Т. И. Резникова

Славянская корпусная лингвистика: современное состояние ресурсов[1]

Последнее десятилетие было отмечено бурным развитием корпусных ресурсов для славянских языков. Всплеск активности был обусловлен прежде всего переосмыслением роли корпуса в лингвистической исследовательской деятельности. Если ранние опыты собрания электронных текстовых коллекций были нацелены в основном на статистический анализ языка и лексикографическую практику (ср., например, разработанный в конце 60-х—начале 70-х гг. польский корпус (500 тыс. словоупотреблений), использовавшийся для составления словаря [Kurcz et al. 1990], одномиллионный корпус русского языка, создававшийся в 1970-е гг. и ставший основой для частотного словаря [Засорина 1977]; корпус хорватского языка М. Могуша (1 млн., 1976–1996) и созданный на его базе словарь [Moguš et al. 1999]) и тем самым оставались на периферии интересов лингвистического сообщества, то с осознанием важности корпуса как эффективного самостоятельного инструмента, коренным образом меняющего исследовательские возможности лингвис-

[1] Настоящая статья является актуализованной и расширенной версией публикации [Резникова 2008].

та в самых различных научных областях, создание корпусов стало актуальной задачей для широкого круга специалистов в разных странах. Такой подход к предназначению корпуса выдвигал новые требования к его основным параметрам: он должен был характеризоваться, во-первых, многомиллионным объемом, во-вторых, наличием лингвистической разметки и, в-третьих, доступностью через Интернет. Эти требования заложили программную основу целого ряда проектов, возникших во второй половине 90-х—первых годах нового века. Результатом их работы стало появление значительного числа лингвистических ресурсов, существенно преобразующих ситуацию в современной славистике.

Задача настоящего очерка—дать общее представление о существующих на сегодняшний день в Интернете корпусах славянских языков, описать принципы их составления, лингвистический аппарат, поисковые возможности. Обсуждаемые корпуса будут представлены по языкам, соответственно, читатель сможет оценить степень корпусной оснащенности интересующего его языка и выбрать ресурс, в наибольшей степени отвечающий его исследовательской задаче. В то же время описание корпусов будет строиться по одной и той же схеме, что позволит читателю сопоставить потенциал разноязычных ресурсов.

За пределами обзора останутся диахронические и параллельные корпуса: создание корпусов обоих типов сопряжено с целым рядом дополнительных трудностей, тем самым их описание требует иного в сравнении с синхронными и одноязычными корпусами набора параметров.

1. Западнославянские языки

Чешский

В 90-е гг. Чехия стала форпостом корпусной лингвистики в славянском мире. Именно здесь был создан первый для славянского языка большой представительный корпус—Чешский национальный корпус, отвечающий мировому стандарту, заданному Британским национальным корпусом, и именно здесь был разработан первый для славянского языка корпус с синтаксической аннотацией—Prague Dependency Treebank. На сегодняшний день

Чешский национальный корпус объединяет в себе ряд подкорпусов, отражающих различные формы функционирования чешского языка и предлагающих широкие возможности поиска и статистического анализа языковых данных, что, безусловно, позволяет говорить о хорошей оснащенности чешского языка корпусными ресурсами.

Чешский национальный корпус (ЧНК). Возникшая в начале 90-х гг. идея создания корпуса обрела институциональный статус в 1994 г.: при Карловом университете Праги был основан Институт Чешского национального корпуса. Разработчики корпуса рассматривают Институт как проект с открытыми временными рамками, призванный постоянно расширять состав корпуса, в том числе за счет вновь появляющихся текстов.

Состав. Корпуса, объединенные под названием ЧНК, распадаются на диахроническую (719 тыс. словоупотреблений) и синхронную части. Интересующая нас синхронная часть в свою очередь подразделяется на корпуса письменного и устного языка: письменная часть, включающая как оригинальные, так и переводные тексты (всего 500 млн. словоупотреблений), объединяет 2 представительных корпуса по 100 млн. (SYN2000 и SYN2005), подкорпус на базе SYN2000—FSC2000 (96 млн.), 2 специализированных корпуса—публицистики (SYN2006PUB, 300 млн.) и частной корреспонденции (KSK, 800 тыс.) и небольшой корпус ORWELL (80 тыс.); устная часть (всего 2,17 млн.) включает 3 корпуса, распределенных по месту записи текстов: Прага (PMK, 675 тыс.), Брно (BMK, 490 тыс.), различные диалектные регионы Чехии (ORAL2006, 1 млн.).

Наименования **SYN2000** (100 млн.) и **SYN2005** (100 млн.) отражают год открытия соответствующего корпуса и тем самым указывают временные различия входящих в их состав публицистических текстов: в SYN2000—это тексты, написанные с 1990 по 1999 гг., в SYN2005—с 2000 по 2004 гг. Две другие составляющие—художественная литература и специализированные тексты—не различаются с точки зрения нижней временной границы включения текстов: специализированная литература в SYN2000 охватывает период с 1990 по 1999 гг., в SYN2005—с 1990 по 2004 гг., основная масса художественных текстов относится к тем же временным промежуткам, хотя незначительную долю образуют более ран-

ние тексты, созданные с 1959 г. Существенно при этом, что два корпуса не содержат никаких одинаковых текстов. Разработчики корпуса очень тщательно подошли к проблеме сбалансированности типов текстов. Для выявления их реальных соотношений в функционировании языка авторы каждый раз проводили новые социолингвистические исследования, которые легли в основу процентных долей типов текстов в корпусах. Интересным образом результаты исследований существенно различаются для корпусов, появившихся с промежутком в 5 лет. Основные типы текстов представлены в SYN2000 vs. SYN2005 соответственно следующим образом: художественная литература (15 vs. 40%), специализированная литература (25 vs. 27%), публицистика (60 vs. 33%).

Корпус **FSC2000** (96 млн.) разрабатывался как основа для частотного словаря [Čermák, Křen 2004]. Он представляет собой несколько улучшенный вариант корпуса SYN2000: для аккуратности статистического анализа были исключены тексты, случайно попавшие в корпус дважды, а также исправлены некоторые ошибки автоматической лемматизации.

Корпус **SYN2006PUB** (300 млн.)—несбалансированный корпус публицистики, включающий тексты с 1989 по 2004 гг., не вошедшие в корпуса SYN2000 и SYN2005. Этот корпус представляет интерес прежде всего для решения исследовательских задач, требующих работы с большим объемом языковых данных.

Корпус **KSK** (800 тыс.) призван отразить последнюю стадию существования традиционного эпистолярного жанра. В него включено 2000 написанных от руки писем, созданных 2000 разных людей в период с 1990 по 2004 гг.

Корпус **ORWELL** (80 тыс.) создавался в рамках международного проекта Multext-East (1995–97 гг.), задача которого состояла в разработке ресурсов для автоматической обработки текста на материале нескольких языков Восточной и Центральной Европы. Одним из основных результатов проекта стало создание параллельного корпуса, в состав которого вошел текст романа Дж. Оруэлла «1984» и его переводы на анализируемые языки. Чешский перевод романа и образует подкорпус ORWELL в составе ЧНК.

Корпус **PMK** (675 тыс.) включает свыше 300 записей устной речи, проведенных с 1988 по 1996 гг., **BMK** (490 тыс.)—250 записей

с 1994 по 1999 гг. Оба корпуса сбалансированы с точки зрения пола, возраста и уровня образования участников, а также типа разговора (формальный, т.е. монологические ответы на вопрос интервьюера, vs. неформальный, т.е. диалоги знакомых друг с другом людей). Корпус **ORAL2006** (1 млн.) содержит 220 записей, проведенных с 2002 по 2006 гг. Все разговоры носят неформальный характер.

Метаразметка текстов. Письменные тексты характеризуются по следующим параметрам: имя автора, название текста, тип текста (для художественных: *роман, рассказ/сборник рассказов…*, для специализированных: *научный, популярно-научный, учебник…*, для публицистики: собственно *публицистика* и «*эфемерные тексты*»), жанр текста (сюда попадает тематика для специализированных текстов—*история, география, право, домашнее хозяйство* и т. п., жанры для художественных—*детектив, фантастика, мемуары* и т. п., а также в ряде случаев цель создания произведения (напр., *развлечение*) или целевая аудитория—*литература для детей*), тип носителя (*книга, журнал, интернет…*), библиографические данные (издательство, год и место издания, ISBN/ISSN). Кроме того, в корпусах SYN2005 и SYN2006PUB проработана зона переводов: указываются язык исходного текста и имя переводчика. Отсутствие подобной информации является определенным недочетом SYN2000: пользователь не может ограничить поиск только оригинальными или только переводными текстами. В целом следует отметить, что метаразметка письменных текстов ЧНК не лишена некоторых недостатков. Так, как можно видеть из приведенного выше перечня, параметр «жанр текста» предполагает классификацию текстов по разным основаниям: указание тематики для специализированной литературы делает невозможным ее распределение по жанрам (ср. *статья, монография, диссертация* и т.д.), цель создания текста и целевая аудитория могут накладываться как на тематику, так и на жанр, ср. вполне естественные комбинации *книга по истории для детей* или *развлечение* как назначение *мемуаров*. Вообще говоря, склеенные здесь характеристики в мировой практике создания корпусов нередко образуют отдельные параметры классификации текстов (ср. в частности рекомендации EAGLES по разметке

корпусов [EAGLES 1996]): цель создания текста и информация об аудитории (помимо возраста включающая также ее предполагаемый размер и ограничения на пол и уровень образования). Кроме того, при метаразметке не учитывается ряд других параметров, существенных для характеристики языковых особенностей текста: возраст автора в момент написания текста (или год его рождения), пол автора, год создания текста (который, особенно в случае художественной литературы, может отличаться от года его издания).

Среди письменных корпусов особая система метаразметки, приближенная к разметке устных текстов, принята в KSK. И в устных корпусах, и в KSK тексты классифицируются по полу говорящего/пишущего, его возрасту (в устных корпусах—до 35 vs. выше 35, в KSK—4 возрастные группы), его уровню образования (высшее vs. невысшее), в KSK и ORAL2006—по диалектной принадлежности говорящего/пишущего (в PMK и BMK этот параметр менее релевантен, т.к. все тексты записаны в одном городе, хотя во внимание можно принимать и тот факт, что диалектные особенности говорящего могли сформироваться в ином месте). Кроме того, в KSK учитываются параметры пола, возраста и уровня образования адресата, а в PMK и BMK—формат разговора (формальный vs. неформальный).

Морфологическая разметка. На основном массиве письменных корпусов (SYN2000, SYN2005, SYN2006PUB и ORWELL) была проведена лемматизация и морфологическая разметка. Процедура осуществлялась автоматически, с использованием статистических методов снятия грамматической омонимии. На материале небольшого подкорпуса ORWELL (80 тыс.) проводилась ручная коррекция ошибок программы автоматического снятия омонимии. Морфологическая разметка для каждой словоформы хранится в виде 16-местной цепочки букв и цифр, каждая позиция в которой соответствует определенному грамматическому признаку с заданным набором возможных значений. В позиции, нерелевантной для данной словоформы (напр., падеж для глагола), ставится прочерк. Отметим, что характеристика глагола по виду была добавлена на более позднем этапе разработки корпуса и отсутствует в корпусе SYN2000. Как уже отмечалось, в FSC2000 по сравнению с SYN2000 усовершенствована лемматизация, однако

отсутствует морфологическая аннотация. В KSK и устных корпусах лемматизация и морфологическая разметка не проводились.

Поиск в корпусе. Для поиска в корпусе используется графический пользовательский интерфейс Bonito программной системы корпусного обеспечения Manatee, разработанной П. Рыхли (Университет им. Масарика, Брно). Наряду с ЧНК эта поисковая система используется в Словацком и Хорватском национальных корпусах. Программа позволяет строить разнообразные запросы с использованием регулярных выражений (специальной системы записи шаблонов для поиска) и логических операторов. Поиск может вестись по любому атрибуту корпуса: словоформе или ее части, лексеме или ее части (в корпусах с лемматизацией), последовательности словоформ/лексем с указанием расстояния между ними или с заданием структурного единства (напр., предложение), в пределах которого заданные единицы должны встретиться, а также по любой комбинации грамматических признаков (в корпусах с морфологической разметкой). При поиске могут учитываться знаки препинания и положение искомой единицы относительно начала/конца предложения.

После получения конкорданса можно осуществить фильтрацию найденных примеров (т.е. удалить часть найденных контекстов). Конкорданс выдается в формате KWIC (key word in context), т.е. искомое слово отображается в центре экрана, что позволяет быстро просматривать его левый и правый контекст. В командном меню предусмотрена опция отображения леммы и/или грамматических признаков при искомом выражении или во всех выданных словах. Возможно упорядочение выданных контекстов по первой или последней словоформе искомого выражения, по левому или правому контексту (с возможностью указания количества учитываемых позиций), а также по любому атрибуту этих словоформ (по лемме или грамматическим признакам в тех корпусах, где эти атрибуты включены в разметку). При сортировке можно задавать комбинацию из нескольких условий, каждое их которых отвечает одной позиции, относительно которой сопоставляются разные строки. Удобной для изучения типов встретившихся в корпусе контекстов представляется также возможность оставить в выдаче по одному примеру из тех, в которых совпадают упорядочиваемые

элементы (словоформы, леммы или грамматические признаки) в заданном интервале. Упорядочить выдачу можно и вручную: по группам, на которые с помощью расстановки соответствующих номеров пользователь расклассифицировал выданные контексты. Максимальный контекст выдачи составляет по 500 знаков или по 50 слов справа и слева от найденного выражения или по 1 предложению справа и слева от того, в котором оно было найдено.

