Прикладная и математическая лингвистика

И. В. Азарова, канд. филол. наук, Санкт-Петербургский государственный университет (Россия)

**Оптимальные параметры синтаксической аннотации  
в корпусах текстов**

При решении общих вопросов создания и обработки корпусов текстов необходимо предусмотреть параметры, которые потребуют систематизации. В частности, одним из важных аспектов лингвистического обеспечения корпусов текстов является аннотация текстов корпуса [Захаров, Азарова, 2012]. Чаще всего рассматривают морфологическую разметку, при этом приоритет отдается международным стандартам, например ME (MULTEXT-East). Однако стремление к универсальности и простоте набора грамматических категорий зачастую приводит к увеличению списка параметров и несовместимости их значений. При этом удлинение параметрической строки и возникновение «запрещенных» цепочек значений категорий усложняет процедуры снятия морфологической неоднозначности [Копотев, 2010].

В докладе планируется расширить предложенные параметры морфологической разметки [Азарова, Гордеев, 2012] до более сложных уровней лингвистического описания. Во-первых, рассмотреть аннотацию линейных синтаксических конструкций, синонимичных морфологическим формам. Во-вторых, предложить необходимые и достаточные параметры и их значения для описания отношений синтаксической зависимости в непредикативных словосочетаниях. В-третьих, показать варианты строения простых предложений в терминах ядра (структурной схемы) и его расширения (распространяющих конструкций). В-четвертых, предложить варианты описания коммуникативной структуры текста: тематических, рематических и иных компонентов организации сверхфразовых единств.

*Литература*

*Азарова И. В., Гордеев С. С.* Оптимальные параметры морфологической аннотации при использовании базы данных форм // Материалы XLI международной филологической конференции. Секция прикладной и математической лингвистики. СПб., 2012. C. 3—27.

*Захаров В. П., Азарова И. В.* Параметризация специальных корпусов текстов // Структурная и прикладная лингвистика: Межвузовский сборник. Вып. 9. СПб.: СПбГУ, 2012. С. 176—184.

*Копотев М. В.* Поиск ошибок в корпусе с помощью MTE-разметки // Международная конференция по компьютерной лингвистике «Диалог 2010». М., 2010. URL: http://www. dialog-21. ru/digests/dialog2010/materials/html/34. htm

MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages. URL: http://nl.ijs.si/ME

I. V. Azarova

**Optimal parameters for syntactic annotation in corpora**

The parameters for the syntactic annotation of corpora are discussed in the paper. The different levels of syntactic representation are in the focus of the study: (1) linear syntactic constructions which are synonymous to morphologic forms; (2) non-predicative syntactic dependencies; (3) structural schemes of clauses and their expansions; (4) communicational units: topics and comments.

Е. Л. Алексеева, канд. филол. наук, Санкт-Петербургский государственный университет (Россия)

**К вопросу о древнерусской тайнописи**

М. Н. Сперанский определяет тайнопись как такой вид письма, «который, отклоняясь от общеупотребительного, имеет целью сделать написанное слово или мысль, положенную на письмо, недоступными для обыкновенного грамотного читателя или, по крайней мере, создать затруднение для прочтения или уразумения написанного» [Сперанский, 1929, с. 2].

Использовалось несколько способов тайнописи: замена кириллических букв буквами других алфавитов, измененное начертание букв, замена одних букв другими по какой-нибудь определенной системе (литорея), изменение порядка букв в слове (анаграмма), акростих, а также счетная или цифирная тайнопись, использующая тот факт, что почти все буквы обладали также и числовым значением.

Нами опубликован ряд статей, посвященных проблемам дешифровки разных видов тайнописи [Алексеева, 2007a; 2007b; 2008; 2009; 2010; 2011; 2012]; предлагаемый доклад представляет собой продолжение этой темы.

*Литература*

*Алексеева Е. Л.* Дешифровка древне­русской тайнописи, использующей условный алфавит // Материалы XXXIX Меж­дународ­ной фи­ло­ло­гической конфе­рен­ции, 15—20 марта 2010 г. Секция при­кладной и математи­ческой лингвис­тики.СПб., 2010 С. 6—14.

*Алексеева Е. Л.* Еще раз о загадке Христофора // Сборник материалов по итогам XLI Международной филологической конференции. Секция прикладной и математической лингвистики. СПб., 2012.

*Алексеева Е. Л.* К вопросу о счетной (цифирной) тайнописи в русских памятниках письменности // Cтpуктуpнaя и пpи­клaднaя линг­виcтикa. Вып. 7. CПб.: Изд-вo CПбГУ, 2008. C. 322—327.

*Алексеева Е. Л.* Криптограмма в «Диоптре» и ее вариан­ты в древнерусской письменности // Материалы XXXVIII Меж­дународ­ной фило­ло­гической конфе­рен­ции, 16—21 марта 2009 г. Секция прикладной и мате­матической лингвис­тики.СПб: Факультет филологии и искусств СПбГУ, 2009. С. 3—9.

*Алексеева Е. Л.* Счетная тайнопись в «Житии Кассиана Угличского»: Атрибу­ция // Прикладная и мате­матическая лингвис­тика: Материалы секции XXXVII Меж­дународ­ной фило­ло­гической конфе­рен­ции, 11—15 марта 2008 г., Санкт-Петербург.СПб: Факультет филологии и искусств СПбГУ, 2008. С. 55—56.

*Алексеева Е. Л.* Тайнопись в Житии Дионисия Глушицкого // Материалы XXXVI Международ­ной фило­логической конфе­рен­ции 12—17 марта 2007 г. Вып. 10: Прикла­дная и математиче­ская лингвистика. СПб.: Филологический факультет СПбГУ, 2007. С. 9—10.

*Алексеева Е. Л.* Тайнопись в несколь­ких рукописях Кирил­ло-Белозерского мона­стыря начала XVII в. // Русская агиография: Исследования. Материалы. Публикации.СПб.: Пушкинский Дом, 2011. Т. II. С. 319—325.

*Сперанский М. Н.* Тайнопись в югославянских и русских памятниках письма. Л.: Изд-во АН СССР, 1929 С. 2.

E. L. Alexeeva

**A Study in Old Russian Cryptography**

The paper reveals some problems encountered while deciphering Old Russian cryptograms.

Л. Н. Беляева, д-р филол. наук, Российский государственный педагогический университет им. А. И. Герцена (Россия)

**Система машинного перевода как часть технологического процесса**

Современный уровень развития лингвистических технологий определяет необходимость уточнения места и функций технического перевода и самого технического переводчика в особой технологической цепочке. Эта цепочка как минимум включает в качестве инструментального ресурса системы машинного перевода, комплекс автоматизированных словарей, предметно-ориентированный корпус текстов, издательские прикладные программы и системы. Оставив в стороне рассуждения о том, является перевод ремеслом или искусством, мы просто вынуждены определить, каковы функции переводчика и терминолога в новой структуре, как должен быть организован обмен информацией в технологической цепочке перевода.

