние десятилетия претерпело ряд существенных изменений. Традиционно корпус рассматривался как совокупность текстов какой-либо определенной категории. Например, корпус текстов (рукописей) Платона. Для подобного понимания корпуса еще не имело значения, является ли доступной эта совокупность текстов как закрытая группа. Такое понимание стало следствием сужения понятия корпуса в лингвистике. Недавно предпринятое учеными новое сужение данного понятия потребовало электронной формы предъявления корпуса. Необходимость рассматривать корпус как закрытую совокупность текстов несет с собой коннотацию о компиляции (составлении) корпуса. Как следствие, возникает проблема достаточности, исчерпанности корпуса. Расширение понятия корпуса на более позднем этапе было обусловлено отказом от понимания объемных частей корпуса как текстов. Части корпуса могут быть представлены языковым материа-

лом любого вида: в лингвистический обиход вошли понятия корпусов предложений, суждений информантов, словарных статей (Lexikoneinträge). Изменение понятия корпуса на дальнейшем этапе заключалось в требовании естественного существования текстов до этапа анализа; компилятор не создавал тексты, его роль заключалась лишь в подборе текстов. Однако современные корпуса включают также языковой материал, специально созданный для них. Исследователь организует языковое употребление таким образом, что в результате образуется закрытое собрание текстов определенной категории, к которому имеется доступ. Требования наличия именно текстов и их существования как закрытых совокупностей сохраняются и сегодня, требования же исчерпанности корпуса и существования текстов, предшествующего их анализу, аннулируются [5, с. 16–17].

Й. Асмуссен трактует корпус как чрезвычайно объемное цифровое (digitalisierte) собрание текстов и текстовых отрывков, служащее репрезентативной выборкой для определенного, ограниченного на основе различных параметров языкового употребления, а в общеязыковом лексикографическом контексте являющегося целевой выборкой из языка в целом [1, с. 123].

Другой ученый, А. Клоза, указывает, что репрезентативность корпуса не трактуется в статистическом смысле: корпус – собрание текстов, составленное специально для проведения лингвистического анализа, которое, как предполагается, репрезентативно для определенного языка. Действенным является требование естественного существования высказываний (чаще в письменной, чем в устной форме), совокупность которых была компилирована (составлена) на основе установленных критериев отбора для достижения определенной цели. Эта совокупность текстов, по мнению А. Клоза, должна репрезентировать естественный язык [3, с. 106–107].

В представлении К. Шерер, корпуса служат отражением языка или его варианта. Под отражением понимается одномоментный языковой срез (eine sprachliche Momentanaufnahme), содержащий определенные тексты, отобранные на основе заранее заданных критериев. Наряду с неизменными, статическими корпусами в исследовании К. Шерер анализируются подверженные изменениям мониторные корпуса, а также аннотированные / неаннотированные, современные / исторические, одноязычные / многоязычные, общеязыковые / специальные корпуса, корпуса письменной / устной речи. Наряду с корпусами, содержащими лишь первичные данные, т.е. чистый текст, имеются корпуса, включающие дополнительную грамматическую и/или структурную информацию, например, о частеречной принадлежности или словоизменении (флексии). К. Шерер называет такие сведения, выходящие за рамки текста в корпусе и кодированные в тексте с помощью специальных средств (маркировок), аннотацией. С помощью нее имплицитная информация, содержащаяся в тексте, переводится в эксплицитную форму, т.е. делается видимой, что облегчает и ускоряет процесс сбора данных. Необходимо, чтобы аннотация не разрушала оригинальный текст, который после удаления маркировок должен предстать в изначальном виде. Такого рода корпуса получили название «аннотированных корпусов». Информация, выходящая за пределы текстов в корпусе, может представлять собой метасведения (Metadaten), т.е. сведения об отдельных текстах либо закодированную таким образом лингвистическую информацию о содержащихся в тексте элементах. Лингвистическая аннотация может быть представлена на различных уровнях языка. Так, на звуковом копируются признаки произношения (фонетическая аннотация), а также ударение и интонация (просодическая аннотация). Такая аннотация – редкий случай разметки корпуса, она характерна для корпуса Баварского архива языковых сигналов (die Korpora des Bayerischen Archivs für Sprachsignale – BAS). На уровне слова добавляется информация о флективных признаках или части речи (морфологическая аннотация), предложения – информация о типе фраз или синтаксических функциях (синтаксическая аннотация). На уровне значения кодируются семантические признаки слов или семантические отношения между элементами в тексте (семантическая аннотация). Аннотация в дискурсе или тексте охватывает такие феномены, как маркировка вежливости, такие лингвистические явления, как пролепсис (предвосхищение) или возобновление, повтор определенных содержательных моментов в тексте. Возможна и аннотация определенных типов ошибок в корпусе, включающем тексты изучающих иностранный язык (проблемно ориентированная аннотация) [8, с. 20–22].