Поиск может вестись как по всему заданному корпусу (т.е. по одному из корпусов в составе ЧНК), так и по определенному пользователем подкорпусу (ограничение может производиться по одному или нескольким из доступных метаатрибутов, т.е., например, по году издания текста, фамилии автора, типу текста и т. п.).

Одной из особенностей системы Manatee являются широкие возможности вычисления различных статистических параметров корпуса. Предусмотрена возможность составления частотных списков для заданных значений одного из доступных атрибутов (т.е. для заданных словоформ, лексем (в корпусах с лемматизацией) или грамматических признаков (в корпусах с морфологической разметкой)). Тем самым, например, можно получить частотное распределение словоформ корпуса по частям речи. Кроме того, для заданной словоформы (леммы, грамматического признака) можно получить частотный список словоформ (лемм, грамматических признаков), в контексте которых (на заданном пользователем расстоянии) она встречается. На выдаче пользователь получает таблицу с указанием для каждой коллокации абсолютной и относительной частотности, а также статистических характеристик T-score и MI-score (взаимная информация).

Доступ к корпусу. Для доступа к ЧНК в полном объеме и к использованию всех предусмотренных поисковых возможностей необходимо пройти регистрацию (для исследовательских целей осуществляется бесплатно). Без этой процедуры пользователь имеет доступ к корпусу SYN2000, однако выдача ограничена 50 контекстами (при этом указывается и общее число имеющихся в корпусе примеров, удовлетворяющих заданному запросу).

Prague Dependency Treebank (PDT). PDT разрабатывается с 1995 г. в Институте формальной и прикладной лингвистики Карлова университета (с 2000 по 2004 гг. при участии Центра ком-

пьютерной лингвистики). PDT представляет особое направление корпусной лингвистики, в рамках которого создаются корпуса, нацеленные не на объем ресурса, а на детальность его лингвистической разметки, предполагающей в значительной степени ручную обработку языковых данных.

Состав. В корпус вошли взятые из ЧНК тексты нескольких ежедневных газет и специализированных журналов, охватывающие период с 1991 по 1995 гг., общим объемом 2 млн. словоупотреблений. Небольшой объем и однородность типов текста в корпусе делает их метаразметку не столь обязательной.

Морфологическая разметка. На всем объеме корпуса была проведена лемматизация и морфологическая разметка. Процедура осуществлялась автоматически с последующим ручным снятием омонимии. Морфологический тэг представляет собой 15-местную цепочку букв и цифр (ср. ЧНК). Следует заметить, что, как и в первой версии ЧНК SYN2000, морфологическая разметка не учитывает глагольную категорию вида (ее значения приписываются только на семантическом уровне аннотации).

Синтаксическая разметка. Синтаксическая аннотация на сегодняшний день осуществлена в подкорпусе объемом 1,5 млн. словоупотреблений. Разметка на синтаксическом (в терминах разработчиков корпуса—аналитическом) уровне предполагает приписывание каждой единице в тексте ее синтаксической функции (тем самым элементы аналитических словоформ трактуются как отдельные единицы) и построение дерева зависимостей для каждого предложения (т.е. указание для каждой текстовой единицы ее порядкового номера в предложении и порядкового номера ее вершины). На основе свыше 19 тыс. построенных вручную деревьев был создан автоматический парсер, работающий с точностью 80%, результаты его анализа проверяются вручную. Синтаксические функции приписываются после построения деревьев автоматически и также проходят ручную постобработку.

Семантическая разметка. Разметка самого глубинного уровня (в терминах разработчиков—тектограмматического) на сегодняшний день осуществлена в подкорпусе объемом 0,8 млн. словоупотреблений. Принципы аннотации основаны на теории функциональной порождающей грамматики П. Сгалла. Разметка

предполагает построение дерева, отражающего глубинную структуру предложения; лемматизацию тектограмматического уровня (несколько отличающуюся от грамматической лемматизации: так, аналитическим формам приписывается общая лемма смысловой составляющей, притяжательным прилагательным—лемма соответствующего существительного и т. д.); приписывание семантических частей речи (также частично отличающихся от лексико-семантических разрядов); указание значений грамматем (тектограмматического коррелята грамматических категорий—их значения могут расходиться со значениями соответствующих морфологических категорий: например, число у существительных pluralia tantum указывается в соответствии с количественной характеристикой денотата; кроме того, здесь учитываются такие семантико-грамматические категории, не включенные в морфологическую разметку, как, например, результативность, итеративность, деонтическая модальность и др.); приписывание тектограмматических функций—семантического аналога синтаксических функций (например, агенс, пациенс, направление, принадлежность и др.); разметку коммуникативной структуры предложения (топик-фокус); указание отношения кореферентности между узлами дерева. Подробнее с данным типом разметки в PDT можно ознакомиться на сайте *http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html*.

Поиск в корпусе. В PDT реализовано две возможности обращения к языковым данным: просмотр синтаксических и тектограмматических деревьев и поиск по заданным параметрам. Пользователь может, во-первых, открыть один из доступных файлов и последовательно просматривать структуры зависимостей с разметкой имеющихся атрибутов (синтаксического или тектограмматического уровня) и, во-вторых, построить запрос при помощи специально разработанной программы Netgraph с удобным графическим интерфейсом. Она позволяет осуществлять поиск по заданной форме дерева и по произвольной комбинации значений атрибутов одного или нескольких узлов дерева (например, по лемме, грамматическим признакам, синтаксическим функциям в предложениях с синтаксической разметкой или по тектограмматической лемме, по грамматемме, семантическим функ-

циям в предложениях с семантической разметкой). Для упрощения работы пользователю предлагаются в поисковом окне списки доступных атрибутов и их возможных значений. Предусмотрена возможность фильтрации результатов предыдущего запроса в соответствии с заданным условием.

Доступ к корпусу. PDT распространяется на платной основе через LDC (Linguistic Data Consortium, *http://www.ldc.upenn.edu*). В сокращенном варианте корпус доступен также в Интернете, однако для его использования необходимо пройти регистрацию.

Словацкий

В распоряжении исследователей словацкого языка в настоящее время имеется один общедоступный ресурс—Словацкий национальный корпус—большой корпус с лемматизацией и морфологической разметкой. В 2005 году была начата работа по созданию корпуса с синтаксической разметкой Slovak Dependency Treebank, во многом опирающаяся на опыт PDT, однако результаты этой работы пока недоступны.

Словацкий национальный корпус (СНК). Проект по созданию представительного корпуса начал свою работу в 2002 г. с открытием отдела Словацкого национального корпуса в рамках Института языкознания Словацкой академии наук. Наряду с одноязычным корпусом в институте ведется разработка двух параллельных корпусов—русско-словацкого (см. [Гарабик, Захаров 2006]) и французско-словацкого (см. [Vasilišinová, Garabík 2007]).

Состав. СНК в отличие от ЧНК включает только синхронные тексты, однако нижней временной границей включения текстов является 1955 г. (ср. 1990 для большинства текстов в ЧНК). На данный момент СНК не ставит и задачу сбора устной речи. Основной корпус **(prim-3.0-public-all)** объемом 339 млн. словоупотреблений образуют оригинальные и переводные тексты в следующей пропорции: публицистика (60,6%), художественная литература (17,5%), специализированные тексты (11,6%), другое (10,3%). Доступен также подкорпус объемом ок. 200 млн. словоупотреблений, считающийся сбалансированным (**prim-3.0-vyv**), доли разных типов текстов в котором однако не столь значительно отличаются от целого корпуса (60% публицистики, 20% художественной ли-

тературы, 20% специализированных текстов). Отдельный подкорпус образуют тексты с ручной морфологической разметкой (**r-mak-2.0**, 511,5 тыс. словоупотреблений).

Метаразметка. В СНК метаразметка сделана с учетом большего числа параметров, чем в ЧНК, и тем самым более адекватно представляет текстовые типы. Она включает среди прочих следующие атрибуты: имя автора, его пол, название текста, год его издания, год его первого издания, оригинал/перевод, язык-источник, имя и пол переводчика, тип текста (*художественный* с подтипами *поэзия, проза, драма*; *информативный*, в т.ч. *публицистика*, *реклама* и др., *профессиональный*, в т.ч. *научный, учебник* и др.; *коммуникация*); жанр текста (*стихи, роман, очерк, статья* и др.); предметная область (тематика для специализированной литературы); тип носителя (*книга, газета, интернет* и др.), вариант языка (*стандартный/нестандартный*). Среди неучтенных параметров здесь можно отметить характеристики целевой аудитории и цель создания текста.

Морфологическая разметка. На основном массиве текстов была проведена автоматическая лемматизация и морфологическая аннотация, основанная на статистических методах снятия грамматической омонимии. В части корпуса (доступной как подкорпус r-mak-2.0, 511,5 тыс. словоупотреблений) осуществлена ручная морфологическая разметка. Система аннотации несколько отличается от ЧНК: для каждой части речи предусмотрена своя схема тэга, в которой учитываются только релевантные для нее категории. Таким образом, тэги представляют собой не длинные цепочки с большим количеством прочерков для обозначения нерелевантных параметров, а более компактные и удобные для прочтения и воспроизведения в запросе последовательности.

Поиск в корпусе. Предусмотрено два вида обращений к корпусу. Первый—непосредственно с сайта СНК—подразумевает несколько ограниченные возможности обработки запросов. Благодаря используемому языку регулярных выражений поиск может вестись по всем тем параметрам, которые описаны выше для ЧНК (словоформа, лексема, их последовательность, различные условия их взаимного расположения, грамматические признаки). Конкорданс выдается в формате KWIC. По команде

пользователя система может отображать леммы и/или грамматические признаки при искомом выражении. Для каждой строки конкорданса предусмотрена возможность просмотра большого контекста (до 100 текстоформ справа и слева от искомых) и метаинформации о тексте-источнике. Однако при данном типе работы с корпусом пользователь не может задавать собственный подкорпус (поиск ведется по всему корпусу или по r-mak-2.0). Эта и многие другие возможности, связанные с фильтрацией контекстов, сортировкой их выдачи, статистической обработкой данных предусматривает второй вид обращения к корпусу, требующий предварительной регистрации. В этом случае работа с корпусом осуществляется при помощи системы Bonito, описанной выше для ЧНК и аналогичным образом функционирующей для СНК.

Доступ к корпусу. Возможен как доступ с предварительной регистрацией (для исследовательских целей осуществляется бесплатно), предоставляющий расширенные возможности обработки данных, так и поиск непосредственно с сайта СНК (об ограничениях см. выше).

Польский

В отличие от чешского и словацкого польский до сих пор не имел собственного национального корпуса, что, безусловно, отражалось на общем уровне развития корпусных ресурсов языка. Однако в настоящее время работа над его созданием уже ведется[2]. Наряду с Институтом польского языка Польской академии наук в Консорциум Национального корпуса польского языка вошли организации, ранее уже разрабатывавшие корпусные ресурсы для польского. Именно на базе этих ресурсов и создается новый—Национальный—корпус (НКПЯ). Однако на настоящий момент возможности исследователей-полонистов в целом все еще ограничиваются этими корпусами–«предшественниками», поэтому о них и пойдет речь ниже. Необходимо тем не менее отметить, что в рамках проекта по созданию Национального корпуса уже были собраны новые текстовые коллекции, и демонстрационные версии НКПЯ предоставляют к ним доступ через поисковые системы

[2] С проектом можно ознакомиться на сайте корпуса: *http://nkjp.pl*

двух ранее разработанных корпусов (IPI и PELCRA). Ниже, при описании этих корпусов, мы будем кратко останавливаться и на характеристиках демо-версий нового корпуса.

Каждый из уже разработанных общедоступных корпусов польского языка по некоторым параметрам не соответствует представлению о современном корпусе как эффективном инструменте исследования определенного языкового состояния. (Собственно, это и побудило их создателей к запуску проекта НКПЯ). Наибольшим потенциалом в этом смысле обладает корпус IPI PAN — большой корпус, снабженный лемматизацией и морфологической разметкой, однако он довольно однороден по своему составу. Корпуса PELCRA и PWN представляют уступающие по объему, но более сбалансированные коллекции, однако в них отсутствует морфологическая разметка (в корпусе PWN проведена только лемматизация).

Корпус IPI PAN. Корпус разрабатывался в Институте основ информатики Польской академии наук в рамках проекта, поддержанного Государственным комитетом научных исследований, с 2001 г. Именно Институт основ информатики в настоящее время является координатором проекта по созданию Национального корпуса польского языка.

