075-Sokolova_E-Kononenko_I:_Layout 1 13.05.2008 22:55 Page 482

ПРОБЛЕМЫ ОПИСАНИЯ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ В ВИДЕ ОНТОЛОГИИ ДЛЯ ПОРТАЛА ЗНАНИЙ[1]

PROBLEMS OF DESCRIBING COMPUTATIONAL LINGUISTICS IN ONTOLOGY OF A KNOWLEDGE PORTAL

Соколова Е.Г. *(minegot@rambler.ru), Российский государственный гуманитарный университет*
Кононенко И.С. *(irina_k@cn.ru),* ***Загорулько Ю.А.*** *(zagor@iis.nsk.su)*
Институт систем информатики СО РАН, Новосибирск

В этой статье мы обсуждаем проблемы, которые возникли в связи с построением предметной онтологии для научного направления, связанного с компьютерным моделированием языка, преобразований текстов и звучащей речи.

Введение

Существует два подхода к моделированию области исследований – на основе автоматического извлечения терминологии и путем ручного построения онтологий. Разработка портала знаний по компьютерной лингвистике базируется на втором подходе, но предполагает и создание терминологического словаря-тезауруса изучаемой области, поскольку тезаурус (терминологический ресурс, реализованный в виде словаря понятий и терминов со связями между ними [6]) служит инструментом поиска информации в Интернете с целью пополнения информационного содержания портала. При разработке портала по компьютерной лингвистике мы исходили из общей концепции портала знаний, примененной ранее к таким областям знаний, как археология и этнография ([2, 3]).

В основе создания портала знаний по компьютерной лингвистике лежат идеи и принципы, изложенные в работах [7, 8], где в качестве информационной модели портала используется онтология. Под онтологией понимается формальная спецификация концептуальной модели предметной области ([1, 13]).

С содержательной точки зрения, онтология портала представляет систему понятий и отношений, необходимых для описания как научной деятельности и научного знания в целом, так и конкретной области знаний. В связи с этим онтология портала разделяется на метаонтологию и предметную онтологию. В этой статье мы обсуждаем проблемы, которые возникли в связи с построением предметной онтологии для научного направления, связанного с компьютерным моделированием языка, преобразований текстов и звучащей речи. Проблем этих оказалось больше, чем при построении онтологии такой классической гуманитарной науки, как археология. Формулирование проблем скорее способствует выявлению особенностей новой науки, чем помогает созданию онтологии.

Ниже в каждом разделе рассматривается одна конкретная проблема, начиная с названия научного направления и кончая свойствами, связанными с формальными требованиями представления на портале.

Название научного направления

Проблемы начинаются с названия научного направления и установления его границ. Наиболее распространенным названием является термин «компьютерная лингвистика» (КЛ) – калька с английского «Computational Linguistics». КЛ обозначает компьютерные исследования в области лингвистики и ее компьютерных приложений[2]. Термин КЛ вытесняет термин «Прикладная Лингвистика» (ПЛ), содержанием которого является использование моделей языка в прикладных целях. ПЛ у́же, чем КЛ, но шире, чем собственно приложения. Для названия учебной дисциплины в системе высшего образования также используется термин Автоматическая Обработка Естественного Языка (АОЕЯ) (калька с английского «Natural Language Processing» (NLP)). Этот термин не получил у нас широкого распространения. NLP обычно используется для обозначения раздела, посвященного ЕЯ системам, на конференциях по искусственному интеллекту (ИИ). Недостаток этого термина по сравнению с КЛ состоит в том, что он не содержит ключевого для данной области слова «компьютер». Роль компьютера в КЛ отличается от его роли в других науках, где он является только средством усиления и интеллектуали-

[1] Работа выполняется при финансовой поддержке РГНФ (проект № 07-04-12149).

[2] Характерно название конференции Диалог, которое в 1995–2001-х годах звучало как «компьютерная лингвистика и ее приложения», а в 2002 г. изменилось на «компьютерная лингвистика и интеллектуальные технологии», включив в сферу КЛ также и приложения, или прикладные системы (applications)

075-Sokolova_E-Kononenko_I:_Layout 1 13.05.2008 22:55 Page 483

зации средств моделирования. В ЛЭС дается определение научной дисциплины Автоматическая Обработка Текста (АОТ) как науки, которая занимается «преобразованием текстов на естественном или искусственном языках с помощью ЭВМ» ([10], стр. 14). Из него следует, что понятие «компьютер» входит в само определение науки. Термин АОТ у́же, чем КЛ, так как не включает исследования по моделированию звучащей речи. В последнее десятилетие появился термин “Linguistic Technology” (LT) и параллельный русский термин «лингвистические технологии».