Система машинного перевода в подобной структуре должна включать большое количество специализированных словарей, ориентированных не только на принятую и стандартизированную терминологию соответствующих отраслей, но и на корпоративную терминологию, которая может существенно отличаться и по составу, и по переводным эквивалентам [Кит, 2010]. Сегодня можно утверждать, что результаты, получаемые разными системами МП, сопоставимы по качеству. Более высокое качество характерно для систем, использующих тексты, написанные по жестко заданным правилам и на контролируемых языках. Однако большинство перерабатываемых системами МП текстов написано на естественных языках и, что очень важно, на языках глобальных. Следовательно, для улучшения качества перевода необходимо учитывать особенности текстов и возможности его предварительной подготовки и редактирования.

Особую часть технологической цепочки перевода составляют лексикографические ресурсы, ориентированные на необходимость выполнения терминологической работы: для терминолога существует насущная необходимость реагировать быстро (и стандартным образом), для того чтобы удовлетворять требования к обработке информации и выделять не зарегистрированные ранее или просто новые терминологические единицы. Различия самих исходных текстов, уровней специализации текстов, целей и профилей конечных пользователей и уровня автоматизации объясняют отсутствие универсальных методов для решения задачи извлечения терминов из текстов. Результаты работы терминолога должны вводиться в систему лексикографических ресурсов до того, как переводчик получает текст и результат машинного перевода. В современной технологической цепочке перевода терминологическая работа не просто является самостоятельным звеном, но осуществляется до собственно перевода.

Соответственно, важным решением для организации технологического процесса является разделение функций терминолога и переводчика, т. е. выбор того, кто, собственно, осуществляет ведение специализированных словарей в системе машинного перевода по результатам работы терминолога и переводчика, осуществляется ли это решение до или после окончательного перевода текста, как учитываются переводчиком результаты работы терминолога.

Естественно, технологическая цепочка переводческого процесса должна включать специальное программное обеспечение, поддерживающее работу всех участников.

*Литература*

*Кит М. С.* О стратегии построения высокоэффективных сетевых словарей (на базе разработки словаря LexSite) // Вестник РГГУ. М.: Издательский центр РГГУ. 2010. № 9. С. 149—160.

L. N. Beliaeva

**Machine Translation System as a Technological Process Component**

Modern state of linguistic technologies determines the necessity to specify the place and functions of both technical translation and translator in the specific technological process and appropriate production string. This string includes as minimum a specialized machine translation system, complex of automated and/or automatic dictionaries, subject-oriented parallel or comparable text corpora, application software. Paying no attention to the problem whether translation is a craft or art, we are really forced to determine, what are translator and terminologist functions in this new structure, how information exchange in this technological process is to be organized.

С. Ю. Богданова, д-р филол. наук, Иркутский государственный лингвистический университет (Россия)

**О разработке учебного мультимодального корпуса текстов**

Мультимодальный (мультимедийный) корпус — это электронный ресурс, предназначенный для изучения звучащей речи, «погруженной» в обстоятельства ее произнесения. Корпус такого рода кроме текстовой составляющей включает видео- или аудиозаписи процесса коммуника­ции с привязкой к тексту. Тексты выравнивают с их расшифровками, что позволяет исследовать не только языковые единицы, но и речевые действия говорящего в различных ситуациях общения, его неречевое поведение (мимику, жесты, позы).

В Иркутском государственном лингвистическом университете идет работа по созданию учебного мультимодального корпуса (УМКО) видеозаписей неподготовленных учебных диалогов носителей и неносителей русского и китайского языков по определенным темам, размеченных в программе ELAN. В качестве подкорпуса создаваемого мультимодального корпуса разрабатывается параллельный русско-китайский корпус, в котором между собой выравниваются тексты на одном из языков, произнесенные носителями и не носителями языка. Выравнивание осуществляется по смысловым блокам внутри диалогов. Например, диалог носителей русского языка на русском языке сопоставляется с диалогом на ту же тему («Знакомство», «Регистрация в аэропорту» и др.) китайцев, изучающих русский язык, на русском языке. Видеоматериалы снабжены метаданными, включающими такие значимые сведения о коммуникантах, как возраст, пол, образование, стаж изучения иностранного языка и др. Координатором создания корпуса УМКО выступила кафедра прикладной лингвистики.

Выполнение данного крупномасштабного проекта относится к сфере одного из восьми приоритетных направлений развития науки РФ «Информационно-телекоммуникационные системы» в руслеинформационных технологий и создания электронных баз данных и преследует несколько важных целей:

1) лингводидактическая цель проекта заключается в разработке новых методик преподавания русского и китайского языков как иностранных на эмпирическом материале нового поколения, в том числе с учетом типичных ошибок при построении устных высказываний на иностранном языке студентами старших курсов. Данный корпус предназначен в первую очередь для выработки методических рекомендаций по обучению китайскому языку русских студентов и обучению русскому языку китайских студентов, поскольку позволяет найти пути устранения ошибок в ходе учебных занятий и самостоятельной работы студентов.

Возможность слышать не просто отдельно произнесенное слово, которую предоставляют в настоящее время многие словари, но слышать его в разных контекстах и с поддержкой видео, благодаря которому можно отследить движение губ, выражение лица и интонацию, позволяет точнее определить значение слова в предложении. Это подразумевает расширение осведомленности обучающихся о языковом разнообразии;

2) лингвистическая цель заключается в анализе различных языковых и сопутствующих им паралингвистических явлений, зафиксированных в подкорпусах, созданных на основе видеозаписей диалогов носителей русского и китайского языков. С помощью корпусов текстов можно не только с большей точностью анализировать отдельные факты реализации языковых единиц, но и выявлять общие закономерности, присущие языковой системе.

Постановка цели создания мультимодального русско-китайского учебного корпуса обладает абсолютной новизной. Аналоги корпуса отсутствуют.

S. Yu. Bogdanova

**Designing the multimodal learners’ corpus**

The first Multimodal Learners’ Corpus of the Russian and Chinese languages is being constructed in Irkutsk state linguistic university. Several topics of everyday use have been chosen and Chinese undergraduate students of Russian and Russian undergraduate students of Chinese were asked to make up dialogues in their own language and in the foreign language. These dialogues have been filmed and annotated with the help of ELAN corpus-manager. Paralinguistic annotation has been added, too. All the dialogues are also represented in the parallel corpus. The corpus will be used in Foreign language teaching.

Д. С. Воевудский, асп., Воронежский государственный университет (Россия)

**Статистическая обработка данных о длине нидерландской лексики**

**(по данным нидерландско-русских словарей)**

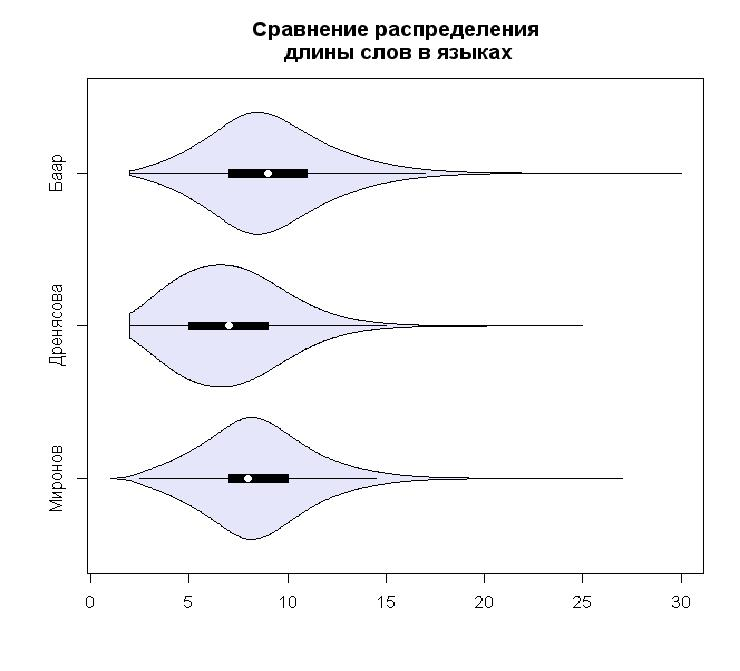
Целью предлагаемого исследования является выявление закономерностей распределения количества слов по длине посредством построения адекватной стохастической модели.