Наиболее распространенная форма аннотации – это грамматическая аннотация на уровнях слова и предложения. Корпуса, аннотированные на уровне предложения, называют древовидными (Baumbank), т.к. информацию на уровне предложения часто представляют в виде структуры дерева. В отличие от чисто текстовых и аннотированных на уровне слова корпусов, в древовидных корпусах в центре внимания исследователя оказывается не слово, а предложение или фраза. Самым известным немецкоязычным древовидным корпусом является TIGER-корпус, совместный проект университетов в Потсдаме, Саарбрюккене и Штуттгарте. В настоящее время он охватывает около 50000 предложений с 900000 текстовых слов (Textwörter). Другими известными немецкоязычными древовидными корпусами являются Саарбрюкский NEGRA-корпус и 3 древовидных корпуса в Тюбингене. NERGA-корпус – предшественник TIGER-корпуса, первый немецкоязычный древовидный корпус. В своей второй актуальной версии он охватывает 20602 предложения с 355000 текстовых слов. Как и TIGER-корпус, NEGRA-корпус базируется на газетных статьях («Frankfurter Rundschau»). Два из трех Тюбингенских древовидных корпусов являются корпусами немецкого письменного языка и основываются на газетных текстах. Третий Тюбингенский корпус (die Tübinger Baumbank des Deutschen/Spontansprache) основывается на устной речи и содержит 38000 транскрибированных предложений и около 360000 текстовых слов, изъятых из собраний диалогов по теме «Договоренность о встрече».

Важным признаком корпусов является языковой медиум, из которого отбираются тексты для корпуса. Число корпусов, базирующихся на текстах письменного языка, намного превышает количество корпусов устной речи. Имеются корпуса, принадлежащие устной речи или объединяющие устные и письменные формы. Корпуса письменного языка доминируют по двум причинам. Затраты времени на создание письменного корпуса существенно меньше, чем на образование устного корпуса. Не менее важен также и доступ к материалу. Письменные тексты существуют в бумажной версии, частично в электронном виде в форме Textdatei (собрание текстов на дискете или магнитофонной ленте). Тексты же устной речи, как полагает К. Шерер, сначала необходимо записать, транскрибировать (т.е. перевести их в письменную форму), и только потом включить в корпус. Только благодаря транскриптам устную речь в большом объеме возможно подвергнуть изучению на основе системных методов [8, с. 21–23].

На современном этапе развития лингвистики при решении научных проблем ученые стремятся использовать результаты количественного и качественного анализов данных в корпусах текстов. Важнейшее различие между этими двумя методами заключается не в том, какие проблемы исследуются, а в том, каким образом они исследуются, утверждает К. Шерер. Например, влияние англицизмов и галлицизмов на немецкий язык исследовали два автора – Шанке (2001) и О'Халлоран (2002). Оба работали с собственными корпусами, составленными из текстов газет и журналов. Корпус Шанке содержит все выпуски немецкой газеты «Handelsblatt» за март 2000 г., а корпус О'Халлоран включает корпус языка моды (выпуски немецкого журнала для женщин «Brigitte», датированные различными годами) и корпус стандартного языка (относимые к разным годам выпуски немецкого журнала «Stern» и немецкой газеты «Berliner Illustrierte Zeitung»). Шанке применяет качественный метод анализа, О'Халлоран – количественный. Цель при использовании качественного метода заключается в том, чтобы выявленные иностранные слова классифицировать по частям речи и отнести их к различным тематическим областям, например, «компьютер», «биржа», «банковское дело». Качественный анализ предполагает выявление, классификацию и интерпретацию феноменов. При количественном анализе изучалось распространение англицизмов и галлицизмов в течение последних ста лет. В результате анализа было установлено, что количество типов иностранных слов в корпусе в целом растет, а именно: с 0,6% в год (1902 г.) до 2,0% в год (1997 г.). Кроме того, доля иностранных слов (точнее, их словоформ) в корпусе языка моды (14%) в любой момент времени превышает долю иностранных слов в корпусе стандартного языка (4%). Цель применения количественного метода состоит в том, чтобы выявить частотность определенных феноменов и сравнить их с целью подведения итогов исследования. По количественным показателям стандартным образом вычисляется величина корпуса, она измеряется в текстовых