Состав. Создавая корпус, авторы включали в него все доступные тексты вне зависимости от их типа или даты возникновения, поэтому в своем полном варианте (250 млн. словоупотреблений[3]) он крайне нерепрезентативен. Основную часть корпуса образуют газетные, юридические тексты и стенограммы парламентских слушаний. С целью создания более представительной (с т. зр. типа текста) коллекции была подготовлена выборка объемом 30 млн. (доступна также ее предыдущая версия объемом 15 млн., включающая газетные тексты (49,3%), художественную литературу (20,3%, в т.ч. классическую конца XIX — начала XX вв. (9,7%)), стенограммы парламентских слушаний (15,5%), научные тексты (10%), юридические тексты (4,9%), см. [Przepiórkowski 2006]; состав 30-миллионной выборки разработчики не указывают, неиз-

[3] Демонстрационная версия НКПЯ через поисковую систему IPI предоставляет доступ к корпусу объемом 430 млн. текстоформ, однако сведений о составе этой текстовой выборки на сайте корпуса нет.

вестно и распределение текстов по дате создания). Отдельный подкорпус составляет также разработанный в 60–70-е гг. корпус, ставший основой для словаря [Kurcz et al. 1990], который в рамках данного проекта был вычищен и снабжен новой разметкой (подкорпус freq, объем 0,5 млн.). В нем по 20% приходится на популярно-научные тексты, художественную прозу, драму, новостные и длинные публицистические статьи.

Метаразметка. Данный тип аннотации включает только 5 атрибутов: имя автора, название произведения, год издания, год первого издания и год создания.

Морфологическая разметка. Как и в ЧНК и СНК, в корпусе IPI сначала осуществлялась автоматическая лемматизация и морфологическая разметка, которая приписывала каждой словоформе все возможные варианты разбора, а затем на основе статистических закономерностей было проведено автоматическое снятие омонимии. Примечательным при этом является то, что в корпусе сохраняются варианты разбора, отвергнутые программой снятия омонимии, так что при желании пользователь может вести поиск по всем вариантам разбора. Такое решение открывает целый ряд дополнительных возможностей: например, позволяет выявлять все омонимичные формы определенного типа или искать ошибки автоматической программы снятия омонимии. Каждой грамматической категории при разметке соответствует отдельный атрибут с заданным набором значений (напр., число (единств./множ.), лицо (1/2/3) и т.д.)—форма, которая в силу своей традиционности является удобной для пользователя. В корпусе freq лемматизация и морфологическая разметка осуществлялись вручную.

Поиск в корпусе. Для поиска в корпусе была специально разработана система Poliqarp, основанная на синтаксисе языка регулярных выражений. Как и в ЧНК и СНК, поиск может вестись по заданному значению любых атрибутов: словоформе или ее части, лексеме или ее части, последовательности словоформ/лексем с указанием расстояния между ними или с заданием структурного единства (предложение, абзац), в пределах которого заданные

единицы должны встретиться, а также по любой комбинации грамматических признаков (как по разметке с автоматически снятой, так и по разметке с неснятой омонимией). При поиске могут учитываться знаки препинания и положение искомой единицы относительно начала/конца предложения/абзаца. При помощи языка запросов можно ограничить поиск заданными значениями метаатрибутов (например, по году создания текста).
Конкорданс выдается в формате KWIC. По команде пользователя возможно отображение леммы и/или грамматических признаков при искомом выражении или во всех выданных словах. Не предусмотрено получение метаинформации об источнике текста. Максимальный контекст выдачи составляет по 20 текстоформ слева и справа от искомого выражения, выбранный контекст может быть расширен до 200 текстоформ. Возможна сортировка выданных контекстов по искомому выражению, а также по первому слову левого или правого контекста (заметим, что упорядочение по началу левого контекста осмысленно только в том случае, если задан размер левого контекста, равный единице; гораздо более удобная система предусмотрена в программе Bonito, где сортировка начинается с ближайшего от искомого слова слева, вслед за которым учитывается второе от него слово слева и т.д.). Предусмотрена и обратная сортировка по искомому выражению, левому или правому контексту (т. е. по концу соответствующего фрагмента).

Доступ к корпусу. Корпус находится в открытом доступе.

Корпус PELCRA. Разработка корпуса ведется с 1996 г. на Кафедре английского языка университета г. Лодзь в рамках совместного проекта с Отделением лингвистики и современного английского языка университета Ланкастера. Наряду с одноязычным польским корпусом ведется работа над созданием англо-польского параллельного корпуса, а также польского учебного корпуса английского языка.

Состав. Структура корпуса строилась во многом по модели Британского национального корпуса. Планируемый объем корпуса—100 млн. словоупотреблений, на сегодняшний день для поис-

ка доступно 93 млн.[4] 90% корпуса образуют письменные тексты (в т.ч. 13,5% художественные, 76,5%—остальные), 10%—устные. Основной массив текстов относится к 1992–2003 гг., нижней временной границей включения текстов является 1989 г., исключение делается только для некоторых художественных текстов. Устный подкорпус состоит из двух неравных частей: осуществленные в рамках проекта записи непубличных разговоров (всего свыше 160, объем—600 тыс. словоупотреблений, планируется довести до 1 млн.) и транскрипции устной речи официального характера (публичные выступления, дебаты, интервью и т. п.)

Метаразметка. Аннотация письменных текстов учитывает довольно мало параметров: автор, название, источник текста, тип текста (*письменный*, *устный—официальный* или *неформальный*), тип носителя (*книга*, *интернет* и т.д.), дата публикации. Для устных текстов размечаются пол, возраст говорящего и уровень его образования.

Морфологическая разметка. Потенциал корпуса значительно снижает фактическое отсутствие лемматизации и морфологической разметки. Единственная опция в области лемматического поиска, которую предполагает корпус,—это выдача по заданной словоформе всей парадигмы в виде списка (с возможностью указания частотности для каждой формы) с последующим поиском по каждой словоформе в отдельности[5].

Поиск в корпусе. Корпус предполагает несколько типов поиска, незначительно различающихся по синтаксису запроса и параметрам выдачи. Ниже обобщаются основные поисковые возможности. Предусмотрен поиск по словоформе или ее части,

[4] Демонстрационная версия НКПЯ через поисковую систему PELCRA предоставляет доступ к корпусу объемом 350 млн. текстоформ, полученному в результате объединения материалов трех ресурсов – корпуса IPI PAN, самого корпуса PELCRA и корпуса PWN (см. ниже), а также добавления ряда новых текстов. Точный состав и процентное соотношение типов текстов в итоговой выборке разработчики не указывают.

[5] Возможности поисковой системы PELCRA были расширены для текстов НКПЯ, доступных в демонстрационной версии нового корпуса: эти тексты прошли лемматизацию, тем самым по запросу пользователь может получать все словоформы заданной лексемы.

нескольким словоформам или их частям (следующим непосредственно друг за другом или находящимся в пределах одного предложения/абзаца, нельзя задать расстояние между единицами). При запросах на словоформы в составе предложения/абзаца доступны также логические операторы ИЛИ и НЕ. Поиск можно ограничить определенным типом и/или носителем текста и годом его публикации для письменных текстов и определенным полом, возрастом и уровнем образования говорящего—для устных. Еще один—необычный—параметр, по которому могут накладываться ограничения в письменных текстах,—это тип предложения (утвердительное, вопросительное, восклицательное)[6].

Формат выдачи—KWIC или обычный текст (в зависимости от типа запроса). Упорядочение выдачи возможно по искомому выражению, первому слову левого или правого контекста (неудобство сортировки по первому слову левого контекста уже обсуждалось выше в связи с корпусом IPI), а также по источнику текста. Максимальный контекст выдачи—1 предложение или 1 абзац (в зависимости от типа запроса). В окне результатов существует возможность расширения выбранного контекста до 3 абзацев (по 1 до и после того, в котором встретилось искомое выражение). Количество выдаваемых контекстов ограничено 250 примерами. При

[6] Несколько иные возможности предоставляет поисковая система PELCRA для текстов НКПЯ: здесь также предусмотрен поиск по словоформе или ее части, нескольким словоформам или их частям, однако в данном случае пользователь может и задавать расстояние между единицами. Кроме того, как уже отмечалось, доступен поиск по лемме или комбинации из нескольких лемм. При формулировании запроса возможно использование логического оператора ИЛИ. По выбору пользователя искомые слова в итоговых контекстах могут располагаться в произвольном порядке или же только в заданной последовательности. Результаты выдаются в формате KWIC. Возможна сортировка получаемых контекстов по ключевому слову, по левому или правому контексту (учитывается сначала ближайшее слово слева или справа от искомого выражения, потом второе и т.д.). При этом, однако, сортируются только результаты в пределах каждой отдельной страницы выдачи (ее объем выбирает пользователь— от 10 до 1000 контекстов). По запросу выдается расширенный контекст — три предложения (по одному слева и справа от того, в котором встретилось искомое выражение), а также метаинформация—автор, название и год создания текста.

этом если поиск ведется по одной словоформе (не по группе), система создает сбалансированный подкорпус в 10 тыс. текстов, на материале которого ведется поиск. При повторном запросе формируется новый подкорпус, который, соответственно, может выдать другие примеры. Для каждого контекста можно просмотреть информацию о его источнике.

Система позволяет получать ряд статистических данных. Кроме уже упомянутой выше частотности словоформ в составе парадигмы возможна выдача списка (до 2000 единиц) самых частотных словоформ в корпусе, частотного списка для выбранного текста, а также списка коллокаций для данной словоформы, упорядоченного по абсолютной частотности (учитывается по выбору левый или правый контекст) или по статистическому параметру MI3 (задается размер контекста, в пределах которого ищутся коллокации).

Доступ к корпусу. Корпус находится в открытом доступе.

Корпус PWN. Корпус разработан Польским научным издательством и служит основой для выпускаемых им словарей.

Состав. Корпус состоит из двух частей—коллекции текстов различных типов (общий объем 22 млн. словоупотреблений, в т. ч. художественная литература 20%, книги non-fiction 21%, газеты и журналы 45,5%, устная речь 4,5%, тексты «эфемерных жанров» 5,5%, тексты из Интернета 3,5%) и текстов из газеты Rzeczpospolita (18 млн.). Временной охват текстов в корпусе PWN довольно широк: с 1925 по 2005 гг. для прессы и с 1903 по 1997 гг. для книг, хотя основной массив корпуса образуют все же тексты, созданные во второй половине века. В подкорпус Rzeczpospolita вошли отдельные номера газеты с 1997 по 2005 гг. В открытом доступе находятся уменьшенные варианты двух составляющих корпуса: 3,7 млн.—общий корпус и 3,6 млн.—Rzeczpospolita. Отметим, что вопреки современному стандарту создания корпусов в общий раздел корпуса PWN включались не целые тексты, а их фрагменты.

Метаразметка. Тексты аннотируются по автору, его полу, возрасту (по принадлежности одной из 7 групп) и уровню образования, по названию, году издания, типу (учитывается 8 типов—*рассказ*, *статья*, *разговор*, *письмо* и др.), однако для пользователя эти данные не имеют практического значения, поскольку

не могут учитываться при поиске. Кроме того, ряд элементов маркируется внутри текста—это, например, единицы иностранного происхождения, неправильные формы (с указанием соответствующей правильной), диалектные формы, цитаты из текстов, значительно отстоящих по времени создания от основного текста, в устных текстах—паузы, наложения реплик и др. По запросу пользователя эта разметка может отображаться при выдаче.

Морфологическая разметка. Корпус лемматизован (грамматическая омонимия не снята), однако морфологическая разметка отсутствует.

Поиск в корпусе. Поиск можно осуществлять по словоформе или ее части, лексеме или ее части (причем по умолчанию поиск ведется именно по лексеме), по нескольким словоформам/лексемам, расположенным на заданном расстоянии друг от друга (в открытой версии корпуса возможен поиск только по следующим друг за другом единицам). Размер контекста выдачи устанавливается пользователем и фактически не ограничен. Пользователь может регулировать и количество выдаваемых примеров. Возможна сортировка контекстов по искомому выражению, первому слову левого или правого контекста (проблема сортировки по левому контексту здесь, как и в корпусах IPI и PELCRA, решена неоптимальным способом). Контексты выдаются в формате KWIC. Для каждого контекста можно получить информацию о его источнике. Доступна версия контекста с метатекстовыми тэгами (т. е., например, с разметкой неправильных форм, диалектных элементов и т. п., см. выше в разделе «Метаразметка»).

Доступ к корпусу. Доступ к полной Интернет-версии корпуса осуществляется на платной основе, в открытом доступе находится небольшая часть корпуса с несколько ограниченными поисковыми возможностями (см. выше).

2. Южнославянские языки

Словенский

Уровень корпусной оснащенности южнославянских языков в целом ниже, чем западнославянских. Пожалуй, наиболее разработанной областью в этом отношении является словенский язык. На

сегодняшний день только для него созданы большие представительные корпуса с лингвистической разметкой—корпус FIDA и его существенно расширенная версия—корпус FidaPLUS. В распоряжении исследователей словенского есть и другой большой ресурс—Nova beseda, однако он нерепрезентативен и не снабжен аннотацией лингвистического уровня. Еще одним инструментом изучения словенского может стать система WWW-Concordance, представляющая маленькую коллекцию специализированных текстов, в части которой была проведена морфологическая разметка. Наконец, следует отметить другой небольшой специализированный корпус, представляющий ответвление проекта FidaPLUS,—корпус KoRP—это морфологически аннотированная коллекция текстов по тематике «Связи с общественностью». В настоящее время ведется также работа по созданию корпуса с синтаксической разметкой—Slovene Dependency Treebank, который строится по образцу PDT. На данном этапе подготовлен небольшой фрагмент этого корпуса.