Из сказанного следует, что наиболее подходящим является термин КЛ, самый широкий, общеупотребительный и содержащий ключевое слово «компьютер». КЛ включает следующие направления: 1) АОТ – первичную для КЛ область преобразования письменных текстов с помощью компьютера, включая приложения; 2) моделирование звучащей речи, включая приложения, 3) мультимодальные приложения, совмещающие обработку текста, звучащей речи и/или изображений; 4) создание языковых ресурсов.

Источники информации

Структуры описания областей знаний обычно создаются в учебниках, которые систематизируют исследования и на этой основе рисуют по возможности полную картину состояния данной области. Они являются выражением опыта авторов учебников, являющихся хорошими специалистами описываемых направлений. Особенность КЛ состоит в том, что она систематизируется не только в виде учебников, но и в виде обзоров, представляющих область КЛ как полную совокупность составляющих ее более узких направлений. Такие обзоры пишутся коллективами авторов – специалистов узких направлений. Для классических наук пишутся только обзоры по узким разделам.

В последние 30 лет исследования по КЛ неоднократно и в разных аспектах описаны в зарубежных учебниках, например, [15, 20, 21, 24], и обзорах [16, 17, 18, 23]. В российских изданиях эта область представлена двумя учебниками – [4] и [12], оба по ПЛ. Российский обзор представлен трехтомником «Искусственный интеллект» [9], содержащим информацию по системам ИИ и КЛ. Западные учебники и обзоры описывают КЛ в разных аспектах – основанный на лингвистике разделяющий теории и приложения подход [23], технологический подход [17], преобладание метода (регулярных выражений) [21], систематизация практических инструментов и методов создания прикладных систем [18].

Изучение учебников и обзоров экспертами предполагает и другие методы сбора информации, необходимые для получения актуальной картины области исследований или ее конкретных разделов. Важнейшим источником информации для эксперта и одновременно источником данных, составляющих информационное пространство портала, является Интернет, где широко представлены информационные ресурсы, отражающие процессы и результаты научных исследований в данной области. К числу наиболее важных результатов в области КЛ относятся лингвистически ориентированные программные средства и лингвистические ресурсы. Информационное наполнение портала такого рода данными невозможно без автоматизации работы эксперта по поиску информационных ресурсов в сети. Эта проблема решается путем создания поискового робота, использующего специальные словари-тезаурусы по КЛ, позволяющие находить, индексировать и предъявлять эксперту странички, относящиеся к данной области, и даже строить гипотезы по их отнесению к тому или иному разделу онтологии КЛ. (см. [5]).

Отличие отечественной парадигмы КЛ от западной

Зарубежные исследования базируются на американских и европейских школах КЛ, для которых характерна разработка формализмов для моделирования лингвистических явлений, а советские и российские исследователи занимались в основном формализацией конкретных языковых явлений. Это противопоставление было отмечено С.А. Шаровым в [14]. Наиболее ярко оно выразилось в создании формализмов для синтаксических анализаторов. Разрабатывались простейшие методы (pattern matching, chunks) и синтаксические формализмы, способные реализовать синтаксический анализ английского языка с постепенным углублением и расширением их возможностей: PSG → трансформационная грамматика (увеличение мощности за счет добавления глубинных структур, в частности, моделирование залоговых трансформаций, преимущественно синтез) → расширенные сети переходов ATN (применимость к анализу) → APSG (возврат к продукциям как преодоление громоздкости ATN) → … ([24]). Процесс завершился изобретением унификационного механизма и созданием унификационных грамматик, которые позволяют моделировать комплексный анализ предложений, используя все формализуемые виды информации, в отличие от предыдущего этапа, основывающегося на сочетаниях синтаксических классов. Из отечественных методов можно отметить два – фильтровый анализ, развивавшийся в работах О.С. Кулагиной и Л.В. Иорданской, и подход на основе продукций, соединяющих различные виды информации (синтагмы) в

075-Sokolova_E-Kononenko_I:_Layout 1 13.05.2008 22:55 Page 484

системе ЭТАП, близкий APSG. Для других реализаций создавались специальные средства, например, язык стандартных операторов (ФРАП).