словах и является важнейшей исходной (основной или условной) величиной для количественного анализа. Если величина корпуса неизвестна, то количественный анализ имеет смысл только в том случае, когда можно сравнить результаты для многих подобных феноменов внутри корпуса. Особый интерес для исследователя представляет количество типов (Type) и реализаций (Token) анализируемого феномена, т.к. обе величины, их соотношение дают представление о том, как часто данный феномен встречается в тексте (Token) и на какие различные типы и в каком количестве (Type) эти реализации распределяются. Если количество реализаций, приходящихся на один тип, высокое, то они представляют собой часто употребляемые выражения, количество же новых, спонтанно образованных по этому типу форм невелико. Если какой-либо тип встречается в корпусе единственный раз, говорят о Hapaxlegomenon. Если корпус содержит многие Hapaxlegomenon и иные редкие типы, то высока вероятность того, что этот языковой образец продуктивно используется говорящими и на его основе могут быть созданы другие образования.

Если исследование корпуса в большей степени ориентировано на уровень предложения, чем на уровень слова, то важно выявить количество предложений в корпусе или тексте. При этом такие данные, как средняя длина предложения, количество предложений с определенной длиной, а также доля предложений с определенным числом слов, могут помочь определить синтаксические характеристики корпуса. Чтобы быть уверенным, что полученные результаты не случайны, рекомендуется подтвердить их статистически – путем применения так называемого «теста на значимость» (Signifikanztest). Однако большинство подобных тестов относятся к области высшей математики, поэтому К. Шерер советует использовать соответствующие статистические программы [8, с. 35–37].

Количественный и качественный методы анализа, отмечает А. Люделинг, применимы как к тексту корпуса, так и к уровням аннотации в корпусе. Любой вид аннотации, по мнению исследователя, является категоризацией и представляет собой неизбежную контролируемую потерю информации. Каждый способ категоризации – это также интерпретация данных. Нередко в больших корпусах каждому уровню аннотации (например, уровню частей речи или значений) сопутствует уровень интерпретации. В последние годы наряду с линейно аннотированными корпусами получили распространение многоуровневые модели корпусов, в которых все уровни аннотации сохраняются независимо от текста. В диахронической лингвистике и диалектологии ученые традиционно использовали данные корпусов текстов, не имея иных источников экспериментального материала. В синхронической лингвистике корпуса текстов, рассматриваемые как банк данных (источник материала), все в большей степени находят применение в теоретических исследованиях. Наряду с использованием корпусов как банка данных при проведении качественного анализа лингвисты все чаще применяют методику количественного анализа: статистические тесты, анализ коллокаций, т.е. сочетаемости лексем, мультивариантные методики. Но в отличие от таких эмпирических областей исследования, как психолингвистика, в корпусной лингвистике разработка стандартов количественного анализа корпусных данных находится в стадии дискуссий. Помимо принципов осуществления подсчета важное значение имеет вопрос о том, что подвергнуто количественной обработке. Не останавливаясь на математических операциях, А. Люделинг заостряет свое внимание на основе любого количественного анализа, а именно на качественном анализе или категоризации данных.

Любой количественный анализ зависит от предшествующего этапа категоризации, последняя же не всегда бесспорна. Нередко в исследованиях, базирующихся на количественном анализе корпусов, отсутствуют сведения о проведенной категоризации, применяемых категориях (Tagsets) и критериях выделения категорий. Еще реже можно встретить сведения, подтверждающие надежность заданных категорий. Без этих сведений результаты количественного анализа не могут считаться достоверными. А. Люделинг считает неправомерным отказ от категоризации даже в случае отсутствия видимых оснований для эксплицитной категоризации, показывает на примере корпусного исследования (Lernerkorpus Falko), насколько значительным может быть влияние категоризации на результаты количественного анализа [7, с. 28–29].