Корпуса FIDA и FidaPLUS. FIDA представляет собой коммерческий продукт, создававшийся с 1997 по 2000 гг. в университете Любляны и Институте им. Йозефа Стефана при поддержке коммерческих организаций DZS и Amebis. Впоследствии на базе этого корпуса был создан новый, существенно расширенный ресурс—корпус FidaPLUS (проект поддержан Министерством образования Словении).

Состав. Разработчики FIDA ориентировались на стандарт, заданный Британским национальным корпусом—ресурс задумывался как 100-миллионный представительный корпус современного языка. Новый корпус FidaPLUS существенно превышает старый по объему: он содержит 621 млн. словоупотреблений. Тексты в FIDA относятся ко второй половине XX в., при этом большинство создано в 90-е гг., основной массив текстов в FidaPLUS охватывает временной интервал с 1990 по 2006 гг. Пропорции типов текстов в двух корпусах в некоторой степени отличаются (поскольку при этом оба корпуса считаются сбалансированными, можно предположить, что различия состава отражают произошедшие между выпуском FIDA и FidaPLUS изменения в функционировании словенского языка, хотя в отличие от ЧНК авторы не сообщают

о социолингвистических исследованиях, позволяющих сделать подобный вывод). Типы текстов, с одной стороны, и типы носителей—с другой, представлены в FIDA и FidaPLUS соответственно в следующих соотношениях: художественные тексты (6 vs. 3,47%), научные (18,5 vs. 10%), другие (75,5 vs. 86,34%); книги (22,7 vs. 8,74%), газеты (46,6 vs. 65,26%), журналы (23,9 vs. 23,26%), тексты из Интернета (электронные тексты) (0,02 vs. 1,24%), другое (в т. ч. незначительная доля устной речи—стенограмм парламентских слушаний) (6,78 vs. 1,5%).

Метаразметка. При параметризации текстов большое внимание уделялось типу носителя (книга, газета, журнал с классификацией последних по периодичности и т.д.). К другим метаатрибутам, учитываемым при поиске, относятся тип текста (художественные—проза, поэзия, драма; научные—гуманитарные и технические; прочие) и год его создания.

Морфологическая разметка. В корпусе проведена лематизация и морфологическая разметка. Аннотация выполнена на основе рекомендаций для словенского языка, выработанных в рамках международного проекта по развитию языковых ресурсов Multext-East (*http://nl.ijs.si/ME*). Грамматический тэг представляет собой цепочку символов, в которой каждая позиция соответствует значению определенной грамматической категории. Для каждой части речи предусмотрена своя схема тэга (тем самым принципы морфологической разметки в FIDA и FidaPLUS сходны с разметкой в СНК). Грамматическая омонимия в корпусах снята частично, процедура осуществлялась автоматически на основании статистических закономерностей. При этом пользователь имеет доступ к разборам, отвергнутым автоматической программой снятия омонимии.

Поиск в корпусах. Запрос может строиться по словоформе/ее части, лексеме/ее части, последовательности словоформ/лексем, находящихся на заданном расстоянии друг от друга или в пределах одного предложения, а также по грамматическим признакам. При формулировании запроса возможно использование логических операторов. Конкорданс выдается в формате KWIC. Контекст выдачи составляет по несколько слов слева и справа от искомого. По команде пользователя высвечивается подробная

информация об источнике текста, а также выдается расширенный контекст—абзац, в котором встретилось искомое выражение. Предусмотрена возможность просмотра контекста с полной морфологической разметкой.

Выдачу можно упорядочить по левому или правому контексту (как и в ЧНК, в общем случае сортировка сначала учитывает ближайшее к искомому слово—по выбору пользователя слева или справа, затем следующее и т. д. Кроме того, пользователь может сам задать позицию—от 1-й до 4-й вправо или влево от искомого выражения, по которой будет осуществляться сортировка). К сожалению, упорядочивание собственно по искомой цепочке не предусмотрено, что при наличии поиска по грамматическим признакам было бы удобной опцией. К другим возможностям обработки полученного конкорданса относится фильтрация найденных примеров (т. е. можно отсеять не подходящие пользователю контексты, в качестве условий фильтрации задаются значения любых атрибутов—определенные словоформы, лексемы, грамматические признаки, которые должны—или же не должны—находиться на заданном расстоянии от искомого выражения).

Поиск можно ограничить по метаатрибутам (типу текста, типу носителя, году создания). Предусмотрены также некоторые функции, связанные со статистической обработкой данных, в частности, выдача для заданной единицы частотного списка коллокаций, включающего значения статистических параметров MI и MI3, с возможностью определения размера учитываемого контекста.

Доступ к корпусам. Доступ к корпусу FIDA осуществляется на платной основе. В демонстрационной версии, находящейся в открытом доступе, по запросу пользователя выдается не более 10 контекстов. Для доступа к FidaPLUS необходимо пройти регистрацию (для исследовательских целей осуществляется бесплатно).

Корпус Nova beseda (NB). Корпус разрабатывается с 1999 г. в Институте словенского языка Словенской академии наук. Исследователи рассматривают нынешний корпус как шаг на пути к созданию Словенского национального корпуса.

Состав. В своем нынешнем виде корпус несбалансирован. Коллекцию объемом 240 млн. словоупотреблений образуют 7 под-

корпусов: тексты газеты DELO за 1998–2007 гг. (70,4%), стенограммы парламентских слушаний 1996–2007 гг. (12,9%), оригинальная и переводная художественная литература (5%), литература non-fiction (0,83%), научная и техническая литература (1,25%), журнальные тексты (8,75%), тексты законодательства Словении (5%).

Метаразметка. Тексты классифицируются по следующим параметрам: автор, название, оригинальный vs. переводной, жанр и тип (*проза, поэзия, драма*—для художественной литературы, *мемуары, эссе* и т.д.—для non-fiction, тематика, т.е. *естественно-, гуманитарно-научные* или *юридические тексты*,—для научно-технической литературы).

Морфологическая разметка. Лемматизация и грамматическая аннотация в корпусе отсутствует.

Поиск в корпусе. Поиск может осуществляться по словоформе, ее начальной части или по нескольким словоформам (их начальным частям), следующим непосредственно друг за другом. Формат выдачи—KWIC. Сортировка контекстов не поддерживается. Максимальный контекст выдачи—по 1 предложению слева и справа от того, в котором встретилось искомое выражение. По запросу пользователя выдается информация об источнике текста. Поиск можно ограничить по любому метапараметру (заметим, что год создания текста, не вынесенный в метаатрибуты, соответственно, не может учитываться при поиске). Другой вид запроса, предусмотренный в NB,—это запрос на список слов, отвечающих определенным параметрам. Здесь поиск может вестись по любым буквенным последовательностям, входящим в состав словоформы, по количеству букв в ее составе, по частотности словоформы в корпусе. Запрос может включать логические операторы.

Доступ к корпусу. Корпус находится в открытом доступе.

Система WWW-Concordance. Система позволяет осуществлять поиск по нескольким специализированным коллекциям текстов разного уровня аннотации.

Состав. Система включает следующие текстовые собрания: перевод на словенский романа Дж. Оруэлла «1984» (корпус, подготовленный и размеченный в рамках международного проекта Multext-East, см. подробнее выше в связи с ЧНК, объем—90 тыс. словоупотреблений), коллекция газетных статей конца 80-х гг.

о Югославской национальной армии (270 тыс.) и записи электронной конференции по горному делу (300 тыс.), DSI—материалы Словенской конференции по информатике за 2003–2007 гг. (1,4 млн., корпус подготовлен Отделом языка Словенского общества информатики как основа для электронного словаря по информатике). Однородность текстовых коллекций не предполагает осуществления метаразметки.

Морфологическая разметка. Лемматизация и грамматическая разметка проведены для двух подкорпусов—корпуса «1984» (осуществлены в рамках проекта Multext-East, о принципах разметки см. FIDA), а также для корпуса DSI. Обратим внимание, что в двух корпусах использовались разные аннотационные формализмы, соответственно, для правильного построения запросов каждая система требует отдельного изучения. Грамматическая омонимия снята.

Поиск в корпусе. При поиске может использоваться мощный аппарат языка регулярных выражений. Запрос может строиться по словоформе/ее части, нескольким словоформам/их частям, находящимся на заданном расстоянии друг от друга или в пределах одного предложения. В корпусах «1984» и DSI все те же типы поиска могут осуществляться и по значениям других доступных атрибутов—лемм и грамматических признаков. Формат выдачи—KWIC или обычный текст. В формате KWIC пользователь может установить размер контекста выдачи—от 10 до 160 знаков справа и слева от искомого выражения, при этом чем больше заданный размер контекста, тем меньше максимально возможное число выдаваемых примеров (при контексте в 10 знаков пользователь получает не более 2000 контекстов, при ограничении в 160 знаков—не более 125). В формате обычного текста количество примеров не ограничено, однако размер контекста составляет примерно по 20 знаков справа и слева от искомого выражения. Предусмотрена также выдача в виде списка слов, отвечающих заданному условию, с указанием частотности для каждого из элементов списка.

Доступ к корпусу. Корпус находится в открытом доступе.

Корпус KoRP. Корпус текстов по тематике «Связи с общественностью» разрабатывается с 2006 г. на социологическом факуль-

тете университета Любляны как основа для терминологического словаря данной предметной области.

Состав. В корпус вошли оригинальные (73,2%) и переводные (26,8%) тексты нескольких типов—научные, специальные и популярные статьи, тезисы конференций, монографии, учебники, интервью, дипломные и магистерские работы, рецензии и под., связанные с изучаемой предметной областью. Временной охват текстов—с 1994 по 2007 гг., основная доля (70%) приходится на 2002–2006 гг. Общий объем корпуса—1,8 млн. словоупотреблений.

Метаразметка. При поиске могут учитываться следующие метапараметры—год создания текста, функциональная сфера (все тексты по этому признаку делятся на научные, специальные и популярные), тип носителя (книга—электронная публикация), исходный язык текста (оригинал—перевод).

Морфологическая разметка. Как уже отмечалось, KoRP является ответвлением проекта FidaPLUS, соответственно, принципы лемматизации и морфологической разметки двух корпусов совпадают (см. описание выше).

Поиск в корпусе. Поисковый интерфейс корпуса KoRP также повторяет систему поиска, реализованную в FidaPLUS, тем самым пользователь KoRP имеет столь же широкий спектр возможностей в области построения запросов и обработки выданных контекстов (см. FidaPLUS).

Доступ к корпусу. Для доступа к KoRP необходимо пройти регистрацию (для исследовательских целей осуществляется бесплатно).

Slovene Dependency Treebank (SDT). Работа над созданием синтаксически аннотированного корпуса ведется с 2003 г. в Институте им. Йозефа Стефана и Институте словенского языка Словенской академии наук.

Состав. Подготовленная на сегодняшний день версия SDT представляет собой фрагмент корпуса, созданного в рамках проекта Multext-East (см. WWW-Concordance), а именно, в SDT вошла первая часть словенского перевода романа Дж. Оруэлла «1984» объемом 30 тыс. словоупотреблений (2 тыс. предложений). В дальнейшем предполагается расширение состава корпуса, в частности,

за счет интернет-текстов как наиболее приближенных к сфере возможного применения будущего корпуса (см. [Džeroski et al. 2006]).

Синтаксическая разметка. Поскольку аннотация морфологического уровня с ручным снятием омонимии была осуществлена уже на этапе проекта Multext-East, то в рамках подготовки SDT вся работа была направлена на синтаксическую разметку. Последняя строилась по образцу PDT, однако пока что разработчики SDT реализуют только аннотацию «аналитического» уровня, не обращаясь к более глубокому «тектограмматическому» слою языковой информации (ср. PDT). На первом этапе разметка осуществляется автоматически, затем построенные таким образом деревья зависимостей проверяются вручную.

Доступ к корпусу. Желающим работать с корпусом предлагается написать электронное письмо с соответствующей просьбой его разработчикам (адрес указан на сайте корпуса).

Хорватский

Хорватский является единственным среди южнославянских языков, для которого на сегодняшний день разработан национальный корпус (ХНК). ХНК характеризуется широкими поисковыми возможностями, но пока что довольно незначительным объемом лингвистически аннотированных текстов и несбалансированностью состава. На основе фрагмента ХНК с 2006 г. ведется работа по созданию синтаксически аннотированного корпуса Croatian Dependency Treebank, который строится по образцу PDT. Результаты этой работы пока недоступны. Кроме ХНК разрабатывается корпус Croatian Language Repository (CLR), нацеленный на отражение стандартного хорватского языка и включающий, соответственно, ограниченный набор типов текстов. Лингвистическая разметка CLR пока не осуществлена.

Хорватский национальный корпус (ХНК). Работа над ХНК ведется с 1996 г. в Институте лингвистики Загребского университета.

Состав. Статус национального определяет тот факт, что ХНК естественно задумывался как сбалансированный: были заранее определены процентные соотношения разных типов текстов в со-

ставе будущего 100-миллионного корпуса. Однако на настоящий момент не все типы текстов собраны в предусмотренном для них объеме, поэтому пользователю временно открыты все имеющиеся коллекции текстов без соблюдения их пропорций в корпусе общим объемом 101 млн. словоупотреблений: это газетные и журнальные тексты с 1990 по 2005 гг. (97 млн., разбиты на несколько подкорпусов по названию издания) и художественная литература с XVI в. (ок. 4 млн., 2 подкорпуса—классическая литература и произведения М. Марулича).

