Slavica Helsingiensia 34
Инструментарий русистики: корпусные подходы, Хельсинки, 2008
Под ред.: А. Мустайоки, М.В. Копотева, Л.А. Бирюлина, Е.Ю. Протасовой

М.В. Копотев, А. Мустайоки
(Хельсинки)

Современная корпусная русистика

„Von dem jedesmal *Gesprochenen* ist die *Sprache*, als die Masse seiner Erzeugnisse, verschieden“
(Humboldt 1836: 60)[1]

Интуиция VS текст — два источника лингвистической информации

Со времен В. фон Гумбольдта языковеды осознают двойственный характер объекта своего научного интереса. Позже эта двойственность была многократно дополнена и уточнена, в частности, Ф. де Соссюром, Э. Косериу, Н. Хомским. Несмотря на общепринятость утверждения о разных формах существования языка, лингвистика, тем не менее, не обладает общепринятой теорией, объясняющей двойственный (или тройственный) характер языка. Не имеет однозначного решения вопрос о том, как это разделение должно отражаться на методике проведения лингвистических исследований. Что, в конце концов, является объектом изучения в лингвистике? То, что находится в сознании носителей языка (*langue*), то, что они говорят (*parole*), или этой дихотомии не существует? Ученые нередко утверждают, что это две формы существования одного и того же явления. В целом, никто не отрицает, что *langue* и *parole* — тесно взаимосвязанные явления. Несмотря на это, они представляют собой разные объекты исследования, требующие разного исследовательского аппарата.

Двойственный характер объекта исследования конкретизируется в базовом вопросе исследовательской практики: на каком основании ученый-лингвист может утверждать, что в языке L существует языковая форма X (лексема, словоформа, синтаксическая структура)? На этот вопрос существует два принципиально разных ответа: единица X существует, поскольку (1) носитель языка считает данную форму правильной (акцент на интуиции); (2) носители языка употребляют ее (акцент на языковом материале). В обоих случаях необходимо определить источник лингвистической информации: мнение *каких* носителей языка принимается во внимание при определении *langue* (интуиция самого лингвиста, близкого круга его знакомых, «носителей литературного языка» в целом и т.д.), и *какие* тексты представляют *parole* (устные ~ письмен-

ные, официальные ~ бытовые ~ художественные тексты) (ср. Мустайоки 1988, 1995). Сравнительно распространенное решение этой дилеммы строится на совмещении подходов: описывается *langue*, но анализ основывается на тщательно отобранном (письменном) языковом материале — *parole*.

В большинстве случаев два указанных подхода (опора на языковой материал и интуицию) ведут к схожим результатам: и тот, и другой подтверждают наличие, например, в русском языке таких слов и словоформ, как *университет* или (*много*) *заводов*. Однако если тот же вопрос касается таких языковых единиц, как *мерчандайзинг* или *(много) апельсин* ответ уже не столь однозначен (часто квалифицированный и образованный носитель языка отрицает существование подобных слов, не замечая их в собственной речи или речи своего окружения)[2]. Конечно, в общих описаниях языка, грамматиках и словарях, строгая шкала «правильно / неправильно» смягчена с помощью стилистических помет типа *разг.*, *прост.* и т.п, но это не снимает вопроса, на чем они основываются — на интуиции составителя или на широком языковом материале.

Уточнение методики лингвистических исследований необходимо не только из практических соображений — это касается базового принципа любой науки: ученый должен четко определить, на чем основываются его суждения об изучаемом объекте. В научной практике в этой связи принято употреблять слово *evidence* (доказательство). Фундаментальная статья М. Пенке и А. Розенбах (Penke & Rosenbach 2004) так и называется: *What counts as evidence in linguistics?* Согласно общим требованиям к научному знанию способы и приемы приобретения научной информации должны быть прозрачными в такой степени, чтобы другие ученые могли проверить достоверность результатов исследования, используя ту же самую методологию. В этом, собственно, и состоит принцип верифицируемости научного знания. Следование этому принципу в лингвистике становится в последнее время существенно более простым, поскольку на помощь лингвистам пришли технологии, сделав возможным создание больших по объему совокупностей языкового материала, которые стали называть **корпусами**, а направление лингвистики, использующее эти инструменты, получило название **корпусная лингвистика**[3].

Как замечают Т. Макинери и А. Вильсон (McEnery & Wilson 2001: 2-4 и далее; см еще Dash 2008), корпусные методы возникли задолго до появления электронных корпусов. Так, например, диахронические исследования опирались и опираются на некоторый, часто весьма ограниченный «корпус текстов», на основе которого делаются выводы о том или ином языковом явлении. Лишенный возможности производить интроспективные выводы, историк языка использует заданную совокупность

текстов. И переход на электронные формы хранения диахронного материала по сути не меняет методологию исследования, поскольку опора на интуицию в указанном смысле при изучении исторического материала по естественным причинам исключена. Такой же подход свойствен, как правило, любым полевым исследователям, например диалектологам, которые работают с естественным, но часто чужим для них языковым материалом, т.е. с речью носителей определенного диалекта.