Основная единица анализа эмпирической лингвистики – это языковое высказывание. Корпуса всегда представляют собой интерпретируемые репрезентации подобных высказываний и отличаются от исходных, не подвергнутых интерпретации. Таким образом, между высказыванием в опубликованной газетной статье и корпусом, в котором оно представлено, проходит разделительная грань. С одной стороны, отличаются возможности реагирования и анализа (например, невозможно направить корпусу письмо читателя, но можно провести статистический анализ), с другой – различия наблюдаются в информации, содержащейся в контексте. Это различие становится очевидным, если подумать об электронных версиях исторических манускриптов или транскрипциях разговорного языка. В любом случае различные формы оригинала должны быть сведены к одинаковым формам в корпусе. Оценка корпусных данных происходит на двух уровнях интерпретации: более низком уровне категоризации и на более высоком уровне аннотации. Решения принимаются на обоих уровнях, исследователь никогда не имеет дело с необработанными данными. Исходное утверждение о том, что интерпретация корпусных данных всегда есть категоризация, т.е. предсказуемая контролируемая потеря информации, имеет силу как при качественной, так и при количественной оценке данных, при работе как со сведениями, не имеющими эксплицитной аннотации, так и с данными, не обладающими таковой, вне зависимости от способа обработки корпусных данных. Так, в целях наглядного объяснения сложного грамматического явления в иностранном языке обучаемому достаточно продемонстрировать ряд подобных примеров. В этом случае корпус используется как «банк данных». Однако соответствующие примеры отбираются из корпуса и группируются на основе эксплицитных «внешних», т.е. выделенных исследователем критериев [7, с. 29–30].

В небольших специальных корпусах, обрабатываемых вручную, категоризация может существенным образом повлиять на результаты анализа. Однако и в объемных корпусах (например, в Институте немецкого языка в Маннхайме или Академии наук Берлин-Бранденбург), а также в корпусах, применяемых в компьютерной лингвистике, категоризация играет важную роль. Подобные корпуса обрабатываются автоматически, что предполагает разметку корпусных данных по частеречной принадлежности, типу основного слова. В большинстве случаев обработка корпуса происходит поэтапно. Сначала выделяется определенное явление (высказывания группируются в соответствии с ним), затем производится группировка выделенных образцов по выделенному слову, осуществляется разметка по частям речи, вслед за этим рассматриваются более крупные единицы, например, предложения.

Возникает вопрос о возможном отличии способов обработки стандартных и специальных корпусов (например, корпуса диалектов, разговорного языка, исторического, учебного корпусов). Для специальных корпусов еще не разработаны стандарты аннотирования. Чем менее стандартизированными оказываются данные, тем в меньшей степени удается использовать имеющиеся программы и методы аннотирования, поскольку последние по большей части содержат статистический компонент и, следовательно, ориентированы на обнаружение закономерностей. Именно поэтому специальные корпуса обрабатываются вручную. Но в этом случае исследователи сталкиваются с серьезной проблемой: зачастую результаты таких разнонаправленных анализов оказываются спорными, противоречивыми (Там же, с. 31–32).

Широкий спектр технических возможностей и безграничное количество примеров в электронных корпусах побуждают лингвистов к поиску точной дефиниции лингвистического феномена «экспериментальные данные». Л.М. Айхингер указывает на возможные методические подходы к решению этой проблемы. С одной стороны, ввиду стремительного развития технических возможностей создания языковых корпусов и кардинального изменения состояния исследований в данной области невозможно оставить незамеченными результаты этого развития, с другой – влияние этих результатов ощущается и в области теоретических работ. Даже те авторы, которые всецело доверяли интроспекции как исследовательскому методу, рассматривая реальные факты как неточные рефлексы (verschmutzte Reflexe) абстрактных принципов, видят в анализе корпусов возможность дальнейшего расширения теоретического знания о языке. В наибольшей степени подобный подход отвечает интересам лингвистов,

прокладывающих путь к эмпирическому лингвистическому познанию посредством подтверждения гипотезы корпусом примеров.

Эти способы приближения к языковым данным в результате интенсивного развития корпусных технологий приобрели статус, позволяющий рассматривать их в равноценном противопоставлении друг другу, как отмечает Л.М. Айхингер. Стремление к документации языковой реальности выглядит при различных подходах к анализу языковых фактов по-разному. Сбор данных, осуществляемый лингвистом для подтверждения выдвигаемой им гипотезы, и в еще большей степени интроспекция, апеллирующая к языковой компетенции, основываются на относительной независимости производства речевых высказываний от говорящего субъекта, обладающего более ощутимой властью над создаваемыми им продуктами, нежели в случае с естественными объектами или артефактами. Корпусные исследования, напротив, представляют собой попытку приблизиться к языковой реальности посредством статистического анализа и математического моделирования в отвлечении от субъекта, его языковой компетенции и метода интроспекции [2, с. 2].