Для достижения поставленной цели были решены следующие задачи: 1) создание электронных баз данных исследуемых словарей; 2) обработка и аппроксимация полученных данных различными видами распределений с помощью GNU R; 3) проверка полученных результатов.

Поскольку звуковая форма является первичной реальностью языка, данные по этому параметру брались в звуках. Для этого показатели длины в буквах были обработаны по правилам чтения нидерландского языка [Берков, 2001, с. 74—75]. Для анализа были взяты три нидерландско-русских словаря различного размера [Дренясова, 1977; Миронов, 2006; Баар, 2012].

Вся обработка данных и их графическое представление произведены с помощью языка статистической обработки GNU R. В отечественной научно-практической литературе вопрос применения языка R для решения задач математической лингвистики остается нераскрытым.

Анализ словарей был проведен с помощью методов описательной статистики, а также посредством визуализации данных на диаграмме типа «скрипка» (violin plot), как это показано на рис. 1. Эта графическая форма представления дает больше информации о характере распределения, чем «ящик с усами» (box-and-whisker plot), т. к. помимо данных о медиане и квартилях отражает еще и показатели ядерной плотности распределения [Hintze, Nelson, 1998].



*Рис. 1.* Распределение в виде скрипки по исследуемым словарям

Из полученного распределения можно сделать вывод о том, что между словарями больших размеров наблюдается больше сходства, чем с малым словарем. Разницу в положении медианы можно объяснить тем, что в словарях большого объема больше представлена специальная лексика и термины, которые обычно обладают большей длиной, нежели общеупотребительные слова. Также можно заметить, что распределение ассиметрично со смещением вправо.

Затем полученные данные были обработаны методом бутстреппинга по методике Каллена и Фрея [Cullen, Frey, 1999], чтобы на основании расчета моментов выяснить, какое распределение следует использовать для последующей аппроксимации. Наиболее близкие результаты дали отрицательное биноминальное распределение и распределение Пуассона. Дальнейший анализ результатов аппроксимации методом максимального правдоподобия показал, что более адекватно в данном случае распределение Пуассона. Оно относится к семейству дискретных распределений и задается следующей функцией вероятности:

, где

λ > 0; *k* = 0,1,2,…; e — основание натурального логарифма.

Проведенные вычисления показали, что при использовании при аппроксимации распределения Пуассона λ ≈ средней длине слова в анализируемом словаре. Так, для словаря ван ден Баара эта величина равна 9,1, для словаря Миронова — 8,5, для словаря Дренясовой — 7,1.

Затем была проведена проверка адекватности аппроксимации распределением Пуассона — были взяты случайные выборки по каждому из словарей в количестве 2000, эти выборки были проведены 10 000 раз и для каждого раза вычислялся критерий согласия Пирсона. Количество успехов, где эмпирическое и теоретическое распределения совпадают, было равно 8441 для словаря Баара, 8508 — для словаря Дренясовой и 8426 — для словаря Миронова.

Таким образом, распределение Пуассона обеспечивает достаточно высокое качество аппроксимации и позволяет выявить закономерность в частотном распределении слов в словаре.

D. S. Voevudsky

**Statistical analysis of the length of Dutch words (based on Dutch-Russian dictionaries)**

In the article the distribution of the length of words in three Dutch-Russian dictionaries is analysed. Based on the analysis, a mathematical model reveals regularity in the distribution of word frequencies of different length depending on the average length of words in dictionaries analyzed by Poisson approximation method of maximum likelihood. The validation of the model by the Monte Carlo method is performed.

*Литература*

*Baar A. H., van den.* Groot Nederlands-Russisch Woordenboek / Большой голландско-русский словарь. Amsterdam: Uitgeverij Pegasus, 2012.

*Cullen A. C., Frey H. C.* Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs / Alison C. Cullen, H. Christopher Frey // Springer, 1999.

*Hintze J. L., Nelson R. D.* Violin Plots: A Box Plot-Density Trace Synergism / Jerry L. Hintze, Ray D. Nelson // The American Statistician. 1998. Vol. 52. P. 181—184.

*Берков В. П.* Современные германские языки / В. П. Берков. М.: Астрель АСТ, 2001.

Большой нидерландско-русский словарь: Оканд. 180 000 сл. и словосочетаний / С. А. Миронов, В. О. Белоусов, Л. С. Шечкова и др.; Под рук. С. А. Миронова. 3-е изд., испр. М.: Живой яз., 2006.

*Дренясова Т. Н., Миронов С. А.* Карманный нидерландско-русский словарь. Около 7000 слов. М.: Русский язык, 1977.

В. П. Захаров, канд. филол. наук, Санкт-Петербургский государственный университет (Россия);

И. В. Азарова, канд. филол. наук, Санкт-Петербургский государственный университет (Россия);

М. В. Хохлова, канд. филол. наук, Санкт-Петербургский государственный университет (Россия)

**Эксперименты по выделению прототипов терминов в корпусах специальных текстов**

Огромное место в лингвистике, в терминоведении, в культурологии занимают специализированные подъязыки. Под специальным текстом, как правило, имеется в виду или текст, относящийся к определенному типу (жанру), или научный текст, относящийся к той или иной области знаний.

Специальные тексты во многом терминологичны, и поэтому должны быть разработаны принципы и методы автоматизированного выделения терминов и терминологических сочетаний (коллокаций), основанные на корпусных данных [Захаров, 2010].

Основные формальные признаки научного термина можно свести к параметрам, фиксируемым на разных уровнях описания языковой структуры: статистические, статистико-комбинаторные, синтаксические, морфологические. В случае применения параметров могут быть выделены лексические единицы, сходные с терминами по своим свойствам, которые можно назвать прототипами терминов [Азарова, Гордеев, 2012; Захаров, Азарова, 2012]. Большая часть подобных единиц представляет собой словосочетания.

При автоматическом выявлении словосочетаний помимо статистических критериев отбора данных должны работать и другие методы, основывающиеся на собственно лингвистических моделях. Нами были проведены эксперименты по выделению прототипов терминов статистическими методами с использованием грамматики лексико-синтаксических шаблонов [Хохлова, 2010] на материале корпусов специальных текстов, в которые вошли: 1) материалы конференций по корпусной лингвистике; 2) публикации по гомеопатии; 3) технические руководства.