Если кроме языкового материала в распоряжении исследователя есть и языковая интуиция, роль первого часто снижается. Закономерно спросить, оправдана ли опора на языковой материал только в том случае, когда нет возможности опираться на собственное чувство языка или вообще на мнение носителей языка? Мнения ученых при ответе на этот вопрос варьируются, однако в целом можно наметить четыре подхода:[4]

1) языковой материал (корпус) как предмет интереса лингвиста целиком отрицается;
2) корпус не только принимается во внимание, но и создание лингвистической теории и системы понятий основывается именно на корпусном подходе;
3) языковой материал используется в изучении отдельных языковых явлений как источник описания *langue*;
4) языковой корпус изучается, считаясь самостоятельным и исчерпывающим объектом исследования.

Рассмотрим кратко, что представляют собой каждый из этих подходов.

(1) Тотальное отрицание значения языкового материала как источника для выводов и обобщений относительно языка можно найти прежде всего у Н. Хомского. Так, в интервью И. Андору он утверждает: “Corpus linguistics doesn’t mean anything. It’s like saying <…> suppose physics and chemistry decide that instead of relying on experiments, what they’re going to do is [to] take videotapes of things happening in the world and they’ll collect huge videotapes of everything that’s happening and from that maybe they’ll come up with some generalizations or insights. Well, you know, sciences don’t do this” (Andor 2004: 97).

Утверждение Н. Хомского, как можно было ожидать, вызвало бурную дискуссию среди специалистов по корпусной лингвистике[5]. Такая реакция понятна, поскольку в своей аргументации Н. Хомский — нарочно или нечаянно — забывает о двух вещах. Во-первых, когда физик повторяет эксперимент, сделанный его коллегой, он исходит из того, что результаты должны обязательно быть теми же самыми. В лингвистике же мнения носителей языка о правильности какой-либо языковой единицы или правила не всегда совпадают. Из-за этого эксперименты лингвистов, основывающихся на своей языковой интуиции (или интуи-

ции других носителей языка), могут привести к различным, даже диаметрально противоположным выводам. Во-вторых, Н. Хомский сужает представление о методологии естественных наук. Физик не только делает эксперименты, но и наблюдает то, что происходит в мире, собирает информацию о действительности. С помощью телескопов физики-астрономы следят за тем, как движутся небесные тела; сложные современные аппараты позволяют собирать и исследовать, например, мельчайшие аэрозольные частицы. Наблюдения над реальностью не только позволяют подтверждать достоверность теоретических рассуждений, но и служат стимулом для построения новых научных концепций.

(2) Как мы сказали, оппонентами Н. Хомского выступили лингвисты, которые делают фундаментальные выводы о языке, опираясь не на интуицию, а на корпусные данные. Во многих концепциях, основывающихся на этом принципе, особое значение уделяется единицам, бóльшим, чем слово в традиционном понимании, но меньшим, чем предложения / высказывания. Этот подход имеет определенное сходство с традицией выделения в русских грамматиках особого уровня словосочетаний, но на практике речь идет о единицах другого типа — речевых штампах, идиомах. Один из ведущих представителей этого направления, Дж. Синклер, уже в 1991 сформулировал принцип идиоматичности: “The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments” (Sinclair 1991: 105). Обсуждение и реализация этого принципа породило целую серию исследований, среди которых можно называть “A Grammar of Speech” (Brazil 1995) и “Pattern Grammar” (Hunston & Francis 2000)[6]. Наконец, в исследовании “Linear Unit Grammar” (Sinclair & Mauranen 2006) вообще отрицается иерархичность языковых структур. Авторы исходят из линейной природы обработки грамматической информации в тексте и даже не используют традиционных терминов классов слов, не говоря уже о глубинных структурах предложения. В целом эти концепции ориентированы на исследование и моделирование сознания человека. Но в отличие от теоретиков-генеративистов, они обращают внимание не на виртуальное конструирование языковых единиц, а на реальный процесс порождения речи. В своих наиболее радикальных подходах сторонники этого направления отрицают существование *langue*, в смысле Ф. де Соссюра.

(3) Третий подход сводится к тому, что корпусные данные используются в качестве доказательства существования в языке (*langue*) того или иного явления. Если принимается положение о том, что *langue* «находится» в сознании человека и на этом уровне может быть регламентирован категориями «можно ~ нельзя», «правильно ~ неправильно», а не

понятиями «употребительно ~ неупотребительно», то, строго говоря, корпус не может служить источником для изучения *langue*. Однако многие ученые не разделяют такого крайнего отношения к этой методологической дилемме. И в пользу более мягкого решения этого вопроса приводятся веские аргументы. Первый аргумент: те языковые элементы, которые сейчас допустимы в *langue*, проявились в большинстве случаев как ошибки, окказионализмы или жанровые изыски, и разграничение *langue* и *parole*, таким образом, нечеткое. Второй аргумент: слишком категоричное отрицание узуса (и корпусов, представляющих узус) в качестве источника для изучения *langue* приводит к субъективизму, поскольку изучение интуитивных преставлений большого количества носителей языка весьма трудоемкая работа. К тому же рядовые носители языка затрудняются отвечать на вопросы лингвистов и последние предпочитают в результате довольствоваться собственными взглядами на язык.

Именно мягкий вариант третьего подхода все чаще можно встретить в докладах и статьях русистов. Конечно, всегда требует уточнения вопрос о том, сколько примеров и в каких текстах нужно для того, чтобы считать какую-либо языковую единицу еще непризнанной (ср. Мустайоки & Пуссинен 2006 о втором родительном падеже) или уже признанной (ср. Мустайоки & Пуссинен 2008 о новообразованиях с приставкой *по-*).