Российские и западные исследования по грамматике плохо сочетаются в одной онтологии. В частности, возникает вопрос о ценности для российского читателя различных стратегий и алгоритмов реализации PSG, которым было посвящено много западных исследований в 70–80-х годах. С другой стороны, современный этап развития науки, основанный на технологиях, в значительной степени обесценивает большой пласт отечественных исследований, представленных в трехтомнике по ИИ и, частично, в учебниках Гердта и Баранова.

По этому же принципу противопоставлены и лексические исследования. На западе они выразились в создании общедоступных интернет-ресурсов, таких как WordNet, FrameNet и других лексико-семантических баз, например, Corelex – база, описывающая регулярную многозначность существительных английского языка. Появление таких ресурсов вызывало к жизни проекты создания аналогичных ресурсов для других языков и закрепление стандартов. Российские исследования либо сохраняли академический статус, например, Новый объяснительный словарь синонимов русского языка ([11]), либо не приводили к созданию ресурсов, т.е. результаты реализовывались такими средствами, которые исключали их использование другими исследователями и в других системах, например, формализованные описания З.М.Шаляпиной, ЭТАП-2. Приходится сравнивать западные ресурсы с российскими системами и теоретическими описаниями, причем первые отличаются формализованностью и широтой, вторые – глубиной. Появление Национального корпуса русского языка – важный шаг к сближению научных тенденций.

Таким образом, создание онтологии и наполнение портала по КЛ означает необходимость сопоставления, отслеживания соответствий и “наведения мостов“ между подходами, технологиями, методами исследования и терминологией КЛ, традиционно сложившимися в отечественной науке и за рубежом, и представления их в едином “межнациональном“ информационном пространстве. Один из путей решения этой проблемы – связывание результатов и предметов исследования ассоциативными отношениями. Например, лексико-семантический ресурс WordNet и Новый объяснительный словарь синонимов русского языка являются результатами исследований, направленных на разные объекты (разные ЕЯ, английский и русский, и совершенно разные объемы языкового материала: 99808 синсетов версии 1.6 WordNet vs. 354 синонимических ряда Нового словаря), но могут быть связаны по сходству предмета исследования (лексико-семантическая система/подсистема ЕЯ).

Изменчивость направления

В отличие от классических наук, КЛ развивалась по этапам, определенным не внутренней логикой развития, а внешними факторами. КЛ неразрывно связана с компьютером, и ее развитие шло параллельно развитию вычислительной техники, в истории которой выделяется несколько этапов (поколений), связанных с используемыми средствами: 40-е годы 20 века – электронные лампы; 50-е – полупроводниковые приборы; 60-е – интегральные микросхемы; 70–90-е – большие интегральные микросхемы – чипы, которые привели к созданию информационного общества (Интернет, массивы цифровых информационных данных, включая корпуса текстов); и, наконец, происходящий в настоящее время переход к нанотехнологиям.

Каждый из этапов характеризуется определенным качеством развития КЛ и определенным фокусом ее интересов: начальный – МП, 50–60-е годы – осознание непригодности традиционных лингвистических описаний и поиски общей теория языка, пригодной для использования в КЛ, 70–90-е – синтаксические формализмы, семантика, моделирование структуры предложения, 90–2000-е – корпуса, статистика, структура текста, NLG. Проблема состоит в выборе того, что должно быть представлено на портале, потому что наряду со знаниями и информацией, с течением времени сохраняющими свою ценность для исследователей и разработчиков, есть информация, привязанная ко времени и интересная преимущественно в историко-научном плане (для изучающих историю данной дисциплины и преподавателей). Например, Джорджтаунский эксперимент 1954 года имеет значение только как первый в истории опыт по машинному переводу.

Движущей силой КЛ являются актуальные приложения, например, сейчас это – традиционный для КЛ, но сохраняющий свою актуальность, машинный перевод, а также порожденные Интернетом полнотекстовый информационный поиск и извлечение информации.

