Секция XX.
Прикладные, компьютерные и квантитативные аспекты моделирования русского языка

Сканирование общественно-политической жизни Украины по данным украинских и русских газет 2004–2006 гг. (тезаурусно-фреймовый подход)

Л. А. Алексеенко, Н. П. Дарчук

Киевский национальный университет им. Тараса Шевченко (Украина)

Лексема, словарь, тезаурус, фрейм, образ

Summary. The scanning of public and political life of the country using the data of social-political discourse can be accomplished by applying of a comprehensive technology. This technology combined the quantitative and qualitative principles of analysis social-political texts and it is reflected in the three dictionaries: the frequency dictionary, frame-thesaurus and the dictionary of imageries for the Ukrainian and Russian languages.

На протяжении последних лет в Украине наблюдается обострение социально-политической ситуации, что активно освещается в СМИ и, как следствие, повышает к ней интерес у языка различных слоев населения. Для современной лингвистики актуальны вопросы взаимодействия публицистического дискурса и власти, роли языка в борьбе людей за власть и влияние этих процессов на язык, язык СМИ. С одной стороны, социоэкономическая и политическая ситуация (тенденция к демократизации общества в Украине) создает определенные условия для политической свободы коммуникации. С другой – масс-медиальный, политический, религиозный и другие типы дискурса влияют на социально-экономическую жизнь страны.

В период изменения идейных ценностей и политических приоритетов государства важным является исследование образов государства, власти, нации, общества и т. д., созданных публицистикой. Положительный имидж этих институций способен консолидировать нацию, воспитать уважение к ней граждан. Отрицательный поляризирует нацию, формирует негативное отношение к государству не только у собственных граждан, но и у граждан других стран.

В языке масс-медиа находят отражение те события внеязыковой действительности, которые релевантны для членов социума, что делает язык СМИ одним из мощных рычагов установления и поддержания властных отношений в обществе. Важным является **кто**, **о чем** и **как** говорит в масс-медиа, мониторинг и оперативный прескриптивный анализ публицистического дискурса по целевым программам и методикам. Современная прикладная и компьютерная лингвистика располагает богатым арсеналом формализированных анализов, которые в сочетании с соответствующим программным обеспечением делают возможным успешное решение этой задачи.

Был создан текстовый корпус по газетным публикациям 2004–2006 гг., издаваемым на украинском и русском языках в четырех регионах Украины (центр, запад, восток, Крым). Для решения целевых задач – систематизации лексики, анализа ее взаимодействия в создании публицистических образов власти, государства, общества в динамике общественно-политических событий была разработана методика, основанная на количественном и качественном анализе, результаты которого отражены в трех словарях: частотном, тезаурусно-фреймовом и словаре образов. Частотный словарь (со статистическими характеристиками – средней частотой, дисперсией, коэффициентом стабильности) дает информацию о закономерно повторяющихся лексемах, которые являются ключевыми для политического дискурса; тезаурусно-фремовый – понятийную классификацию лексики от концепта до лексических способов его вербализации, что приводит к средоточению в ее пределах семантически близких лексических единиц, которые с определенными содержательными вариациями передают одно понятие; словарь образов даст возможность выяснить, как у человека, члена социума, формируется, конструируется образная система под влиянием СМИ, каковы речевые механизмы создания образов власти, государства, политики, политиков и т. п.

Работа методики обеспечивается пакетом программ:

- автоматического морфологического и синтаксического анализа русского и украинского языков;
- автоматического создания частотного словаря слов и словоформ со статистическими характеристиками для них;
- создания конкорданса;
- автоматизированного контроля за значением лексической единицы по электронному толковому словарю (фиксация номера значения);
- создания электронного тезауруса (на данном этапе – существительных) с частотными характеристиками к каждой лексеме для русского и украинского языка отдельно;
- автоматизированного словаря образов.

Полученные результаты будут размещаться на портале www.mova.info, а также использоваться в постоянном мониторинге публицистического дискурса, в сканировании событий общественно-политической жизни Украины.

Применение латентно-семантического анализа при кластеризации слов на основании их контекстов

А. А. Антонова

Компания «Cognitive Technologies», Москва

Е. А. Ильюшина, А. В. Прохоров

Московский государственный университет им. М. В. Ломоносова

Пары глагол-существительное, разреженные данные, латентно-семантический анализ, сингулярное разложение

Summary. This paper is concerned with word clustering and Latent Semantic Analysis, methods that deal with a general problem of evaluating word similarity. Latent Semantic Analysis (LSA) is a statistical method based on a matrix decomposition technique for extracting the contextual-usage meaning of words from a large corpus of data. We argue that applying LSA transformation to the data before a clustering procedure can improve the results impressively. Using that technique, we conduct an experiment with clustering words according to their distribution in verb-object pairs.

Надежные данные о лексической сочетаемости необходимы при проведении автоматического анализа текста. Основным препятствием при получении таких данных является их разреженность. Вероятность каждого отдельного события, не попадающего в выборку, мала, однако всего событий, не попавших в выборку, может оказаться около 25%. В работе

предлагается метод оценки близости слов с помощью сведений об их контекстах, позволяющий учитывать разреженные данные, не прибегая к процедуре сглаживания. В методе используется алгоритм латентно-семантического анализа в сочетании с процедурой агломеративной кластеризации. На экспериментальном материале проводится сравнение результатов, полученных с помощью обычной кластеризации (3) и с применением латентно-семантического анализа.

Рассматриваются два первичных класса слов глаголы и существительные с заданным для них отношением «транзитивный глагол + прямое дополнение». Данными для анализа являются все пары из этих классов и частота, с которой каждая пара представлена в выборке. По выборочным данным о совместной встречаемости строится оценка распределения и далее некоторая разновидность тезауруса, т. е. система, отражающая относительную близость между словами. Материалом послужили три независимые выборки словосочетаний (объемы – 139746, 296484, 1657135 словосочетаний) из коллекции документов, принадлежащей компании «Cognitive Technologies».

Наиболее результативная процедура кластеризации проводится одновременно в двух пространствах, в одном из них (пространстве контекстов) точками являются слова, расстояние между которыми измеряется на основании близости их контекстов, в другом – пространстве слов точками являются контексты. Если на каком-то шаге два близких контекста объединяются, то расстояние между словами, которые различались этими контекстами, становится меньше (аналогично – для слов). Таким образом, на каждом шаге уменьшается размерность матрицы данных. При обработке используются два метода кластеризации, метод наименьших расстояний и информационный метод, основанные на соответствующих мерах близости.

Латентно-семантический анализ также приводит к уменьшению размерности матрицы данных, которая является следствием отбрасывания меньших сингулярных значений матрицы и сохранения больших (SVD-разложение). Метод предполагает, что в результате наилучшим образом выявляются скрытые отношения между словами, при этом исходные данные не могут быть полностью восстановлены по матрице сокращенной размерности. Параметром латентно-семантического анализа является число векторов (столбцов или строк), которыми аппроксимируется исходная матрица данных. В эксперименте матрица 871×848 была приближенно представлена 20-ю векторами. Существенное отличие от факторного анализа состоит в возможности пренебречь и достаточно большими собственными значениями.

Результаты эксперимента показали, что использование латентно-семантического анализа приводит к большей структурированности иерархии (большей глубине поддеревьев) и большей связанности слов внутри поддеревьев. Характерно, что на самых нижних уровнях иерархии (1–2-й шаги кластеризации) применение латентно-семантического анализа не дает улучшения.

Методы формального представления текстовых документов для анализа в автоматических системах

Е. Н. Бендерская, С. В. Жукова

Санкт-Петербургский государственный политехнический университет

Представление документов, кластеризация, частота, распределение, структура текста

Summary. Actual problems of Russian language computer analysis are discussed. Methods of formal text representations used in automatic systems of text processing are shown. A new method of text representation is proposed and the results of its application are given.

Среди актуальных задач компьютерного анализа русского языка посредством автоматической обработки русскоязычных текстов можно выделить, помимо задач определения авторства, выявления психологического состояния автора в момент написания текста, выявления формального смысла текста (о чем или к чему относится текст) и т. п., задачи, связанные с обнаружением изменений использования различных языковых конструкций как в пределах одного жанра или предметной области, так и в русском языке в целом.

Для решения этих задач необходимо определение взаимосвязи слов, а также формальной структуры текста и выражение их в количественном измерении для выявления меры подобия текстов и / или частоты использования словосочетаний.

Результат качественного анализа может быть выражен с помощью нечеткой логики и затем переведен в количественную оценку. Недостатком этого подхода является необходимость в построении экспертных оценок для получения функций принадлежности к разным лингвистическим переменным.

Многие методы по обработке текстовой информации опираются на формальное представление текста в виде некоторого набора признаков, предварительно сформированных методами автоматической предобработки текстов или человеком. В случае наличия большого количества примеров (текстов) и возможности однозначного их отнесения к предварительно выделенному набору классов можно использовать решающие правила, сформированные по принципу обучения с учителем (вероятностные классификаторы, нейронные сети прямого распространения).

Наиболее общим является случай, когда заранее неизвестно распределение элементов по отдельным группам (классам) и ставится задача построения решающего правила для отнесения новых элементов к тем или иным классам. При этом признаковое пространство может быть как избыточно, так и четко не определено. Тогда решающее правило должно выполнять еще и выделение значащих признаков из множества имеющихся.

В любом случае важным моментом является определение способа представления текстовой информации на вход системы автоматической обработки, т. к. одна из причин их неудовлетворительной работы связана с недостатками в используемых способах представления информации. Существуют различные группы методов представления текстовой информации, а именно:

- частотные методы;
- алгебраические методы;
- методы сжатия.

В информационно-поисковых системах в основном используются методы первых двух типов. Частотные методы основываются на вычислении частот ключевых слов в документах, т. е. подсчитывается сколько раз встречается в документе каждое из ключевых слов. Алгебраические методы фиксируют наличие или отсутствие ключевого слова в документе (численный вектор, описывающий документ будет содержать нули и единицы: 0 – есть ключевое слово, 1 – нет ключевого слова). Частотное представление документов имеет ряд недостатков, связанных с усреднением частоты встречаемости каждого из слов и как следствие потерей значимой информации о внутренней структуре документа.

С одной стороны в силу того, что в правилах построения предложений является нежелательным использование одного и того же термина несколько раз подряд (считается, что лучше использовать синонимы), то общая частота использования отдельного термина в отдельном документе может быть не очень высокой. С другой стороны очень частое употребление отдельного ключевого слова не гарантирует того, что документ посвящен теме, которая интересует.

Как показывают результаты активно проводимых в настоящее время исследований в области компьютерной лингвистики, смысл текста коррелирует со структурой текста – распределением употребления отдельных слов и словосочетаний пар, троек и т. д. слов в пределах всего текста. Близкие по содержанию и уровню изложения тексты должны обладать схожими внутренними структурами, а не только усредненными частотами употребления слов.

В данной работе предлагается новый подход представления документов, основанный на распределении ключевых слов по документам. Для этого каждый документ разбивается на M групп слов. Число групп M определяется в соответствии с правилом Стерджеса, т. е. правилом выбора числа интервалов разбиения для построения гистограммы распределения (для каждого документа будет найдено свое M). Далее будет рассчитана частота встречаемости каждого из ключевых слов в каждой из M групп. Чем больше документ, тем больше слов в группе.

Преимущество распределенного частотного представления текста заключается в том, что при минимальных вычислительных затратах при формальном представлении документа на вход системы кластеризации происходит минимальная потеря информации о специфике документа. При этом, рассматривая одновременно распределение нескольких ключевых слов по тексту, будет представлена и информация о частоте употребления пар, троек и т. д. ключевых слов без дополнительных вычислительных затрат. Таким образом, выявляя корреляцию употребления отдельных групп слов в тексте, удастся выявить и внутреннюю структуру текста и значит в какой-то мере его содержание-смысл.

Проверка предлагаемого метода проводилась на примере решения задачи кластеризации текстовых документов из трех различных тематик одной предметной области с использованием хаотической нейронной сети. Анализ полученных результатов позволяет сделать заключение, что предложенный метод распределенного частотного представления ключевых слов из словарей тематик дает лучшие результаты по сравнению с классическим частотным способом. Результаты правильной кластеризации выше на 5–7%. Причем при увеличении числа документов это влияние в сторону улучшения качества кластеризации будет больше.

Гипертекстовое пространство политической лексики: параметризация и принципы описания

Л. Е. Бессонова, Е. Ю. Чепик

Таврический национальный университет им. В. И. Вернадского, Симферополь (Украина)

Компьютерная лексикография, электронный словарь, корпус текстов

Summary. In article opportunities of research of a political word in a computer lexicography are considered. As an actual material the Case of Russian national language is used.

В лексикографической практике последних лет обозначается новое конструктивное направление – компьютерная лексикография, которая открывает перспективы для инноваций в науке о словарях. Исследователями сегодня обсуждаются такие вопросы, как структура и объем электронных словарей, принципы системной организации лексики, методы лингвистического программного обеспечения, параметры систематизации информационного материала (Ю. Д. Апресян, А. Н. Баранов, В. В. Богданов, А. С. Герд, Б. Ю. Городецкий, Ю. Н. Караулов, Е. А. Карпиловская, А. А. Поликарпов, В. А. Широков, И. М. Кульчицкий, В. Б. Черницкий, А. В. Зубов, И. И. Зубова, Н. Н. Леонтьева, Ю. Н. Филиппович, А. В. Прохоров и др.).

Концепция разрабатываемого проекта предполагает полевый подход при построении модели экспериментального электронного словаря политической лексики. Так, весь словарный корпус представляет собой организованную систему, рассматриваемую как семантическое макрополе, состоящее из 5 микрополей:

1) *поле авторского толкования,* которое строится не по всей семантической парадигме слова, а лишь по ее основной части, тематически связанной с политической сферой употребления. Каждая словарная статья включает в себя: зону индексации (служит для удобного поиска в базе данных); зону грамматической информации; зону этимологической справки (этимона); зону толкования и зону иллюстраций;

2) *поле иллюстраций,* необходимое для верификации дефиниции, что позволяет выявить семантику политического слова в определенном контекстуальном окружении. Важно отметить, что в словаре данного типа иллюстрации из текстов различных жанров используются и в традиционном виде как оправдательный контекст (это особенно важно для метафорических значений) и как лексикографический прием, помогающий создать пространственную, временную и социально-историческую перспективы;

3) *поле толковых словарей,* включающее словарные статьи толковых словарей;

4) *поле энциклопедических сведений*, направленное на построение семантического концептуального пространства, заданного определенной лексемой;

5) *поле истории слова,* в котором отражена динамика функционирования данной дефиниции.

Диалог пользователя с электронным словарем политической лексики организовано через разветвленную систему последовательных меню. Наличие нескольких входов в словарь делает его удобным в использовании и более информативным.

Важно отметить, что взаимоотношения между частями словарной статьи не являются линейными и структурирование в словаре такого типа подчиняется законам гипертекста. Словарная статья имеет четкую логическую структуру с иерархическими связями между элементами. Каждая информационная категория занимает здесь строго фиксированное место или зону.

В словаре все информационные блоки, заложенные в словарную статью, самодостаточны, но в то же время могут функционировать в качестве ценного дополнения по отношению друг к другу. Таким образом, в электронном словаре становится возможным вариант лексикографического моделирования, при котором появление на экране той или иной зоны словарной статьи осуществляется по выбору пользователя.

Компьютерное обеспечение словаря складывается из двух основных составляющих:

1) базы данных, определяющейся системой управления базами данных Access. Представленная база состоит из индексов, предназначенные для поиска в базе данных, вокабул, толкований (дефиниций), экспликаций, толкований рассматриваемой лексемы, указаний источников, а также данных о времени появления данного слова в рассматриваемой предметной области;

2) программы-оболочки, структура которой дает возможность управлять этой базой, а также вносить в нее изменения, не открывая ее непосредственно в Access. Все это позволяет структурировать последовательность получения данных в зависимости от потребностей пользователя. Программная оболочка написана на языке программирования Delphi, что позволяет осуществлять задачи автоматического поиска слов, добавления слов в базу данных, озвучивания заголовочного слова, экспортирования в HTML, обработку информации в пользовательском интерфейсе, а также удобной навигации внутри словаря.

Таким образом, описываемый словарь представляет собой систематизированный алфавитный корпус ядерной политической лексики. При электронной организации такого корпуса работают модели лексикографического моделирования, позволяющие выявить основные тенденции в развитии концептуального пространства политической лексики.

Большой электронный словарь словосочетаний и семантических связей в русском языке

И. А. Большаков

Национальный политехнический институт, Мехико (Мексика)

igor@cic.ipn.mx

Электронный словарь, словосочетания, семантические связи, приложения

Summary. A large dictionary of Russian collocations and semantic links is developed. Its core is ca. 820,000 collocations of *noun* / *adjective* / *verb* / *adverb* → *modifier*, *verb* → *complement*, *verb* → *subject*, and *noun* → *noun complement* types. The user can access collocations from both their sides. The semantic links are of synonymy, antonymy, hyponymy, meronymy, and semo-derivation types. The applications are text editing, study of Russian by foreigners, syntactic parse, word sense disambiguation, malapropism detection / correction, and linguo-steganography.