(4) Любой корпус можно изучать и просто как корпус, представляющий только самого себя. При таком подходе любое зафиксированное явление признается легитимным, независимо от степени его «правильности», нормативности. Такой подход, в частности, характерен для анализа разных идиолектов. Так например, говоря о *языке* А.С. Пушкина, обычно имеют в виду не абстрактную языковую систему или представления автора о правильности / неправильности в языке, а то, как автор *использовал* русский язык, то есть авторский идиолект. Естественно, А.С. Пушкин — это особый случай, поскольку он существенно повлиял на нормы современного русского языка, но, изучая идиолект писателя, мы, в общем и целом, используем ограниченный набор текстов, вышедших из-под пера автора. При таком подходе приходится признать, что существует не один *langue*, а практически бесконечное количество разных индивидуальных типов *parole*, закрепленных в авторских текстах.

Итак, в современной лингвистике получает распространение особая методология, которая и формирует содержание корпусной лингвистики. Заметим, что употребление этого термина требует особой оговорки. Дело в том, что сам по себе он имеет два значения. Это, во-первых, теория и методика создания корпусов и, во-вторых, корпусные исследования, т.е. исследования языка с помощью корпусных методов. Впрочем, четкой границы между ними не существует, и практически все создатели

корпусов проводят в то же время и собственно лингвистические исследования. В целом, *корпусная лингвистика* в первом значении более технологична и предполагает совместную работу лингвистов и специалистов по компьютерным технологиям, тогда как вторая задача — дело лингвистов, в том числе и специалистов по статистической обработке языка. Говоря о *русской корпусной лингвистике*, чаще имеют в виду второе значение, но необходимо помнить, что использование термина в первом значении широко распространено в мире и институализировано в виде множества исследовательских центров и специализированных журналов (см., например, журнал *Corpus Linguistics*), и без первого, строго говоря, не существовало бы и второго.

Итак, современная корпусная лингвистика, несмотря на относительно короткую историю существования, является хорошо разработанным направлением языкознания, тесно связанным с компьютерной и когнитивной лингвистикой. С первой она связана технологией и инструментами обработки языкового материала, со второй совпадает в базовой предпосылке: как когнитивная, так и корпусная лингвистика интересуется речевой деятельностью, представленной в бесконечном числе текстов (Gonzalez-Marquez et al. 2007). В определенном смысле корпусная лингвистика меняет приоритеты исследования: объектом изучения становится речь, несводимая к языковой абстракции, нормам литературного языка, суждениям о правильности / неправильности в языке, основанным исключительно на интуиции образованного исследователя. Вторым важным теоретическим следствием корпусных исследований можно считать то, что соссюровская дихотомия *langue-parole* заменяется представлением о первичности речевой деятельности с плавной шкалой генерализаций от речевого штампа до грамматического правила.

Наконец, следует помнить, что корпусная лингвистика — при всей революционности тех возможностей, которые она открывает, — всего лишь часть из обширного методологического инструментария современной лингвистики. Последнее можно проиллюстрировать следующими словами Ч. Филлмора: “I don’t think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore. <…> (but) every corpus I have had the chance to examine, however small, has taught me facts I couldn’t imagine finding out any other way” (Fillmore 1992: 35). Таким образом, любой большой корпус удивляет нас неожиданными открытиями, трудно улавливаемыми без обращения к реальному языковому материалу, с другой стороны, даже самые крупные корпуса не в состоянии отразить все возможное в языке.

История русских корпусов: от коллекции текстов к многоуровневой аннотации

Адам Килгариф в одной из своих лекций обозначил этапы развития автоматического анализа текста: лемматизация (т.е. автоматическое сведение словоформ к начальной форме — *лемме*) → частеречная разметка (то есть (полу)автоматическое приписывание словоформе морфологических признаков) → парсинг (то есть (полу)автоматическое приписывание синтаксической единице определенных признаков) → создание тезаурусов («семантическая разметка») → создание семантических сетей[7].

Каждый новый этап в развитии машинной обработки языкового материала открывает новые возможности сначала для создателей корпусов, а затем и для лингвистов, осуществляющих исследования на основе существующей разметки. Создание частотных словарей и индексов является одним из самых распространенных приложений корпусных исследований и тесно связано с отмеченными этапами в развитии обработки и представления языковых данных. Очевидно, этому предшествует «эра до аннотирования» (электронные корпуса, представляющие собой просто коллекцию текстов, напр., Машинный фонд русского языка или Упсальский корпус русского языка). История создания русских корпусов в основном следует этим этапам.

Значительное число корпусов создается и уже создано для многих языков. Они активно используются как для лингвистических исследований, так и в прикладных целях. В настоящее время стандартом являются корпуса, имеющие морфологическую аннотацию, введенную полностью автоматически или с частичной ручной постобработкой. Что касается славянских языков, то уже существуют корпуса чешского, польского, болгарского и других славянских языков (см. список на www.aclweb.org /aclwiki; Резникова 2008). Не отстает в этом отношении и корпусная русистика. Начав свое развитие в 1980-ые годы и пережив некоторый спад в 1990-ые, в настоящее время это направление активно развивается в 2000-ые и уже достигло существенных результатов, которые большей частью доступны в интернете. Ниже кратко представлены основные русскоязычные корпуса (подробнее см. Шаров 2003; Копотев & Резникова 2005; Копотев & Янда 2006).