Метаразметка. Для корпуса разработана типология текстов, учитывающая тип носителя, тематику, жанр и др. (см. [Tadić 2002]), однако она в полном объеме не включена в разметку корпуса.

Морфологическая разметка. В небольшой части корпуса (подкорпус текстов газеты Croatia Weekly за 2000 г., cw2000, объем 118 тыс. словоупотреблений) была проведена лемматизация и морфологическая разметка с последующим ручным снятием омонимии. Аннотация выполнена на основе рекомендаций для хорватского языка, выработанных в рамках международного проекта по развитию языковых ресурсов Multext-East, тем самым разметка ХНК сходна с реализованной в FIDA.

Поиск в корпусе. Для поиска используется обсуждавшаяся выше программа Bonito (см. ЧНК), соответственно, пользователю предоставляется широкий потенциал поисковых возможностей, настройки параметров выдачи и статистической обработки информации. Нужно, однако, иметь в виду, что в связи с особенностями разметки ХНК все типы запросов, основанные на лемме или грамматических признаках, доступны пока только в маленьком подкорпусе cw2000. Не поддерживается пока и поиск с ограничениями по метаатрибутам.

Доступ к корпусу. В период разработки корпус находится в открытом доступе.

Croatian Language Repository (CLR). CLR разрабатывается с 2005 г. в Институте хорватского языка и лингвистики при поддержке Министерства образования, науки и спорта.

Состав. Нацеленность проекта на отражение *стандартного* хорватского языка определяет особенности его состава. В него

включаются только письменные тексты—в первую очередь художественная и публицистическая литература, переводные тексты выдающихся переводчиков, научные тексты разной тематики, учебники, интернет-журналистика. Нижней временной границей включения текстов определена середина XIX в. (в рамках проекта планируется создать также корпуса древне- и среднехорватского языков). Объем корпуса на настоящий момент составляет 71 млн. словоупотреблений, планируется его расширение до 180 млн.

Метаразметка. Для аннотации текстов используется незначительное число параметров: автор, название, год создания, объем текста, язык оригинала и некоторые библиографические сведения (место и год публикации, издательство).

Морфологическая разметка. На данном этапе лемматизации и морфологической разметки в корпусе нет.

Поиск в корпусе. Запросы осуществляются по всему корпусу или отдельно по художественным и газетным текстам. Поиск может вестись по словоформе или ее части, последовательности словоформ, находящихся на заданном расстоянии друг от друга или в пределах одного предложения/абзаца. Возможно использование регулярных выражений. Поиск с учетом знаков препинания не поддерживается. По заданной последовательности можно найти близкие по буквенному составу словоформы с указанием частотности для каждой (это в некоторой степени заменяет поиск словоформ в составе парадигмы, но в отличие от опции, реализованной в корпусе PELCRA, в данном случае в списке, безусловно, окажутся лишние формы и могут потеряться нужные). Формат выдачи—KWIC (5 текстоформ слева и 7 справа от искомого) или обычный текст (примерно по 40 текстоформ слева и справа). По запросу можно просмотреть расширенный контекст (вплоть до 3 страниц исходного печатного текста или 3 абзацев для газетных статей). Возможна сортировка контекстов по искомому выражению, соседнему левому или правому слову, а также по метаданным (автору, названию текста, году создания). Поиск можно ограничить любыми метаатрибутами. Для изучения роли слова в коммуникативной структуре предложения предусмотрена возможность поиска словоформы отдельно в начальной, конечной или срединной части клаузы.

Корпус позволяет проводить различные типы статистического анализа данных. Кроме частотных списков словоформ для всего корпуса и каждого из входящих в его состав текстов можно по заданной словоформе получить ее распределение по различным метахарактеристикам текста, т.е. изучить ее встречаемость (абсолютную или относительную к общему числу слов) у разных авторов, в разных текстах, в разные периоды времени (при этом временной интервал распределения может составлять от одного года до века). Кроме того, предусмотрена возможность получения для заданной словоформы списка 100 самых частотных коллокаций с указанием размера учитываемого контекста (во избежание получения случайных коллокаций при поиске могут не учитываться 120 самых частотных слов корпуса).

Доступ к корпусу. Корпус находится в открытом доступе.

Боснийский

Корпусные ресурсы для боснийского языка на сегодняшний день довольно ограниченны. Разработанный в Осло Корпус боснийских текстов, открывшийся в Интернете в 1998 г., был одним из первых среди славянских языков общедоступных ресурсов. С тех пор корпус не претерпел значительных изменений. Неудивительно поэтому, что с точки зрения современных стандартов этот корпус несколько устарел: он характеризуется небольшим объемом и отсутствием лингвистической разметки. Маленький корпус устной речи (Корпус боснийских интервью), созданный в рамках исследовательского проекта в университете г. Тюбинген, будет рассмотрен вместе с другими корпусами, разработанными по тем же принципам, в разделе о сербских корпусах.

Корпус боснийских текстов (КБТ). КБТ разрабатывался с 1996 г. в университете Осло в рамках совместного проекта Отделения восточноевропейских исследований и Лаборатории по обработке текстов. В настоящее время ресурс, по всей вероятности, не развивается.

Состав корпуса. В корпус вошли следующие типы текстов: художественная литература (43%), эссеистика (29,6%), публицистика (16,9%), книги для детей (6%), религиозные тексты (2,8%), юридические тексты (1,5%), фольклор (0,2%). Большинство тек-

стов относятся к 90-м гг. XX в. Общий объем корпуса составляет 1,5 млн. словоупотреблений.

Метаразметка. Тексты в корпусе классифицируются по автору, названию, году издания и типу (типы соответствуют составляющим корпуса—худ. литература, эссеистика и т.д.)

Морфологическая разметка. В корпусе отсутствуют лемматизация и морфологическая разметка.

Поиск в корпусе. Поиск может вестись по словоформе, ее части или по последовательности словоформ, находящихся на заданном расстоянии друг от друга (но запрос не может строиться с учетом структурных единств, так как в корпусе не размечены границы предложений и абзацев). Поддерживается поиск с учетом знаков препинания. Формат выдачи—KWIC или обычный текст. Максимальный общий размер выдаваемого контекста—500 знаков или 200 слов. Сортировка контекстов не поддерживается.

Доступ к корпусу. Для доступа к корпусу необходимо пройти регистрацию (для исследовательских целей осуществляется бесплатно).

Сербский

Ситуацию в сербской корпусной лингвистике можно оценивать двояко: с одной стороны, существует текстовая коллекция—Корпус сербского языка, снабженная подробной лингвистической разметкой и предназначенная для размещения в Интернете, с другой—планировавшаяся вывеска так и не состоялась, и проект в настоящее время, по-видимому, не развивается. Тем самым корпус не представляет практического интереса для пользователя. Ниже будут кратко охарактеризованы основные параметры разработанного корпуса, а также представлены маленькие специализированные корпуса (в т.ч. Новосадский корпус устной речи и Сербский корпус комиксов), созданные в Тюбингенском университете.

Корпус сербского языка (КСЯ). В основу корпуса легла текстовая коллекция, собранная в 1957–62 гг. в Институте экспериментальной фонетики и патологии речи под руководством Д. Костича. Работа над электронным корпусом была начата в 1996 г. в рамках совместного проекта Института с Лабораторией экспериментальной психологии Белградского университета.

Состав. Общий объем корпуса составляет 11 млн. словоупотреблений. Его образуют 5 подкорпусов: корпус современного языка (включающий художественную литературу, публицистику, научные тексты общим объемом 7 млн.) и 4 исторических подкорпуса литературы XII-XIX вв., разбитых по хронологическому принципу (4 млн.).

Морфологическая разметка. КСЯ был вручную лемматизован и снабжен подробными грамматическими пометами. Наряду с этим для каждой словоформы указывалось количество ее букв и слогов и фонологическая структура.

Поиск в корпусе. Как уже указывалось, собственно поиск по корпусу недоступен. Единственная текстовая информация, к которой пользователь имеет доступ—это образцы разметки (объемом по 500 словоформ) для каждого из пяти подкорпусов.

Тюбингенские боснийско-сербско-хорватские корпуса (ТБСХК). Данные корпуса разрабатывались с 1999 по 2001 гг. в рамках проекта по исследованию дейктических элементов. Тематика проекта определяет специфику вошедших в их состав текстов и их разметки.

Состав. Данную группу корпусов образуют три подкорпуса: *Сербский корпус комиксов* (57 тыс. словоформ), *Новосадский корпус разговорной речи* (включает записи спонтанных разговоров, 25 тыс. словоформ), *Корпус боснийских интервью* (интервью с беженцами из Боснии, среди которых есть как этнические боснийцы, так и сербы и хорваты, 45 тыс.). Тем самым все коллекции ориентированы на представление устной речи (тексты комиксов, будучи письменными, призваны имитировать нормы разговорного языка).

Разметка. К элементам аннотации в корпусе относится минимальная информация о говорящем, позволяющая идентифицировать высказывания одного и того же человека (социолингвистические данные отсутствуют), маркируется также язык высказывания. Лингвистическая разметка проводится только для дейктических элементов, которые подразделяются на временные, локативные и т. д.

Поиск в корпусах. Поиск может вестись по словоформе или элементам разметки. Для построения комплексных запро-

сов используется язык XMLQUERY, характеризующийся довольно сложным синтаксисом.

Доступ к корпусам. Корпус находится в открытом доступе.

Болгарский

Ситуация в болгарской корпусной лингвистике характеризуется отсутствием доступных через Интернет аннотированных корпусов, снабженных механизмом поиска. На будущие изменения позволяют надеяться несколько сообщений о ведущихся в настоящий момент работах по созданию корпусов для болгарского языка. Одним из центров этих разработок является Институт болгарского языка, в котором хранится электронный текстовый архив, призванный стать основой национального корпуса. Другой проект осуществляется в Лаборатории лингвистического моделирования Института параллельной обработки информации при Болгарской академии наук. Его цель—создание синтаксически аннотированного корпуса, основанного на формализме HPSG (BulTreeBank). В рамках проекта был собран архив объемом 72 млн. слов, включающий тексты разных типов и жанров. Небольшой фрагмент этой коллекции доступен на сайте проекта в простом текстовом формате. Кроме того, была разработана программа автоматического снятия грамматической омонимии, и с ее помощью корпус объемом 2600 предложений (примерно 53 тыс. текстоформ) получил морфосинтаксическую (т. е. частеречную) разметку. Размеченный корпус находится в открытом доступе в виде одного файла в формате XML (соответственно, поисковый интерфейс не предусмотрен). Для работы с остальными ресурсами, созданными в рамках проекта BulTreeBank, пользователю необходимо направить запрос разработчикам корпуса (для исследовательских целей материалы высылаются бесплатно). Речь идет о двух текстовых коллекциях—во-первых, это морфологически аннотированный корпус объемом 214 тыс. текстоформ (разметка осуществлялась на основе рекомендаций, выработанных в рамках международного проекта по развитию языковых ресурсов Multext-East, ср. корпуса FIDA и FidaPLUS, ХНК и др., см. [Simov et al. 2004]) и, во-вторых, это синтаксически аннотированный корпус объемом 196 тыс. текстоформ (на данном этапе работы,

по крайней мере в доступной версии корпуса, разметка строится в терминах деревьев зависимости).

Среди созданных ранее коллекций болгарских текстов следует отметить одномиллионный представительный корпус, созданный Болгарской ассоциацией по компьютерной лингвистике по модели Брауновского корпуса: он включает 500 текстовых фрагментов по 2000 слов. К сожалению, корпус недоступен в Интернете. Целый ряд текстовых коллекций представлен на сайте Отделения болгарского языка и литературы университета Осло (*http://www.hf.uio.no/east/bulg/mat*). Это прежде всего два собрания текстов устной речи, включающие разговоры в семейном кругу, а также в бытовых ситуациях в различных общественных местах. Там же размещены коллекции стенограмм парламентских дебатов и электронной переписки в чате. Все эти коллекции доступны только в простом текстовом формате, но сложность сбора записей устной речи и, соответственно, редкость такого типа ресурсов делает их ценным материалом, который, хотелось бы надеяться, войдет в будущий большой корпус болгарского языка.

Македонский

Македонский язык на сегодняшний день, к сожалению, не имеет общедоступных корпусных ресурсов. Однако совсем недавно в сфере его электронного обеспечения произошло событие, которое, возможно, изменит эту ситуацию к лучшему: речь идет об открытии в Интернете Архива македонского языка (см. *http://damj.manu.edu.mk/index.html*). В настоящий момент на сайте размещены различные тексты, посвященные македонской лингвистике (в т. ч. грамматики и словари, самый старый из них относится к 1875 г.). Разработчики полагают, что нынешний ресурс ляжет в основу будущего Национального корпуса македонского языка.