Резюме. Разработан электронный словарь словосочетаний и семантических связей в русском языке, не имеющий аналога по характеру и объему хранимой информации.

Статья словаря описывает лексему прилагательного или наречия либо граммему (подпарадигму) существительного (отдельно ед. и мн. число) или глагола (отдельно личные формы + инфинитив, причастия, деепричастия по двум видам раздельно). Именная статья может соответствовать словосочетанию типа *точка зрения, уровень жизни*, *сельское хозяйство*. Адъективная статья может включать причастия и обороты типа *с маслом* или *как бархат,* наречная статья – деепричастие и обороты типа *как лед, в особой степени* или *исчерпывающим образом*. Ограничений на область привлекаемых слов нет, это – наука, техника, политика, экономика, бизнес, гуманитарные науки, медицина, быт, включая бранную лексику без мата. Общий объем словаря – около 130 тыс. В нем выделено 1730 омонимических групп из 2–6 элементов с общим числом разных смыслов около 4100.

Связи между **статьями** (т. е. словосочетания, или **коллокации**) бинарны и доступны с обеих сторон. На полюсах – синтаксически связанные и семантически совместимые знаменательные слова. Между полюсами могут быть вспомогательные слова, обычно, предлоги. Устойчивость коллокаций оценивается частотами их вхождения в интернет, целиком и по частям. Привлекаются любые удовлетворяющие требованиям словосочетания, включая идиоматические и свободные.

Пометы при связях и отдельных словах берутся двух типов: **идиоматичность**: {неотмеч.; фигур. (*сесть в галошу*); м. б. фигур. (*сесть в лужу*)} и **стиль**: {неотмеч.; устарелое / книжн. / специальн.; просторечное; бранное; неправильн. (*оплатить за проезд*)}.

Наиболее многочисленны следующие типы коллокаций:

- Определительная пара ***существительное – адъектив*** или ***глагол / адъектив / наречие – наречие*** (*краснокочанная капуста, высказаться резко, ужасно страшно*) – 345 тыс.;
- ***Глагол*** – его прямое / косвенное / предложное ***дополнение*** включая ходовые обстоятельства (*рассмотреть вопрос*, *ковырять в носу, остаться ... из-за погоды*) – 207 тыс.;
- ***Подлежащее – сказуемое*** (*самолет вылетел, внимание привлечено*) – 125 тыс.;
- ***Существительное*** – подчиненное ему ***существительное*** (*сердце матери, отличия в произношении, борьба против терроризма*) – 137 тыс.

Иные привлеченные типы коллокаций менее объемны:

- ***Адъектив*** – его прямое / косвенное / предложное ***дополнение*** (*красный от стыда, покрытый навозом, сощущий степень*);
- ***Глагол*** – его ***инфинитивное дополнение*** (*собраться поехать*, *хотеть перекусить*);
- ***Существительное*** – его ***инфинитивное дополнение*** (*соблазн сказать, желание уйти*);
- ***Глагол*** – его ***адъективное дополнение*** (*вернуться здоровым, найти мертвым*);
- Устойчивые ***сочиненные пары*** (*автобусы и троллейбусы, авторитетный и уважаемый, поматросить и бросить, дешево и сердито*) и др.

Семантические связи между элементами словаря включают:

- ***Синонимы.*** 17,4 тыс. синон. групп по 5,73 элементов; односторонних связей – 1,6 млн;
- ***Семантические дериваты.*** Группы типа {*извлечение; извлекать, извлечь; извлеченный, извлекший; извлекая, по извлечении, путем извлечения*}; односторон. связей – 0,82 млн;
- ***Меронимы / холонимы,*** односторонних связей – 20,8 тыс;
- ***Гипонимы / гиперонимы,*** односторонних связей – 13,8 тыс;
- ***Антонимы,*** односторонних связей – 10,6 тыс.

Взаимодополнение коллокаций и семантических связей. Семантические связи помогают понять смысл титульного слова статьи, что важно из-за отсутствия толкований. При демонстрации связей некоторых слов ряд его (пока) отсутствующих коллокаций порождается автоматически, исходя из связей синонимии и гипонимии, например, (*букет цветов*) & (*астры* IS_A *цветы*) → (*букет астр*). Стопроцентная правильность таких коллокаций не обеспечена, что указывается пользователю словаря.

Иные лингвистические средства в словаре включают:

- ***Паронимы***, ***буквенные*** (*кадка*: *кака, каска, качка, кашка, кладка) и* ***морфемные*** (*бегающий, беглый, беговой, бегучий…* всего 28 элементов);
- ***Морфологические парадигмы*** практически для каждого изменяемого элемента словаря;
- ***Английские переводы*** элементов русского словаря (в настоящее время только для 40%), формирующие отдельный словарь-вход.

Важнейшие приложения включают **интерактивные**:

- ***Справки при редактировании текстов*** русскоязычным автором за компьютером;
- ***Изучение русского иностранцем*** на продвинутом этапе, когда пассивный словарный запас уже значителен, но сочетаемость не освоена (можно использовать англ. вход);

и неинтерактивные (одна прикладная программа обращается за справкой к другой):

- ***Автоматическое распознавание смысла слов*** (используются коллокации и семантические связи отдельных омонимов и выбирается омоним, для которого в контексте найдено наибольшее число коллокаций и семантических связей);
- ***Помощь в автоматическом синтаксическом анализе*** (используются коллокации, встреченные в тексте, и, чем больше обнаружено связанных согласно данному разбору пар, являющихся коллокациями, тем вероятнее этот вариант разбора);
- ***Автоматизированное обнаружение и исправление малапропизмов***: обнаруживаются синтаксически связанные пары, не являющиеся коллокациями (напр., *истерический центр*), просматриваются паронимы для обоих слов пары, дающие коллокацию (здесь находится только *исторический*) и они предлагаются для исправления ошибки;
- ***Лингвистическая стеганография***. Коллокации и синонимы слов, встреченных в тексте, используются для замены одних синонимов другими, и в этих заменах кроется информация, которая и передается тайно несущим текстом без изменения его смысла.

Дистанционные технологии в обучении русскому языку: слово в системе и тексте

Е. Ю. Булыгина, Т. А. Трипольская

Новосибирский государственный педагогический университет

Электронный учебник, система навигации, системный и коммуникативно-прагматический аспекты изучения слова

Summary. The contemporary level of lexicology assumes multifold vocabulary study contributed by system-structural, functional-semantic, communicative-pragmatic, sociolinguistic and other approaches. The principal idea of multimedia textbook is aimed to aspectly differentiated study of lexis.

Создание и использование электронных средств обучения обусловлено потребностями современного образовательного процесса: доля самостоятельной работы в системе вузовского обучения постоянно возрастает, кроме того, существенную роль играет дистанционное освоение теоретического и практического материала. Разработка электронного учебника (далее ЭУ) – это создание обучающей среды, в которую погружен студент, модели, на которой он проверяет свои собственные решения, системы контроля его знаний.

Электронный учебник позволяет поставить ряд задач, решение которых не может обеспечить «бумажный» вариант учебного пособия: сформировать у студента навыки работы с соответствующими информационными системами; вести поиск необходимого учебного материала с использованием системы навигации, что дает возможность учащемуся выстраивать собственную образовательную траекторию, и др.

Особого пояснения требует понятие навигации. При создании гипертекстового информационного пространства «ключевой становится проблема организации прочтения гипертекста» [Кедрова, Дедова, Потапов 2004]. В настоящем ЭУ мы попытались совместить две возможности работы с гипертекстом: свободную и заданную авторами траекторию поиска информации студентом. В учебнике представлены макро- и микронавигационые системы: первая позволяет студенту перемещаться от одного содержательного блока к другому (от лекции к словарям или к хрестоматии и т. д.), а вторая предлагает систему шагов при изучении конкретной темы. Так, студент может начать работу над темой непосредственно с выполнения практических заданий; если ему не удается решить предлагаемые задачи, он может выбрать необходимый источник информации: лекции, справочные и вспомогательные материалы, хрестоматию и др. С другой стороны, последовательность выполнения тех или иных упражнений внутри содержательного блока является оправданно регламентированной и отражает логику освоения конкретной лексикологической проблемы.

ЭУ «Лексикология: современные подходы к изучению слова» состоит из нескольких компонентов: курса лекций, практических заданий, лабораторных работ, словарных материалов, хрестоматии – и предназначен для студентов, изучающих русский язык на филологических специальностях, а также для изучающих русский язык как иностранный на факультетах славистики.

ЭУ включает материалы словарей русского языка разных типов; фрагменты разножанрового и разностилевого современного русского дискурса; дидактический материал, отражающий динамические процессы в современном словаре и тестовые задания.

В настоящем докладе осмысляются теоретические и дидактические подходы к представлению в ЭУ слова в системном и коммуникативно-прагматическом аспектах.

Если рассматривать лексику под таким углом зрения, то следует констатировать, что задача объяснить «поведение» языковых единиц с опорой на общие антропоцентрические механизмы является совершенно органичной, не противоречащей положению дел в семасиологии, так как опирается на высокий уровень структурно-семантического описания.

Изучение лексикологического материала в рамках антропоцентрического подхода, включающее такие исследовательские области, как лексическая организация текста, лексическое воплощение языковой личности, коммуникативно-прагматическое описание разных типов дискурса (в его лексическом воплощении), социолингвистическое изучение словаря, динамические процессы в современном языке и др., практически не представлено в учебно-методической литературе: нет ни учебных пособий, ни методических рекомендаций к практическим и лабораторным занятиям (за исключением материалов к спецкурсу «Антропоцентрические аспекты изучения лексики». СПб, 1994 и учебника по лексикологии русского языка Н. Е. Сулименко).

Одной из наиболее сложных и неоднозначно трактуемых проблем современной лексикологии и современной вузовской дидактики является анализ системных отношений в лексике, а также исследование их текстового воплощения (выявление коммуникативного потенциала слова).

Разработка содержательной стороны ЭУ предполагает предварительное детальное изучение системных отношений лексических единиц (языковых и текстовых парадигм) – таким образом формируется блок ответов (желательно однозначных и непротиворечивых) к каждому этапу задания. Разноаспектный анализ семантических групп, кроме собственно дидактических задач, позволяет уточнить, а иногда и снять противоречия в лексикографической интерпретации лексических единиц. Иными словами, создание дидактического блока для ЭУ требует специального семантического и коммуникативно-прагматического исследования языкового материала.

Специфика подготовки дидактического материала для ЭУ состоит в следующем: 1) выделение в традиционном задании по лексикологии нескольких этапов, что определяется спецификой работы с особым типом учебника; 2) пошаговое выполнение каждого этапа задания; 3) подготовка справочных материалов.

Пошаговая работа с языковым материалом позволяет взглянуть на семасиологическое явление с разных сторон и получить адекватное представление об изучаемом объекте. Кроме того, происходит интеграция полученных ранее знаний (о методике проведения компонентного анализа; о типологии сем и др.) с новой информацией, например, о синонимах русского языка.

С помощью дистанционных технологий удается существенно расширить объем активно используемой информации, а также сформировать у студентов умение самостоятельно создавать «исследовательскую модель» для описания того или иного фрагмента словаря.

Кроме того, использование ЭУ позволяет успешно осваивать изучающим русский язык как иностранный все многообразие единиц лексической системы и осуществлять коммуникативно оправданный выбор слова.

Документационное обеспечение управления: лингвистический аспект

Е. Н. Вакулова

Северо-Западная академия государственной службы, Санкт-Петербург

Документ, стандарт, официально-деловой стиль, гипертекст, грамматикализация

Summary. Text of official document is highly formal and standard one, so it can be understood traditionally as a text itself (the main part of a document) and a whole document, including all its informative parts and also as text plus other texts, which are its appendixes and its part. Another typical feature of documental text is tendency to grammatical liason (ties): many words and expressions which are not expected to be connected in a document text, tend to be grammatically connected.

Документационное обеспечение управления (ДОУ), или делопроизводство, в качестве объекта лингвистических исследований представляет интерес в связи с наличием в тексте служебных документов (СД) экстралингвистических и лингвистических черт, его унифицированности и стандартизированности ГОСТ Р 6.30-2003 «Унифицированная систе-

ма организационно-распорядительной документации. Требования к оформлению документов» и ГОСТ Р 51141-98 «Делопроизводство и архивное дело. Термины и определения» устанавливают правила оформления СД,, причем текст в них признается одним из реквизитов, а его лингвистическая сторона не рассматривается.

Реквизиты, со временем вынесенные за пределы собственно текста, обособлялись, переносились на оборотную сторону документа (визы согласования), даже за пределы листа (резолюция, что демонстрирует высокую степень **аналитичности** документа: *Присутствовали: 46 человек (список прилагается), Приглашенные: 18 человек (список прилагается)*; *СЛУШАЛИ: Иванову В. В. (текст доклада прилагается); ВЫСТУПИЛИ: Петров А. А. (стенограмма выступления прилагается)*; *Приложение: на 3 л. в 2 экз.*), а сопроводительные письма, являясь основным документом чисто формально, на самом деле имеют вспомогательную функцию (облегчение процедуры отправки и дальнейшей работы с пересылаемым документом-приложением). Такая «экстракция» текстовых частей как проявление **гипертекстуальности**, наблюдается также у актов, справок, отчетов и пр.

Одним из вопросов, связанных с текстом СД, является проблема его **границ**:

1) текст как таковой – по ГОСТу, один из 30 реквизитов;

2) текст в расширенном понимании – все, что оформлено на стандартном листе формата А4 или А5 в процессе составления документа, но не входящее в него (резолюция, отметки на документе),

3) текст в широком смысле, некий **гипертекст** – т. е. с включением в его состав приложений, тесно связанных с ним, но вынесенных за его пределы.

Формализованность текста СД может противоречить традиционным видам связности, что проявляется в **тяготении к связности грамматической** – «текстовости», грамматикализации как более устойчивому и традиционному виду (напр., при оформлении реквизитов «Наименование учреждения», «Гриф утверждения», «Наименование вида документа» и др.).

Высокая степень формализации текста СД проявляется в:

- вынесении за пределы текста связанных с текстом реквизитов);
- вынесении за пределы документа значимых частей (списков, приложений, смет, графиков и т. д., поясняющих, уточняющих, дополняющих текст СД);
- отходе от архаичных, устаревших, традиционных форм выражения;
- «деграмматикализации» – отказе от грамматической связности и знаков препинания в формализованных частях текста.

Грамматическая связность побеждает при оформлении числительных (*2-ой, 5-ому, 18-тый, На 2-х листах в 1-ом экземпляре*) что, очевидно, связано со стремлением придать официальному тексту точность, однозначность, исключить возможность инотолкований (смешения числительных количественных, порядковых и собирательных, обозначаемых на письме цифрами) и грифа утверждения.

Отмечен отказ от стандартных для СД выражений в пользу четкости, логичности, простоты: *Дана настоящая такому-то в том, что он…* (справка); *Мы, нижеподписавшиеся, составили настоящий акт в том, что…* (акт); *Довожу до Вашего сведения, что…* (докладная записка). Традиционный предложный падеж из «Повестки дня» протоколов заменяется именительным. Все эти формы, оставаясь в употреблении, начинают устаревать и возможно, вскоре архаизируются.

Компьютерная поддержка формирования словарей общенаучной и терминологической лексики*

В. Д. Гусев, Н. В. Саломатина

Институт математики СО РАН, Новосибирск

Подборка текстов, L-грамма, совместный частотный спектр, словарь

Summary. The computer support system of dictionaries formation of common scientific and terminological lexicon on text's group using the *L*-gram approach is described.

1. *Цели исследования*. Задача формирования различных словарей для анализа текстов на естественном языке не теряет своей актуальности ввиду *эволюции языка* (особенно в быстро развивающихся научных областях) и совершенствования самих словарей (их функциональной дифференциации, интеллектуализации и т. п.). Весьма трудоемкий процесс формирования словарей можно *частично автоматизировать*, освободив экспертов от полного просмотра текстов «обучающей» выборки. Описывается трехуровневая (текст – группа текстов – подборка из разнотипных групп текстов) система компьютерной поддержки формирования словарей и дифференциации лексики на общеупотребительную, общенаучную и терминологическую, основанная на *L*-граммном представлении текстов, учете частот и характера распределения *L*-грамм по текстам обучающей выборки, а также формализации понятия «устойчивая цепочка слов» [1].