1. Тюбингенский корпус (ТК)

www.sfb441.uni-tuebingen.de/b1/rus/korpora.html

В основе корпуса лежит старейший общедоступный русскоязычный Упсальский корпус русских текстов (www.slaviska.uu.se/korpus.htm), к материалам которого были добавлены тексты интервью. Ресурс стал первым морфологически аннотированным корпусом по русскому языку, появившимся в интернете в открытом доступе. В настоящее время работа над корпусом завершена.

2. Корпус газетных текстов (КГТ)
www.philol.msu.ru/~lex/corpus
Компьютерный корпус газетных текстов русского языка конца 20-го века был подготовлен в течение 2000-2002-го гг. в Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ. В настоящее время в интернете доступен небольшой тестовый фрагмент корпуса, более полная версия готовится к представлению.

3. Хельсинкский аннотированный корпус (ХАНКО)
www.slav.helsinki.fi/hanco
Корпус задуман как составная часть проекта «Функциональный синтаксис русского языка» и предназначен прежде всего для учебных целей. В интернете доступны результаты морфологической и синтаксической разметки. Отличительной чертой корпуса является возможность использовать поиск аналитических морфологических форм (*будет читать, читал бы*; см. Мустайоки & Копотев 2004) и точность ручной постобработки. В настоящее время проект продолжается.

4. Национальный корпус русского языка (НКРЯ)
www.ruscorpora.ru
Ресурс является крупнейшим сбалансированным корпусом русского языка, сопоставимым с национальными корпусами других языков. Коллекция включает как письменные тексты, так и расшифрованные записи устной речи[8]. Лингвистическая разметка включает морфологическую, синтаксическую и семантическую аннотацию. Примерно 4 % (около 6 млн. слов) от общего объема корпуса составляет подкорпус со снятой омонимией. В настоящее время работа над корпусом продолжается.

5. Национальный корпус русского литературного языка (НКРЛЯ)
www.narusco.ru
Корпус задуман как морфологически аннотированная коллекция текстов. По завершении корпус будет максимально репрезентативным, представляя весь лексический состав современного русского *литературного* языка. Для этого предполагается довести объем корпуса до 100-150 млн. словоупотреблений.

6. Система баз данных Интегрум
www.integrum.ru
Коммерческий интернет-ресурс, который включает большинство выходящих в настоящее время публицистических текстов (включая радиопередачи), законодательные документы, справочники, а также некоторое количество художественных текстов. В силу достаточно хорошо развитого языка запросов и удобного разделения материала по типам источников эта поисковая система вполне может быть использована как ежедневно пополняемый мониторинговый корпус. Использование Интег-

рума в исследованиях разного рода представлена в книге (Никипорец-Такигава 2006). Следующая таблица обобщает основные особенности описанных корпусов[9].

	ТК	**КГТ**	**ХАНКО**	**НКРЯ**	**НКРЛЯ**	**Интегрум**
Состав	публицистика (1996-2002 гг.); худож. лит-ра 19-20 вв.	газетные тексты (1997 г.)	журнальные тексты (2001 г.)	1) разл. жанры 18-21 вв.; 2) диалект. корпус.; 3) поэт. корпус; 4) корпус разг. речи	тексты, представляющие рус. лит. язык втор. пол. 20-21 вв.	публицистика (1990-ые - нач. 21 вв.); худ. тексты (19-20 вв.); правовые документы
Объем	ок. 25 млн. слов.	ок. 100 тыс. слов.	100 тыс. слов.	ок 150 млн. слов.	ок 1 млн. слов.	> 400 млн. **текстов**
Типы разметки:						
Морф.	автоматическая (в части корпуса, 2,3 млн.)	автоматическая	автоматическая + ручная	автоматическая + частично ручная	автоматическая	—
Словообр.	—	+	—	элементы	—	—
Синтаксич.	элементы	элементы	+	—	—	—
Семантич.	—	элементы	—	+	—	—
Поиск по:						
форме слова	+	+	+	+	+	+
лемме	—	+	+	+	—	+
неск. словам	+	—	+	+	—	+
грамм. приз.	+	+	+	+	—	—
семант. приз.	—	+	—	+	—	—
пункт. знакам	+	+	+	+	—	—

Таким образом, в распоряжении лингвиста, изучающего русский язык, имеется ряд возможностей обращаться к разным текстовым материалам, обладающим своими преимуществами и недостатками. «Интегрум» несопоставимо больше по объему, чем все остальные ресурсы (например, слово *корпус* в разных значениях встречается в нем свыше 2 млн. раз; в Национальном корпусе — ок. 4 тыс., в ХАНКО — 3 раза). Однако Интегрум не предназначен специально для изучения русского языка и содержит только сплошные тексты без морфологической разметки. В Национальном корпусе можно осуществлять поиск на представительной выборке текстов XVIII-XXI вв., используя морфологические и семантические параметры и богатую систему жанровых и функциональных признаков текста. ХАНКО лучше подходит для целей преподавания, поскольку содержит более качественное и традиционное аннотирование.