*Литература*

*Азарова И. В., Гордеев С. С.* Оптимальные параметры морфологической аннотации при использовании базы данных форм // Материалы XLI Международной филологической конференции, Санкт-Петербург, 26—31 марта 2012 г.. Секция прикладной и математической лингвистики. СПб., 2012. С. 3—27.

*Захаров В. П.* Тезаурус по корпусной лингвистике // Информационные технологии и письменное наследие. El'Manuscript-10. Материалы Международной научной конференции. Уфа, 2010. С. 95—98.

*Захаров В. П., Азарова И. В.* Параметризация специальных корпусов текстов // Структурная и прикладная лингвистика: Межвузовский сборник. Вып. 9. СПб.: СПбГУ, 2012. С. 176—184.

*Митрофанова О. А., Захаров В. П.* Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27—31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009. С. 321—328.

*Хохлова М. В.* Исследование сочетаемости и устойчивости лексических единиц автоматическими методами // Структурная и прикладная лингвистика. Выпуск 8. СПб., 2010. С. 206—218.

V. P. Zakharov, I. V. Azarova, M. V. Khokhlova

**Experiments on Term Prototypes Extraction in Special Text Corpora**

Special texts are characterized by a large number of terms. Thus, there should be elaborated principles and methods for automatic term and collocation extraction based on corpus data. The papers discusses the results of statistical experiments on prototype term extraction from a number of special text corpora — scientific texts of various fields.

А. В. Зубов, д-р филол. наук, Минский государственный лингвистический университет (Республика Беларусь)

**Автоматическая статистическая оценка количественных характеристик учебников иностранных языков**

В последние десятилетия необыкновенно возрос интерес к изучению иностранных языков. И число таких языков постоянно растет. Соответственно, появляется большое число учебников и учебных пособий, предназначенных для обучения одному и тому же иностранному языку, как в школах, так и в вузах.

Вполне естественно возникает вопрос: какой учебник выбрать для обучения иностранному языку на определенном этапе обучения некоторой специальности? Однако учебный текст с учетом всех дидактических и методических принципов его организации является достаточно сложной системой, коммуникативная оценка которой нуждается в последовательном экспериментальном анализе. Эффективность такой системы будет определяться степенью согласованности составляющих этой системы, а именно уровнем знаний обучающегося и сложностью учебника, возможностями обучающегося к восприятию новой информации и способами введения ее в учебник, закономерностями человеческой памяти и организацией материала, подлежащего изучению.

Для оценки такой системы могут быть выделены два типа факторов:

1. Количественные факторы.
2. Качественные факторы.

К числу первых можно отнести:

1. Количество незнакомой лексики, подлежащей усвоению за один учебный год.

2. Темп введения новых слов, т. е. количество незнакомой лексики, вводимой за один урок или за одно задание.

3. Повторяемость слова в пределах одного задания (урока).

4. Повторяемость слова во всем учебнике.

5. Длина интервала между повторениями одного и того же слова.

Эти и ряд других подобных параметров могут быть вычислены с использованием статистических методов и современных компьютеров.

Качественные характеристики учебников могут быть оценены с использованием специальных 12 параметров.

Учет всех этих требований при отборе лексики в учебники проводится в определенной степени субъективно, в зависимости от методического опыта, интуиции, «языкового чутья» составителей учебника. Отсюда общеизвестные недочеты в отборе и систематизации лексического и грамматического материала, приводимого в учебниках.

Поэтому возникает задача каким-то образом сравнивать существующие и вновь создаваемые учебники иностранных языков, с тем чтобы выбрать тот из них, который лучше удовлетворяет задачам обучения иностранному языку.

Рассматривая отмеченные выше 5 количественных факторов, можно отметить, что первые два фактора в определенной степени конкретизированы в зависимости от срока обучения и возраста обучаемых.

Три следующих параметра не находят единого количественного выражения. Можно предложить следующий подход к их оценке. Каждое слово текста должно быть оценено с применением трех статистических параметров *D*, *U* и *V*, где:

*D* — коэффициент распространенности слова в учебнике;

*U* — коэффициент употребительности слова в учебнике;

*V* — коэффициент вариации для каждого слова учебника.

Для вычисления коэффициентов *D*, *U* и *V* составляются частотные словари по каждому уроку учебника.

Для принятия решения о том, какой из учебников *А* и *В* лучше удовлетворяет исследуемым трем показателям, проводится сравнение полученных инвентарей слов этих учебников с заданным лексическим минимумом данного курса.

В докладе приводятся данные о результатах применения данного метода к анализу 4 тем в двух учебниках для обучения французскому языку студентов 1-го курса вузов.

A. V. Zubov

**Automatic statistical estimation quantitatives characteristics of foreign language textbooks**

The author propase some quantitatives characteristics for evaluation of foreign language textbooks. All these characteristics uses the frequencies of words in texts and lexical minimum of students.

Е. А. Каневский, канд. техн. наук, Санкт-Петербургский экономико-математический институт РАН (Россия);

К. К. Боярский, канд. физ.-мат. наук, Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Россия)

**Особые слова в русскоязычном тексте**

Вопросам компьютерной морфологии посвящено множество работ, однако эта проблема до сих пор не решена окончательно. Еще большие трудности возникают при синтаксическом разборе предложений русского языка. Известно, что при автоматическом разборе предложений и построении синтаксического дерева зависимостей прежде всего возникает проблема снятия морфологической неоднозначности.

Как показывает наш опыт создания и обкатки морфолого-лексического анализатора TextAn [Каневский, Боярский, 2012а] и семантико-синтаксического анализатора SemSin [Каневский, Боярский, 2012б], в текстах на русском языке часто встречаются особые слова, которые нецелесообразно подвергать стандартному разбору.

К первой группе таких слов следует отнести токены, которые для правильного построения дерева должны анализироваться особым образом. Прежде всего это различные аббревиатуры и единицы измерения, например: *кв., корп., рис., тел., сек., км., руб., долл*. Далее следует выделить частицы, подключаемые к предшествующим словам (*б, бы, ли, ль, же, хотя бы*) и частицы в составе ФИО: *аф, бен, ван, де, де ла, де-ла* и др. Затем следуют составные числительные, например: *тысяча двести двадцать два* — и латинские цифры.

Ко второй группе относятся слова, имеющие по два значения, одно из которых встречается так редко, что им можно пренебречь. Это деепричастия *благодаря, для, зря, мая, моря, некая* и *хотя*, повелительная форма глаголов *весь, всей, почти, при, пять, секретарь, три* и *цель*, а также инфинитив *честь*. На первый взгляд не все значения кажутся такими, чтобы ими можно было пренебречь. Такова, например, повелительная форма *три*. Кажется, что довольно часто можно сказать: «*три* какую-либо вещь». Однако если взять конкурсный текст [Толдова, 2012], состоящий из 695 тысяч словоформ (130 тысяч различных слов), то в нем слово *«три»* встречается 394 раза, и всегда это число. Подозрительно также деепричастие *благодаря*. Однако в том же тексте слово *«благодаря»* встречается 190 раз, и всегда это предлог. Остальные слова вызывают еще меньше подозрений.