2. *Система L-граммного представления* иллюстрируется на примере группы текстов $T = (T_1, T_2, \ldots, T_m)$, где m – число текстов. *Частотной характеристикой L-го порядка группы текстов T* назовем совокупность всевозможных связных *L*-словных цепочек ($L = 1, 2, \ldots$), представленных в текстах из T, с указанием частот их встречаемости и распределения по отдельным текстам: $\Phi_L(T) = \{\varphi_{L1}(T), \varphi_{L2}(T), \ldots, \varphi_{LM_L}(T)\}$, где каждый элемент $\varphi_{Li}(T)$ $(1 \le i \le M_L)$ есть четверка: $\langle$ *i*-я *L*-грамма x_i; $F_1(x_i)$ – число текстов из T, в которых встречается x_i; $F_a(x_i)$ – абсолютная частота встречаемости x_i в T; вектор числа вхождений x_i в каждый из текстов группы T: $f(x_i) = (f_1(x_i), f_2(x_i), \ldots, f_m(x_i))\rangle$. Параметр M_L определяет число различных *L*-грамм в T. Вектор $f(x_i)$ позволяет оценить степень равномерности (неравномерности) распределения x_i по текстам из T в виде среднеквадратичного отклонения $\sigma_{f(x_i)}$ значений $f_k(x_i)$ $(k = 1 \div m)$ от $f(x_i) = F_a(x_i) / m$.

Совокупность частотных характеристик $\Phi(T) = \{\Phi_1(T), \Phi_2(T), \ldots, \Phi_{L_{\max}}(T)\}$, где $L_{\max}(T)$ – длина максимальной цепочки слов, общей хотя бы для пары текстов из T, назовем *совместным частотным спектром* группы текстов T. Иными словами, в $\Phi(T)$ включается такой набор частотных характеристик, который содержит информацию о связях между текстами в виде общих цепочек слов с длинами $L \le L_{\max}(T)$.

L-граммы в $\Phi_L(T)$ могут быть упорядочены лексикографически или по убыванию: а) текстовой частоты $F_1(x_i)$; б) абсолютной частоты $F_a(x_i)$; в) показателя неравномерности распределения по текстам σ_f. Выбор упорядочения может играть существенную роль *в дифференциации лексики на отдельные «пласты»* при всей условности подобного разбиения [2]. Аналогичную цель преследует и совместная обработка групп текстов из разных предметных областей (или жанров), при которой фиксируются лишь *L*-граммы, общие для обеих групп. Например, если T_α и T_β – подборки научных текстов из разных предметных областей, а T_γ – подборка другого жанра (например, общественно-политические тексты), то можно ожидать, что общенаучная лексика будет широко представлена в пересечении *L*-граммных спектров для T_α и T_β, но в гораздо меньшей степени в $T_\alpha \cap T_\gamma$ и $T_\beta \cap T_\gamma$. Общеупотребительная лексика будет в равной мере представлена во всех пересечениях, а терминологическая – в дополнениях $\overline{T_\alpha}$ и $\overline{T_\beta}$ к пересечению $T_\alpha \cap T_\beta$.

* Работа выполнена при финансовой поддержке РФФИ (проект № 06-06-80467).

3. *Апробация*. Рассматривались 3 группы текстов: труды конференции по компьютерной лингвистике (Диалог'2005, число текстов $m = 101$, суммарный объем $N = 257$ тыс. словоупотреблений), по распознаванию образов и анализу изображений (PRIA, $m = 219$, $N \sim 262$ тыс.), а также подборка Политических Комментариев (ПК, $m = 931$, $N \sim 792$ тыс.). Схема обработки включала: выделение слов, их нормализацию, вычисление L-граммных спектров по каждой из групп текстов, их фильтрацию (выделение «устойчивых» цепочек слов), упорядочение, а также сравнительный анализ разных групп (Д'05∩ПК, Д'05∩PRIA и т. п.). Об объемах словарей отобранных L-грамм, представляемых для просмотра эксперту, можно судить по приводимой ниже таблице:

L	Д'05	PRIA	ПК	Д'05∩PRIA	ПК∩PRIA	Д'05∩ПК
1	24491	19007	34465	3841	3497	5364
2	7388	7501	26680	1085	888	1341
3	1420	1661	7204	145	107	191
4	253	306	1400	11	12	16
5	68	64	222	1	1	1
6	18	25	47	–	–	–

Заметим, что значительные объемы словарей 1-грамм в подборках Д'05 и PRIA объясняются спецификой научного текста (богатая терминология, аббревиатуры, даты, иноязычные вкрапления и т. п., трактуемые как отдельные слова). В подборке ПК к этим факторам добавляется еще один – большой объем. Словари пересечений (Д'05∩PRIA и др.), содержащие основной пласт общеупотребительной и общенаучной лексики, существенно меньше по объему. В первом приближении приемлемый по полноте и качеству (см. требования, сформулированные в [3]) подсловарь общенаучной лексики может быть получен уже на основе анализа пересечения Д'05∩PRIA. L-граммы в нем ранжированы по убыванию суммы текстовых частот $F_{T\Sigma} = F_T(x| \; x \in$ Д'05$) + F_T(x| \; x \in$ PRIA$)$. Приведем примеры L-грамм ($L = 1 \div 4$), возглавляющих списки, с указанием значения $F_{T\Sigma}$ и места в упорядочении (ранг r). Для сравнения будем указывать и их ранг $r_{чс}$ в частотном словаре Л. Н. Засориной. $L = 1$: *являться* ($F_{T\Sigma} = 302$; $r = 14$; $r_{чс} = 261$); *литература* (соответственно, 287; 20; 662); *работа* (278; 26; 98); *результат* (268; 28; 603); *значение* (265; 31; 545); *задача* (263; 33; 279) и т. д. $L = 2$: *в* / *качество* ($F_{T\Sigma} = 191$; $r = 2$); *такой* / *образ* (185; 4); *в* / *вид* (178; 5); *на* / *основа* (159; 7); *с* / *помощь* (158; 8) и т. д. $L = 3$: *в* / *этот* / *случай* ($F_{T\Sigma} = 87$; $r = 1$); *в* / *данный* / *работа* (84; 2); *в* / *настоящий* / *время* (76; 3); *с* / *точка* / *зрение* (58; 4) и т. д. $L = 4$: *в* / *то* / *же* / *время* ($F_{T\Sigma} = 32$; $r = 1$); *в* / *связь* / *с* / *этот* (29; 2) и т. д. Многие L-граммы (особенно длины 3) могут быть объединены в группы, описываемые единым шаблоном («образец» с переменным параметром). Самая крупная группа описывается образцом вида: *в* / *X* / *случай*, где переменная *X* может принимать любое значение из множества {*этот*, *данный*, *общий*, *первый*, *большинство*, *противный* и т. д.}.

Описанная система может также найти применение для тематической классификации, выявления заимствований, анализа стиля.

Литература

1. *Гусев В. Д., Саломатина Н. В.* L-граммное представление текстов на естественном языке и его возможности // Квантитативная лингвистика: исследования и модели. Материалы всерос. научн. конф., Новосибирск, 2005. С. 256–270.
2. *Остапенко В. Е.* Выделение и классификация терминов с помощью элементарных квантитативных моделей // НТИ. Сер. 2. 1989. № 11. С. 24–28.
3. *Большакова Е. И.* О принципах построения компьютерного словаря общенаучной лексики // Труды Международного семинара Диалог'2002 по компьютерной лингвистике. Протвино. Т. 1. С. 19–23.

Формализация содержания русского лирического стихотворения

А. В. Зубов

Минский государственный лингвистический университет (Беларусь)

Динамический, слово, содержание, статический, текст, строфа, формула

Summary. In this paper the author submits to conside the content of a text in static and dynamic aspects. The main criterion of the first aspect is the frequency of a word of the text. The dynamic aspect of a text is the linear succession of logical and semantic contents of the text strophes.

Под стихотворением понимается разбитые на строфы последовательности семантически связанных предложений. Поль Валерии заметил, что «стихотворение создается не из идей и не из чувств, оно создается из слов» [1, 275]. Естественно, что без идеи и чувств создать стихотворение невозможно, но когда речь идет об анализе уже готового стихотворения главным становится слово.

Известно, что каждый текст выражает индивидуальные лингвистические и экстралингвистические знания автора. Лингвистические знания связаны со спецификой употребляемых автором слов, словосочетаний, типов предложений и их последовательностей (абзацев, строф). Экстралингвистические особенности автора проявляются в выборе манеры изложения, в способе представления ситуаций окружающей действительности, выражающемся структурной организацией абзаца или строфы, порядке следования в тексте этих укрупненных единиц и т. п.

Выделяя в содержании каждого текста две составляющие – статическую и динамическую – можно определенным образом смоделировать заложенные в текст лингвистические и экстралингвистические знания автора текста.

Статическая составляющая текста может быть представлена в виде таблицы основного статического содержания, в которую путем статистического анализа текста включены наиболее важные для данного текста знаменательные слова – главные и второстепенные опорные слова. Они, с учетом всех словарных и контекстуальных синонимов текста, отражают главные действующие лица текста, место и время действия их, зафиксированные в исследуемом тексте.

Динамическая составляющая каждого текста может быть описана некоторой логико-семантической формулой этого текста. Если говорить о тексте стихотворения, то предполагается, что каждая строфа содержит описание некоторой микроситуации, т. е. определенного числа субъектов, объектов реального мира с их свойствами и взаимоотношениями.

Основной смысловой единицей динамической составляющей содержания стихотворения является предметно-логическое содержание строфы. Под предметно-логическим содержанием строфы понимается дифференцированное по типам главных и второстепенных субъектов, объектов, мест и времени действия ситуации, представленной в стихотворении, перечисление фактов объективной действительности (явлений, событий, состояний, признаков и т. п.), связанных с содержанием всего стихотворения. Например: «Констатация некоторого действия главного субъекта и описание природы». Или: «Описание состояния и действий автора» и т. д. В текстах каждого автора можно выделить конечное число таких предметно-логических составляющих. Давая им определенные буквенные обозначения (например, М001, М002 и т. д.), можно построить логико-семантические формулы таких текстов.

Например: T_{01} = Т001&Т012&Т006&Т021&Т005.

Предлагаемый подход к описанию содержания текстов демонстрируется в докладе на примере более 100 русских стихотворений [2].

Литература

1. *Моруа А.* Литературные портреты. М., 1970.
2. *Зубов А. В.* Всю страсть души… Стихотворения. Минск, 2001.

Моделирование семантики ритмического текста стиха и ее вербальных соответствий

М. А. Красноперова

Санкт-Петербургский государственный университет

Ритмический текст, семантика, модель, лексика

Summary. The problem of the rhythmical text semantics is under consideration. It has been studied on the basis of the author's model of perception and generation of the rhythmical structure of a text. Two types of rhythmical semantics are exposed both the basis and superstructure ones. The question of the relations between rhythmical and lexical semantics is raised and peculiarities of the former are shown as well.

Основная задача предлагаемой работы состоит в том, чтобы описать представление о семантическом потенциале ритмического текста стиха на базе разработанной автором модели его порождения и восприятия (МПВ – [1], ср. рец. [2]) и показать отношения между ритмической и лексической семантикой текста.

Основными компонентами модели являются механизм рецепции и механизм генерации. Первый из них осуществляет прием ритмических структур стихотворных строк (ритмических строк) при порождении и восприятии текста, второй отражает и контролирует процессы, имеющие место в механизме рецепции, и выполняет порождающие функции.

Основными процессами, происходящими в рамках механизма рецепции, являются реализация ритмических строк и накопление выделенностей слогов. В процессе реализации осуществляется взаимодействие текущей ритмической строки и состояния оперативной памяти (ОП) модели, образовавшегося на основе ранее воспринятых строк текста. В результате этого взаимодействия возникают ритмические эффекты и ритм текста. В процессе накопления в ОП возникает последовательность выделенных и невыделенных зон, называемых сильными и слабыми местами, и формируется внутренний метр.

Ядром механизма генерации является языковая система метра (ЯСМ). Ее главные компоненты представлены внешним метром и механизмом ипостас. Внешний метр – это скандирующий механизм, способный маркировать сильные и слабые позиции стихотворной строки и контролировать соответствует ли их расположению ритмические строки. Механизм ипостас – это устройство, способное запоминать процедуры осуществления ритмических эффектов, возникающих при восприятии текстов, и воспроизводить их при порождении с соответствующими вероятностями. В процессе порождения ЯСМ в сочетании с факторами, воздействующими на нее извне, создает предварительный ритм, ориентируясь на который, система, включающая модель, подбирает соответствующее словесное наполнение.

На основе предлагаемой модели разрабатывается теория семантики ритмического текста Принципы семантического описания основаны на следующих соображениях. Модель исходит из ограниченного набора элементарных ритмических эффектов. Каждому из них можно поставить в соответствие наименование, выражающее подходящее к нему ощущение («отягчение», «облегчение», «нагнетание», «распад», «перемещение», «появление», «увеличение» и др.). Каждому ритмическому эффекту, не являющемуся элементарным, соответствует определенная система таких описаний. При этом вступают в силу количественные, позиционные признаки, признак направления и др.: для облегчения / отягчения – «вес», для перемещения – расстояние, перемещение влево или вправо и т. д. Эти признаки, а также отношения между ними порождают, в свою очередь, новые семантические качества (например, отягчение» / «облегчение» → «контраст» и т. п.).

Сравнительное исследование семантических связей ритмических структур, образованных началами соседних стихотворных строк, на материале произведений русских поэтов XIX–XX вв. позволяет предполагать наличие у них довольно устойчивой системы семантических предрасположенностей ([3]). Данные, полученные при исследовании семантических тенденций этих структур в «Евгения Онегина» А. С. Пушкина, дают основание предполагать также, что они участвуют в формировании структуры содержания текста на уровне, предшествующем его словесному наполнению [4]. Это дает эмпирическую поддержку развиваемой теории.

В рамках теории различаются понятия базовой и надстроечной семантики ритмического текста. Под базовой семантикой понимаются ритмические эффекты, возникающие при осуществлении процедур взаимодействия между ритмическим словом и состоянием ОП на участке его реализации Под надстроечной семантикой понимается смысл слов и их сочетаний, описывающих реалии базовой семантики.

Дается общее представление о ядерном языке надстроечной системы: единицах конкретной и абстрактной лексики его словаря, правилах преобразования словарных единиц и их сочетаний. Единицы конкретной лексики представляют собой названия однородных (однородными называются ритмические эффекты, состоящие из одинаковых элементарных) ритмических эффектов (уменьшение, увеличение, перемещение, нагнетание и т. п.), их отрицаний и сочетаний этих элементов: Единицы абстрактной лексики должны включать указания на отношения между ритмическими эффектами или компонентами типовых ситуаций в сфере действия данной единицы: количественные – меньше / больше, легче / тяжелее; позиционные – перед / после, близко / далеко / примыкание / пересечение, над / под, наложение (например, слова на ОП), части / целого; подчинения (безударных слогов – ударному, района безударного сильного места – району ударного); изменения, повторения, противопоставления и другие. Единицы абстрактной лексики либо являются названиями самих отношений, либо названиями более сложных образований, включающих указания и на ритмические эффекты, и на определенные отношения.

Все правила надстроечной системы не являются правилами сугубо языкового типа. Они ориентируются на реалии базовой системы, которые описывают свойства ритмических, а не вербальных сущностей. Так, например, элемент *исчезновение* в системе ритмической семантики влечет за собой семантический признак *облегчение,* в то время как языковое значение соответствующего слова не обязательно сопряжено с этим признаком.

Так как в МПВ языковая система метра отражает и запоминает процедуры формирования ритмических эффектов в своих операторах, то при достаточной степени развития она содержит в себе и тот семантический потенциал ритмики данного размера, который выражает опыт восприятия прошлых текстов. В процессе порождения этот потенциал будет передаваться через предварительный ритм и его фрагменты во вновь создаваемые тексты. В этом процессе он может также обогащаться не зафиксированными в предыдущем состоянии ЯСМ ритмическими эффектами.

Зафиксированные с помощью ЯСМ ритмические эффекты и их сочетания являются с позиций развиваемой теории одним из источников внешних связей МПВ. Соответствующие им ощущения и более сложные образования могут возникать не только в составе ритма, но и под влиянием других факторов. Они могут участвовать в механизме мышления и внутренней речи и могут быть формой реакции на воздействия внешней действительности. Можно представить, что подобные стимулы, пришедшие извне, способны привести в действие те реалии, которые соответствуют определенным операторам механизма ипостас или их комбинациям и активизировать таким образом остальную систему метра. Отличительная особенность стихотворного ритма состоит в том, что они являются его строительными элементами и потому находят здесь систематическое выражение.
В процессе порождения текста они могут получить вербальную или сопутствующую ей семантизацию.

Описанная система ориентирована на моделирование собственных свойств ритмической семантики. Она не касается тех семантических связей ритмики, которые могут возникнуть в опыте восприятия текстов.