Корпусная лингвистика: сферы применения

Исследования русского языка, основанные на современных корпусных данных, уже имеют определенные традиции. В разных странах мира публикуются материалы, посвященные как созданию русскоязычных корпусов, так и исследованиям с помощью корпуса (конференции «Корпусная лингвистика» в Санкт-Петербурге, «Мегалинг» на Украине, «Диалог» в Москве, а также сборники (Плунгян 2005; Никипорец-Такигава 2006). Все это позволяет говорить о становлении нового направления — корпусной русистики. В то же время необходимо сказать, что корпусная лингвистика, как дисциплина, имеющая свою методологию и, как мы указали выше, активно формирующуюся теорию, нередко подменяется простым поиском иллюстративного материала в собрании электронных текстов. Безусловно, это важный и необходимый элемент использования любого корпуса, однако было бы неверным сводить все многообразие корпусных методов к простой задаче быстрого поиска подходящего примера. В настоящей статье нет возможности подробно останавливаться на всем многообразии методов этого направления (см. Tognini-Bonelli 2001 (гл 3-4); Hunston 2002; McEnery et al. 2006; Grondelaers et al. 2007). Однако ниже ряд конкретных примеров демонстрирует широту сфер применения корпусных подходов в современной лингвистике.

1. Использование корпусов в **грамматических и лексикологических исследованиях** стало уже обычным в современной исследовательской практике. Приведем лишь один показательный пример. В «докорпусных» исследованиях, описывающих конструкции типа *Лодку унесло ветром*, было сделано много ценных и точных наблюдений. Однако исследователи оперировали буквально двумя десятками примеров, не представляющих, как выяснилось, всего спектра употреблений. Исполь-

зование корпуса (в данном случае Интегрума) позволило расширить список примеров до более чем двух тысяч и точнее описать эту конструкцию (Мустайоки & Копотев 2005).

2. **Частотные списки и списки ключевых слов** активно создавались и использовались задолго до создания современных электронных корпусов. Эти исследования в большинстве случаев представляли частотные характеристики лексем (точнее, лемм). Корпусные методы позволяют сделать такие исследования более аккуратными и тонкими. Так, например, в исследовании (Коваль 2006) анализируются частотные характеристики омонимичных форм исходя из их реального употребления в современном русском языке. По данным исследователя, причастные формы большинства частотных русских глаголов практически не употребляются (см. фрагмент таблицы из указанной работы), и это означает, что омонимия существует лишь потенциально.

Форма	**Всего**	**Личных форм**	**Кр. причастий**
видим	79	78	1
делаем	37	37	
понимаем	31	31	
любим	28	22	6
просим	26	26	
читаем	24	24	
начинаем	23	23	

3. **Исследование коллокаций** (то есть сочетаний лексем) является в настоящее время одной из самых популярных тем корпусных исследований. Однако кроме этого решение более сложных задач осуществляется с опорой на исследование *коллигаций* (англ. *colligation*; сочетание лексем и / или грамматических признаков; см. Hunston & Francis 2000; Hunston 2001). Так, в работе (Guo 2005) исследуется сочетаемость модальных глаголов и среди прочего демонстрируется, что служебная идиома *as well* часто сочетается с формами сослагательного наклонения *might* и условной клаузой, вводимой союзом *if*. Таким образом, можно говорить о лексико-грамматическом комплексе (коллигации) *if ...might as well*.

4. **Исследование нормы / узуса**. Хотя исследование нормы обычно не входит в задачу корпусных лингвистов, множество острых, востребованных обществом языковых вопросов может быть решено на основе не субъективных оценок, а с привлечением статистически более представительного материала. Так, например, анализ сочетаний употребления второго родительного падежа типа *много народу* позволил выявить, что указания нормативных грамматик не соответствуют действительному употреблению этих форм (Мустайоки & Пуссинен 2006).

5. Корпусные методы с самого возникновения активно использовались в **социолингвистических исследованиях**. В качестве иллюстрации современных исследований приведем данные английских исследователей (McEnery & Xiao 2004). По их данным, в Британском национальном корпусе (BNC) употребление английского глагола *fuck* различается по возрастным группам (см. таблицу).

возраст	<15	16-25	26-35	36-45	46-60
fuck	6,07	16,5	8,86	0,7	3,48

6. Ошибочно считать, что корпусная лингвистика работает только с письменными текстами. Отдельной и активно разрабатываемой областью корпусной лингвистики стало **создание и изучение корпусов устной речи**. Так, в крупнейшие национальные корпуса (BNC, НКРЯ и др.) включены транскрипты записей устной речи. Самой значительной коллекцией устных текстов (включая аудио- и видеозаписи) является, безусловно, проект CHILDES, объединяющий около 130 корпусов детской речи для более чем 20 языков, в том числе и для русского. В качестве примера исследований в рамках «детской» корпусной лингвистики укажем на статью (Protassova & Voeikova 2007), в которой с опорой на корпусные данные (частично доступные в CHILDES) демонстрируется употребление русских диминутивов в детской речи.

7. Корпусная лингвистика с самого своего возникновения была тесно связна с **преподаванием языка** в иностранной аудитории. Известно, что 50 самых частотных английских лексем покрывают 60 % английской разговорной речи (Nation 1990). И этот факт, безусловно, должен учитываться в подборе лексики для изучающих язык. Корпусные исследования такого рода давно проводятся и стали основой множества учебных словарей и грамматик (Oxford Learner's English Dictionary, Collins Cobuild Student's Dictionary, Collins Cobuild English Grammar и др.). К сожалению, такого рода пособия по русскому языку еще не созданы, а существующие работы (напр., Морковкин 2003) опираются на устаревшие частотные словари и не учитывают современные корпусные данные.