Третью и довольно многочисленную группу составляют омонимичные словоформы, значение которых зависит от контекста. К ним относятся, например, *королева, сорок, по, про, потом, перед, включая, исключая, прочую, души* и др. Так, *королева* может являться фамилией или титулом, *сорок* — числительным или птицей, *перед* — предлогом или существительным, *прочую* — прилагательным или глаголом, *души* — существительным или повелительной формой глагола.

В отличие от других омонимов интерпретация таких слов, как правило, однозначно определяется их ближайшим окружением в тексте. Поэтому оказалось целесообразным выделить их обработку в особый блок, предшествующий остальному анализу.

*Литература*

*Каневский Е. А., Боярский К. К.* Об одном опыте создания морфолого-лексического анализатора // Прикладна лiнгвiстика та лiнгвiстичнi технологii: MegaLing-2011. К.: Довiра, 2012а. С. 170—180.

*Толдова С. Ю. и др.* Оценка методов автоматического анализа текста 2011—2012: синтаксические парсеры русского языка // **Компьютерная** лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 10 (17). Т. 2. М.: Изд-во РГГУ, 2012. С. 78—92.

*Каневский Е. А., Боярский К. К.* Семантико-синтаксический анализатор  
SemSin. К.: 2012б. URL: http://www.dialog-21.ru/digest/2012/?type=doc

E. A. Kanevsky, K. K. Boiarsky

**The special words in Russian text**

The report is devoted to questions of the removal of morphological ambiguity when syntactic analysis of texts in Russian language. It is proposed to allocate three groups of special words, a preliminary analysis of which it is expedient to produce separately, outside of the General rules of text processing. Gives a list of suggested similar words and recommendations for data processing, confirming the validity of this approach.

А. В. Клементьева, Институт филологических исследований (Россия); Т. Г. Скребцова, доц., Санкт-Петербургский государственный университет (Россия);

С. Г. Суворов, Институт филологических исследований (Россия)

**Графематический анализатор текста (опыт разработки)**

В докладе описываются принципы действия графематического анализатора русскоязычных текстов, созданного научным коллективом в рамках проекта фундаментальных исследований по приоритетным программам развития СПбГУ. Графематический анализ составляет необходимый этап успешной автоматической обработки текста. Представленный анализатор является компонентом разрабатываемого лингвистического процессора, направленного на автоматическое преобразование текстов на естественном языке в метаязыковую структуру данных. В докладе обсуждаются теоретические предпосылки предлагаемого анализатора (и процессора в целом), обусловливающие его отличия от других алгоритмов графематического анализа.

A. V. Klementjeva, T. G. Skrebtsova, S. G. Suvorov

**A Practical Approach to Graphemic Text Analysis**

The paper outlines a practical approach to the problem of automatic graphemic analysis of Russian texts, put forward by a group of scholars in the framework of a larger project aimed at converting natural-language texts into metalinguistic structures. Graphemic analysis forms an indispenseable component of any linguistic processing system, being its starting point. The paper focusses on theoretical backgrounds of the proposed program as well as principal features which distinguish it from the bulk of similar algorithms.

Т. М. Ландо, Н. А. Остапук, ООО «Яндекс» (Россия)

**Томита-парсер:** **инструмент извлечения фактов из текста**

Томита-парсер — это инструмент для извлечения структурированных данных (фактов) из текста на естественном языке. Извлечение фактов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов.

В основе программы лежит алгоритм GLR (generalized left-to-right) парсинга, описанный Масару Томита в 1984 году [Tomita, 1984; 1985]. GLR-парсер является парсером семейства LR (left-to-right), но в отличие от большинства подобных парсеров он оптимизирован, чтобы работать с естественно-языковой неоднозначностью. Парсер был назван в честь создателя алгоритма.

Томита-парсер позволяет по написанным пользователем шаблонам, контекстно-свободным грамматикам (КС-грамматикам) выделять из текста разбитые на поля цепочки слов или факты. Например, можно написать шаблоны для выделения адресов. Здесь фактом является адрес, а его полями — «название города», «название улицы», «номер дома» и т. д.

Парсер включает в себя три стандартных лингвистических процессора: токенизатор (разбиение на слова), сегментатор (разбиение на предложения) и морфологический анализатор (mystem).

Основные компоненты парсера: газеттир, набор КС-грамматик и множество описаний типов фактов, которые порождаются этими грамматиками в результате процедуры интерпретации.

Газеттир — словарь ключевых слов, которые используются в процессе анализа КС-грамматиками. Каждая статья этого словаря задает множество слов и словосочетаний, объединенных общим свойством. Например, «все города России». Затем в грамматике можно использовать свойство «является городом России». Слова или словосочетания можно задавать явно списком, а можно «функционально», указав грамматику, которая описывает нужные цепочки.

Грамматика — множество правил на языке КС-грамматик, описывающих синтаксическую структуру выделяемых цепочек. Парсер запускается всегда на одном предложении. Перед запуском терминалы грамматики отображаются на слова (или словосочетания) предложения. Одному слову может соответствовать много терминальных символов. Таким образом, парсер получает на вход последовательность множеств терминальных символов. Например, в нашей грамматике есть всего два термина: Verb и Noun, а входное предложение — «Мама мыла стекло». Тогда парсер получит на вход такую последовательность: {Noun}, {Verb, Noun}, {Verb, Noun}. На выходе получаются цепочки слов, распознанные этой грамматикой.

Факты — таблицы с колонками, которые называются полями фактов. Факты заполняются во время анализа парсером предложения. Как и чем заполнять поля фактов, указывается в каждой конкретной грамматике. Это называется интерпретацией. Типы фактов описываются на специальном языке в отдельном файле.

Алгоритм работы парсера на одном предложении и одной грамматике:

1. Ищутся вхождения всех ключей из газеттира.
2. Из всех найденных ключей газеттира отбираются те, которые упоминаются в грамматике.
3. Парсер пытается покрыть предложение непересекающимися ключевыми словами так, чтобы как можно большие куски предложения были охвачены ими.
4. Линейная цепочка слов и словосочетаний подается на вход GLR-парсеру. Терминалы грамматики отображаются на входные символы.
5. На последовательности множеств терминалов GLR-парсер строит все возможные варианты. Из всех построенных вариантов также отбираются те, которые как можно шире покрывают предложение.
6. Затем парсер запускает процедуру интерпретации на построенном синтаксическом дереве. Он отбирает специально помеченные подузлы, а слова, которые им соответствуют, записываются в порождаемые грамматикой поля фактов.

*Литература*

*Tomita Masaru*. LR parsers for natural languages. COLING. 10th International Conference on Computational Linguistics. 1984. P. 354—357.

*Tomita Masaru*. An efficient context-free parsing algorithm for natural languages. IJCAI. International Joint Conference on Artificial Intelligence. 1985. P. 756—764.