Что же представляет собой лексикография, опирающаяся на корпус (korpusgestützte Lexikographie)? Исследователи указывают на необходимость дифференциации автоматически созданных информационных словарных систем и лексикографически обработанных словарей. Под лексикографической обработкой понимается любой способ интеллектуальной обработки человеком автоматически подготовленных данных, начиная с их перепроверки и сортировки вплоть до комментирования. Подобным лексикографически обработанным напечатанным или электронным словарям посвящено исследование А. Клоза. Автоматически же произведенные словарные информационные системы в ее работе только упоминаются, поскольку, по мнению ученого, их становление связано с дальнейшим развитием электронных собраний текстов и развитием корпусных и информационных технологий. Корпусом может называться совокупность текстов в электронной форме, доступ к которой обеспечивается тщательно разработанными исследовательскими компьютерными программами поиска и анализа (Recherche- und Analysesoftware). Сегодня от такой компьютерной программы при проведении лексикографических исследований можно ожидать следующих функций:

– дифференциация основного слова и словоформы, обнаружение основного слова (Lemmatisierungsprogramm);

– составление листов частотности основных слов или словоформ, отсортированных по частотности употребления или по алфавиту;

– составление так называемых KWICs (Key-Word-in-context-Zeilen), т.е. строчек контекстов с основным словом без привязки их к началу или концу предложений, которые могут получить дефиницию и быть отсортированы на основе различных критериев (в алфавитном порядке или по предыдущему/последующему слову);

– анализ условий совместной встречаемости различных классов слов, в результате которого на основе математически-статистических методов выявляются важные закономерности употребления комбинаций слов.

Подобная программа была подготовлена Институтом немецкого языка в Маннхайме [3, с. 106–107].

В лексикографической практике главенствует положение о том, что для исследуемого корпуса в идеальном случае должен быть составлен собственный корпус. Наряду с электронным корпусом в основе словаря могут лежать и другие первичные источники, в частности, традиционные карточки с примерами, находящиеся в специальном архиве. Такой архив может быть представлен в электронном виде, как это сделано, например, в редакции Duden. Цифровая форма архива повышает степень его полезности, т.к. становятся возможными поиск полных текстов путем рассмотрения всех текстовых примеров и соотнесение отдельных примеров реализаций с основными формами слов. Однако электронный архив примеров не идентичен корпусу в понимании А. Клоза. При проведении лингвистического анализа обнаруживается, что тексты примеров являются лишь условно репрезентативными, хотя вы-

бор текстов, составляющих основу архива примеров, и был репрезентативным. Все дело в том, подчеркивает А. Клоза, что люди, отбирающие материал, склонны скорее к тому, чтобы фиксировать на карточках необычное как нормальное, поэтому при подготовке первичных источников для словаря следует обратить особое внимание на проблему соответствующего отбора материала как для архива примеров, так и для корпуса. И корпус, и архив должны быть надежным отображением лексемного состава и употребления слов анализируемого языка, а это не всегда оказывается достижимой целью.

Наряду с корпусами и архивами примеров (в качестве первичных источников) лексикографы используют в своей работе электронные собрания текстов, например, собрание сочинений известного автора или цифровые газетные архивы на CD-ROM, тексты в Интернете. Но в этом случае речь не идет о корпусах, поскольку эти источники не отвечают важному критерию: они не были отобраны с целью проведения анализа на основе определенных критериев и не могут быть интерпретированы посредством применения специально разработанных компьютерных программ. Запросы в Интернете имеют другую цель: разыскать документы в Сети [3, с. 107–109].