8. Относительно новой областью является создание **корпусов ученических текстов**, которые позволяют классифицировать типы ошибок и учитывать их в процессе преподавания. Сведения такого рода учитываются в некоторых из указанных выше англоязычных учебных словарях. На русском материале эта работа только началась, и, насколько нам известно, сколько-нибудь представительных корпусов еще не существует. В качестве первых работ такого рода укажем на исследования (Соснина 2006; Pavlenko & Driagina 2007). В последнем, например, с помощью контрастивного анализа ученических корпусов (Contrastive learner cor-

pus analysis) показано, что носители английского языка предпочитают адъективные формы выражения эмоций, тогда как носители русского — глагольные конструкции. Вместе с тем, американцы, изучающие русский язык на продвинутом этапе, постепенно заменяют английские модели на русские.

9. Тесно связанной с различными педагогическими задачами, однако имеющей и собственно лингвистическое значение является **создание многоязычных параллельных корпусов**. Эта область корпусной русистики активно развивается, и в настоящее время созданы или создаются русско-английский, -немецкий, -японский, -финский, -словацкий корпуса. Отметим, что первая диссертация по корпусной русистике была посвящена именно параллельному (русско-финскому) корпусу (Михайлов 2003).

10. Наличие электронных текстов, принадлежащих одному автору, дает возможность расширить круг задач, традиционно решаемых **стилистикой и авторской стилеметрией**. Так, анализ употребления частотных существительных в текстах Ф.М. Достоевского не позволяет определить специфику авторского употребления (*человек*, *дело*, *время* и др.). Однако внимательный анализ коллокаций этих десемантизированных единиц в текстах разных периодов творчества позволяет сделать определенные выводы о развитии взглядов писателя в сторону конкретики частного дела и человеческой индивидуальности (Копотев 2003).

11. Еще одна задача, которая успешно решается с помощью корпусных методов, это **установление плагиата и скрытого цитирования**. Надо сказать, что эта задача шире, чем поиск скрытых цитат в студенческих и диссертационных работах. На это указывает и расширение тематики исследований (Johnson 1997; Turell 2004), и появление специализированных программ для выявления плагиата (см. Technical Review, 2001). Напомним также, что, при всей критике подхода группы Г. Хьетсо по установлению авторства «Тихого Дона», эта работа стала одной из первых попыток решения подобных задач на корпусном материале (Kjetsaa et al. 1984).

12. Наконец, корпусные методы применяются для решения задач **судебно-лингвистической экспертизы**. Очевидно, самым известным случаем такого рода является дело Дерека Бентли, осужденного в 1953 году за участие в убийстве полицейского и помилованного (посмертно) 45 лет спустя. В ходе повторного судебного разбирательства целый ряд аргументов защиты был связан с интерпретацией языковых фактов. Одним из существенных доказательств невиновности стали данные корпусного исследования, проведенного Р. Култардом. Исследователю удалось доказать, что продиктованное обвиняемым признание было существенно пе-

реработано человеком, привыкшим писать полицейские протоколы (Coulthard 2000).

В целом, для корпусных методов характерно:

- смещение исследовательской стратегии с изучения нормы («как правильно») на изучение узуса («как говорят / пишут»);
- автоматическое извлечение информации с помощью поисковых запросов, что может приводить к получению объемного и не всегда релевантного материала;
- распространенность «формально-морфологического» подхода, при котором поиск примеров основывается на морфологической (или просто на буквенной) форме;
- использование квантитативных методов, позволяющих учитывать частотные характеристики исследуемых единиц, и замена интроспективных оценок материала точными количественными данными об употреблении;
- опора на автоматическое аннотирование, не лишенное, с точки зрения традиционной лингвистики, определенных неточностей и упрощений;
- внимание к контексту в широком смысле (исследование коллокаций, ключевых слов, конструкций предполагает учет окружения исследуемой единицы);

Приведенные примеры исследований не преследуют цели очертить круг всех возможных сфер применения корпусных методов. Они лишь показывают широту применения и перспективность корпусной лингвистики — раздела языкознания, сугубо прикладного в момент возникновения, но развившегося в самостоятельную дисциплину, предлагающую в настоящее время как новые теоретические решения, так и конкретные исследовательские и педагогические инструменты для работы с языком.

Примечания

[1] «Язык как совокупность своих порождений отличается от отдельных актов речевой деятельности» (пер. В. В. Бибихина).

[2] Наиболее часто остракизму подвергается обсценная лексика, исключенная из целого ряда словарей, несмотря на очевидную частотность этих слов. К известным историям, связанным с редактированием словарей В.И. Даля и М. Фасмера, добавим еще судьбу Большого русско-финского словаря под ред. М.Э. Куусинена, изданного одновременно в Финляндии (без купюр) и в России (с купюрами).