Восточнославянские языки

Русский

Русский язык долгое время оставался неохваченным разработками в области современной корпусной лингвистики. Ситуация существенно изменилась за последние несколько лет, когда прак-

тически одновременно в Интернете появился целый ряд корпусных ресурсов. Прежде всего следует назвать Национальный корпус русского языка (НКРЯ)—большую представительную коллекцию, снабженную подробной метаразметкой и богатой лингвистической аннотацией. Поскольку различным аспектам функционирования НКРЯ посвящено большинство статей настоящего сборника, здесь мы не будем на нем подробно останавливаться (основные параметры НКРЯ приведены в обзорной таблице в Приложении). Первым по времени появления в открытом доступе русскоязычным корпусом стал ресурс, разработанный в университете Тюбингена (Тюбингенский корпус). Одним из его достоинств является эффективный язык запросов. К сожалению, объем лингвистически размеченного языкового материала в нем довольно невелик. Совсем небольшой корпус ХАНКО был разработан в университете Хельсинки, к его отличиям относится тщательная ручная разметка. Специализированный Корпус русских газет подготовлен в МГУ им. М. В. Ломоносова. Наконец, относительно недавно в Интернете появился еще один ресурс—Национальный корпус русского литературного языка, нацеленный прежде всего на охват стандартной формы языка, однако пока что этот корпус сильно ограничен в своих возможностях.

Тюбингенский корпус (ТК). Тюбингенский корпус русского языка создавался в рамках проекта по исследованию форм обращения и вежливости в славянских языках с 1999 по 2004 гг.

Состав. ТК разрабатывался в условиях отсутствия каких-либо открытых ресурсов для русского языка, поэтому тексты собирались во многом по принципу доступности. Тем самым корпус представляет собой набор разнородных коллекций. В основу ТК лег знаменитый Уппсальский корпус, который благодаря тюбингенскому проекту стал доступен онлайн и получил лингвистическую разметку (1 млн. слов, 600 текстовых фрагментов, примерно в равной пропорции распределенных между художественной прозой, созданной с 1960 по 1988 гг., и публицистикой 1985–88 гг.). Следующая коллекция отражает специальные исследовательские интересы создателей корпуса—это тексты интервью из различных журналов и газет, а также транскрипции радиоинтервью (с 1996 г.,

290 тыс.). К остальным подкорпусам относятся тексты журнала «Огонек» (1996–2002 гг., 9,19 млн.), собрание детективных романов и другие коллекции художественной литературы XIX и XX вв., разделенные по авторам (более 14 млн.). Общий объем корпуса—более 25 млн. слов.

Метаразметка. Кроме разбиения текстов на подкорпуса классификация текстов не производилась.

Морфологическая разметка. Морфологической аннотацией в ТК снабжены 3 подкорпуса: Уппсальский и коллекции текстов М. А. Булгакова и И. С. Тургенева. Общий объем морфологически аннотированных текстов—2,3 млн. словоупотреблений. Разметка осуществлялась при помощи статистического морфологического анализатора. Однако даже в морфологически размеченной части корпуса отсутствует лемматизация.

Поиск в корпусе. Поиск может вестись по словоформе или ее части, последовательности словоформ/их частей, находящихся на заданном расстоянии друг от друга или в пределах одного предложения, а также—для корпусов с морфологической разметкой—по грамматическим признакам. При построении запросов используется язык регулярных выражений, характеризующийся мощным поисковым потенциалом. Однако существенное неудобство для пользователя составляет отсутствие в открытом доступе списка атрибутов, используемых при грамматической разметке, и их возможных значений.

Формат выдачи—обычный текст. Сортировка контекста возможна только по искомому слову. Максимальный контекст выдачи—по 120 слов или по 6 предложений слева и справа, соответственно, от самого искомого выражения или предложения, в котором оно встретилось. При поиске по морфологически аннотированному корпусу существует возможность отображения при каждом слове в выдаваемом контексте его грамматических характеристик.

Доступ к корпусу. Корпус находится в открытом доступе.

Корпус ХАНКО. Работа над ХАНКО ведется на Отделении славянских и балтийских языков и литератур Хельсинкского университета с 2001 г. Одним из основных принципов построения корпуса является его направленность на максимальный охват граммати-

ческой информации, а не на объем материала. В настоящее время в корпусе проведена морфологическая и синтаксическая разметка, планируется осуществление подробной семантической аннотации (подробнее см. [Мустайоки и др. 2005]).

Состав. В корпус вошли все крупные статьи из журнала «Итоги» за январь 2001 г. Общий объем корпуса составляет 100 тыс. словоупотреблений.

Метаразметка. Будучи довольно однородным по текстовым параметрам, ХАНКО содержит минимальную метаинформацию: номер журнала и тип текста (статья, рецензия, интервью), однако эти параметры не предназначены для задания подкорпуса.

Морфологическая разметка. Корпус снабжен лемматизацией и морфологической разметкой. Процедура осуществлялась автоматически с последующим ручным снятием омонимии. Нацеленность проекта на детальность аннотации и небольшой объем определяют более подробную и аккуратную по сравнению с остальными корпусами русского языка систему морфологической аннотации. Это проявляется, например, в разметке аналитических форм, составных и дробных числительных, разрывных форм местоимений (*ни от кого*) и др.

Синтаксическая разметка. В основу разметки положена система синтаксического анализа, традиционная для грамматических описаний русского языка. В корпусе учитываются следующие типы синтаксической информации: параметры предложений (простое или сложное с дальнейшим делением по типам связи—сочинительной, подчинительной, бессоюзной), параметры клауз (роль—самостоятельная, главная или зависимая; структура—одно- или двусоставная, фразеологизированная; эллиптическая), функция слова в предложении (подлежащее; сказуемое; части именного сказуемого—связочная и присвязочная; главный член односоставного предложения; дополнение; определение; обстоятельство; слово, не являющееся членом предложения—обращение).

Поиск в корпусе. Поиск может вестись по словоформе или ее части, по лексеме или ее части, последовательности словоформ/лексем, находящихся на заданном расстоянии друг от друга, по морфологическим и синтаксическим признакам. В случае запроса

по морфологическим или синтаксическим атрибутам пользователь может получить для каждого из них список всех возможных значений и выбрать интересующие его параметры (тем самым система поиска здесь сходна с реализованной в НКРЯ). Возможен поиск с учетом знаков пунктуации. Формат выдачи—обычный текст. По запросу пользователь получает расширенный контекст (по 5 предложений слева и справа от того, в котором встретилось искомое выражение), а также информацию о грамматических признаках словоформ и синтаксических параметрах членов выданного предложения.

Доступ к корпусу. Корпус находится в открытом доступе.

Корпус газетных текстов (КГТ). «Компьютерный корпус газетных текстов русского языка конца XX века» был подготовлен в течение 2000–2002 гг. в Лаборатории общей и компьютерной лексикологии и лексикографии филологического факультета МГУ им. М. В. Ломоносова.

Состав корпуса. В КГТ вошли полные тексты избранных номеров ряда российских газет на русском языке, опубликованных в 1994–1997 гг. При отборе материала авторы ставили задачу создания репрезентативной выборки с учетом периодичности издания, его политической направленности, аудитории (центральные vs. региональные, общие vs. профессиональные). Общий объем корпуса—свыше 11 млн. словоупотреблений, однако доступная в Интернете версия существенно отличается от исходной: она насчитывает 200 тыс. слов, планируется ее увеличение до 1 млн.

Метаразметка. Метаописания включают название газеты, дату ее выпуска, а также жанр в терминах детальной жанровой классификации статей. На основе анализа материала был выявлен круг основных жанрообразующих факторов, характеризующих предмет сообщения, его коммуникативную цель и композиционно-стилевую форму. По этим параметрам было выделено 9 жанровых типов (собственно информационные, информационно-публицистические, собственно публицистические, художественно-публицистические, рекламные жанры и др.), которые распределяются между 96 конкретными жанрами. Использование такой подробной жанровой классификации представляется небесспорным. Во-первых, для исследования лингвистических особенностей то-

го или иного жанра необходимо, чтобы каждому из них соответствовало значительное количество статей в корпусе. Очевидно, что при нынешнем числе статей (446) разбиение на 96 жанров не имеет практического смысла для пользователей. Во-вторых, в этом случае, как кажется, трудно избежать произвольных решений при отнесении той или иной статьи к конкретному жанру. Так, например, не вполне понятно, можно ли провести четкую границу между жанрами «Очерк проблемный + Репортаж» и «Репортаж + Очерк проблемный» или «Статья аналитическая» и «Статья аналитическая + Статья проблемная». Кроме того, метаразметка КГТ не учитывает ряд параметров, традиционно используемых для классификации текстов; и если, например, характеристика по полу и возрасту автора действительно не столь существенна в применении к газетным текстам, то тематика статьи (политика, спорт и т. п.) в некоторой степени определяет ее лингвистические особенности.

Морфологическая разметка. Лемматизация и морфологическая разметка осуществлялась автоматически на основе оригинальной системы аннотирования, разработанной авторами КГТ. Процесс приписывания словоформам грамматических показателей соответствует в этой системе, как правило, их разбиению на непересекающиеся классы. Так, признак *см* приписывается существительным мужского, *сж*—женского и *сс*—среднего рода. При этом признак *с* получают не все существительные, а только существительные с неустановленным родовым оформлением. Аналогичным образом трактуются и омонимичные формы. Им приписываются особые кластерные признаки. Так, например, дескриптор *е-ив* получают имена, у которых совпадают формы именительного и винительного падежа единственного числа. При этом дескриптор *е-и* присваивается только тем именам, у которых форма именительного падежа единственного числа не омонимична какой-либо другой или же была однозначно распознана как таковая. К сожалению, сайт КГТ не содержит подробного описания системы морфологической разметки, что существенно затрудняет работу с корпусом для неподготовленного пользователя.

Синтаксическая и семантическая разметка. Помимо морфологической в КГТ включены некоторые элементы лин-

гвистической аннотации других языковых уровней. На синтаксическом уровне размечаются предложные группы (предлог + именная группа в заданном падеже, с определением существительного по признаку одушевленность/неодушевленность, всего выделяется 109 типов таких сочетаний). Словообразовательная разметка состоит в приписывании каждой лемме морфемной модели, т. е. схемы с заполненными аффиксальными позициями и переменной для корня. Аннотация семантического уровня включает, во-первых, присвоение некоторым леммам семантических признаков на основании таксономической классификации лексики (при этом, однако, из выделенных 70 классов 60 образуют имена, обозначающие лиц и животных), во-вторых, разметку синонимических отношений между отдельными лексическими единицами.

Поиск в корпусе. Поиск может вестись по словоформе, лексеме, грамматическим признакам, а также атрибутам других уровней разметки (по предложным группам определенного вида, по заданной морфемной модели, семантическим признакам или по синонимам к заданной лексеме). Главным недостатком системы поиска в КГТ является невозможность построения запроса на последовательность словоформ или лексем. Все формулируемые при запросе условия (в том числе грамматические признаки) могут относиться только к одной единице текста или же к нескольким, разделенным логическим оператором ИЛИ. Формат выдачи—обычный текст. Максимальный контекст выдачи—по 30 слов справа и слева от искомого. Ограничено и количество выдаваемых контекстов—не более 30. Поиск может вестись по всему корпусу или по подкорпусу, сформированному на основе заданных пользователем значений метаатрибутов.

КГТ позволяет пользователю осуществлять различные виды статистической обработки данных, правда, его потенциал ограничен небольшим объемом. Все лексемы в корпусе распределены по 20 группам, соответствующим рангам их частотности, что позволяет, например, ограничивать поиск какого-либо лингвистического явления словами определенного уровня частотности. Кроме того, предусмотрена возможность просматривать частотное распределение заданного значения любого из доступных атрибутов (словоформ, лексем, лингвистических и металингвистических

признаков) по значениям любого другого атрибута, например, определенного существительного по типам предложных групп или словоформ, характеризующихся определенными грамматическим признаками, по жанровым типам статей.

Доступ к корпусу. Интернет-составляющая корпуса находится в открытом доступе.

Национальный корпус русского литературного языка (НКРЛЯ). НКРЛЯ разрабатывается с 2001 г. сотрудниками С.-Петербургского университета и Института лингвистических исследований Российской академии наук, однако в открытом доступе он находится только с 2006 г. Как и CLR, НКРЛЯ нацелен на отражение стандартного письменного языка.

Состав. В настоящий момент Интернет-версия корпуса включает свыше 1 млн. словоупотреблений, в том числе беллетристику (33,7%), публицистику (28,8%), драматургию (18,6%), научно-популярную литературу (18,9%). Временной охват текстов—с середины XX в. по настоящее время. Отметим, что в корпус помещаются не целые тексты, а их фрагменты.

Метаразметка. Тексты в корпусе классифицируются по четырем типам, соответствующим составляющим корпуса (беллетристика, публицистика и т. д.).

Морфологическая разметка. Лемматизация и морфологическая разметка в корпусе пока отсутствуют.

Поиск в корпусе. Поиск может вестись только по одной словоформе. Формат выдачи—обычный текст. Выдаваемые примеры сортируются по типам текста. Все словоформы в выдаче акцентуированы. Возможен также запрос на частотное распределение заданной словоформы по типам текста.

Доступ к корпусу. Корпус находится в открытом доступе.

Украинский

Для украинского языка на сегодняшний день еще не создано общедоступного электронного корпуса текстов. Из работ в этой сфере следует отметить деятельность, осуществляемую в Украинском языково-информационном фонде НАН Украины под руководством В. А. Широкова. К основным задачам фонда относится создание различных электронных словарей—грамматических, синоними-

ческих, фразеологических и др. (см. *http://lcorp.ulif.org.ua/dictua*), и базой для этих словарей служит разрабатываемый сотрудниками фонда Национальный корпус, см. [Широков 2005]. Однако у широкого пользователя доступа к этому корпусу пока что нет.