Литература

1. *Красноперова М. А.* Основы реконструктивного моделирования стихосложения. На материале ритмики русского стиха. СПб., 2000.
2. *Златоустова Л. В.* О новом направлении в стиховедении: Теория реконструктивного моделирования стихосложения // Вестник Моск. ун-та. Сер. 9. Филология. 2004. № 5. С. 122–125.
3. *Красноперова М. А., Шлюшенкова Т. Б.* О семантических отношениях ритмообразующих единиц в русской поэзии XIX–XX веков // Славянский стих: VII. Лингвистика и структура стиха. М., 2004. С. 319–338.
4. *Красноперова М. А., Шлюшенкова Т. Б.* Ритмика и лексика романа «Евгений Онегин» в тезаурусе Роже // Studia Metrica et Poetica. Сборник статей памяти П. А. Руднева. СПб., 1999. С. 91–109.

Принципы организации электронной базы диалектных текстов*

О. Ю. Крючкова, В. Е. Гольдин, И. А. Батраева

Саратовский государственный университет

Диалект, база данных, текстовый корпус

Summary. The paper deals with the principles of organization of the dialectical textual corpus considered as the main source for studying cultural and communicative specific of dialects.

Корпусы диалектной речи, отражающие коммуникацию на диалекте в том или ином конкретном населенном пункте и сохраняющие в машиннообрабатываемой форме значительные массивы связной речи, являются основным источником изучения коммуникативной специфики диалектов. Использование материалов таких корпусов дает возможность не ограничиваться отдельными примерами, а переходить к выявлению общих принципов, тенденций, действующих в диалектной коммуникации. Текстовый диалектный корпус должен служить моделью традиционной сельской коммуникации на диалекте, а если он включает текстовые материалы одного говора, то – моделью коммуникации в конкретных условиях жизни данного речевого коллектива.

В лингвистике уже разработаны принципы построения текстовых корпусов как коммуникативных моделей (Британский национальный корпус, Национальный корпус русского языка и др.). Электронные базы диалектных корпусов должны отличаться от других корпусов вследствие специфики своих задач и материала.

Особенностями создаваемого в Саратовском государственном университете диалектного текстового корпуса являются его ориентация на один конкретный говор – говор с. Белогорное (бывш. Самодуровка) Вольского района Саратовской области, решение задач не только собственно лингвистического, но и социокультурного характера, учет специфики диалектной коммуникации.

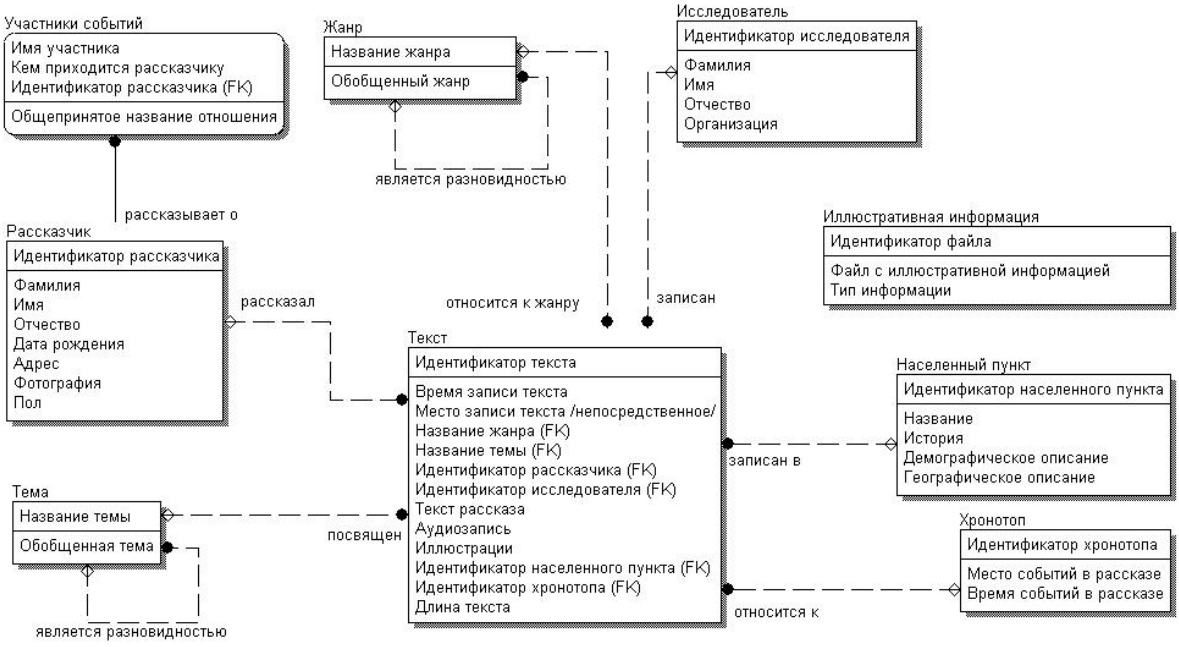
Единицей хранения в корпусе текстов с. Белогорное является «запись» – расшифровка магнитофонной фиксации непрерывного фрагмента общения независимо от числа его участников и тематического варьирования. В корпус включаются тексты, записанные начиная со 2-й половины XX в. от лиц с различными социолингвистическими характеристиками (возраст, пол, образование, род деятельности, вероисповедание, место рождения и др.). Кроме записей от постоянных жителей села, в корпус войдут и записи речи информантов, связанных с изучаемым говором в детские и юношеские годы. Тексты представлены в упрощенной транскрипции.

Говор с. Белогорное расположен на территории позднего заселения, сложился в результате взаимодействия речи различных групп поселенцев и является вторичным. Считается, что основателями села были бежавшие из центральных областей России старообрядцы. Поэтому социолингвистический параметр «вероисповедание» имеет особую актуальность для создаваемого корпуса: в селе традиционно существуют группы, принадлежащие различным конфессиям, в том числе старообрядцы (брачники и безбрачники).

Наряду с символьным представлением текстов корпус включает аудиозаписи, видеозаписи, графические материалы (фотографии, схемы, карты), часть из которых соотнесена с расшифровками текстов, другая часть образует отдельный подкорпус.

В корпусе используется многоаспектная метаразметка записей: сведения об информантах (фамилия, имя, отчество, год рождения, место рождения, родственные связи, образование, род занятий), сведения о времени, месте записи, о конкретной ситуации общения, об адресатах речи, сведения о тематике текста, его хронотопе, упоминаемых лицах и событиях, о жанре и функциональных особенностях общения.

Схема базы диалектных текстов



* Работа выполнена при поддержке Российского фонда фундаментальных исследований (РФФИ), проект № 06-06-80428-а.

Параметрами собственно лингвистической разметки являются лемма, морфологическая характеристика словоформы, контекстная семантика, диалектная фонология.

Планируемый корпус должен отразить особенности коммуникации на диалекте, современное состояние изучаемого говора, характер и факторы варьирования диалектной речи жителей с. Белогорное, основные прецедентные явления и тексты, культурно-речевую специфику отдельных групп населения и своеобразие картины мира носителей диалекта.

Компьютерная версия «Энциклопедия академического журнала “Вопросы языкознания”»

С. В. Лесников

Сыктывкарский государственный университет

Энциклопедия, гипертекст, вопросы языкознания, гизаурус

Summary. The computer version «the Encyclopedia of the academic journal “Questions of linguistics”» for the period of 1952–2006 actually reflects all peripetias of development of linguistics in Russian for last half-centuries. The encyclopedia of materials «Questions of linguistics» is designed as hypertext in the form of the electronic arch – directory and allows to find any material which has been published in academic journal «Questions of linguistics».

Компьютерная версия «Энциклопедия академического журнала “Вопросы языкознания”» (ВЯ) за период 1952–2006 гг. фактически отражает все перипетии развития языкознания на русском языке за последние полвека.

Кроме фактологической систематизации материалов ВЯ (следуя принципам: декомпозиции, единства системы и среды, иерархичности, интеграции, контринуитивного проектирования, контролируемости, обратной связи, полиморфизации, преемственности, реализуемости, синергетического эффекта, системности, совместимости, согласованности, управляемости и т. д.), энциклопедия обеспечивает: – возможность оперативного поиска требуемой филологу (и любому другому ученому) информации; – получение справочной информации по определенной теме, разделу языкознания; – реализацию разнообразных интерактивных режимов работы с материалами ВЯ (анализ, аннотирование, библиографирование, печать, поиск, просмотр, реферирование, рецензирование, цитирование, чтение).

Атрибутированный поиск в энциклопедии может осуществляться по следующим параметрам: 1) рубрикаторы (АПУ, АСВИЯ, ББК, ГАСНТИ, ГСК, ЕКЛ, ИСБН, ОКСТУ, СИБИД, СКС, УДК и др.); 2) автор (редактор, рецензент, составитель, издатель и писатель, творчество которого рассмотрено в языкознании, а также др. субъекты); 3) название (книга, статья, материал, лингвистическое событие); 4) реферат (аннотация, содержание, структура); 5) ключевые слова (дефиниция, определение, понятие, словосочетание, термин, фраза, цитата); 6) библиография (в т. ч. индекс цитирования); 7) раздел журнала (статья, обзор, хронология, рецензия и т. д.); 8) год выпуска; 9) номер журнала.

Энциклопедия материалов «ВЯ» сконструирована в виде гипертекстовой хрестоматии в форме электронного свода-справочника и позволяет найти любой материал, который был опубликован в академическом журнале «Вопросы языкознания» в период с 1952 по 2006 годы. Часть фрагментов энциклопедии доступна в Интернете по адресам www.lsw.ru и http://vault.syktsu.ru/www.lsw.ru.

Психолингвистическое исследование «незаконченное предложение»

Д. С. Лесникова, С. В. Лесников

Санкт-Петербургский государственный университет / Сыктывкарский государственный университет

Мониторинг, тест, эксперимент, психолингвистика, система отношений

Summary. The purpose of ours psycholinguistic experiment is check of the following hypothesis: at students from younger rates to grown-ups changes not only system of mutual relations with world around, but also the stock of active lexicon is essentially enriched.

В экспериментально-психологической практике метод «незаконченные предложения» и в лингвистической практике ассоциативный метод используют давно и в разных вариантах. Данное психолингвистическое исследование базируется на варианте метода «незаконченное предложение», разработанном Саксом и Леви.

Целью нашего психолингвистического эксперимента является проверка следующей гипотезы: у студентов от младших курсов к старшим изменяется не только система взаимоотношений с окружающим миром, но и существенно обогащается запас активной лексики.

60 незаконченных предложений разделены на 15 групп, характеризующих систему отношений испытуемого: к себе, к отцу, к матери, к подчиненным, к будущему, к вышестоящим лицам, к друзьям, к своему прошлому, к лицам противоположного пола, к семье, к сотрудникам. Кроме этого, другие группы предложений отражают отношение к испытываемым человеком страхам и опасениям, к имеющемуся у индивидуума чувства осознания собственной вины, затрагивают собственные жизненные цели и нереализованные возможности. Для каждой группы предложений определена характеристика, определяющая данную систему отношений как положительную, отрицательную или безразличную, что дает возможность выявить у испытуемого элементы дисгармонической системы отношений, при этом наиболее важно качественное психолингвистическое изучение дополненных предложений непосредственно психолингвистом.

В качестве испытуемых выступали студенты первого, третьего и пятого курсов разных факультетов (филологический, юридический, физкультуры и спорта, финно-угорский, гуманитарный, психологии и социальной работы, экономический, математический, информатики и информационных систем) Сыктывкарского государственного университета. Эксперимент проводился в компьютерной форме. Для получения искренних, естественных ответов тестирование предлагалось на персональном компьютере, с возможностью студенту-испытуемому «скрыться» в таком случае за псевдонимом («ником»). Результаты эксперимента относительно самого себя каждый студент мог получить по электронной почте. Для студентов, которые не скрывались за псевдонимами, было проведено также в компьютерной форме более подробное анкетирование.

В целом в эксперименте приняли участие свыше 500 студентов.

Особый интерес представляет ассоциативная составляющая, т. е. то, какой именно лексикой пользуется студент для отражения своих отношений с окружающим (в том числе и внутренним) миром. Результаты анкетирования позволяют также выявить и конкретизировать молодежную, студенческую лексику.

Важным аспектом данного экспериментального исследования является осуществляемый мониторинг испытуемых, т. к. эксперимент проводится среди студентов одного университета, что позволяет в дальнейшем продолжать исследование, а именно при переходе студентов на старшие курсы.

Эксперимент продолжается в системе Интернет. Тест выставлен по адресам:

- www.lsw.ru;
- http://vault.syktsu.ru/www.lsw.ru.

Однозначные словосочетания-синонимы многозначных слов

Н. В. Лукашевич

Московский государственный университет им. М. В. Ломоносова

Разрешение лексической многозначности, синтаксические синонимы

Summary. The paper describes one of additional sources of information for lexical disambiguation. We show that for many senses of ambiguous words there exist unambiguous synonymic multiword expressions, which are useful for disambiguation and therefore needed to be collected in linguistic resources intended for natural language processing.

Одной из серьезных проблем систем автоматической обработки текстов является разрешение лексической многозначности. Предложено множество методов выбора значения многозначных слов при автоматической обработке текстов, базирующиеся на семантически размеченных корпусах, словарных статьях электронных словарей, формализованных моделей управления, использовании структурных лингвистических ресурсов типа тезауруса Роже или тезауруса WordNet и др. ([1]). Однако проблема далека от решения и требует дальнейшего развития комплексных методов и сбора необходимых источников информации.

Одним из факторов, который может быть учтен при процедуре разрешения лексической многозначности, является существование у значительного числа многозначных слов синонимичных однозначных словосочетаний, включающих в свой состав либо само это многозначное слово либо его дериват (обычно также многозначный).

Известными примерами таких словосочетаний являются, описанные в [2], словосочетания с использованием родовых понятий вида Q(C0) + Gener (C0), где Q(C0) – обозначает некоторый синтаксический дериват от C0, например, *республика = республиканское государство* [C0 = *республика,* Gener (C0) = *государство*, Q(C0) = *республиканский*].

Известным видом словосочетаний, синонимичных глаголам и также часто являющихся однозначными, являются фразеологические синонимы, включающие лексические функции Oper1, Oper2 – *оказать помощь = помочь, оказать сопротивление = сопротивляться, принимать решение = решать*.

На самом деле, словосочетания, синонимичные значениям многозначного слова, весьма разнообразны. Они образуются из исходного слова или его деривата и из наиболее значимого слова из толкования.

Например, в [3] первое значение слова *агрессия* толкуется следующим образом: «вооруженное нападение государства или группы государств на какое-то государство…». Как синоним этого значения слова *агрессия* активно употребляется словосочетание *вооруженная агрессия.*

Часто у каждого из значений многозначного слова имеется свой однозначный синоним-словосочетание.

Например, слово *болид* имеет два значения в [3]: «1. Очень яркий крупный метеор 2. Гоночная машина со сверхмощным двигателем». Соответственно достаточно употребительны словосочетания *космический болид* как синоним к первому значению слова и *гоночный болид* как синоним ко второму значению.

Если рассматривать основные типы структур словосочетаний-синонимов к многозначным существительным, то подавляющее большинство таких словосочетаний представляют собой следующие конструкции (исходное слово C0):

- A(C0) + Gener(C0): *авангард$_3$ = авангардное искусство, архив$_1$ = архивное учреждение, авиация$_2$ = авиационная техника, экология$_2$ = экологическая система*;
- Gener(C0) + C0 в родительном падеже: *авангард$_3$ = искусство авангарда, авангард$_4$ = произведения авангарда, экспедиция$_2$ = отдел экспедиции, чай$_3$ = настой чая*. Такие конструкции становятся возможными из-за метонимической связи между значениями: внутри словосочетания многозначное слово обычно употребляется в значении, отличном от значения целого выражения;
- C0 + (существительное в родительном падеже) или прилагательное + C0. Зависимые существительные и прилагательные могут в таких словосочетаниях выражать достаточно широкий спектр характеристик значения слова, например, его целое (*бородка$_2$ = борода ключа*), происхождение (*болид$_1$ = космический болид, челюсть$_2$ = искусственная челюсть*), назначение (*блок$_1$ = подъемный блок, бревно$_2$ = гимнастическое бревно*), типы его актантов (*арест$_2$ = арест имущества, адаптация$_2$ = адаптация текста*), а также другие значимые характеристики (*карьер$_{II}$ – открытый карьер,* брак$_1$ – *зарегистрированный брак*).

Реже встречаются конструкции с предлогами, которые обычно передают назначение предмета: *экран$_2$ = экран для показа, штопор$_1$ = штопор для бутылок.*

Предложные конструкции синонима-словосочетания также могут основываться на метонимии значений слова*: шахматы$_1$ = игра в шахматы, шерсть$_4$ = ткань из шерсти.*

Таким образом, явление активного употребления однозначных словосочетаний-синонимов для многозначных слов достаточно распространено. При этом для каждого конкретного многозначного слова нельзя точно предсказать, существуют ли для его значений однозначные синонимы-словосочетания. Их существование приходится проверять по корпусам и в сети Интернет.