[3] О краткой истории этой дисциплины свидетельствует и то, что ударение и морфологические формы русского термина и его производных еще не устоялись: *ко́рпусы — корпуса́, ко́рпусная — корпусна́я*. По наблюдениям одного из авторов этой статьи, на конференции по корпусной лингвистике большинство специалистов предпочитало формы *корпуса́, корпусна́я* и т.д. Лишь один докладчик последовательно

употреблял формы *ко́рпусный, ко́рпусы* и т.д. Письменная норма менее стабильна: в пяти русскоязычных сборниках по корпусной лингвистике встретилось 26 форм им. пад. *корпуса* и 37 — *корпусы* (благодарим В.П. Захарова за эти данные). По данным Интегрума, запрос «национальные корпус**а** + лингвистика» выдает 37 результатов, а «национальные корпус**ы** + лингвистика» — 1.

[4] В рамках подходов, признающих корпус необходимым инструментом, в западной лингвистике сложилось три вида деятельности, которые по-русски можно назвать исследованиями, «использующими корпус», «основанными на корпусе» и «инспирированными корпусом»: “Specifically, the spectrum of corpus linguistics ranges from ‘corpus-informed’ research (in which the corpus is used as a collection of natural examples) to ‘corpus-based’ approaches (in which corpora are analysed quantitatively and qualitatively on the basis of theoretical preconceptions, possibly in conjunction with noncorpus data) as well as to ‘corpus-driven’ approaches (in which an entirely corpus-generated model of language is envisaged)” (Mukherjee 2004: 115; см еще Tognini-Bonelli 2001: 65-100; Aarts 2002).

[5] См. обсуждение в электронной рассылке Corpora-List: www.uib.no/mailman/public/corpora/2007-June/004723.html и далее.

[6] Как пример ранних подходов к использованию языкового материала для определения единиц языка стоит упомянуть работы Н. Д. Андреева, который в 1960-ые годы, т.е. еще до эры современной корпусной лингвистики, поставил задачу выделить классы слов (части речи) на основе анализа сплошного текста с помощью созданного им «статистико-комбинаторного метода» (см., например, Андреев 1967).

[7] Лекция 6.10.2007, Хельсинкский университет.

[8] Подкорпус представляет уникальный материал устной речи 1930-2000 гг., собранный как в России, так и в Финляндии (у носителей «досоветской» нормы); см. подробнее (Гришина 2005).

[9] Таблица с некоторыми изменениями воспроизводится по (Копотев & Резникова 2005). Отметим еще два корпуса, представляющие материалы по истории русского языка: **Регенсбургский диахронический корпус русского языка (**www-korpus.uni-r.de/diakorp/index.php) и проект **«Манускрипт»** (manuscripts.ru). В последние годы активно обсуждается и представление об интернете, как своеобразном корпусе, который можно изучать, используя как общедоступные (Google, Яндекс), так и специализированные поисковые системы (см. Kilgarriff & Grefenstette 2003; Беликов 2004; Захаров 2005; Kilgariff 2007).

Литература

Aarts, J.: 2002, ‘Does corpus linguistics exist? Some old and new issues’, *From the COLT’s mouth . . . and others’: Language Corpora Studies in Honour of Anna-Brita Stenström*, Amsterdam, 1-17.

Andor, J.: 2004, ‘The master and his performance: An interview with Noam Chomsky’, *Intercultural Pragmatics*, 1:1, 93-111.

Brazil, D.: 1995, *A Grammar of Speech (Describing English Language)*, Oxford.

Coulthard, R.M.: 2000, ‘Whose text is it? On the linguistic investigation of authorship’, S. Sarangi and R.M. Coulthard (eds.), *Discourse and Social Life*, London, 270-287.

Dash, N., S.: 2008, *Language Corpora: Past, Present and Future*, Kolkata, [In press].

Fillmore, Ch.: 1992, 'Corpus linguistics or computer-aided armchair linguistics', *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August, 1991*, Berlin, 35-60.

Gonzalez-Marquez, M., Mittelberg, I., et all. (eds.): 2007, *Methods in Cognitive Linguistics*, Amsterdam / Philadelphia.

Grondelaers, S., Geeraerts, D., Speelman, D.: 2007, 'A case for a cognitive corpus linguistics', *Methods in Cognitive Linguistics*, Amsterdam / Philadelphia, 149-169.

Guo, X.: 2005, 'Modal auxiliaries in phraseology: a contrastive study of learner English and native speaker English', *Proceedings from the Corpus Linguistics Conference Series*, [электронный ресурс] www.corpus.bham.ac.uk/PCLC/CL%202005%20xiaotian% 20guo.doc.

Humboldt, W. von.: 1836, *Über die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluss auf die geistige Entwickelung des Menschengeschlechts*, Berlin.

Hunston, S.: 2001, 'Colligation, lexis, pattern, and text', Scott, M. and G. Thompson (eds.), *Patterns of Text*, 13-33.

Hunston, S.: 2002, *Corpora in Applied Linguistics*, Cambridge.

Hunston, S. & Gill F.: 2000, *Pattern Grammar: a Corpus-driven Approach to the Lexical Grammar of English*, Amsterdam.

Johnson, A.: 1997, 'Textual kidnapping — a case of plagiarism among three student texts?', *Forensic Linguistics: The International Journal of Speech Language and the Law*, 4:2, 210-226.

Kilgarriff, A.: 2007, 'Googleology is bad science', *Computational Linguistics*, 33:1, 147-151.

Kilgarriff, A. & Grefenstette, G.: 2003, 'Introduction to the special issue on web as corpus', *Computational Linguistics*, 29:3, 333-347.