Руководство разработчика. URL: http://api.yandex. ru/tomita/

T. Lando, N. Ostapuk

**Tomita-parser: the instrument for fact extraction from natural language text**

Tomita-parser is a natural language processing tool designed to extract structured data (facts) from unstructured text. The process of fact extraction is based on context-free grammars and thesauruses. The software is based on GLR (generalized left-to-right) algorithm, created by Masaru Tomita in 1984 [Tomita, 1984; 1985]. GLR parser is an extension of an LR parser algorithm to handle natural language ambiguity. Parser extracts facts according to the rules created by a user, the context-free grammars. The facts are represented as tuple filled with words, word chains or other fact parts.

О. А. Митрофанова, канд. филол. наук, Санкт-Петербургский государственный   
университет (Россия)

**Автоматическое выделение ключевых слов и моделирование тематики текстов на основе алгоритма LDA**

В докладе обсуждаются современные системы автоматического выделения ключевых слов и моделирования тематики текстов, а также анализируются результаты экспериментов, проведенных с их помощью. Рассматривается ряд программных продуктов, находящихся в открытом доступе. Особое внимание уделяется системе TMT (http://code. google. com/p/topic-modeling-tool/, разработчики Дэвид Ньюмэн и коллеги, Калифорнийский университет), в которой производится обработка текстов с помощью алгоритма LDA (Latent Dirichlet Allocation, латентное размещение Дирихле). Данный алгоритм позволяет описывать содержание текста в виде тем, с которыми ассоциируются наборы ключевых слов. Эксперименты, показывающие возможности данного инструмента, проводятся на текстах разной жанровой принадлежности. Полученные данные свидетельствуют о том, что системы указанного типа позволяют осуществлять семантическую свертку текстов и могут быть использованы в исследовании тематической структуры текстов.

O. A. Mitrofanova

**Automatic keyword extraction and topic modeling of texts based on LDA algorythm**

The talk deals with the discussion of contemporary tools for automatic keyword extraction and topic modeling, as well as the analysis of results achieved in experiments. Our attention is focused on TMT system (http://code. google. com/p/topic-modeling-tool/), which performs text processing based on LDA (Latent Dirichlet Allocation). This algorithm allows to describe text content as topics associated with groups of keywords. Experiments illustrating the possibilities of TMT were carried out for texts of various genres. The data show that the tools of the given type perform semantic compression and may be used in the study of conceptual structure of texts.

О. А. Митрофанова, канд. филол. наук, Санкт-Петербургский государственный университет (Россия);

О. Н. Ляшевская, канд. филол. наук, Национальный исследовательский университет Высшая школа экономики (Россия)

**Структурная интерпретация данных для каталога русских лексических конструкций (на материале НКРЯ)**

Доклад отражает новые результаты, полученные в ходе исследования, проводимого совместными усилиями коллективов Национального корпуса русского языка (НКРЯ, http://ruscorpora.ru) и кафедры математической лингвистики. Цель исследования — разработка технологии автоматического распознавания в тексте конструкций, связанных с той или иной лексической единицей, и применение этой технологии в создании Каталога русских лексических конструкций. Выделение конструкций предполагает использование потенциала многоплановой лингвистической разметки НКРЯ (прежде всего лексико-семантической). В докладе обсуждаются: 1) новые эксперименты по автоматическому выделению и кластеризации конструкций с помощью усовершенствованного компьютерного инструмента WSD и CxI; 2) уточнение информации о конструкциях, реализующих значения исследуемых слов, с использованием модуля визуализации данных; 3) возможные подходы к структурированию данных для пилотной версии Каталога русских лексических конструкций.

O. A. Mitrofanova, O. N. Lyashevskaya

**Structural interpretation of data for building the catalogue of russian lexical constructions**

Our research aims at automatic identification of constructions associated with particular lexical items and its subsequent use in building the Catalogue of Russian lexical constructions. The study is based on the data extracted from the Russian National Corpus (RNC, http://ruscorpora.ru). The main accent is made on extensive use of morphological and lexico-semantic data drawn from the multi-level corpus annotation. In our talk we focus our attention on: 1) the new experiments on automatic construction identification and clustering with the help of WSD&CxI toolkit, 2) the use visualization module for improving representation of constructions, 2) structuring the information on constructions for the pilot version of the Catalogue of Russian lexical constructions.

Л. Н. Панько, канд. филол. наук, Санкт-Петербургский университет (Россия)

**Метамодель динамичной организации языка**

**(на материале планирования в лингвометодике РКИ)**

Область русского языка как иностранного (далее — РКИ) — сфера лингвистики, в которой в режиме реального времени реализуется процесс становления динамичной системы русского языка. В изначально свободном сознании инофона целенаправленно и контролируемо лингвистами-практиками моделируется система, подобная (в идеале — тождественная) системе носителя языка. Совокупность описанных теоретической лингвистикой фрагментов используется в РКИ как «рабочий материал» структурирования доступных наблюдению и регистрации разномасштабных рече-языковых отрезков.

Анализ планирования и контроля языковой компетенции инофона позволяет обобщить этапы последовательности формирования динамичной системы и выделить девять взаимосвязанных уровней: 1) фонем, 2) морфем, 3) слова, 4) формы слова, 5) словосочетания, 6) синтаксических групп предикативной основы предложения, 7) простого предложения, 8) высказывания, 9) текста. Лингвометодические материалы РКИ свидетельствует, что в сложившейся обучающей модели реализуется принцип соответствия *масштаба* и *объекта*.

Практика формирования языковых умений в фрагменте отдельно взятого отрезка времени (одного «урока») предусматривает *процесс планирования* от «большего» к «меньшему», в то время как *процесс презентации* подлежащих усвоению языковых единиц обратный — от «меньшего» к «большему». Использование «обратимости» процессов движения языковых элементов — естественное состояние учебной практики, регламентированное принятым коммуникативным подходом.

Каждый отдельно взятый уровень представляет собой зеркальное единство языка-речи, обладающее свойством двусторонности и подчиняющееся закону обратной пропорциональности. Данное явление удобно представить как встречное движение однородных левонаправленных потоков:

1 2 3 4 **5** 6 7 8 9

9 8 7 6 **5** 4 3 2 1

Целеполагание урока — выход в монологическую или диалогическую речь — обеспечивает моделирование обучающего процесса как совокупности элементов *взаимодействующих* уровней. Так, например, наблюдаемые примеры моделирования учебного текста (как единицы 9-го уровня) свидетельствуют, что внутренний механизм обеспечения динамичности системы обеспечивается отношениями трех близлежащих уровней. Например: 1 — 2 — 3; 2 — 3 —4; 3 — 4 — 5 и т. д.

Наблюдаемые в учебных аудиториях процессы обеспечивают понимание организации языковой системы, приближая к решению проблем ненаблюдаемой уровневой организации языка.

L. N. Panjko

**Dynamic organization of language meta model**

**(based on planning in lingvomethodics RCT)**

Contemporary principles of forming Russian language system in the minds of inofona allowed toallocate 9 hierarchically levels. The algorithm of relations between the contact and the

distant levels. The meta model language is offered as a dynamic system.

Рогозина Е. А. Санкт-Петербургский государственный университете (Россия)

**Возможности электронных исторических корпусов: по итогам конференции International Conference on “Historical Corpora 2012”**

Данный доклад направлен на то, чтобы рассказать о достижениях зарубежных коллег в области создания исторических электронных корпусов и работы с ними. Доклад составлен по результатам посещения конференции International Conference on “Historical Corpora 2012”, приуроченной к 25-летию проекта TITUS.