Поскольку знание таких синонимов-словосочетаний кажется значимым фактором при автоматическом разрешении многозначности, то мы, разрабатывая Тезаурус русского языка РуТез ([4]), предназначенный для автоматической обработки текстов, специально ищем такие однозначные словосочетания и добавляем их в синонимические ряды соответствующих значений. Критерием добавления служит нахождение около 100 интернет-страниц, в которых упомянуто такое словосочетание.

Такие однозначные синонимы-словосочетания могли бы быть полезными и для лексико-семантической разметки корпусов ([5]).

Литература

1. *Кобрицов Б. П.* Методы снятия семантической многозначности // Научно-техническая информация. Сер. 2. 2004. № 2.
2. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл-текст». М., 1974.
3. Большой толковый словарь русского языка. СПб., 1998.
4. *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог’2002 / Под ред. А. С. Нариньяни. М., 2002. Т. 2. С. 338–346.
5. *Рахилина Е. В., Кобрицов Б. П., Кустова Г. И., Ляшевская О. Н., Шеманаева О. Ю.* Многозначность как прикладная проблема: лексико-семантическая разметка в корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог’2006 / Под ред. А. С. Нариньяни. М., 2006. Т. 2. С. 445–450.

Русский язык и некоторые проблемы новых информационных технологий

Ю. Н. Марчук

Московский государственный университет им. М. В. Ломоносова

Современная автоматическая обработка текстов включает такие проблемы, как автоматический анализ и синтез текстов с помощью лингвистических алгоритмов, построенных на чисто формальных признаках, и автоматических словарей – собраний лексики с определенными формальными характеристиками. Из практических применений новых информационных технологий остановимся на машинном переводе.

В Интернете и на рынке информационных технологий имеется достаточное количество систем машинного перево-

да (см. обзор П. Н. Хроменкова – [4]). Однако качество перевода таково, что без пост-редактирования он может быть использован только в чрезвычайно редких случаях.

Одной из главных причин низкого качества перевода является неправильный перевод лексики. Мы усматриваем два основных положения в основе такого явления: 1) нерешенный вопрос о взаимодействии различных словарей в переводе и 2) отсутствие четких правил и алгоритмов контекстного разрешения лексической неоднозначности в современных системах, предлагаемых на информационном рынке.

Принято различать лексику основного словарного фонда и лексику периферийных слоев – в том числе терминологию разных областей знания. Хорошо известно, что в нынешней информационной ситуации действуют новые принципы философии языка, которые, в частности, определяют роль языка среди других знаковых систем ([3]). Границу между словарным составом основного фонда языка и терминологическими областями часто трудно провести достаточно четко, вследствие постоянного развития терминологии, появления новых терминов и новых значений у уже имеющихся слов. (Пример: *интерфейс* и *материнская плата*). Активные процессы в современном русском языке (см. [1]) делают эту проблему весьма актуальной. Решением проблемы в автоматической обработке текстовой информации является правильный выбор иерархии словарей, которая, однако, не может быть построена априори, поскольку в этом случае дает ошибки, а должна базироваться на тщательном исследовании особенностей лексики конкретного микроподъязыка. В таком исследовании применяются методы корпусной лингвистики, особенно квантитативные методы.

Второй вопрос – о роли контекста в разрешении лексической многозначности. Как можно убедиться из рассмотрения примеров машинного перевода современными системами, в большинстве случаев терминологические словари и словари общеупотребительной лексики не учитывают контекста употребления слов и поэтому дают ошибки. Разрешение лексической многозначности происходит с помощью автоматически действующего контекстологического словаря ([2]). Однако работа по составлению такого словаря требует больших усилий и времени даже у многочисленного коллектива разработчиков, поэтому большинство систем МП, представленных на рынке, не имеют в своем составе контекстологических словарей или подобных устройств, разрешающих неоднозначность лексических единиц и их сочетаний в конкретных контекстах, пусть даже достаточно обобщенных.

Теория контекстного определения и разрешения неоднозначности, как лексической, так и лексико-грамматической, требует своей дальнейшей разработки и создания соответствующей эффективной методики. До этого, в современном положении с машинным переводом, приходится полагаться либо на явно выраженную однородность переводимых машинным способом текстов, либо на работу пост-редакторов, причем последнее практически сводит на-нет преимущества собственно автоматического перевода.

Литература

1. *Валгина Н. С.* Активные процессы в современном русском языке. М., 2003.
2. *Марчук Ю. Н.* Основы компьютерной лингвистики. М., 2002.
3. *Рождественский Ю. В.* Философия языка. Культуроведение и дидактика. Современные проблемы науки о языке. М., 2003.
4. *Хроменков П. Н.* Современные системы машинного перевода. М., 2005.

Лексическая база данных как основа для компьютерного тренажера по лексике

М. Н. Михайлов

Тамперский университет (Финляндия)

Лексический тренажер, веб-интерфейс, лингвистический подход к обучающим программам, многоязычная база данных

Summary. SysMLL is a web-based software application for the purposes of language learning in the academic framework. SysMLL is developed so that it can store multilingual data. The main storage unit of the database is the meaning, not the lexeme. This makes it possible to effectively store multilingual correspondences, to distinguish between different meanings of the same lexeme, to store synonyms under same meaning. The system automatically generates exercises using lexical database, which makes the system flexible and infinite.

Мечта сделать компьютер надежным помощником преподавателя в учебном процессе достаточно стара. Причем именно преподавание языков оказывается одной из самых важных сфер компьютеризации обучения. Однако действительно широкого распространения обучающие программы до сих пор не получили, хотя работы в этом направлении ведутся достаточно активно (см., напр., [6]).

Один из пионеров в этой области, Дж. Хиггинс, критиковал компьютерные реализации традиционных «бумажных» сборников упражнений как бесперспективные и призывал использовать весь потенциал вычислительной техники ([4], [3]). К сожалению, за почти двадцать лет, прошедшие с тех пор, кардинальных изменений в ситуации не прослеживается. Современные обучающие программы в большинстве своем – сборники упражнений и тестов либо обучающие игры с замаскированными языковыми заданиями. Существуют и готовые программные оболочки для создания упражнений и тестов, например HotPotatoes. Промежуточный тип – лексические тренажеры. Этот класс программ ориентирован на разучивание новых слов и выражений, назову в качестве примера Vtrain и KvocTrain (см. ссылки в конце статьи). Сильная сторона лексических тренажеров состоит в том, что эти программы предоставляют новый вид упражнений, отличный от предлагаемых в традиционных учебниках. Слабость этих программ в том, что практически все известные нам тренажеры ориентированы на работу со списками практически без использования лингвистически релевантной информации (семантика слова, толкование, синонимы / антонимы и т. п.). Описываемая в настоящих тезисах обучающая программа является лексическим тренажером нового поколения, усиленным в плане лингвистических данных и реализующая «индивидуальный подход» к пользователю.

Это SysMLL (Multilingual System for Language Learning), обучающая система по иностранным языкам, создаваемая в Тамперском университете (Финляндия) силами нескольких подразделений (Институт современных языков и переводоведения, Центр обучения иностранным языкам, Отделение вычислительной техники, Виртуальный университет). Цель системы – расширение, активизация и тестирование словарного запаса студентов.

Система может поддерживать несколько языков, но в настоящее время активно вводятся данные только по двум языкам – русскому и финскому. Позднее к этим языкам будет добавлен английский, возможно пополнение системы и другими языками. Доступ к системе осуществляется с помощью веб-интерфейса.

Главное отличие этой системы от существующих состоит в том, что в ее основе – лингвистический подход. Иначе говоря, сердцевину программы составляет не шаблон для ввода упражнений / тестов, и не списки слов и выражений на двух языках (возможно, с картинками), а большой структурированный словарный массив, используя который система сама генерирует тесты и упражнения разных типов. Это позволяет сделать работу с тренажером «бесконечной». Задача преподавателя сводится лишь к вводу лексики для активизации и редактированию существующих записей (уточнение толкований, комментариев, ввод дополнительных примеров употребления). Пользователи системы также получают возможность вести свои рабочие глоссарии, часть этого материала может позднее включаться администратором базы данных в основной словарь.

Основной единицей хранения в базе данных SysMLL является лексико-семантический вариант, а не лексема, что позволяет гибко решать проблему многозначности, омонимии и отсутствия точного эквивалента в другом языке. Зна-

чения слов описываются с помощью наборов английских ключевых слов, например ‘director film production person’ = рус. *режиссер, режиссер-постановщик*, англ. *director*, фин. *elokuvaohjaaja*. Наш подход в некоторой степени напоминает подход WordNet, FrameNet и др. (см. напр. [1], [2], [5]), но в сильно упрощенном виде.

Группы записей с одинаковым значением объединяются в гнезда, как в только что приведенном примере. Таким образом тренажер может получать из базы данных группы переводных эквивалентов и (квази)синонимические ряды.

База данных и программная оболочка размещены на одном из серверов Тамперского университета и доступны через Интернет по адресу: https://www11.uta.fi/laitokset/kielikeskus/sysmll/ (для входа в систему необходим логин и пароль, которые можно получить у администратора базы данных). В настоящее время идет тестирование программы на небольших группах пользователей.

В базе данных на настоящий момент около 3500 записей, значительную часть которых составляют русские и финские слова и выражения.

Пакет веб-приложений генерирует различные упражнения: просмотр межъязыковых соответствий, подбор переводных эквивалентов, подбор слов и выражений по толкованию. Планируется создание новых типов упражнений и языковых игр (кроссворды, «Виселица» и т. п.), а также работа со звуком и графикой.

У каждого пользователя свой профиль, позволяющий выбирать языковую пару, уровень сложности и темы для работы. Программа запоминает ошибки пользователя и постоянно заставляет его делать «работу над ошибками», повторяя те задания, с которыми пользователь не справился.

Более подробную информацию о проекте можно получить на сайте http://hamppu.uta.fi/~lomimih/sysmll_project/.

Литература

1. *Boas H. C.* Semantic Frames as Interlingual Representation for Multilingual Lexical Databases // International Journal of Lexicography. 18(4). 2005. P. 445–479.
2. *Fellbaum C.* WordNet: an Electronic Lexical Database. Cambridge, Mass., 1998.
3. *Higgins J.* Language, Learners, and Computers. London; New York, 1988.
4. *Higgins J., Johns T.* Computers in Language Learning. Collins ELT. Addison-Wesley, 1984.
5. *Janssen M.* Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA // International Journal of Lexicography. 17(2). 2004. P. 137–155.
6. *Mohan B.* Models of the Role of the Computer in Second Language Development // Pennington M. C., Stevens V. (eds.). Computers in Applied Linguistics: an International Perspective. Clevedon; Philadelphia; Adelaide, 1992. P. 110–127.

Интернет-ссылки

1. ABBYY Lingvo: http://www.lingvo.ru/multilingual/.
2. Hot Potatoes: http://web.uvic.ca/hrd/halfbaked/.
3. KVocTrain: http://edu.kde.org/kvoctrain/.
4. Open Book: http://www.vinidiktov.ru/openbook.htm.
5. Vtrain: http://www.paul-raedle.de/vtrain/home-ru.htm.

Компьютерные технологии для русистики

Г. Никипорец-Такигава

Токийский университет иностранных языков (Япония)

Nikiporets-takigawa@tufs.ac.jp

Корпусные ресурсы, интернет-поисковые машины, статистический анализ, электронные частотные словари, лексикология

Summary. The paper describes the history of the creation of Russian databases, their parameters, operating principles and the goals they were created for. The author emphasizes the fundamental differences between searching in databases and in the Internet and focuses on the peculiarities of information retrieval systems. The conclusion of the paper stresses the need of better knowledge of different computer instrument to ensure a proper choice of tool for various research goals.

Интерес к электронным библиотекам и компьютерным возможностям поиска и анализа исследуемых явлений возник в конце 1990-х. Сначала Интернет сулил перспективу объять необъятное – анализировать множество документов, напечатанных по-русски, быстро отыскивая в них необходимое. Скоро такое представление оказалось иллюзией: стали видны недостатки использования Интернета в научной работе: Интернет не гарантирует качество материала; Интернет-поисковые машины (далее ИПМ) не позволяют ограничить область поиска на этапе запроса и превращают поиск данных в трудоемкую процедуру их отбора из многостраничной выдачи; в Интернете отсутствуют полнотекстовые архивы средств массовой информации; ИПМ не имеют сложного языка запросов для точного поиска, не справляются с проблемой омонимии; выдача по запросу в Интернете содержит большое количество повторов одних и тех же документов.

По сравнению с Интернетом корпусы предлагают определенные средства поиска и гарантируют некий баланс текстов, который репрезентативен для современного русского языка. Однако большинство корпусов по сравнению с Интернетом имеют существенный недостаток – в них заложены научно целесообразные и тщательно проработанные принципы, но они реализованы пока на малом количестве текстов (Тюбингенский корпус – 25 млн. словоупотреблений; Хельсинкский аннотированный корпус русского языка – 100 000 словоупотреблений; Национальный корпус русского языка – 120 000 000 словоупотреблений). Самым крупным корпусом русского языка (если понимать корпус как совокупность текстов в электронном виде) является «Интегрум» – в его базах сейчас 350 миллионов документов. «Интегрум» сразу поставил перед собой задачу преодоления недостатков Интернета. «Интегрум» ищет в собственных качественных базах данных, позволяет ограничить поисковое пространство по нескольким параметрам. Как и Интернет, «Интегрум» дает возможность исследовать русскоязычные массивы информации, состоящие из сотен миллионов документов, однако, в отличие от Интернета, содержит полные архивы тысяч СМИ, большинство из которых начинается с середины 90-х годов XX века. «Интегрум» имеет многофункциональную информационно-поисковую систему «Артефакт», располагающую разветвленным языком запросов со сложным синтаксисом. Повторы документов в «Интегруме» незначительны. Кроме объединения достижений корпусной лингвистики и Интернета «Интегрум» реализовал и другую задачу – изобрел специальные инструменты для статистического анализа данных и «Частотный словарь русского языка».

Сравнительный анализ возможностей применения разных современных компьютерных ресурсов и баз данных можно продемонстрировать на примере исследований различный лингвистических феноменов. Среди них, исследование вторичных заимствований («К экстравагантным платьям выбираем духи *агрессивные* и тяжелые»; «*Амбициозность* великолепна!»; «Меня *приятно шокировало* то, что я увидела»; «*Шок*» как название шоколада).

Прежде всего, следует выяснить, являются ли подобные примеры единичными или частотными настолько, что позволяют говорить о семантическом сдвиге в русском языке. Для этого необходимо собрать картотеку примеров нового употребления исследуемых слов. Очевидно, что задача вычленения строго определенной группы примеров перед ИПМ поставлена быть не может. Почти все корпусы русского языка для решения такой задачи тоже непригодны, так как содержат малое количество текстов. «Интегрум» нашел 389 примеров употребления словосочетания «приятно шокировать» со всеми дериватами и 1383 примера употребления идиомы «шок – это по-нашему».

Помимо репрезентативного количества примеров для обоснования гипотезы о возникновении нового явления в язы-

ке следует статистически исследовать явление на временном отрезке, большем, чем пять лет, чтобы отличить проявления языковой моды от явления языка. Представляется, что сервисы «Сравнительной и относительной статистики» дают надежный метод различения, так как позволяют анализировать существование в языке того или иного слова на отрезке до пятнадцати лет.

Еще один источник информации о частотности – частотный словарь. На сегодняшний день исследователь располагает несколькими частотными словарями, среди которых «Частотный словарь русского языка» под редакцией Л. Н. Засориной и три электронных: «Машинный словарь» С. А. Шарова, «Частотный словарь Полистилевого корпуса русского языка», Частотный словарь «Интегрума». Частотный словарь «Интегрума» является самым представительным, так как сделан на материале 9 миллиардов слов, тогда как «Частотный словарь русского языка» под редакцией Л. Н. Засориной делался на материале миллиона слов, корпус «Машинного словаря» С. А. Шарова – около 40 миллиона слов.

Использование современных компьютерных ресурсов целесообразно при решении многих методических и научных задач. Однако каждый из ресурсов в современном виде предназначен для решения разных научных задач.

Интернет-ресурсы полезны изучающим русский язык студентам, которые привыкли к жизни в киберпространстве и располагают временем для длительных поисков не столько точной, сколько любопытной информации.

Применение корпусных ресурсов перспективно для разных научных направлений. Для лингвиста это изучение лексики современного русского языка в динамике развития: истории слов конца XX начала XXI веков, лексических изменений, изменений лексической семантики (семантических преобразований, сдвигов и колебаний), инноваций (их первого вхождения в русский язык и этимологии), влияния экстралингвистических факторов на лексику и лексическую семантику, проявлений языковой моды, нормативного аспекта языка. При помощи «Интегрума» решаются сложные лексикографические задачи: фиксация вхождения слова в русский язык и умение отличать явления языковой моды от явлений языка. Статистический сервис «Интегрума» вместе с его же Частотным словарем, «Машинным словарем» С. А. Шарова, «Частотным словарем Полистилевого корпуса русского языка», позволяют за секунды составить словник реального современного русского языка, хотя в традициях даже развитой западной лексикографии на это уходили годы.