А. Клоза разрабатывает классификацию методов анализа корпуса при составлении словарей. Первая оппозиция «korpusgestützt (опирающийся на корпус) – korpusgebunden (привязанный к корпусу)» получает следующее толкование: в отличие от словаря, опирающегося на корпус, привязанный к корпусу словарь создается исключительно на базе корпуса словаря без дополнительных первичных, вторичных и прочих источников. Вторичными источниками являются все словари, к которым обращаются при лексикографической работе как вспомогательному средству, к третьим источникам относятся все прочие используемые языковые материалы, например, лингвистические монографии и грамматики. Привязанный же к корпусу словарь точно отражает языковую реальность, которую репрезентирует сам корпус, составляющий основу словаря. Примером подобного словаря служит, например, словарь одного автора или словарь определенного текста. Метод опоры на корпус при разработке словаря нашел применение в практике составления толкований значений, например, в Кембриджском международном словаре английского языка. Такой словарь не дает точного отображения языковой реальности корпусом, составляющим его основу, но дополняет картину, полученную на основе анализа корпуса. Опирающаяся на корпус лексикография преследует и другую цель: корпус является не только первичным источником, образующим основу словаря, но и предметом изучения с помощью методов корпусной лингвистики. Различия между словарями двух типов главным образом методические (Там же, с. 110–111).

В настоящее время в лексикографической науке во всех странах преимущество отдается лексикографии, опирающейся на корпус, использующей наряду с первичным источником – корпусом – иные, вторичные и прочие источники. Как представляется, пользователю важнее получить полную картину происходящего в языковой реальности, нежели копировать саму языковую реальность.

В качестве второй А. Клоза предлагает оппозицию «korpusvalidierend (определяющий ценность корпуса) – korpusgesteuert (управляемый корпусом)». В лексикографии, опирающейся на корпус (korpusgestützte Lexikographie), это два подхода к анализу корпусных данных. Под термином «определяющий ценность корпуса» понимают метод, использующий корпус для того, чтобы изложить теорию, провести описание, проконтролировать данные, сделать корпусные данные наглядными. В противоположность этому подходу термин «управляемый корпусом» означает, что корпус используется для совершенно иных целей, нежели для поиска примеров, подтверждающих определенные лингвистические тезисы или определяющих ценность теоретических гипотез. В формулировке авторов словаря Wahrig суть этого подхода определена следующим образом: эмпирически все примеры и все случаи употребления слов защищены (подтверждены) выравниванием с цифровым текстовым корпусом Wahrig, собранием 500 млн слов в электронной форме, отражающим актуальное употребление слов современного немецкого языка (Там же, с. 111–112). Видимо, главное в подходе, управляемом корпусом, – очевидность языкового факта. Исследователь не вносит в анализ свои гипотезы, не опирается на метод интроспекции, а руководствуется корпусом, выделяя реальные закономерности.

Лексикография, опирающаяся на корпус, получила свое развитие в области создания учебных словарей современного языка при описании значения слов и их применения. Расширение области использования данных методов в плане создания специальных словарей (словарей профессиональных жаргонов, диалектов, антонимов, региональных словарей) требует дальнейших исследований. Применение и оценка корпуса как метода исследования играет важную роль при создании листа опорных слов (Stichwortliste), представленных в словаре. При переработке уже имеющихся словарей и подготовке новых изданий системное использование корпусов может способствовать выявлению новых слов-кандидатов для включения в лист опорных слов, а также обнаружению в словаре устаревших образований. Для создаваемых словарей могут быть разработаны путем системной оценки корпусных данных совершенно новые листы опорных слов. Важно отметить, что традиционно слова включаются в этот список в своей основной назывной форме, т.е. в форме именительного падежа единственного числа. Но в корпусах текстов такие формы слов встречаются очень редко. В текстовом корпусе обнаруживаются преимущественно словоформы в косвенных падежах, которые путем анализа или использования компьютерной программы для выявления основной формы слов (Lemmatisierungsprogramm) могут быть сведены к одной форме. Дж. Синклер задается вопросом, должно ли быть так, что основная форма, зафиксированная в качестве опорного слова в словаре, почти не имеет подтверждения в корпусе. Не может ли иная словоформа, подтвержденная многочисленными примерами в корпусе, стать опорным словом в словаре? Однако большинство словарей придерживаются традиционной трактовки опорного слова. Безусловно, это сделано ради удобства пользователей в соответствии с общепринятой методикой составления списка опорных слов. При разработке совершенно нового списка опорных слов для проекта словаря можно идти двумя путями: выбрать уже имеющиеся листы с опорными словами, произвести их оценку, использовать их в модифицированном виде или составить на базе корпусов новые списки опорных слов. Неизбежно возникает проблема необходимой редакционной перепроверки слов-кандидатов в лист опорных слов создаваемого словаря. Такой редакционный подход в принципе соответствует методу, управляемому корпусом, т.е. количественному методу анализа корпусных данных. При этом лексикограф может проанализировать, оценить, отсортировать данные корпуса и в заключение описать картину, представленную корпусом [3, с. 114–115].