Особенность теоретической базы

В отличие от классических наук, которые движутся от общего к частному, от общей теории к разработке конкретных ее разделов, развитие теоретической базы КЛ в течение нескольких десятилетий определялось развитием компьютеров и параллельным ему состоянием общественных потребностей. В 70–80-х годах были созданы общие модели языка – наиболее известные у нас трансформационная грамматика Н. Хомского и модель

075-Sokolova_E-Kononenko_I:_Layout 1 13.05.2008 22:55 Page 485

«Смысл-Текст» И.А. Мельчука, – а также практически неизвестная у нас системно-функциональная грамматика М. Хэллидея. В настоящее время модель Хомского развилась в лингвистическую теорию «управления – связывания»; модель Мельчука была реализована в виде системы МП ЭТАП-2, систем генерации текстов фирмы CoGenTex и ресурса языковой реализации RealPro. Модель Хэллидея была реализована в виде генератора Penman и затем в виде среды для разработки многоязыковых генераторов текстов KPML. Однако эти системы пока остаются в разряде экспериментальных, не обеспечивая создания эффективных приложений, что является главной целью КЛ. Одна из причин состоит в том, что это – модели языка, а языковые теории и модели не объясняют структуры текста и закономерностей обмена информацией, а именно этот аспект важен для приложений.

Как реакция на углубляющиеся потребности возникли модели структуры текста, в частности, RST и модели, основанные на DRT. Они продвинули понимание структуры текста, связали ее со структурой предложения и коммуникативной целью автора текста. Но они описывают способы подачи информации, а не саму информацию и знания. Пробел в области знаний отчасти компенсируется лингвистическими ресурсами на основе предметных онтологий – WordNet (имеются толкования, иерархия толкований), FrameNet (имеется структура типов процессов и их иерархия). В явном виде знания моделируются в виде общих онтологий, таких как CYC, SUMO, Sowa's ontology.

В 80-х годах возникло новое, эмпирическое, направление в КЛ, основанное на статистических исследованиях на базе корпусов текстов и звуковых корпусов. Его эмпиризм состоит в том, что статистическая модель прямо зависит от материала, на котором она построена и нет явных данных, чтобы объяснить ее действие. (Заметим, что корпуса являются не только основой статистических методов, но и средством верификации созданных вручную ресурсов и теоретических описаний КЛ и традиционной лингвистики.) Статистические методы представляют собой отдельное направление КЛ и описываются в отдельных учебниках, например, [22], но в приложениях в последние 15 лет происходит их слияние с традиционными методами лингвистического анализа «по правилам». Примером этого процесса могут служить вероятностные порождающие грамматики, гибридные методы в системах машинного перевода, использование вероятностных характеристик в алгоритмах морфологического анализа, разрешения неоднозначности, в том числе референциальной, и обработки дискурса.

Процесс интеграции инструментария и методов находит отражение в литературе: [18, 23]. Учитывая это обстоятельство, наиболее общая классификация методов исследования в онтологии КЛ строится не от используемого научного инструментария (логические, статистические, лингвистические), а исходя из предмета/задачи исследования (методы синтаксического анализа, методы разрешения лексической многозначности и т.д.).

В настоящее время, когда в результате развития вычислительной техники внешние факторы, определявшие развитие КЛ в течение нескольких десятилетий исчезли, КЛ впервые получает возможность развиваться по своим внутренним законам и реализовать тягу к научно-обоснованным решениям наряду с эмпирикой.

Мультимодальность

Текст и звучащая речь – основные два вида передачи информации, которые в течение некоторого периода были двумя не зависимыми друг от друга объектами исследования. Речевые технологии развивались как чисто инженерная дисциплина, главным образом направленная на решение практических задач, таких как распознавание голоса диктора. Сейчас в КЛ разрабатываются системы, в которых виды передачи информации на ЕЯ комбинируются: например, в системах общения с компьютерами. Создаются системы, обучающие глухих детей говорить, обучающие читать и понимать прочитанное, системы-гиды и др. Характерно, что в них, кроме текста и звучания, моделируется и зрительная информация. Компьютерная система имеет лицо, мимику, жесты. Последнее вызвало к жизни новое течение КЛ – моделирование невербальной коммуникации, – в котором, кроме жестов, сопровождающих общение людей, исследуются и формализуются языки немых. Таким образом, мультимодальность КЛ проявляется не только в соединении текстовых и речевых технологий, но также текстовых и графических, например, исследования, связанные с поиском на изображении объекта, упомянутого в сопутствующем изображению тексте, см. [19].