Информационные параметры русского текста

Р. Г. Пиотровский

Российский государственный педагогический университет им. А. И. Герцена, Санкт-Петербург

1. Мало кто из молодых лингвистов слышал о знаменитом афоризме Н. С. Трубецкого: «Язык лежит вне “меры и числа”» [3, 12]. Языковеды старшего поколения, стараются не вспоминать об этой формуле: классикам ведь свойственно иногда ошибаться... Тем не менее применение информационно-статистической технологии в экспериментально-доказательной парадигме современной лингвистики занимает пока скромное место. Причины этого понятны: получение надежных информационно-статистических данных связано с крайне трудоемким (даже при использовании вычислительной техники) статистическим и психолингвистическим экспериментом.

2. Именно поэтому для получения основных информационных параметров русского письменного текста понадобились многие годы проведения и осмысления результатов психолингвистического эксперимента. В основе исследовательской процедуры лежала выдвинутая еще в середине 1950-х гг. К. Шенноном ([4, 669–668]) идея, смысл которой состоял в том, что по результатам «игры» с носителем (или носителями) языка в последовательное предсказание букв неизвестного ему (им) текста можно получить количественные оценки *статистической информации* (СтИ), содержащиеся в тексте и его отдельных фрагментах.

3. В международной научной группе «Статистика речи» (СтР) были разработаны и использованы две методики проведения эксперимента – индивидуальная, коллективная и соответственно две математических модели обработки получаемых результатов (см. [1, 143, 146–148]). С помощью этих процедур к настоящему времени обработано ок. 200 связных текстов длиной в среднем по 28 словоупотреблений каждый, взятых из художественной прозы, записей разговорной речи, научно-технических и публицистических текстов, а также около 800 отдельных словоформ (с / ф) и словосочетаний (с / с). В экспериментах по угадыванию участвовало ок. 4 тыс. тысяч испытуемых – носителей русского языка, в основном учащихся гуманитарных университетов в возрасте от 18 до 25 лет.

4. В итоге реализации указанных процедур получены измерения *синтактической информации* (СтИ) для различных участков текста, с / с и с / ф. В этих измерениях суммируются оценки комбинаторики единиц текста и различные виды содержащейся в нем *смысловой* (прагматической, семантической и коннотативной) *информации* (СмИ). Разумеется, наибольший интерес представляют здесь оценки СмИ, поэтому с конца 1990-х гг. шли интенсивные поиски приемов измерения этой последней. В настоящее время такая методика создана и получены первые результаты замеров смысловой информации, содержащейся в с / ф и с / с ([2, 105–116]).

5. Анализ указанных выше замеров русских текстов и их сопоставление с аналогичными измерениями в других индоевропейских и некоторых тюркских и финно-угорских языках приводит нас к следующим выводам:

1) избыточность (R) русского текста находится на том же уровне (от 70 до 96%) что и избыточность других языков; это определяется биосоциальной природой языка: такой уровень R служит защитой речевой коммуникации от физических и психолингвистических помех;

2) колебания избыточности связаны с изменением тематики, профессиональной и стилистической ориентации текста: разговорные и художественные тексты показывают низкий уровень избыточности, напротив, рост нормализованности текста в публицистической и научно-технической речи сопровождается возрастанием величины R, которая достигает 95–96% по ГОСТ’у в переговорах ‘земля-воздух’ и ‘земля-вода’;

3) некоторые отклонения от нормы избыточности в сторону ее уменьшения отмечено у больных, страдающих речемыслительными расстройствами;

4) русский текст дает квантовое распределение информации; это свидетельствует о том, что речь генерируется, воспринимается и перерабатывается нашей памятью не непрерывно, а путем ритмической отдачи накопленных квантов информации, в качестве которых выступают морфемы;

5) основная часть СтИ сосредоточена в начале русских с / ф, их концы и особенно середины несут немного информации, флексии иногда оказываются вообще избыточными;

6) русская с / ф несет больше смысловой информации, чем с / ф в аналитических языках (французском и болгарском), но заметно меньше, чем это имеет место в типично синтетических языках (например, в агглютинирующем эстонском);

7) количество СмИ, извлекаемой испытуемым из с / ф зависит от богатства его тезауруса и лингвистической компетенции, так студенты педагогических университетов извлекают из с / ф в среднем 11,4 двоичных единиц., в то время как учащиеся техникумов получают только 9,8 двоичных единиц.

Литература

1. *Пиотровский Р. Г.* Лингвистический автомат (в исследовании и непрерывном обучении). СПб., 1999.
2. *Пиотровский Р. Г.* Синергетика текста. Минск, 2005.
3. *Трубецкой Н. С.* Основы фонологии. М., 1960.
4. *Шеннон К.* Предсказание и энтропия печатного английского текста // Работы по теории информации и кибернетике / Пер. с англ. М., 1963. С. 669–668.

Фрагментация предложений корпуса параллельных текстов

С. Б. Потёмкин

Московский государственный университет им. М. В. Ломоносова

potemkin@philol.msu.ru

Summary. This paper presents novel approaches to phrase alignment for example-based machine translation. We use matching of delimiters instead of word matching while determining fragment borders. Than we construct the best critical path in the two-dimensional bilingual space using dynamic programming. We follow a monotonic machine translation approach, for which we develop an efficient and flexible partial reordering that allows introducing different reordering constraints. Than we merge fragments to avoid mistakes connected to inversion.

Правильная фрагментация параллельных двуязычных текстов представляет существенную и первую по очередности проблему для решения комплексной задачи машинного перевода на основе примеров (EBMT). Устойчивые фрагменты предложений, иногда называемые фразеологизмами, могут встречаться в параллельных текстах достаточно часто, уступая только пословным соответствиям. В отличие от известных методик, в качестве элементов матрицы смежности рассматриваются не отдельные сопоставленные слова, а интервалы от одного пробела в предложении до следующего. Это дает возможность сопоставлять словарные словосочетания из одного предложения – слову или словосочетанию из другого. Затем в исходном предложении (SS) выявляются и обрабатываются фрагменты, которые могут быть инверсными относительно предложения-перевода (TS), затем методом динамического программирования выполняется поиск наилучшей фрагментации. Полученные фрагменты анализируются относительно ограничений по относительной длине и по вхождению переводов всех слов SS в TS и наоборот.

В качестве координатных отсчетов пространства билингвы будем принимать не слова как таковые, а разделители (пробелы) между соседними словами ([1]). При таком подходе отображение слова исходного предложения на слово целевого предложения представляет собой отрезок с координатами начала и конца слова SS по x и начала и конец слова TS по y. Теперь возможно ставить в соответствие не только однословные эквиваленты, но также эквиваленты типа словосочетаний. На рис. 1 показано отображение SS на TS, выполненное с учетом встретившегося в словаре словосочетания (*вдруг = all at once*) Коллизия возникает, когда некоторые отображающие отрезки перекрываются по горизонтали или по вертикали (т. е. отображение не однозначное). Например, слово исходного текста *к* отображается на слова целевого текста *at, to, with.* Чаще всего в коллизии участвуют служебные слова, а также знаки препинания ([2]).

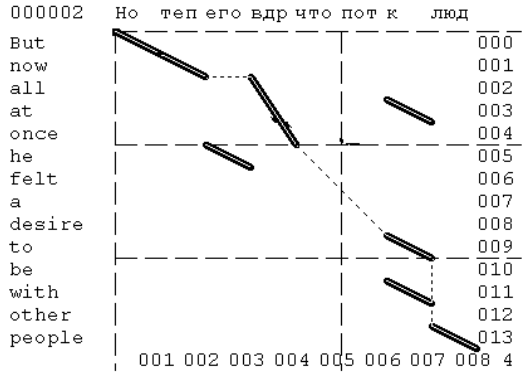
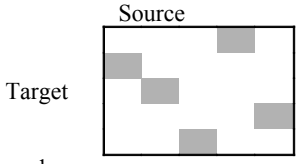


Рис. 1. Отображение SS *Но теперь его вдруг что-то потянуло к людям* на TS *But now all at once he felt a desire to be with other people* (Ф. М. Достоевский, перевод Dostoevsky, Fyodor, Crime and Punishment, Translated by Garnett, Constance Black, Publisher: Dover Pubns, Mineola, New York, USA). Пунктиром показан критический путь.

Построив пословное отображение, можно переходить непосредственно к фрагментации, то есть к отображению интервалов SS на интервалы TS, которые лежат между уже построенными отображающими отрезками. Слова двух предложений можно сопоставлять в разной последовательности. Одно и то же предложение может быть переведено как с прямым, так и с обратным порядком слов и оба перевода будут правильными. Более общий случай – когда одни группы слов переведены в прямом направлении, другие – в инверсном и сами эти группы сопоставляются хаотически.



Число вариантов фрагментации можно оценить как O(n!), где n-число слов SS или TS следовательно, перебор вариантов для сколько-нибудь длинного предложения практически невозможен. Однако в случае, если мы рассматриваем только монотонные отображения (т. е. считаем порядок слов исходного и целевого предложения по большей части совпадающим), задача попадает в класс задач динамического программирования. Действительно, последовательность отображающих отрезков фрагментации можно считать путем из точки 0 в точку (m, n), где m и n длина исходного и целевого предложения соответственно. Тогда путь наибольшего веса будет соответствовать наилучшей фрагментации.

Как правило, исходное предложение и его перевод, даже имеющие в целом совпадающий порядок слов, содержат фрагменты с инверсией, напр. {*изредка только; only occasionally*}. Такую частичную инверсию желательно включить в критический путь, но приведенный алгоритм этого не допускает. Предлагается формировать фиктивное отображение для инверсных фрагментов заданной длины (напр., не больше 3 слов), которые включаются в общий набор отображающих отрезков и участвуют в алгоритме поиска критического пути.

Вернемся теперь к предложению рис. 1.

Критический путь разбивает исходную пару предложений на следующие фрагменты:

1. *Но теперь = But now*
2. *теперь его вдруг = now all at once*
3. *вдруг что-то потянуло к = all at once he felt a desire to*
4. *к людям = to be with other people*

Фрагмент 1 не вызывает возражений. Фрагмент 2 SS содержит местоимение *его*, которого нет во фрагменте TS. Наоборот, фрагмент 3 TS содержит местоимение *he*, которое отсутствует во фрагменте SS. Если мы объединим фрагменты 2 и 3, результат будет более осмысленным. При решении вопроса об объединении фрагментов мы придерживаемся двух критериев:

а) отношение длин фрагментов SS и TS не должно сильно отличаться от 1 и

б) переводы всех слов, содержащихся в SS должны содержаться в TS и наоборот.

Фрагмент 4 является конечным или хвостовым, и уже не может быть объединен со следующим и его оценка по любому из критериев а) или б) не производится. Поэтому доверие к хвостовому фрагменту заведомо ниже, чем к начальным.

Представленный выше алгоритм опробован на тексте 1-й части романа Ф. М. Достоевского. Для каждого предложения построен путь с наибольшим весом и сопоставлены фрагменты исходного и целевого предложения. Выделены слова SS, не нашедшие перевода в TS. Эксперименты проводились также на корпусе параллельных текстов юридического содержания, на котором метод дает гораздо лучшие результаты. В развитие данного подхода будет составлен автоматический словарь фрагментов для использования в системе автоматического перевода, основанного на примерах (EBMT) ([3]).

Литература

1. *Потемкин С. Б., Кедрова Г. Е.* Автоматическая оценка качества машинного перевода на основе семантической метрики // Вестник Луганского НПГУ им. Т. Шевченко. 2005. 15(95). С. 35–41.
2. *Потемкин С. Б., Кедрова Г. Е.* Семантическое разделение омонимов с использованием двуязычного словаря и словаря синонимов // II Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и современность»: Труды и материалы. М., 2004.
3. *Ralf D. Brown.* Example-Based Machine Translation in the Pangloss System // Proceedings of the Sixteenth International Conference on Computational Linguistics (COLING-96). Vol 1. Copenhagen, 1996. P. 169–174.

Использование корпуса параллельных текстов для пополнения специализированного двуязычного словаря

С. Б. Потёмкин, Г. Е. Кедрова

Московский государственный университет им. М. В. Ломоносова

potemkin@philol.msu.ru, kedr@philol.msu.ru

Фрагментация, перевод, эквиваленты, законодательство, кодекс

Summary. Special dictionaries are necessary for perfect translation in any particular subject area. Amendment and extension of such dictionary is performed on the base of parallel text corpora comprising a number of RF Codices with English translations. Extraction of potential equivalents involves fragmentation technique based on a large Russian-English dictionary, Dictionary of Russian paradigms and extended English-English Dictionary of Synonyms. The critical path is searched in two-dimensional lexical space using dynamic programming. Exemplary equivalents are discussed.

В процессе перевода юридических текстов возникает необходимость использования тех эквивалентов и тех языковых конструкций, которые уже приобрели характер «узаконенных» благодаря их фактическому использованию (узусу) в официальных переводах нормативных документов. Такие стандартизованные переводы содержатся в Корпусе параллельных текстов, который включает ряд законодательных актов РФ, в том числе основные кодексы – Гражданский, Уголовный, Уголовно-процессуальный, Налоговый, Таможенный, Жилищный и др. с их переводами на английский язык, сделанными сотрудниками Юридической Академии и выходивших в специализированных издательствах. Для решения задачи выбора эквивалентов необходимо прежде всего произвести выравнивание параллельных текстов, состоящее в сопоставлении крупных фрагментов текстов в масштабах статей, параграфов и предложений. Далее решается более сложная задача – сопоставление отдельных фрагментов предложений. В идеальном случае каждое предложение разбивается на двухсловные фрагменты: (ra rb) ~ (ea eb), где ra ~ ea, rb ~ eb, т. е. пары русско–английских эквивалентов, зафиксированных в общих или специализированных словарях. Очевидно, что реальное положение вещей здесь может оказаться очень далеким от идеала. Задача осложнена:

- различным порядком слов в парах предложений;
- наличием в каждом из предложений элементов, нормально не имеющих эквивалентов в его переводе (артикли, некоторые предлоги, знаки препинания и др.);
- наличием в каждом предложении слов, имеющих несколько потенциальных эквивалентов в его переводе;
- наличием словосочетаний, переводимых одним словом или другим словосочетанием при отсутствии эквивалентов для каждого из членов словосочетания;
- наличием в каждом из предложений непереведенных слов.

Последний случай встречается достаточно часто и может быть обусловлен, в первую очередь, проявление неполноты используемых для перевода двуязычных словарей. При этом, как правило, каждое из непереведенных слов зафиксировано в словарях вместе с его эквивалентами, но сами эти эквиваленты отсутствуют в переводе данного предложения. Рассмотрим способы отыскания возможного эквивалента для данного непереведенного русского слова R. Первоначально можно предлагать в качестве кандидатов для такого непереведенного слова перечислить все непереведенные слова Ei из английского предложения. Получаем $n \times m$ вариантов перевода, где n и m – число непереведенных слов в каждом из предложений. Очевидно, что в достаточно длинных предложениях это число вариантов оказывается большим и не позволяет вручную анализировать каждый из предполагаемых эквивалентов. Для ограничения числа кандидатов предлагается использовать расширенный словарь синонимов ([1]). Ограничим рассматриваемые кандидаты (Rj ~ Ei) теми, для которых словарь синонимов содержит пару (Ej ~ Ek) и имеется зафиксированная в словаре пара (Rj ~ Ek). Эксперименты показывают, что в этом случае число кандидатов эквивалента снижается на порядок, т. е. пропорционально SQRT ($n \times m$).

Дальнейшее ограничение вариантов перевода достигается в результате фрагментации параллельных предложений. Фрагментация выполняется по следующей схеме ([2]):

- Лемматизация каждого слова каждого двух предложений.
- Поиск по словарям всех пар однословных эквивалентов для данной пары предложений. Поиск производится для исходной словоформы и для всех ее возможных лемм.
- Поиск по словарям словосочетаний в одном предложении и его эквивалентов в другом предложении.
- Определение весов пар эквивалентов по лексической базе данных с наложенной семантической метрикой ([3]).
- Построение матрицы смежности для двудольного графа, вершинами которого являются слова каждого из предложений.
- Поиск критического пути из начала предложения – точка (0, 0) в конец предложения – точка (N, M), где N, M – число слов в каждом из предложений. При построении критического пути применен метод динамического программирования, учитывающий веса пар эквивалентов и расстояния между ними (число слов, их разделяющих).
- Фрагмент определяется как подстрока каждого предложения, ограниченная точками на критическом пути.
- Расширение фрагментов в результате: наложения ограничений на относительную длину русского / английского фрагмента; вхождение в английский фрагмент переводов всех слов, принадлежащих русскому фрагменту и наоборот; включения словосочетаний в каждый из фрагментов целиком.