Итак, исследование с опорой на корпус способствует ускорению анализа: удается быстро выявить слова-кандидаты в опорные слова и создать их объемные списки. Однако сложно оценить, будет ли происходить лексикографическая перепроверка слов-кандидатов и разработка списка действительных опорных слов быстрее, чем это происходит в результате компиляции списков опорных слов из уже существующих словарей. Следует также иметь в виду: если в основу словаря положен динамический корпус, то список опорных слов должен обновляться через регулярные промежутки времени. Будут ли списки опорных слов качественнее, если они разрабатываются с опорой на корпус, сказать сложно. Эти списки отражают корпус и окажутся хорошими, если корпус соответствует цели словаря. В противном случае список опорных слов должен быть дополнен редакцией и откорректирован. Если такой список компилируется из других источников, следует обратить внимание на критерии отбора и полноту, избыточность списка опорных слов (Там же, с. 116).

Современная лексикография находится в чрезвычайно напряженной фазе своего развития: она неизбежно должна и далее анализировать методы и возможности корпусной лингвистики, которая в свою очередь осваивает методы и возможности лексикографии.

Литература

1. Asmussen J. Korpuslinguistische Verfahren zur Optimierung lexikalisch-semantischer Beschreibungen // Sprachkorpora – Datenmengen und Erkenntnisfortschritt (Hrsg. von W.Kallmeyer, G.Zifonun). Institut für Deutsche Sprache. Jahrbuch 2006. Berlin – N.J. : Walter de Gruyter, 2007. S. 123–151.

2. Eichinger L.M. Linguisten brauchen Korpora und Korpora Linguisten // Sprachkorpora – Datenmengen und Erkenntnisfortschritt. (Hrsg. von W. Kallmeyer, G. Zifonun). Institut für Deutsche Sprache. Jahrbuch 2006. Berlin – N.J. : Walter de Gruyter, 2007. S. 1–8.

3. Klosa A. Korpusgestützte Lexikographie: besser, schneller, umfangreicher? // Sprachkorpora – Datenmengen und Erkenntnisfortschritt. (Hrsg. von W.Kallmeyer, G.Zifonun). Institut für Deutsche Sprache. Jahrbuch 2006. Berlin – N.J. : Walter de Gruyter, 2007. S. 105–122.

4. Langancke U. Das Hypermedia-Online-Wörterbuch an der Schnittstelle zwischen Philologie, Kognition und Informatik. Am Beispiel des Wortschatz-Lexikons // Abgründe und Brücken. Festgabe für Regina Hessky. (Hrsg. von R. Brdar-Szabo und E. Knipf-Komlosi). Frankfurt am Main : Peter Lang, 2004. S. 379–393.

5. Lehmann Chr. Daten. Korpora. Dokumentation // Sprachkorpora – Datenmengen und Erkenntnisfortschritt (Hrsg. von W. Kallmeyer, G. Zifonun). Institut für Deutsche Sprache. Jahrbuch 2006. Berlin – N.J. : Walter de Gruyter, 2007. S. 9–27.

6. Lexikalische Semantik und Korpuslinguistik. (Hrsg. von W. Dietrich). Tübinger Beiträge zur Linguistik. 1 Aufl. Tübingen : Narr, 2006. Bd. 490.

7. Lüdeling A. Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik // Sprachkorpora – Datenmengen und Erkenntnisfortschritt. (Hrsg. von W.Kallmeyer, G.Zifonun). Institut für Deutsche Sprache. Jahrbuch 2006. Berlin – N.J. : Walter de Gruyter, 2007. S. 28–48.

8. Scherer C. Korpuslinguistik // Kurze Einführungen in die germanistische Linguistik. (Hrsg. von J.Meibauer u. M.Steinbach). Heidelberg : Universitätsverlag Winter, 2006. Bd. 2.

9. Sprachkorpora – Datenmengen und Erkenntnisfortschritt. (Hrsg. von W.Kallmeyer, G.Zifonun). Institut für Deutsche Sprache. Jahrbuch 2006. Berlin – N.J. : Walter de Gruyter, 2007.

Case linguistics and lexicography

There is characterized the case linguistics – a new philological school, which is actively developing in Germany.

Key words: *case linguistics, definition, lexicography, text, interpretation, method.*