* * *

Завершая обзор славянских корпусных ресурсов (их основные параметры в кратком виде представлены в таблице в Приложении), хотелось бы обозначить ряд вопросов и задач, актуальных на нынешнем этапе их развития. Первый комплекс проблем очевиден: он связан с оснащением корпусами языков, для которых они еще не созданы или находятся в стадии разработки (украинский, белорусский, сербский, болгарский, македонский), а также совершенствованием (а в некоторых случаях и значительной доработкой) уже существующих ресурсов. Пути этого совершенствования вытекают, в частности, из сопоставления различных корпусов. Ряд параметров, относящихся к разным аспектам структуры и функционирования корпуса, можно признать бесспорно положительными характеристиками данного вида ресурсов, соответственно, их отсутствие в той или иной степени уменьшает эффективность корпуса для пользователя. Если речь не идет о специализированных коллекциях, это, конечно, объем и репрезентативность состава (хотя отражение картины реального употребления и представляет самостоятельную исследовательскую задачу для каждого языка, отдельную для разных периодов его функционирования (ср. [Шимкова 2005]), тем не менее в ряде корпусов (например, IPI, ХНК, NB) проблема сбалансированности является заведомо не решенной. Следующей областью параметризации корпуса является метаразметка, которая в больших корпусах, безусловно, может и должна становиться инструментом социолингвистических, стилистических и—в случае достаточного временного охвата—исторических исследований. В этом смысле бесспорно полезным кажется учет при аннотации таких однозначно определяемых параметров, как имя, пол и возраст автора, год создания текста, характеристики целевой аудитории, тип носителя. Между тем в большинстве рассмотренных корпусов, в том числе таких, которые включают достаточно подробную метаразметку (например, в ЧНК), некоторые из этих

параметров не учитываются. Исследовательски более творческую задачу представляет собственно типология текстов, включающая их распределение по типам, жанрам, тематике и т. п. Здесь на материале славянских корпусов (тех из них, которые вообще учитывают этот параметр, ср. обратное, напр., в CLR, IPI) можно проследить различные решения—от чрезмерно обобщающего деления, ср. FIDA, где выделяются только художественные (проза, поэзия, драма), научные (гуманитарные, естественные) и прочие тексты, до классификации по 5 различным параметрам (ср. НКРЯ)—сфера функционирования, тематика, хронотоп, жанр, тип,—со значительным набором конкретных значений для каждого из атрибутов. Как кажется, в этой области предпочтительным является решение в пользу увеличения параметров, на основе которых строится типология текстов. В обратном случае разметчик оказывается вынужденным каждый раз произвольным образом выбирать, какое значение из разнородного набора признаков следует приписывать данному типу текстов—например, статья или физика, публицистика или эссе и т. д. (ср. выше обсуждение метапараметра «жанр» для ЧНК). Вообще говоря, чем больше метаинформации помещается в корпус, тем шире его потенциал при решении различных лингвистических задач. В то же время следует иметь в виду, что дробность классификации приобретает практическое значение только при больших объемах языковых данных (ср. КГТ).

Очевидной необходимостью для корпуса является наличие морфологической разметки. В этом отношении, к сожалению, многие славянские корпуса (и в худшем положении здесь находятся южнославянские языки) требуют качественных изменений.

Ряд желательных функций связан и с параметрами выдачи. Сортировка различного типа (как алфавитная—по искомому выражению, левому и правому контексту, так и по метаатрибутам, например, по времени создания текста), возможность получения данных об источнике текста и просмотра лингвистической разметки для каждой из единиц выдаваемого примера—все эти опции, безусловно, являются нужными для пользователя и требуют внесения в систему тех корпусов, где они пока отсутствуют. Наконец, еще одним направлением совершенствования корпуса является внесение в него информации, относящейся к статистическому анали-

зу языковых данных, и возможности построения статистических запросов.

Таким образом, очерченный выше круг проблем носит, так сказать, привативный характер: речь шла в основном о таких элементах или свойствах корпуса, наличие которых повышает его эффективность. Второй комплекс вопросов соотносится скорее с эквиполентной оппозицией: анализ характеристик различных ресурсов позволяет выявить ряд спорных решений и противоположных тенденций в принципах создания корпусов, при этом каждый из подходов имеет свои положительные и отрицательные стороны. Ниже будут обозначены некоторые проблемные зоны.

Широко известно, что потенциал корпуса как инструмента лингвистических исследований тем выше, чем полнее и разнообразнее его разметка. Очевидным кажется и один из постулатов аннотирования корпусов, сформулированный Дж. Личем, согласно которому схема разметки должна основываться на общепринятой классификации языковых данных, не связанной с какой-либо конкретной теорией [Leech 1993: 275]. Объединение этих установок таит в себе противоречие: об общепризнанной типологии признаков можно говорить, пожалуй, только применительно к морфологии. Уже на синтаксическом уровне разработчики корпусов вынуждены жертвовать или детальностью разметки, или ее теоретической нейтральностью (ср. [Резникова, Копотев 2005]). Два противоположных в этом смысле подхода можно проследить на материале синтаксически аннотированных корпусов русского языка. Одну тенденцию представляет корпус ХАНКО: его создатели ориентировались прежде всего на то, чтобы разметка была понятна как можно большему числу пользователей, поэтому в ее основу и была положена известная по школьной программе классификация по членам предложения (см. выше). Другой подход реализован в синтаксическом подкорпусе НКРЯ: здесь под разметкой понимается построение для каждого предложения его синтаксической структуры в виде дерева зависимостей, в котором все связи получают имена соответствующих им синтаксических отношений. Всего используется около 80 таких отношений, их перечень представляет собой существенно расширенную версию списка, предложенного в теории И. А. Мельчука «Смысл⇔Текст»

(см. [Апресян и др. 2005]). Тем самым очевидно, что эту разметку никак нельзя признать теоретически нейтральной, и действительно, пользователю, незнакомому с теорией Мельчука, потребуется немало времени, чтобы освоить разработанную классификацию синтаксических отношений и применять ее для своих поисковых задач. Но столь же очевидно, что данный тип разметки включает в себя гораздо более детальный анализ явлений синтаксического уровня, чем аннотация, реализованная в ХАНКО.

При дальнейшем движении вглубь языковых уровней разработчик корпуса еще неизбежнее сталкивается с необходимостью выбора формализма, в рамках которого должна строиться схема аннотации. Отдавая предпочтение какой-либо теории, автор тем самым значительно ограничивает возможности применения созданной разметки для исследователей, работающих в рамках других научных парадигм. Показательным здесь является пример PDT: с одной стороны, подробная семантическая информация, вносимая авторами, представляет собой ценный лингвистический материал, с другой стороны, обращение к нему пользователя может быть продиктовано скорее интересом к теории функциональной порождающей грамматики, чем необходимостью решения независимой исследовательской задачи. В этом смысле любопытно, как будут развиваться опирающиеся на опыт PDT проекты по созданию глубоко аннотированных корпусов для других славянских языков—приведут ли они к созданию аналогичных ресурсов, претерпят ли при этом принципы разметки какие-либо изменения и не выработается ли при этом новый стандарт семантического аннотирования корпуса.

Можно предположить, что в области семантической разметки сформируются два различных направления корпусных разработок, как это уже фактически имеет место в сфере грамматического аннотирования. Неизбежно сталкиваясь с дилеммой «объем корпуса vs. точность его обработки», создатели корпусных ресурсов или делают выбор в пользу большого объема и автоматических программ морфологической разметки, или ограничиваются небольшим количеством данных, подвергая их тщательной ручной обработке на морфологическом и синтаксическом уровнях (ср., например, ХАНКО). Материал славянских корпусов обозначает

возможность такого расхождения и для разработок семантического уровня. Опыту детального аннотирования PDT можно противопоставить менее сложную семантическую разметку НКРЯ, выполненную в автоматическом режиме на материале многомиллионного корпуса. При этом принципы разметки НКРЯ, апеллирующей к понятным широкому кругу пользователей таксономическим категориям, могли бы в свою очередь заложить основу развития стандарта для семантического аннотирования больших корпусов.

Следующей зоной расхождения славянских корпусов является подход к грамматической омонимии на больших массивах текста. В ряде корпусов (напр., ЧНК, СНК) грамматическая омонимия снимается при помощи статистических программ, обученных на размеченных вручную текстах, в других корпусах (напр., НКРЯ, FIDA) грамматическая омонимия не снимается или снимается лишь частично. Тем самым в первом случае при поиске пользователь получает большую долю отвечающих его запросу примеров и незначительное количество «шума», при этом незначительная доля подходящих под запрос контекстов окажется потерянной вследствие неправильных разборов, во втором случае пользователь получает значительно большее количество «шума», но не рискует потерять какие-либо соответствующие запросу контексты. Эффективность того или иного поискового метода определяется исследовательской задачей пользователя, поэтому кажется естественным, чтобы именно ему был предоставлен выбор той или иной стратегии. В этом отношении чрезвычайно интересным видится решение, реализованное в корпусе IPI: в нем сохраняются и открыты для поиска все разборы, отвергнутые автоматической программой снятия омонимии.

Наконец, славянские корпуса обнаруживают разные принципы организации пользовательского интерфейса и языка запросов. Здесь можно выявить две тенденции: системы, ориентированные в первую очередь на удобство широкого круга пользователей, и системы, характеризующиеся мощностью языка запросов. Как ни странно, эти дополняющие друг друга принципы оказываются отчасти противоречащими друг другу. Первый тип систем представляют, например, НКРЯ и ХАНКО: пользователь должен

самостоятельно вводить только искомые словоформы и лексемы, остальные атрибуты и их значения предлагаются ему в виде списка, из которого он может выбрать нужные ему признаки. Для определения различных параметров поиска (например, расстояния между искомыми единицами) предусмотрены специальные окна, при которых имеются соответствующие комментарии. Второй тип систем реализован, например, в корпусах, использующих программу Bonito (ЧНК, СНК, ХНК): здесь имеется одна поисковая строка, в которой пользователь в соответствии с синтаксисом языка запросов задает поисковые параметры. Тем самым человеку, который обращается к корпусу, необходимо предварительно освоить принципы построения запроса, ознакомиться с системой используемых в корпусе атрибутов и их значений, изучить соответствующие им аббревиатуры, а также способ их представления. Все дополнительные параметры поиска (например, расстояние между искомыми единицами или ограничение запроса определенным типом текста) задаются в рамках того же формализма.

Очевидно, что системы первого типа более удобны для работы с корпусом неподготовленного пользователя. В то же время язык запросов, применяемый в системах второго типа, часто основан на использовании аппарата регулярных выражений, которые позволяют накладывать некоторые дополнительные ограничения на условия поиска и тем самым решать более сложные и разнообразные исследовательские задачи. Попытку соединить удобство пользовательского интерфейса и мощность языка запросов представляет поисковая система НКРЯ: с одной стороны, как уже отмечалось, НКРЯ организован по принципам систем первого типа, с другой—за последние годы корпус пополнился новыми поисковыми функциями (например, построение запросов на конструкции с повторами лексем и/или определенных грамматических значений), что сблизило его поисковый потенциал с возможностями систем второго типа. И все же эффективность последних остается выше. Дело в том, что в системах первого типа жесткая структура, при которой пользователю предлагается выбор из заданного списка параметров, по-видимому, просто не может вместить в себя все мыслимые комбинации типов запрашиваемой информации и ограничений на их выдачу. Между тем в системах, основанных на

языке регулярных выражений, необходимые признаки можно совершенно произвольно комбинировать посредством логических операторов. Так, запрос на повторы в НКРЯ ограничен, во-первых, содержательно—искаться могут конструкции с дублированием лексемы, части речи, падежа, числа и т. д., но не семантических признаков, во-вторых, структурно—под повторами понимается только отношения между двумя непосредственно следующими друг за другом словами, но не конструкции со «вставными элементами» между тождественными единицами, ср. *сказать-то он сказал*. Понятно, что в НКРЯ ради сохранения удобства интерфейса приходится выбирать из всех возможных типов запросов те, которые, скорее всего, будут в наибольшей степени востребованы пользователем, иначе мы будем иметь дело с необозримыми перечнями всех возможных комбинаций (ведь, напомним, все поисковые параметры, кроме конкретного лексического наполнения, в НКРЯ задаются списками). Между тем в системах второго типа таких сложностей не возникает: например, в данном случае отношение тождества накладывалось бы на любые элементы и любые признаки, которые учтены в разметке.

Таким образом, раз удобство пользования корпусом заставляет отчасти жертвовать мощностью языка запросов, решение этой дилеммы, как и проблемы снятия грамматической омонимии, могло бы лежать в объединении обоих типов поиска в системе корпуса, с предоставлением пользователю возможности выбора между ними.

Итак, сопоставление различных корпусных ресурсов, разработанных к настоящему времени для славянских языков, позволяет выявить спектр исследовательских подходов к методике их создания, очертить круг возможностей, реализованных в разных системах, и тем самым обозначить потенциал развития как для каждого из ресурсов в отдельности, так и для славянской корпусной лингвистики в целом.