После построения фрагментов для дальнейшего рассмотрения допускаются только те кандидаты в эквиваленты, которые принадлежат фрагментам, соответствующим друг другу. В результате число кандидатов на роль англо-русских эквивалентов сокращается, согласно эксперименту, до 1–2 в каждом предложении.

Выявленные кандидаты выделяются внутри соответствующих предложений и представляются для экспертной оценки профессиональному переводчику.

В докладе приводятся примеры из файла, полученного при обработке Жилищного кодекса РФ с кандидатами эквивалентов, выделенными жирным шрифтом и дается их предварительный анализ.

Литература

1. *Потемкин С. Б., Кедрова Г. Е.* Homograph disambiguation with the use of bilingual dictionary and dictionary of synonyms // 32-я Конференция Европейских лингвистов. Лион (Франция), 02–07.09.2003.
2. *Потемкин С. Б., Кедрова Г. Е.* Автоматическая оценка качества машинного перевода на основе семантической метрики // Труды II Международной научно-практической конференции, посвященной Европейскому Дню языков. Луганск, 26–28 сентября 2005.
3. *Потемкин С. Б.* Лексическая база данных с наложенной семантической метрикой // II Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и современность»: Труды и материалы. М., 2004.

Корпус текстов и речевая деятельность – проблемы подобия

В. В. Рыков

Московский физико-технический институт

Корпус текстов, речевая деятельность, подобие

Summary. The corpus of texts by its definition should reflect in its structure and in its word stuff a complex semiotic object – speech activity. Only then, investigating it this or that way, for any purposes the scientist can be sure that the received results are applicable and are fair for speech activity as a whole.

Корпус текстов (КТ) по своему определению должен отразить в составе своего речевого материала сложный семиотический объект – речевую деятельность человека. Только тогда, исследуя тем или иным способом, для тех или иных целей КТ можно быть уверенным, что полученные результаты применимы и справедливы для речевой деятельности в целом.

Следовательно, КТ должен представлять собой сложно спроектированную систему, имеющую в своем составе набор специально отобранных текстов, имеющую сложную структуру и другие системные свойства, которые должны быть достаточно четко определены. Одной из отличительных особенностей КТ может служить одно из его системных свойств – доминанта. Доминанта представляет собой характерное функциональное свойство системы, для реализации которой адаптируются все остальные ее элементы, их взаимосвязи и свойства.

Итак, доминантой КТ изначально декларируется отражение в составе своих текстов всего (или достаточно большой части) разнообразия речевой деятельности общества. Этой системной доминанте должны быть подчинены все другие свойства КТ – его структура, функции, состав. Отдельные части КТ (входящие в его состав тексты) и связи между ними (их структура) должны быть подчинены, согласованы или поставлены в соответствие с доминантой всей системы в целом – КТ.

Такое понятие доминанты дает нам возможность сопоставлять такие разнородные характеристики КТ, состав входящих в него текстов и жанров, их свойства, связи и отношения. Важно также то, что исходными объектами, их свойствами, связями и отношениями из которых конструируется КТ, являются тексты, их жанры, связи между ними – то есть сама речевая деятельность (РД). КТ, как сконструированная из них система, должен быть максимально адаптирован к РД, которую он призван отразить. При достижении этого свойства произойдет переход из количества (множества текстов на машинном носителе) в качество – сложно организованную систему, которой является КТ.

Существование такой системной доминанты неизбежно влечет за собой выводы о свойствах текстов и составе жанров, представленных в этом КТ. Они действительно должны отражать разнообразие речевой деятельности, происходящей ежесекундно в обществе. Это разнообразие может не соответствовать требованиям изящной словесности и даже элементарной культуры речи. В устной речи и особенно сейчас в Интернете, общении по мобильным телефонам существуют жанры или функциональные стили, которые не вполне соответствуют этим требованиям. Но они существуют, следовательно должны быть включены в КТ. При составлении корпуса следует иметь в виду, что в соответствии со вкусом составителя формируются не КТ, а такие собрания текстов на машинных носителях как электронные библиотеки. Если же составить КТ (отобрать в него тексты, задать структуру отношений между ними) согласно вкусам составителей, то научные результаты, полученные на речевом материале такого корпуса, будут отражать соответственно вкусовые предпочтения его составителей.

Нетрудно видеть, что неуклонное следование описанным выше принципам в отношении к речевому материалу задает новую парадигму отношения к речи, где в каком-то смысле равные права имеют самые разные виды речи – устной, письменной и на машинных носителях.

Сформированный таким образом корпус представляет собой (или должен представлять) словесное и системное единство. Это единство отражает в себе состав речевого материала, при помощи которого реально осуществлялась коммуникация в устном, печатном или в другом роде словесности. Печатный или любой другой текст только тогда может быть представлен в КТ, если он был прочитан массовым читателем. Это должно быть выявлено и подтверждено в соответствующих исследованиях. Например, в Японии уже много лет ежегодно проводятся исследования состава текстов, используемых для общения японцами из различных социальных слоев в реальной повседневной жизни в рамках национальной программы «языкового существования» (*генгу сейкацу*).

Только тогда мы сможем наблюдать явление, давно известное в философии и системном анализе – переходе количество в качество. Правильно (в описанном выше смысле) составленный КТ будет своеобразным зеркалом, отражающим реальную повседневную речевую деятельность.

Критерием этого, как в оптическом зеркале, так и в радиотелескопе должна быть взаимозаменяемость и даже допустимая потеря отдельных частей без ущерба для выполнения основной функции (отражения). Нельзя ничем заменить Пушкина или Диккенса (которые и должны входить в электронную библиотеку, а не в корпус). Но можно легко и без ущерба заменить или выбросить любой текст из правильно составленного корпуса. Для того, чтобы отразить РД в КТ, нужно прежде всего создать простейшую модель – т. е. представить себе онтологию РД. Это значит, что нужно описать:

1) какие объекты (в нашем случае тексты) входят в ее состав;

2) какие их свойства (прежде всего интересующие нас свойства) должны быть отражены в КТ;

3) какие связи между этими объектами.

Тогда мы рассматриваем РД с целью отразить в КТ самые общие, онтологические ее свойства и нас прежде всего интересует вопрос – какие ее свойства должны быть в нем отражены. Действительно, теперь, после того, как были рассмотрены проблемы соотношения корпуса как системы с внешней средой, которая должна быть отражена, то есть речью, следует перейти к некоторым внутренним свойствам КТ как системы – прежде всего его структуре и составу. Структура КТ представляет собой множество определенным образом связанных между собой групп текстов. Каждая группа текстов корпуса отражает некоторую часть РД. Структурные связи между группами текстов в корпусе также отражают (по замыслу его составителей) структуру РД.

В разных корпусах и в разных лингвистических традициях эти группы могут называться по-разному. Например, самая элементарная группа текстов в Брауновском корпусе (БК) называется жанром. Составляющие эту группу или жанр тексты должны отразить в своем речевом материале все речевые свойства соответствующего фрагмента РД. И это должно быть четко определено словесной формулой. Для БК это определение звучит так – прозаическая печатная речь США, состоящая из текстов, впервые напечатанных в 1961 году, авторы которых родились в США.

Обычно принято делить РД по признаку общности условий коммуникации. Принято вместе группировать репортажи и отдельно передовые статьи в газетных текстах. Их вполне могут читать разные группы читателей, не говоря уже о газетной и литературной традиции деления таких текстов на жанры. Существенно то, что выбраны должны быть самые массово представленные жанры, прочитанные «широкой читательской массой», а не шедевры, отмеченные вниманием литературоведов.

Внутрикомплексные и межкомплексные квантитативные характеристики классификационных категорий

А. Г. Сильницкий

Смоленский государственный университет

Кластерная классификация, квантитативный аспект, разноуровневые признаки

Summary. The theses are dedicated to a quantitative clusteral classification of verbs expressing various types of economic relations ("economic" verbs) on the basis of their multilevel semantic features. Two clusters of semantic features are differentiated. The quantitative aspect of this type of classification consists in defining the degrees of integration of various clusters, distances between clusters, "nuclear", "peripheral" and differentially marked features.

Одной из фундаментальных проблем лингвистики (как и любой индуктивной науки) является разработка методологии классификация исследуемого материала. Различаются *качественный* аспект классификации, состоящий в распределении рассматриваемых элементов по группам, подгруппам и т. д., и ее *количественный* аспект, определяющий в квантитативных терминах степени интегрированности выделенных комплексов различных иерархических уровней и степени их удаленности друг от друга в общей классификационной схеме, т. е. расстояния между ними.

Таблица 1. Семантические признаки «экономических» глаголов.

ПРИЗНАКОВЫЙ ПАРАМЕТР	ПРИЗНАК		ПРИМЕР
ЭВЕНТУАЛЬНЫЙ – количество ситуаций и темпорально-мотивационные отношения между ними	*РЕТРОСПЕКТИВНЫЙ* ***(РСП)***	Импликация действия в прошлом	*штрафовать*
	ПРОСПЕКТИВНЫЙ ***(ПСП)***	Импликация действия в будущем	*подкупать*
	СИНХРОННЫЙ ***(СИН)***	Отсутствие импликации действия в прошлом или будущем	*платить*
ИЕРАРХИЧЕСКИЙ – преобладающая роль субъекта / объекта в реализации действия	*ДОМИНАНТНЫЙ* ***(ДОМ)***	Определяющая роль субъекта при реализации действия	*нанимать*
	СУБОРДИНАТИВНЫЙ ***(СУБ)***	Определяющая роль объекта при реализации действия	*наследовать*
ХРОНОСТРУКТУРНЫЙ – количество, функциональный тип и соотношение состояний в составе значения	*КАУЗАТИВНЫЙ* ***(КАУ)***	Результативное воздействие субъекта на объект	*покупать*
	ОПЕРАЦИОННЫЙ ***(ОПР)***	Нерезультативное воздействие субъекта на объект	*торговаться*
ИНТЕРАКТИВНЫЙ – согласованность / рассогласованность воли субъекта и воли объекта при реализации действия	*КОНВЕРГЕНТНЫЙ* ***(КНВ)***	Совпадение интересов субъекта и объекта при реализации действия	*дарить*
	ДИВЕРГЕНТНЫЙ ***(ДИВ)***	Несовпадение интересов субъекта и объекта при реализации действия	*конфисковать*

Методика квантитативного аспекта классификации иллюстрируется на материале разноуровневых семантических признаков «экономических» глаголов, т. е. глаголов, обозначающих различные типы экономических процессов, действий и отношений. Каждый глагол характеризуется по бинарному принципу наличием или отсутствием 9 семантических признаков «эвентуального», «иерархического», «хроноструктурного» и «интерактивного» параметров (таблица 1).

Соответствующая база данных подвергается компьютерной обработке (программа Statistica for Windows, Release 4.3) методом кластерного анализа по критерию Уорда (Ward). Различные опции данного метода позволяют не только отнести каждый объект / признак к определенному кластеру, но также выявить количественные соотношения между комплексами.

Рассматриваемые семантические признаки «экономических» глаголов сгруппировались в два противопоставленных друг другу кластера, К1 и К2 (см. табл. 2).

Количественный аспект кластерного описания основан на учете *бинарных расстояний* между любыми двумя признаками, определяемых посредством опции Euclidean Distances. Расстояние между двумя признаками обратно пропорционально степени их связанности между собой.

Бинарные расстояния позволяют установить различные типы *внутрикомплексных* и *межкомплексных* расстояний.

Таблица 2. Внутрикомплексные бинарные расстояния между признаками.

К1						*К2*						
	ПСП	**КАУ**	**КНВ**	**ДОМ**	СВР		**СИН**	**РСП**	**ОПР**	**ДИВ**	**СУБ**	СВР
ПСП	.0	15.8	**14.1**	16.4	15.4	**СИН**	.0	17.4	**12.6**	**13.5**	13.9	14.4
КАУ	15.8	.0	**13.2**	**12.6**	**13.9**	**РСП**	17.4	.0	14.8	**11.7**	14.0	14.5
КНВ	**14.1**	**13.2**	.0	16.7	**14.7**	**ОПР**	**12.6**	14.8	.0	**13.1**	**12.5**	**13.3**
ДОМ	16.4	**12.6**	16.7	.0	15.2	**ДИВ**	**13.5**	**11.7**	**13.1**	.0	**12.1**	**12.6**
						СУБ	13.9	14.0	**12.5**	**12.1**	.0	**13.1**

Среднее внутрикомплексное расстояние (СВР) каждого отдельного признака, исчисляемое как арифметическое среднее его бинарных расстояний со всеми остальными признаками данного комплекса, отображает степень его «ядерности» (центральности / периферийности) в иерархической структуре данного комплекса; степень ядерности признака обратно пропорциональна величине его СВР. Признаки, СВР которых ниже ОВР соответствующего комплекса, являются *ядерными* (выделены жирным шрифтом), признаки с противоположной квантитативной характеристикой – *периферийными*. Так, в составе кластера 1 ядерными признаками являются КАУ и КНВ, в составе кластера 2 – ОПР, ДИВ, СУБ. Следовательно, в обоих кластерах ядерными являются признаки, относящиеся к *хроноструктурному* и *интерактивному* семантическим параметрам.

Обобщенное внутрикомплексное расстояние (ОВР) – арифметическое среднее всех бинарных расстояний признаков данного комплекса между собой – отображает степень внутренней интегрированности данного комплекса, обратно пропорциональной величине ОВР. Кластер 1 (4 признака, ОВР = 13,6) является несколько более интегрированным, чем кластер 2 (5 признаков, ОВР = 14.8).

Среднее межкомплексное расстояние (СМР) признака – арифметическое среднее бинарных расстояний отдельного признака одного комплекса от всех признаков другого комплекса – отображает обобщенную степень удаленности данного признака от другого комплекса в целом.

Обобщенное межкомплексное расстояние (ОМР) между двумя комплексами определяется как арифметическое среднее бинарных расстояний всех признаков одного комплекса от всех признаков другого и, таким образом, отображает обобщенную степень противопоставленности этих двух комплексов друг другу, прямо пропорциональную величине ОМР.

Таблица 3. Межкомплексные расстояния между признаками К1 и К2.

	СИН	РСП	ОПР	ДИВ	СУБ	*СМР*
ПСП	**19.0**	**18.9**	15.7	16.8	15.0	*17.1*
КАУ	**18.7**	17.1	**22.2**	**18.4**	**19.0**	***19.1***
КНВ	17.5	15.7	**18.2**	**18.2**	15.5	*17.0*
ДОМ	17.5	17.4	**18.5**	**18.1**	**21.8**	***18.7***
СМР	***18.2***	*17.3*	***18.7***	*17.9*	*17.8*	***18.0***

ОМР между кластерами 1 и 2 равно 18.0.

Выделенные признаки, показатели СМР у которых превышают 18.0, являются *дифференциально маркированными*, т. е. играют основную роль в противопоставлении сопоставляемых двух кластеров. В данном случае такую роль выполняют признаки КАУ и ДОМ в составе К1, ОПР и СИН в составе К2, т. е. признаки *хроноструктурного*, *иерархического* и *эвентуального* семантических параметров.

К вопросу о систематизации в области русской орфографии в прикладных целях

В. П. Синячкин, Н. А. Красс, М. А. Брагина

Российский университет дружбы народов, Москва

Орфография, методика преподавания русского языка как родного, функциональная грамотность, систематизация

Summary. The thesis is devoted to the analysis of the level of student's knowledge graduated from the school in 2006 year and their results after testing at the university.

В 2005 году Постановлением Правительства РФ № 833 от 29.12 была утверждена Федеральная целевая программа «Русский язык (2006–2010 годы)». Разработчики Программы, характеризуя современное состояние уровня грамотности, указывают на то, что в последнее время наблюдается снижение уровня владения русским языком как родным, особенно среди представителей молодого поколения, искажение литературных норм и культуры речи в среде политических деятелей, государственных служащих, работников культуры, радио, телевидения и в конечном итоге снижение профессионального уровня работников различных сфер.

Подтверждением этому могут служить результаты диагностического тестирования по основам орфографии, проведенного среди студентов различных специальностей подготовительного факультета ИИЯ РУДН.

Среднее количество ошибок в группе от 10 до 20 человек в тесте из 176 слов по основным правилам (правописание О / Ё после шипящих; ЦЫ / ЦИ; Ъ и Ь; Ы / И после приставок; Н / НН в различных частях речи; окончания глаголов; суффиксы наречий; слитные, дефисные, раздельные написания, в том числе с пол- / полу-) колеблется от 30 до 40.

Наибольшее количество ошибок допустили студенты экономических, юридических, медицинских специальностей, а также будущие специалисты в области международных отношений, бизнеса и туризма.

Как показывает статистика, по-прежнему вызывают трудности написание слов с О / Е после шипящих, Н / НН в различных частях речи, слитные, дефисные, раздельные написания, окончания глаголов. Так, среди 24 слов на правописание О / Ё после шипящих допускается от 1 до 7 ошибок студентами-международниками, от 1 до 14 студентами-экономистами, от 1 до 11 студентами-юристами. При написании 41 слова с Н / НН международники допускают от 4 до 20 ошибок, экономисты – от 3 до 22, юристы – от 1 до 23. Из 68 слов на случаи слитного и дефисного написания неправильно пишется до 21 слова у международников, до 40 – у экономистов, до 37 – у юристов.