Форма «портал знаний», степень детализации и виды размещаемой информации

Форма «портал знаний» накладывает определенные требования на способ представления онтологии КЛ, степень ее детализации и виды размещаемой информации (см. рис.1). Так, формальное деление на классы и экземпляры понятий (объекты), нередко обусловленное дополнительными соображениями удобства навигации и поиска информации на портале, вынудило представить иерархию разделов КЛ с использованием не только отношения таксономии между разделами, представленными как классы, но и отношения партономии между разделами-объектами. Например, показанные на рисунке объекты раздела «Моделирование языка и языковой деятель-

075-Sokolova_E-Kononenko_I:_Layout 1 13.05.2008 22:55 Page 486

ности» связаны в иерархию отношением партономии: раздел «Речевые технологии» включает «Распознавание речи» и «Синтез речи», а раздел «Автоматическая обработка» текста включает «Понимание текста» и «Генерацию текста».

Существуют трудности, связанные с определением необходимого и достаточного уровня развития/проработки и степени детализации онтологии, которая, не являясь полным справочником по данной предметной области, должна представить и систематизировать ее основные элементы и дать возможность указать информационные источники (ресурсы), которые содержат более подробную информацию.

Возникают также проблемы с представлением «научной деятельности», подробностью отражения конференций и семинаров, персональных данных, а также проектов, не завершившихся общедоступными результатами.

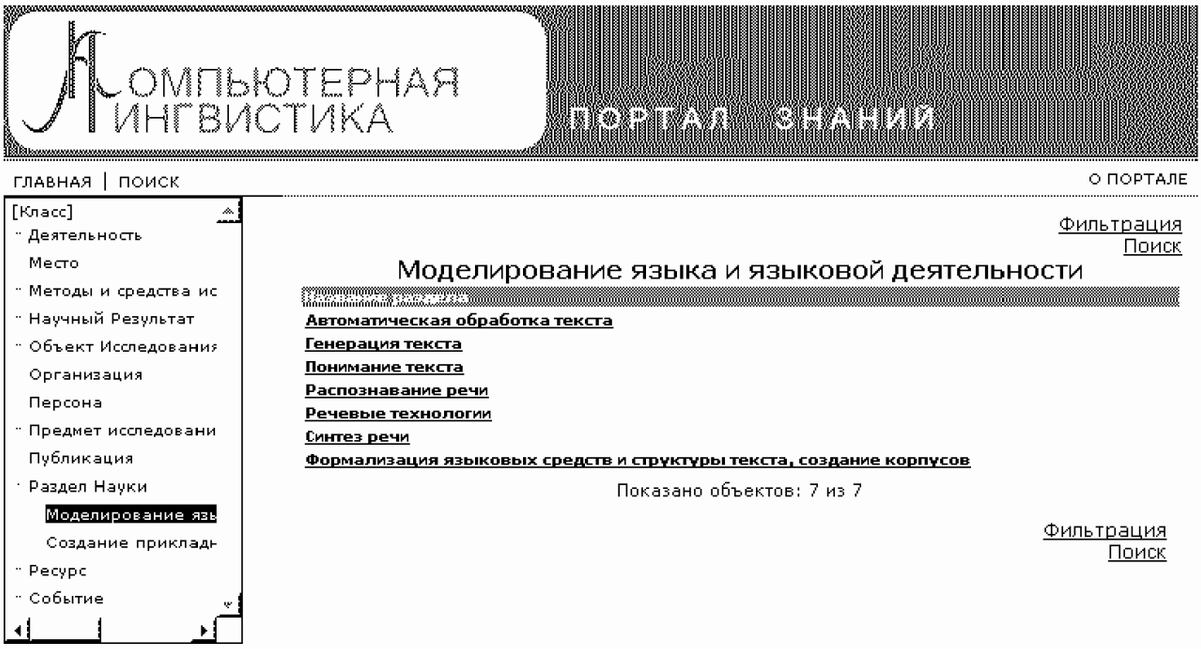

Рис.1. Фрагмент онтологии КЛ на портале знаний

Заключение

Особенности предметной области КЛ, в частности, ее сложность и разветвленность, требуют при разработке онтологии КЛ привлечения квалифицированных экспертов и организации совместной работы специалистов в отдельных направлениях КЛ. По сути, создавая онтологию для портала знаний по КЛ, мы попытались наметить ядро, которое включает наиболее общие классификации и нуждается в развитии и дополнении. В частности, в онтологии в ее текущем виде очень эскизно представлены речевые технологии и очень фрагментарно – обширная область статистических методов, едва намечены методы, связанные с применением нейронных сетей и т.д.

В связи с вышеизложенными проблемами мы нуждаемся в обсуждениях и участии коллег. Поэтому мы сочли возможным представить эскиз портала знаний, создаваемого на основе онтологии КЛ, в Интернете по адресу http://speedy.iis.nsk.su/cl.