Список литературы

Апресян, Ю. Д.; Богуславский, И. М..; Иомдин, Б. Л.; Иомдин, Л. Л.; Санников А. В.; Санников В. З.; Сизов В. Г.; Цинман, Л. Л. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005.—193–214.

Гарабик, Р.; Захаров, В. П. Параллельный русско-словацкий корпус // Труды международной конференции «Корпусная лингвистика—2006». СПб.: Изд-во С.-Петербургского университета 2006.—81–87. *http://korpus.juls.savba.sk/publications/block1/2006-garabik-russian-slovak-corpus/2006-garabik-zacharov-paralelnij.pdf*

Засорина Л. Н. (ред.) Частотный словарь русского языка. Л.: Наука, 1977.

Кустова, Г. И.; Ляшевская, О. Н.; Падучева, Е. В.; Рахилина, Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. М.: Индрик, 2005.—155–176.

Мустайоки А.; Копотев М. В.; Гурин, Г. Б.; Саломатина М. С. Принципы синтаксической разметки Хельсинкского аннотированного корпуса русских текстов ХАНКО // Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей». СПб., 2005.—С. 90–95.

Резникова Т. И. Корпуса славянских языков в интернете: Обзор ресурсов // Die Welt der Slaven LIII, 2008.

Резникова Т. И., Копотев М. В. Лингвистически аннотированные корпуса русского языка (обзор общедоступных ресурсов) // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. М.: Индрик, 2005.—С. 31–61.

Шимкова М. Репрезентативность корпуса как лингвистическая проблема // Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей». СПб., 2005.—С. 130–139.

Широков В. А. (отв. ред.) Корпусна лінгвістика. Київ: Довіра, 2005.

Čermák, F.; Křen, M. Frekvencní slovník češtiny. Praha 2004.

Džeroski, S.; Erjavec, T.; Ledinek, N.; Pajas, P.; Žabokrtský, Z.; Žele, A. Towards a Slovene Dependency Treebank // Proceedings of Fifth International Conference on Language Resources and Evaluation, LREC'06, 24–26 May 2006. Genoa. *http://nl.ijs.si/sdt/bib/SDT-LREC06.pdf*

EAGLES (Expert Advisory Group on Language Engineering Standards). Preliminary recommendations on text typology. [EAGLES Document EAG-TCWG-TTYP/P], 1996. *http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html*

Kurcz, I., A. Lewicki, J. Sambor, K. Szafran, and J. Woronczak. Słownik frekwencyjny polszczyzny współczesnej. Kraków: Wydawnictwo Instytutu Języka Polskiego PAN, 1990.

Leech G. Corpus annotation schemes // Literary and Linguistic Computing, 1993.—8/4.—Pp. 275–281.

Moguš, M.; Bratanić, M.; Tadić, M. Hrvatski čestotni rječnik. Zagreb: Zavod za lingvistiku, Školska knjiga, 1999.

Przepiórkowski, A. The potential of the IPI PAN corpus // Poznań Studies in Contemporary Linguistics, 2006.—Vol. 41.—31–48.

Simov, K.; Osenova, P.; Slavcheva, M. BTB-TR03: BulTreeBank Morpho-syntactic Tagset. BulTreeBank Project Technical Report № 03, 2004. *http://www.bultreebank.org/TechRep/BTB-TR03.pdf*

Tadić, M. Building the Croatian National Corpus // Proceedings of the Third Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain, ELRA, 2002.—Pp. 441–446.

Vasilišinová, D.; Garabík, R. Parallel French-Slovak Corpus // Computer Treatment of Slavic and East European Languages. Proceedings of the conference Slovko 2007. Eds. J. Levická, R. Garabík. Brno: Tribun 2007. *http://korpus.juls.savba.sk/~garabik/publications/22/french_slovak_parallel_corpus.pdf*

Список корпусов

КБТ Корпус боснийских текстов (Осло)
http://www.tekstlab.uio.no/Bosnian/Corpus.html

КГТ Корпус газетных текстов русского языка
http://www.philol.msu.ru/~lex/corpus

КСЯ Корпус сербского языка
http://www.serbian-corpus.edu.yu/indexie.htm

НКПЯ Национальный корпус польского языка
http://nkjp.pl

НКРЛЯ Национальный корпус русского литературного языка
http://www.narusco.ru

НКРЯ Национальный корпус русского языка
http://ruscorpora.ru

СНК Словацкий национальный корпус
http://korpus.juls.savba.sk

ТБСХК Тюбингенские боснийско-сербско-хорватские корпуса
http://tusnelda.sfb.uni-tuebingen.de/tusnelda-query.html#b8

ТК Тюбингенский корпус русского языка
http://www.sfb441.uni-tuebingen.de/b1/korpora.html

ХАНКО Хельсинкский аннотированный корпус русского языка
http://www.ling.helsinki.fi/projects/hanco

ХНК Хорватский национальный корпус
http://www.hnk.ffzg.hr

ЧНК Чешский национальный корпус
http://ucnk.ff.cuni.cz

BulTreeBank
Bulgarian Treebank
http://www.bultreebank.org

CDT Croatian Dependency Treebank
http://hobs.ffzg.hr/default_en.html

CLR Croatian Language Repository
http://riznica.ihjj.hr

FIDA Корпус словенского языка FIDA
http://www.fida.net

FidaPLUS
: Корпус словенского языка FidaPLUS
http://www.fidaplus.net

IPI
: Корпус Института основ информатики Польской академии наук
http://korpus.pl

KoRP
: Корпус словенского языка (тематика текстов—«Связи с общественностью»)
http://www.korp.fdv.uni-lj.si

NB
: Корпус словенского языка Nova beseda
http://bos.zrc-sazu.si/a_beseda.html

PDT
: Prague Dependency Treebank
http://ufal.mff.cuni.cz/pdt

PELCRA
: Polish and English Language Corpora for Research and Applications
http://korpus.ia.uni.lodz.pl

PWN
: Корпус польского языка издательства PWN
http://korpus.pwn.pl/szukaj.php

SDT
: Slovene Dependency Treebank
http://nl.ijs.si/sdt

WWW-Concordance
: Корпус словенского языка
http://nl2.ijs.si/index-mono.html

Корпуса славянских языков в Интернете: основные параметры

язык	корпус	содержание	объем корпуса (в млн. словоупотреблений)	Типы разметки		
				морфологическая	снятие грамматической омонимии (автоматическое/ручное)	синтаксическая
чешский	ЧНК — подкорпуса письменного языка	коллекция сбалансированных и специализированных корпусов (1990–2004)	500	+	а (весь корпус)/р (0,08)**	–
	ЧНК — подкорпуса устной речи	записи устной речи из разных регионов Чехии	2,3	–		–
	PDT	газеты и журналы (1990–95)	2	+	р	+ (1,5)**
словацкий	СНК	письменные тексты разных типов (1955–2006)	339	+	а (весь корпус)/р (0,5)**	–
польский	IPI PAN	несбалансированная коллекция текстов нескольких типов	250	+	а (сохраняются все варианты разбора)	–
	PELCRA	письменные и устные тексты разных типов (1989–2003)	93	–		–
	PWN	фрагменты письменных текстов различных типов, устная речь (1903–2005)	22/3,7*	только лемматизация	–	–

Типы разметки		Поисковые возможности: поиск по:						Параметры выдачи				
семантическая	метаразметка	словоформе	лексеме	последовательности словоформ	грамматическим признакам	синтаксическим структурам	семантическим признакам	максимальный контекст	ограничения на количество контекстов	сортировка выдачи	фильтрация выдачи (поиск в найденном)	статистическая обработка запроса
–	+	+	+	+	+	–	–	≈1000 знаков/ 100 слов/ 3 предл.	нет	+	+	+
–	+	+	–	+	–	–	–		нет	+	+	+
+ (0,8)**	–	+	+	+	+	+ (1,5)**	+ (0,8)**	1 предл.	нет	–	+	–
–	+	+	+	+	+	–	–	≈200 слов	нет	+	+	+
–	+	+	+	+	+	–	–	200 слов	нет	+	–	–
–	+	+	–	+	–	–	–	3 абзаца	250	+	–	+
–	±	+	+	+	–	–	–	не ограничен	нет	+	–	–

язык	корпус	содержание	объем корпуса (в млн. словоупотреблений)	Типы разметки		
				морфологическая	снятие грамматической омонимии (автоматическое/ручное)	синтаксическая
польский	PWN - Rzeczpospolita	статьи газеты Rzeczpospolita (1997–2005)	18/ 3,6*	только лемматизация	–	–
словенский	FIDA	сбалансированный письменный (1990–1997)	100	+	частичное (а) (сохраняются все варианты разбора)	–
	FidaPLUS	сбалансированный письменный (1990–2006)	621	+		–
	NB	несбалансированная коллекция текстов нескольких типов	162	–		–
	WWW-Concordance	несколько разноплановых текстовых коллекций	2,1	+ (0,09)**	р (0,09)**	–
	KoRP	тексты по тематике «Связи с общественностью» (1994–2007)	1,8	+	частичное (а) (сохраняются все варианты разбора)	–

Типы разметки		Поисковые возможности: поиск по:						Параметры выдачи				
семантическая	метаразметка	словоформе	лексеме	последовательности словоформ	грамматическим признакам	синтаксическим структурам	семантическим признакам	максимальный контекст	ограничения на количество контекстов	сортировка выдачи	фильтрация выдачи (поиск в найденном)	статистическая обработка запроса
–	±	+	+	+	–	–	–	не ограничен	нет	+	–	–
–	+	+	+	+	+	–	–	1 абзац	нет	+	+	+
–	+	+	+	+	+	–	–		нет	+	+	+
–	+	+	–	+	–	–	–	3 предл.	нет	–	–	±
–	–	+	(0,09)** +	+	(0,09)** +	–	–	≈330 знаков	125 (KWIC)/ нет	–	–	±
–	+	+	+	+	+	–	–	1 абзац	нет	+	+	+

язык	корпус	содержание	объем корпуса (в млн. словоупотреблений)	Типы разметки		
				морфологическая	снятие грамматической омонимии (автоматическое/ ручное)	синтаксическая
хорватский	ХНК	газеты, журналы (1990–2005), худ. лит-ра с XVI в.	101	+ (0,118)**	р (0,118)**	–
	CLR	письменные тексты, отражающие стандартную форму языка (с XIX в.)	71	–		–
боснийский	КБТ	письменные тексты разных типов (90-е гг.)	1,5	–		–
сербский	КСЯ	худ. лит-ра с XII в., публицистика, научные тексты XX в.	11	+	р	–
босн./серб./хорв.	ТБСХК	устная речь, комиксы	0,127	–		–

Типы разметки		Поисковые возможности: поиск по:						Параметры выдачи				
семантическая	метаразметка	словоформе	лексеме	последовательности словоформ	грамматическим признакам	синтаксическим структурам	семантическим признакам	максимальный контекст	ограничения на количество контекстов	сортировка выдачи	фильтрация выдачи (поиск в найденном)	статистическая обработка запроса
–	+/–	+	(0,118)** +	+	(0,118)** +	–	–	не ограничен	нет	+	+	+
–	+	+	–	+	–	–	–	3 страницы/ 3 абзаца	нет	+	–	+
–	+	+	–	+	–	–	–	500 знаков/ 200 слов	нет	–	–	–
–	–	Поиск в корпусе невозможен										
для дейкт. элементов	–	+	–	+	–	–	по характеристикам дейктических элементов	1 реплика	нет	–	–	–

язык	корпус	содержание	объем корпуса (в млн. словоупотреблений)	Типы разметки		
				морфологическая	снятие грамматической омонимии (автоматическое/ ручное)	синтаксическая
русский	НКРЯ	сбалансированный корпус с 1950 г. (в т.ч. устные тексты), худ. лит-ра, научн. тексты и публицистика с сер. XVIII до сер. XX вв.	163	+	р (6)**	+ (0,5)
	ТК	Упсальский корпус; публицистика (1996–2002); худ. лит-ра XIX-XX вв.	25	+ (2,3)** (нет лемматизации)	а (2,3)**	–
	ХАНКО	журнальные тексты (2001 г.)	0,1	+	р	+
	КГТ	газетные тексты (1994–1997)	11/ 0,2*	+	–	±
	НКРЛЯ	фрагменты письменных текстов, отражающих стандартную форму языка (с сер. XX в.)	1	–		–

Типы разметки		Поисковые возможности: поиск по:						Параметры выдачи				
семантическая	метаразметка	словоформе	лексеме	последовательности словоформ	грамматическим признакам	синтаксическим структурам	семантическим признакам	максимальный контекст	ограничения на количество контекстов	сортировка выдачи	фильтрация выдачи (поиск в найденном)	статистическая обработка запроса
+	+	+	+	+	+	+ (0,5)	+	7 предл.	нет	+	–	–
–	–	+	–	+	+ (2,3)**	–	–	≈240 слов/ 13 предл.	нет	±	–	–
–	–	+	+	+	+	–	–	11 предл.	нет	–	–	–
±	+	+	+	–	+	±	±	≈60 слов	30	–	–	+
–	±	+	–	–	–	–	–	≈40 слов	нет	–	–	±

* Формат записи объема X/Y применяется для тех корпусов, в которых общий объем корпуса (X) отличается от объема общедоступного корпуса (Y).

** Число в скобках после значения параметра соответствует объему текстов в миллионах словоупотреблений, на которых реализован данный тип разметки или доступен данный тип поиска.