Большинство слов, написание которых регулируется данными правилами, составит в будущем лексическую основу, активно используемую в письменной речи специалистов в области экономики, юриспруденции, международных отношений, т. е. тех областей знаний, где достаточно много терминов, сложных слов, глаголов и глагольных форм.

Таким образом, налицо проблема снижения общего уровня грамотности, слабого ориентирования в системе русского правописания.

Причинами данного явления считаем следующие: 1) отсутствие систематического подхода при изучении норм орфографии в школьной программе и учебниках; 2) сложность, запутанность, немотивированность объяснения большинства правил; 3) нарушение принципов частотности и типичности при изучении орфографических тем; 4) недостаточное использование современных технологий, в том числе информационных, в практике преподавания русского языка в школе.

На наш взгляд, необходимо начать работу по систематизации в области методики преподавания русской орфографии и пунктуации, ориентированную на функциональную грамотность, с целью повышения уровня грамотности школьников, студентов, популяризации русского языка и повышения интереса к нему, а также для формирования и развития языкового мышления учащихся.

Документная лингвистика как наука о деловой письменной коммуникации

О. П. Сологуб

Новосибирский государственный технический университет

Деловая коммуникация, документная лингвистика, документный дискурс, официалема

Summary. The objective of this work is to substantiate a new theoretical and applied discipline called document linguistics.

Деловая речь в современных условиях стала занимать все более значимое положение в коммуникативной практике общества и отдельного человека, что связано с расширением и углублением деловых контактов как внутри страны, так и на международном уровне, с внедрением новых форм деловой коммуникации, с актуализацией вопросов эффективного речевого поведения в деловой сфере.

Признание данной сферы деятельности нашло свое внешнее выражение. Во-первых, появилось огромное количество пособий, справочников, журналов, содержащих, помимо публикаций, посвященных специальным проблемам работы с документами, материалы о специфике письменной и устной деловой коммуникации, практические рекомендации по оформлению документов (традиционных и электронных, на русском и иностранных языках). Во-вторых, во многих вузах страны открывается специальность «Документоведение и документационное обеспечение управления»; на разных факультетах осуществляется обучение навыкам делового общения; проводятся многочисленные семинары и тренинги. В-третьих, появилась обширная литература методического характера, в которой предлагаются материалы по организации обучения специалистов по работе с документами – учебные планы, программы учебных курсов; обсуждаются методики проведения занятий, посвященных вопросам работы с документами. Указанные факторы *экстралингвистического характера* требует теоретического осмысления, что во многом обусловило возросший интерес лингвистов к сфере деловой коммуникации.

С другой стороны, интерес к проблемам деловой коммуникации обусловлен и *собственно лингвистическими* фак-

торами, а именно, особенностями современной парадигмы, предполагающей рассмотрение языка в его отнесенности к человеку, в контексте его частной и социальной деятельности (в качестве одной из ее составляющих является и производственная, предпринимательская деятельность). Функциональный и антропоцентристский подходы к языку тесным образом связаны с широким выходом лингвистики в смежные отрасли знания. Данная тенденция обнаружила свое влияние и в сфере изучения деловой речи, в которой особо тесная связь установилась с такой дисциплиной, как *документоведение*. Эта связь возникла и развивается на основе общих задач дисциплин – путем выявления сущностных свойств документа выработать пути эффективной деловой письменной коммуникации. С одной стороны, лингвисты опираются на результаты деятельности документоведов, разрабатывающих на основе принципа целесообразности и эффективности формы документов. С другой стороны, положения документоведов, содержащие рекомендации по языковому оформлению документов, опираются на лингвистические данные.

Результатом расширения подходов при изучении деловой речи, ее многостороннего исследования, с одной стороны, и специализации лингвистики в данной сфере, сохранения интереса к рассматриваемому функциональному слою языка (особенно в его письменном варианте), а также объединения усилий двух отраслей знания – лингвистики и документоведения, с другой стороны, явилось становление новой «сдвоенной» дисциплины – *документной лингвистики*. Считаем, что говорить о новой науке возможно только сейчас, хотя сам термин «документная лингвистика» был введен в научный оборот еще в 70-х годах XX века (он был предложен проф. К. Г. Митяевым в связи с выходом в свет ряда пособий по нормализации языка «канцелярской письменности»).

Стимулом для формирования документной лингвистики послужил заказ от госаппарата на проведение практических работ по совершенствованию управления и дальнейшей унификации и стандартизации документов в связи с автоматизацией процессов документооборота, что потребовало углубленного изучения документа, приведения его к единообразию. Результатом такой работы должно было стать создание Единой государственной системы делопроизводства. Для выполнения поставленных задач были необходимы специально подготовленные кадры, и в 1964 г. в Московском государственном историко-архивном институте открывается факультет документоведения (в учебный план по подготовке документоведов включается дисциплина «Документная лингвистика»), а в 1965 г. – Всесоюзный научно-исследовательский институт документоведения и архивного дела, в структуру которого входит и сектор документной лингвистики.

Таким образом, документная лингвистика, призванная удовлетворить определенный социальный заказ, возникла и развивалась как дисциплина *прикладного* характера.

Серьезное теоретическое осмысление деловой речи осуществляется в рамках функциональной стилистики; новый импульс эта традиция получает в период неофункционализма: внимание исследователей привлекают особенности коммуникации в различных ситуациях делового общения, производится анализ речевого поведения деловых коммуникантов, активно исследуются деловые речевые жанры, описывается языковая личность государственного служащего, чиновника, создаются их речевые портреты и т. д. Все это свидетельствует о становлении и развитии документной лингвистики и как *теоретической* дисциплины.

Основными ее задачами в настоящий момент следует признать анализ специфики деловой письменной коммуникации, создание теории делового текста, анализ процессов официализации / деофициализации, позволяющих выявить сущностные аспекты механизма генезиса и функционирования деловой письменной речи и др.

Объектом документной лингвистики в рамках современной парадигмы следует признать *документный дискурс* – корпус текстов документов, включенных в процесс делового общения, управленческой деятельности.

В целях конкретного исследования процессов официализации и результатов их реализации в текстах необходимо ввести операциональную единицу языкового анализа, являющуюся показателем степени официальности текста. Такой единицей предлагаем считать *официалему*, которую определяем как типовую языковую единицу, объединяющую в себе, как правило, комплекс признаков, коррелирующих с нормами официальной деловой речи и свидетельствующих тем самым об официальном характере данной языковой единицы. Официалемами могут признаваться единицы разных уровней – от низших (лексика, морфемика) к более высоким уровням (генристика, прагматика) с выходом в лингвоперсонологию – к типам языковых личностей.

Термины латинского происхождения в гуманитарных науках: особенности компьютерного терминографирования

Р. А. Хасанова

Казанский государственный университет

Термины латинского происхождения, компьютерный терминологический фонд, терминографированные параметры

Summary. In work features of computer representation of terms of the Latin origin concerning the humanities are considered. The basic attention is given to revealing of set of the parameters describing borrowed terms and also to reflection of the selected parameters in computer terminological stock.

На современном этапе одной из составляющих лингвистических информационных ресурсов являются компьютерные терминологические фонды, представляющие собой некоторое количество терминологических словарей, которые включают в свой состав слова и словосочетания, обозначающие понятия специальных сфер научного знания. Банки данных подобного типа способны обеспечить информационную поддержку процесса обучения и последующей профессиональной деятельности специалистов.

Как известно, в научной терминологии, в том числе и терминологии дисциплин гуманитарного цикла, большой удельный вес имеют термины латинского происхождения. Компьютерная форма реализации материала позволяет не только представить всю совокупность терминов-латинизмов, функционирующих в рамках терминосистем некоторых гуманитарных наук (лингвистики, культурологии, психологии, социологии и др.), но также отразить их многоаспектную характеристику, в разном объеме и с разной степенью глубины зафиксированную в обычных словарях (терминологических, этимологических, энциклопедических, толковых), учебниках и научных трудах.

Традиционные терминографические источники обычно приводят этимологическую информацию, включающую толкование термина, название языка-источника (редко – языка-посредника), слово-прототип с указанием его значения. Однако даже эти сведения приводятся не всегда последовательно. Для создаваемого компьютерного терминологического фонда отобран широкий круг параметров, характеризующих латинские заимствования в этимолого-хронологическом и функциональном аспектах. В ходе терминографирования появляется возможность сравнения того, как одни и те же данные фиксируются в разных источниках. При этом выявляются некоторые различия в подаче материала и даже отдельные ошибки. Следует отметить, что вся параметрическая информация, включенная в компьютерный терминологический фонд, дается со ссылками на традиционные источники. Случаи расхождения данных и приведения ошибочной информации отмечаются специальными комментирующими текстами.

Для удобства пользовательской работы все терминографированные параметры распределены по информационным уровням.

Первый уровень содержит базовые параметры этимологического характера: данные о языке-источнике, языке-посреднике и слове-прототипе с указанием заимствованного значения. Здесь же приводится толкование терминологических единиц. В процессе компьютерного описания терминов выясняется, что из перечисленных параметров больше всего различий и неточностей в информации, приводимой «бумажными» источниками, имеет слово-прототип. Анализ латинского материала позволил в ряде случаев скорректировать сведения, касающиеся обозначенного параметра.

Следующий информационный уровень включает данные о различных системных связях терминов в пределах своих терминосистем, об их смысловых отношениях разного типа. В рамках этого уровня представлены сведения о вариантных формах, о синонимах, антонимах, гиперонимах, гипонимах, согипонимах, о сочетаемостных возможностях терминов, их словообразовательной активности. Совокупность параметров второго уровня позволяет судить об особенностях функционирования терминов-латинизмов в русском языке.

Наконец, третий уровень представляет блок дополнительной информации с хронологическими данными, сведениями об изменении значения заимствованных терминов (в случае их фиксации в терминологических изданиях), информацией о сфере распространения и характере переоформления иноязычных терминов в других языках и некоторыми другими параметрами.

В компьютерном терминологическом фонде, способном реализовать возможность подачи материала по любому количеству заданных параметров, имеет смысл эксплицитно представить полный перечень включенных характеристик латинских заимствований, чтобы в случае необходимости пользователь мог выбрать один определенный параметр, характеризующий термин (например, его толкование), или совокупность нескольких параметров в любой комбинации (например, время фиксации термина в письменных памятниках или словарях, его этимологию и соответствия в других языках).

Поскольку разрабатываемый фонд представляет собой открытую систему и оперативно может пополняться новой информацией, а некоторые сведения могут корректироваться и даже изыматься из терминологической общности, если утратят в силу каких-либо причин свою актуальность, то компьютерное представление материала позволит объективно отражать состояние терминосистем гуманитарных областей знания на каждом этапе их развития. В качестве примера можно привести культурологию, понятийный аппарат которой в настоящее время находится в стадии формирования.

Распределение информации по уровням никак не будет ограничивать свободы пользователя в работе. В соответствии со своими потребностями он может работать в пределах одного уровня и попутно обращаться к информационным зонам любого другого уровня, используя при этом общий перечень терминографированных параметров. При этом поиск необходимой информации можно значительно упростить, а время, затраченное на работу, существенно сократить, если информационные ресурсы, представленные в компьютерном терминологическом фонде, организовать в виде гипертекста.

Разработка словарей типа WordNet для русского языка

С. А. Яблонский

Петербургский государственный университет путей сообщения / ЗАО «РУССИКОН»

Онтологии, WordNet

Работа над словарем Princeton WordNet (PWN – wordnet.princeton.edu/) начата в начале 1980-х годов и продолжается сегодня. Сейчас доступна версия 2.1. PWN. Существующая версия WordNet охватывает более 120 тысяч слов общеупотребительной лексики современного английского языка. В настоящее время созданы многоязычные варианты WordNet – EuroWordNet (http://www.illc.uva.nl/EuroWordNet) и BalkaNet (http://www.ceid.upatras.gr/Balkanet), объединяющие основные европейские, болгарский, турецкий, чешский, румынский и сербский языки.

Установление отношений между WordNet и версиями WordNet на других языках в EuroWordNet велось через специальный межъязыковый индекс (Interlingual Index – ILI (PWN 1.5)), а в BalkaNet межъязыковой индекс базировался на PWN 2.0 – ILI (PWN 2.0).

Одновременно ведутся работы по установлению соответствия между синсетами WordNet и понятиями различных онтологий, при котором каждый синсет WordNet:

- или напрямую сопоставляется с понятием онтологии,
- или является гипонимом для некоторого понятия,
- или является примером понятия онтологии.

Так построено отображение PWN 1.6 на онтологию SUMO – Standartized Upper Merged Ontology (http://ontology.teknowledge.com) и частично на онтологию OpenCyc (http://www.opencyc.org). Онтологии позволяют точно и эффективно описывать семантику для предметной области и представляют систему понятий, для которых описаны отношения и заданы правила вывода.

В настоящее время разрабатывается worldwide wordnet grid (WG – http://www.globalwordnet.org/gwa/gwa_grid.htm). WG будет построен на основе множества наиболее общих концептов WordNet – Common Base Concepts, которые будут выражены через синсеты Wordnet и понятия SUMO. WG будет представлен на всех языках. Сейчас существует первая версия WG на английском языке, состоящая из 4689 Common Base Concepts.

Известно о нескольких реализациях WordNet-подобных лексических баз данных для русского языка:

- проект RussNet, разрабатывается с 1999 года на филологическом факультете СПбГУ (http://www.phil.pu.ru/depts/12/RN/). Методика и принципы построения словаря проекта RussNet ориентированы на длительный процесс разработки ресурса группой лингвистов без автоматизации процесса построения и связи с исходным PWN;
- проект тезауруса RuThes, используемого в УИС РОССИЯ ([5]). Закрытый коммерческий ресурс.

Рассматриваемая в работе реализация русской версии WordNet (Russian WordNet – RWN – http://www.pgups.ru/WebWN/wordnet.uix) ориентирована на формирование ядра RWN (более 100 тыс. слов) за счет:

- привязки RWN к PWN;
- использования доступных русских, англо-русских и русско-английских словарей.

Разработка RWN включает:

- построение русской версии WordNet, достаточно полно (100–120 тыс. лексических единиц) описывающей лексику русского языка и сопоставимой по числу лексических единиц с английской версией. Для этого используются морфологический анализатор и лексические ресурсы ЗАО «РУССИКОН», словари, свободно распространяемые в Интернете (http://www.slovarik.ru/, http://www.artint.ru/projects/frqlist.asp) и ряд печатных изданий;
- интеграцию с другими лексическими системами на основе использования технологии Semantic Web;
- программное построение межъязыкового индекса, определяющего соответствие между синсетами PWN и RWN, на основе использования электронных версий словарей издательства Oxford Press, ряда доступных в Интернете англо-русских и русско-английских словарей, WordNet-Domains.

В рамках технологии Semantic Web консорциум W3C разрабатывает стандарт представления WordNet, в котором WordNet представляется как фрагмент описания онтологии (http://www.w3.org/2001/sw/BestPractices/WNET/wordnet-sw-20040713.html). Для этого PWN представляется в формате описания RDF.

Формат OWL принят в качестве основного для экспорта и импорта данных в / из базы данных в проекте Russian WordNet. OWL / RDFS-схема Russian WordNet соответствует основным рекомендациям W3C-консорциума. OWL-схема проверена на корректность с помощью RDF-анализатора.

Литература

1. *Fellbaum C.* WordNet: an Electronic Lexical Database. Cambridge, MA, 1998.
2. *Miller G. et al.* Five Papers on WordNet // CSL-Report. Vol. 43. Princeton, 1990. ftp://ftp.cogsci.priceton.edu/pub/wordnet/5papers.ps.
3. *Vossen P.* EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht, 1998.
4. *Сухоногов А. М., Яблонский С. А.* Разработка русского WordNet // Электронные библиотеки: перспективные методы и технологии, электронные коллекции // Тр. Шестой Всерос. науч. конф. RDCL'2004. Пущино, 2004. С. 113–117.
5. *Сухоногов А. М., Яблонский С. А.* Словари типа WordNet в технологиях Semantic Web // Девятая Национальная конф. по искусственному интеллекту с международным участием КИИ–2004: Труды конф.: В 3 т. Т. 2. М., 2004. С. 557–564.
6. *Сухоногов А. М., Яблонский С. А.* Использование WordNet при поиске и индексации текстов // Междунар. конф. «Корпусная лингвистика – 2004»: Тезисы докл. СПб., 2004. С. 84–89.
7. *Balkova V., Suhonogov A., Yablonsky S. A.* Russia WordNet. From UML-notation to Interne / Intranet Database Implementation // Proceedings of the Second International WordNet Conference. GWC 2004. Brno, 2004. P. 31–38.