Kjetsaa et al.: 1984, *The Authorship of The Quiet Don*, Oslo.

McEnery, T. & Wilson, A.: 2001, *Corpus Linguistics*, Edinburgh.

McEnery, A.M. & Xiao, Z.: 2004, 'Swearing in modern British English: the case of *fuck* in the BNC', *Language and Literature*, 13:3, 235-268.

McEnery, T., Xiao, R., Tono, Yu. 2006: *Corpus-based Language Studies: An Advanced Resource Book*, London.

Mukherjee, J.: 2004, 'The state of the art in corpus linguistics: three book-length perspectives', *English Language and Linguistics*, 8:1, 103-119.

Nation, I.S.P.: 1990, *Teaching and Learning Vocabulary*, New York.

Pavlenko, A. & Driagina, V.: 2007, 'Russian emotion vocabulary in American learners' Narratives', *The Modern Language Journal*, 91:2, 213-234.

Penke, M. & Rosenbach, A.: 2004, 'What counts as evidence in linguistics?: An introduction', *Studies in Language*, 28:3, 480-526.

Protassova, E. & Voeikova, M.: 2007, 'Diminutives in Russian at the early stages of acquisition', I. Savickienė, W. U. Dressler (eds.), *The Acquisition of Diminutives: A cross-linguistic perspective*, Amsterdam, 43-72.

Sinclair, J., : 1991, *Corpus, Concordance, Collocation*, Oxford.

Sinclair, J. & Mauranen, A.: 2006, *Linear Unit Grammar; Integrating Speech and Writing*, Amsterdam.

Technical Review of Plagiarism Detection Software Report, 2001, [электронный ресурс], http://turnitin.com/static/pdf/luton.pdf.

Tognini-Bonelli, E.: 2001, *Corpus linguistics at work*, Amsterdam.

Turell, M.T.: 2004, 'Textual kidnapping revisited: the case of plagiarism in literary translation', *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 11:1, 1-26.

Андреев, Н.Д.: 1967, *Статистико-комбинаторные методы в теоретическом и прикладном языковедении*, Ленинград.

Беликов, В.И.: 2004, 'Yandex как лексикографический инструмент', *Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог 2004»*, Москва, 39-46.

Гришина, Е.А.: 2005, 'Устная речь в Национальном корпусе русского языка', *Национальный корпус русского языка: 2003-2005. Результаты и перспективы*, Москва, 94-110.

Захаров, В.П.: 2005, 'Веб-пространство как языковой корпус', *Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2005»*, Москва, 166-171.

Коваль, С.А.: 2006, 'Роль корпуса в создании реалистичных моделей словоизменительной морфологии', *Труды международной конференции «Корпусная лингвистика — 2006»*, Санкт-Петербург, 148-158.

Копотев, М.В.: 2003, 'Из наблюдений над публицистикой Ф. М. Достоевского (*человек* и его *дело*)', *Slavic Almanac: The South African Year Book for Slavic, Central, and East European Studies*, 9:12, 153-164.

Копотев, М.В. & Резникова, Т.И.: 2005, 'Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов)', *Национальный корпус русского языка: 2003-2005. Результаты и перспективы*, Москва, 31-61.

Копотев, М.В. & Янда, Л.: 2006, 'Рецензия (Национальный корпус русского языка)', *Вопросы языкознания*, 5, 149-155.

Михайлов, М.: 2003, *Параллельные корпуса художественных текстов*, Tampere.

Морковкин, В.В. (ред.): 2003, *Система лексических минимумов современного русского языка*, Москва.

Мустайоки, А.: 1988, 'О предмете и цели лингвистических исследований', *Язык: система и функционирование*, Москва, 170-181.

Мустайоки, А.: 1995, 'О лингвистических экспериментах', *Язык — система, язык — текст, язык — способность*, Москва, 155-160.

Мустайоки, А. & Копотев, М.: 2004, 'К вопросу о статусе эквивалентов слов типа *потому что, в зависимости, к сожалению*', *Вопросы языкознания*, 3, 2004, 88-107.

Мустайоки, А. & Копотев, М.: 2005, '*Лодку унесло ветром*: условия и контексты употребления русской «стихийной» конструкции', *Russian Linguistics*, 1, 1-38.

Мустайоки, А. & Пуссинен, О.: 2006, ‘Почему народу много, или новые наблюдения над употреблением второго родительного падежа в современном русском языке’, *Integrum: точные методы и гуманитарные науки*, Москва, 50-75.

Никипорец-Такигава, Г. (ред.): 2006, *Integrum: точные методы и гуманитарные науки*, Москва.

Плунгян, В.А. (ред.): 2005, *Национальный корпус русского языка: 2003-2005. Результаты и перспективы*, Москва.

Резникова, Т.И.: ‘Корпуса славянских языков в интернете: обзор ресурсов’, *Die Welt der Slaven*, LIII: 1, 10-38.

Соснина, Е.П.: 2006, ‘О разработке и использовании российского учебного корпуса переводов’, *Труды международной конференции «Корпусная лингвистика — 2006»*, Санкт-Петербург, 365-373.

Шаров, С. А.: 2003, ‘Представительный корпус русского языка в контексте мирового опыта’, *Научно-техническая информация*, Сер.2, № 6, 9-18.