TITUS — это проект Франкфуртского университета им. Гете, который разрабатывался совместно с несколькими университетами других стран: Чехии, Дании и Испании. Проект представляет собой электронный многоязычный корпус исторических текстов, предназначенных для дальнейшего изучения и исследования. Часть текстов находится в открытом доступе, другие тексты доступны только участникам проекта.

Рабочая группа проекта сотрудничает с университетами других стран, что позволяет создавать совместную обширную базу исторических текстов, охватывающих не только индоевропейские, но и северокавказские, уральские, алтайские и сино-тибетские языки. Кроме того, такое сотрудничество позволяет обмениваться опытом в области автоматизации обработки текстов, создания технических инструментов и интерфейсов пользователя для работы с рукописями.

В рамках данного доклада прежде всего хотелось бы остановиться именно на технических средствах, которые могут использоваться для работы с текстами на самых разных языках.

Тексты проекта TITUS, находящиеся в открытом доступе, представлены в формате HTML (UTF-8) и доступны для просмотра через веб-интерфейс проекта. Для текстов предусмотрена возможность аннотирования отдельных слов с указанием адреса вхождения, нормализованной формы слова и информации о морфологии слова. В системе предусмотрено три варианта поиска: поиск без уточнений по текстам всех языков, поиск словоформ по текстам заданного языка и поиск по текстам одного языка, возвращающий словоформы с контекстом до 10 строк. Аннотации и система поиска доступны для всех посетителей веб-страницы проекта.

Кроме этого в рамках проекта создаются библиотеки рукописей в форме цифровых изображений. Подобные материалы пока доступны только участникам проекта. Для работы со снимками страниц рукописей предусмотрены инструменты, позволяющие довольно подробно аннотировать не только рукопись в целом, но и каждую страницу и даже отдельные участки изображений. При аннотировании страниц можно указать информацию о месте хранения рукописи, о времени ее создания, авторстве, степени сохранности, особенностях оформления и т. п. Инструмент более подробного аннотирования позволяет выделять внутри изображения отдельные области (например, словоформы) и вводить пояснения в заранее заданные поля, например вводить нормализованную форму слова, указывать морфологическую и другую информацию. По различным полям аннотации можно выполнять поиск. В данный момент ведется работа над созданием веб-интерфейса для многоуровневого аннотирования, который позволит исследователям из разных учреждений одновременно осуществлять совместную работу над одними и теми же текстами.

E. A. Rogozina

**Tools for Digital Historical Corpora: Overview of the International Conference on “Historical Corpora 2012”**

The present paper sums up the results of International Conference on “Historical Corpora 2012” concerning possibilities of digital multilingual corpora of historical texts — TITUS project of Goethe University of Frankfurt. The report focuses mainly on technical instruments used for text processing, data mining, manuscript annotations and linguistic research.

В. И. Рубинер, Санкт-Петербургский государственный университет (Россия)

**К вопросу о естественности стилевой классификации текста**

Безусловно, в процессе познания мира одной из главных задач является выделение некоторых признаков и построение на их базе классификации, которая бы соответствовала задачам исследователя и при этом максимально полно отражала суть рассматриваемых явлений. Именно поэтому в научном познании выделяют естественные и искусственные классификации.

Как правило, любой исследователь начинает с составления *искусственной классификации*: отбираются произвольные признаки, с учетом этих признаков строится классификации, при этом отбор признаков для классификации производится с точки зрения их полезности для выполнения непосредственных исследовательских задач. Примером такой классификации может считаться ботаническая классификация К. Линнея.

Следующим шагом становится выделение существенных признаков, т. е. признаков, адекватно описывающих суть исследуемых явлений. Классифицирование в соответствии с существенными признаками и приводит к созданию *естественной классификации*, которая позволяет не только довольно односторонне описать объект, но и выделить закономерности между объектами, а в некоторых случаях даже установить или подтвердить некоторые законы природы. К естественным классификациям, без сомнения, можно отнести периодическую таблицу элементов Д. И. Менделеева, которому удалось в рамках своей классификации предсказать существование еще неоткрытых химических элементов.

В языкознании процесс классифицирования также является одним из существенных методов познания. Стоит упомянуть, например, генетическую (основной признак — генетическая схожесть) и географическую (признак — общее географическое распространение) классификации языков или вычленение различных типов текстов. В текстовой типологии учитываются самые различные признаки, при этом остается не совсем ясным, существуют ли среди этих классификаций естественные.

Особо в рамках данной статьи хотелось бы остановиться на функциональной классификации текстов. В качестве трех основных параметров классификации выделяются: *референциальный* (способность текста отсылать читателя к каким-либо объектам), *интерперсональный* (способность служить целям коммуникации), *формальный* (идея о том, что текст является вербально структурированным объектом). Так, Питер Ньюмарк в соответствии с этими параметрами предлагает выделять три типа текстов: тексты с экспрессивной функцией, тексты с информативной функцией, тексты с апеллятивной функцией [Newmark 1995, 23]. С другой стороны, И. Р. Гальперин пишет о том, что именно социальная обусловленность дает возможность провести классифицирование текстов и является определяющим параметром [Гальперин, 1973, 16]. Поэтому при помощи формальных подходов (кластерный анализ, машинное обучение) будет предпринята попытка установления, какие параметры являются действительно существенными для классификации текстов и какие текстовые классификации можно считать естественными.

*Литература*

*Newmark P*. Approaches to Translation. New-York, London, Toronto, 1995.

*Гальперин И. Р*. О понятиях «стиль» и «стилистика» // Вопросы языкознания. 1973. № 3.

V. Rubiner

**On Natural Text Classifications**

This article deals with some peculiarities concerning natural and artificial text classifications. Formal approaches are applied to determine which parameters should be kept as the most relevant for the text classification.

А. К. Филиппов, канд. филол. наук, Институт лингвистических исследований  
Российской академии наук (Россия)

**Применение процедуры автоматического кластерного анализа для исследования структуры лексико-семантической группы глаголов положения в пространстве**

Доклад освещает фрагмент исследования в рамках русскоязычного компьютерного тезауруса RussNet, посвященного исследованию структуры лексико-семантической группы (ЛСГ) глаголов положения в пространстве. Под глаголами положения в пространстве понимается обширная неоднородная группа, включающая как статичные глаголы (*находиться*, *стоять*, *сидеть*, *лежать*, *висеть*), так и глаголы с динамичной семантикой — фазовые глаголы (*сесть*, *лечь*, *встать*, *окружить*), каузативные глаголы (*положить*, *поставить*, *вытащить*) и глаголы, семантика которых связана с переходом из одного пространственного положения в другое, например: *бросить*, *падать*, *опуститься*, *подняться*; общее число исследованных глаголов — 57.

Единицы исследования — не глаголы, а пары «глагол — значение» (составлены на основе контекстов из корпуса кафедры математической лингвистики, вручную классифицированных на основе дефиниций МАС, исследовано 67 пар).