Список литературы

1. Александровский Д.А., Кормалев Д.А., Кормалева М.С., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Развитие средств аналитической обработки текста в системе ИСИДА-Т //Труды 10-й национальной конференции по искусственному интеллекту с международным участием КИИ’2006. М.: Физматлит, 2006. Т. 2. С. 555-563. http://www.raai.org/resurs/papers/kii-2006/doklad/Alexandrovsky.doc

075-Sokolova_E-Kononenko_I:_Layout 1 13.05.2008 22:55 Page 487

2. Андреева О.А., Боровикова О.И., Булгаков С.В., Загорулько Ю.А., Сидорова Е.А., Циркин Б.Г., Холюшкин Ю.П. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии //Труды 10-й национальной конференции по искусственному интеллекту с международным участием КИИ'2006. М.: Физматлит, 2006. Т.3. С.832-840.

3. Археологический портал знаний http://www.sati.archaeology.nsc.ru/classarch2/

4. Баранов А.Н. Введение в прикладную лингвистику. Серия «Новый лингвистический учебник» //М.: Эдиториал УРРС, 2001.

5. Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний //Труды международной конференции Диалог'2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. С. 65-70.

6. Журавлев С.В., Добров Б.В. УИС «РОССИЯ». Автоматическое тематическое индексирование полнотекстовых документов //Материалы научно-практической конференции «Проблемы обработки больших массивов неструктурированных текстовых документов», 2001. http://www.fep.ru/text/dataarrays03.html

7. Загорулько Ю.А., Боровикова О.И. Методологические проблемы построения и использования онтологий в портале научных знаний //Труды IX Международной конференции «Проблемы управления и моделирования в сложных системах». Самара: Самарский Научный Центр РАН, 2007. С. 447-454.

8. Загорулько Ю.А., Боровикова О.И., Загорулько Г.Б. Организация содержательного доступа к информационным ресурсам на основе онтологий //Труды 9-й Всероссийской научной конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" RCDL'2007. Переславль-Залесский, 2007. Т. 1. С. 217-224.

9. Искусственный интеллект. Справочник в 3 кн. . Под ред. Попова Э.В //М.: Радио и связь, 1990.

10. Лингвистический энциклопедический словарь. Под ред. Ярцевой В.Н. //М.: Советская энциклопедия, 1990.

11. Новый объяснительный словарь синонимов русского языка. Под общим руководством акад. Апресяна Ю.Д. //Второе издание. Москва, Вена: Языки славянской культуры: Венский славистический альманах, 2004.

12. Прикладное языкознание. Учебник. (ред. Гердт А.С.) //СПб., 1996.

13. Смирнов С.В. Онтологии в прикладных интеллектуальных системах: прагматический подход //Труды 9-й национальной конференции по искусственному интеллекту с международным участием КИИ'2004. М.: Физматлит, 2004. Т.3. С.1059-1067.

14. Шаров С.А. Средства компьютерного представления лингвистической информации //1996. http://www.ksu.ru/eng/science/ittc/vol000/002/

15. Allen J. Natural Language Understanding //Benjamin Cummings, 1995.

16. Butter, C.S. (ed.) Computers and written text //Blackwell, Oxford (UK) and Cambridge (USA), 1992.

17. Cole Ronald (ed.) Survey of the state of the Art in Human Language Technology //1996. (http://cslu.cse.ogi.edu/HLTsurvey/).

18. Dale R., Moisi H., Somers H. (eds.) Handbook of Natural Language Processing //Marcel Dekker, New York, 2000.

19. Deschacht K., Moens M-F. Text analysis for automatic image annotation //The 45th Annual meeting of the Association for Computational Linguistics, Prague, June 2007. http://acl.ldc.upenn.edu/P/P07/P07-1126.pdf

20. Grishman Ralph. Computational linguistics. An introduction //Cambridge, 1986.

21. Jurafsky Daniel, Martin James H. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics //Prentice Hall, 2000.

22. Manning Ch., Schütze H., Foundations of Statistical Natural Language Processing //MIT Press. Cambridge, MA, May 1999. http://nlp.stanford.edu/fsnlp/

23. Mitkov Ruslan (ed.) The Oxford handbook of computational linguistics //N.Y.: Oxford university press, 2003.

24. Salton Gerard. Automatic text processing. The transformation, analysis, and retrieval of information by computer //Addison-Wesley Publishing Company, Inc., 1989.