Структура ЛСГ исследовалась с помощью метода, известного как использование маркеров локального контекста: между дистрибуциями глагольных контекстов автоматически вычислялась мера сходства на основе формальных морфологических параметров окна анализа [-5; +3] (размер окна анализа для данного типа исследований экспериментально установлен А. С. Мариной [Марина, 2005]). Данный метод, таким образом, дает информацию о степени близости глаголов друг к другу, которая получает наглядное выражение в виде кластерного анализа; в ходе него глаголы с высоким значением сходства объединяются в один кластер раньше, с низким — позже. Таким образом, исходная совокупность представляется в виде иерархической структуры.

Согласно исследованиям [Leacock, Chodorow, 1998], минимальное число контекстов, необходимое для репрезентативности при данном методе, — 10, оптимальное — не менее 25. Первому критерию соответствовало 37 глагольных значений, второму — 17.

Результаты кластерного анализа для 17 глагольных значений свидетельствуют, что в иерархических классификациях, получаемых путем кластеризации, находят регулярное отражение способы усложнения базовой семантики ЛСГ — статичность, фазовость, каузативность; вместе с тем их влияние не настолько сильно, чтобы детерминировать процедуру кластерного анализа.

Кластерный анализ для 37 значений (с использованием менее строгого критерия отбора) показывает, что более высокочастотные значения легче объединяются в кластеры, нежели низкочастотные, в дистрибуциях которых оказывается слишком мало признаков для противопоставления.

Кластерный анализ с направленным сокращением списка глагольных значений позволяет глубже понять закономерности, стоящие за автоматической классификацией на основе сравнения морфологических дистрибуций. Противопоставление глаголов исследуемой ЛСГ по их основным семантическим признакам не находит регулярного отражения в результатах кластеризации. Напротив, способ усложнения базовой семантики ЛСГ является одним из основных факторов, влияющих на формирование итоговой иерархической структуры; при этом наиболее ярко противопоставлены друг другу дистрибуции значений статичных и каузативных глаголов; их взаимоотношение со значениями фазовых глаголов носит более сложный характер.

*Литература*

*Leacock C., Chodorow M.* Combining Local Context and WordNet Similarity for Word Sense Identification // WordNet: An Electronic Database. 1998.

*Марина А. С.* Автоматизированная классификация контекстов при подготовке данных для компьютерного тезауруса RussNet. СПб, 2005.

A. K. Filippov

**Using automatic cluster analysis for studying the structure of the verb group of space location**

The presentation describes a research within the framework of RussNet, a computer lexicon for the Russian language, studying the structure of the group of verbs denoting space location. The group’s structure is analyzed via comparison of morphological distributions for verbal meanings in their local context (window [-3… +5]). The results are expressed as a hierarchy of clusters. The resulting hierarchies show that one of the main factors influencing the cluster analysis is whether classified verbs are static, phase, or causative.

Е. В. Ягунова, д-р филол. наук, А. Н. Савина, Санкт-Петербургский государственный университет (Россия)

**Фразеологизмы-коллокации в женском детективе**

Введение

В докладе описывается метод исследования фразеологизмов как подмножества коллокаций (биграмм), представляющий собой комбинацию статистического  
и корпусного подходов. Набор предполагаемых фразеологических единиц заранее  
не может быть определен, но может быть получен с помощью статистической оценки текстов корпуса.

Основная задача исследования — анализ фразеологизмов-коллокаций русского женского детектива.

1. Материал и методика исследования

Материалом в данном исследовании послужили тексты иронических детективов, написанные двумя русскими авторами — Дарьей Донцовой и Галиной Куликовой.

Выбор материала обусловлен стилистическими особенностями данного жанра. Предположительно наиболее типичные авторы данного жанра используют большое количество фразеологизмов.

Данное исследование проводилось на материале двух коллекций, однородных  
по автору.

Методика исследования включала в себя применение статистических мер,  
в частности *MI* и *DICE*, основанных на ограничениях на совместную встречаемость:

,

,

где *f(x)* и *f(y)* — частота встречаемости слов *x* и *y* в коллекции, *f(x,y)* — частота совместной встречаемости слов *x* и *y*, а *N* — объем коллекции.

Важно отметить, что ведущей мерой в данном исследовании выступала мера *DICE* [Daudaravicius, 2010]. В соответствии со значением этой меры определяется степень связанности рассматриваемых сочетаний слов.

2. Результаты

Выбранные статистические меры дают прекрасные результаты для выделения фразеологических оборотов, характерных для стиля автора [Ягунова и Пивоварова, 2012].

В рамках исследования были проанализированы наборы двухсловных сочетаний, полученные на материалах двух коллекций разных авторов.

В качестве примера приведем некоторые фразеологические единицы, выделенные из коллекций текстов разных авторов.

В текстах Дарьи Донцовой были выделены следующие фразеологические единицы: «*сложа руки*», «*ничтоже сумняшеся*», «*свернуться калачиком*», «*шельму метить*», «*лясы точить*», «*ковать железо*», «*клином вышибать*», «*угрызение совести*», «*слава Богу*» и др.; в текстах Галины Куликовой — «*ничтоже сумняшеся*», «*<накрыться> медным <тазом>*», «*свернуться калачиком*», «*пруд пруди*», «*сбавить обороты*», «*<до> белого каления»*, «*раздуть ноздри»* и др.

Важно отметить, что ряд фразеологизмов встречается в текстах всех рассматриваемых авторов, например: «*ничтоже сумняшеся*», «*свернуться калачиком*», «*пруд пруди*».

Корпус детективов Дарьи Донцовой содержит 29% фразеологизмов и фразеологических единиц, 46% (из 174) — предикативные конструкции.

Корпус детективов Галины Куликовой содержит 32% фразеологизмов и фразеологических единиц, 33% (из 142) — предикативные конструкции, 13% —предикативные с глаголом в вершине.

Заключение

В докладе планируется привести подробные результаты сопоставительного анализа, полученные на материале текстов женских детективов разных авторов.

Результатом такой работы будут данные о фразеологических единицах, типичных для данного жанра, стилевых характеристиках и особенностях их употребления: как для жанра женского детектива в целом, так и для конкретных авторов.

*Литература*

*Daudaravicius V.* Automatic identification of lexical units // Computational Linguistics and Intelligent text processing CICling-2009, Meksikas, Meksika. 2010.

*Manning C., Schutze H.* Collocations // Manning C., Schutze H. Foundations of Statictical Natural Language Processing, 2002. P. 151—189.

*Ягунова Е. В., Пивоварова Л. М.* Экспериментально-вычислительные исследования художественной прозы Н. В. Гоголя // XLII Виноградовские чтения в МГУ «В. В. Виноградов о художественном тексте»: Материалы. М., 2012. URL: http://www.webground.su/data/lit/pivovarova\_yagunova/Experimentalno-vychislitelnyie\_issledovaniya\_prozy.pdf

E. V. Yagunova, A. N. Savina

**Phraseologismes-collocations in female detective story**

This paper presents a method based on the statistical measures to the study  
of phraseological units, which are typical of Russian female detective story discourse. The main purpose of the research is the extraction of collocations — potential phraseological units — and their analysis. Our decisions were illustrated by the results of this research on the basis of two corpora: texts of (1) Daria Dontsova and (2) Galina Kulikova.