нашу планету (1021 куб. м), Солнечную Систему (1037 куб. м), Галактику, в которой мы живем и даже ближайшую галактику Андромеда (свет от нее идет 2 млн лет); Наудачу выбрать нужную песчинку в этой невообразимо грандиозной куче можно как раз с той же вероятностью 10-177. Понятно, что практически подобный результат невозможен. Следовательно, можно определенно сделать вывод о неслучайности совпадений. Сомневаться в этом равносильно непониманию написанного в статье русским языком текста.

Выводы. Ключевые слова и словосочетания темы А.С. Пушкина обладают весьма примечательным аномальным свойством квантованности: их числовое содержание кратно числу 6666.

Весьма малая вероятность случайности подобных совпадений убедительно свидетельствует об участии еще неизвестных науке могущественных сил в формировании языка людей.

Великий и могучий русский язык, язык Пушкина был выбран специально для данной миссии: только в нем есть необычайное количество совпадений.

Библиографический список:

1. *Некрасов, С.А.* Аномалия гематрии на тему А.С. Пушкина [Электронный ресурс] / С.А. Некрасов // Портал научно–практических публикаций. – Режим доступа: http://portalnp.ru/2013/11/1217 (дата обращения: 12.04.2014).

2. *Некрасов, С.А.* Хронологические, информационные и биофизические аномалии системы [Текст] / С.А. Некрасов. – ЕТА: Изд-во Palmarium Academic Press (Германия), 2012. – 259 с.

3. *Большая* Советская энциклопедия [Текст]. – М.: Советская энциклопедия, 1977.

4. *Большая* энциклопедия Кирилла и Мефодия [Текст]. – 7-е изд. – М., 2003.

5. *Кирпичников, А.И.* Биография А.С. Пушкина [Текст] / А.И. Кирпичников // Энциклопедический словарь Брокгауза и Ефрона. – М., 1890-1907.

6. *А.С. Пушкин* (биография) [Электронный ресурс] // Википедия. – Режим доступа: https://ru.wikipedia.org/wiki/Пушкин (дата обращения: 12.05.2014).

7. *Последняя* дуэль и смерть А.С.Пушкина [Электронный ресурс] // Википедия. – Режим доступа: https://ru.wikipedia.org/wiki (дата обращения: 12.05.2014).

8. *Корн, Г.* Справочник по математике [Текст] / Г. Корн, Т. Корн. – М.: Наука, 1978. – 832 с.

9. *Nekrasov, S.A.* Numerically–linguistic Anomalies of Gematria. Sociology and Anthropology / S.A. Nekrasov. – 2013. – Vol. 1(3). – P. 158-163.

УДК 81’33
ББК 81.1

В.А. Яцко

ПРЕДМЕТНАЯ ОБЛАСТЬ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

Дается определение термина «компьютерная лингвистика». Рассматриваются основные понятия предметной области, приводится классификация лингвистического программного обеспечения. Особое внимание уделяется соотношению терминологии компьютерной лингвистики и теоретической лингвистики.

Ключевые слова: *обработка единиц естественного языка; алгоритмы и программы; классификация; компьютерная лингвистика; терминология*

V.A. Yatsko

TOWARDS A DEFINITION OF THE SUBJECT OF COMPUTATIONAL LINGUISTICS

Definition of the term «computational linguistics» is given. Main notions of its subject field are described, and classification of linguistic software is suggested. The emphasis is made on correlation between terminology in computational linguistics and in theoretical linguistics.

Key words: *natural language processing; algorithms and programs; classification; computational linguistics; terminology*

Характерная особенность нашего времени – проникновение во все сферы жизни общества технологий обработки единиц естественного языка в рамках глобального процесса информатизации. Миллионы пользователей во всем мире посылая запросы в информационно-поисковые системы, отдавая голосовые команды телефонам, выполняя автоматическое

реферирование текстов, не подозревают, что это стало возможным в результате развития предметной области, в рамках которой проводятся исследования и разработки алгоритмов и программ обработки текстов на естественном языке.

Для обозначения этой предметной области используются разные термины. В зарубежной науке наиболее распространен термин *компьютерная лингвистика* (*computational linguistics*). В настоящее время функционирует *Ассоциация компьютерной лингвистики* (*Association for Computational Linguistics*) – международная организация, объединяющая специалистов, которые исследуют проблемы автоматической обработки естественного языка [Association, 2014]. Ассоциация проводит международные научные конференции, совместно с Массачусетским институтом технологий издает журнал *Computational linguistics* – одно из ведущих научных изданий предметной области. В университетах Европы и США созданы факультеты, кафедры, лаборатории и программы компьютерной лингвистики, проводятся исследования и готовятся специалисты соответствующего профиля, защищаются магистерские и докторские диссертации. В Германии в Штутгартском университете открыта кафедра основ компьютерной лингвистики [Chair, 2012], в Гейдельбергском университете – Институт компьютерной лингвистики [Institut, 2014]; в Дюссельдорфском университете – кафедра компьютерной лингвистики [Philosophische fakultät, 2014]. Лаборатории компьютерной лингвистики работают в США (в Корнельском, Техасском, Калифорнийском университетах), Канаде, Германии, Японии, Италии, Корее, Таиланде [TCL, 2005]. В Йельском и Стэнфордском университетах созданы исследовательские группы по проблемам компьютерной лингвистики. В Стэндфордском университете работает группа автоматической обработки естественного языка (The Stanford natural language processing group), члены которой разрабатывают алгоритмы машинного перевода, вероятностного аннотирования и синтаксического анализа, извлечения биомедицинской информации, создания грамматик, разработки вопросно-ответных систем [Computational linguistics, 2014]. В Массачусетском институте технологий студентам предлагается курс *Advanced natural language processing* [Advanced, 2001-2014].

Термин *обработка естественного языка* (*natural language processing*) часто используется в зарубежной литературе в одном ряду с термином *компьютерная лингвистика*. Как мы полагаем, данным термином можно обозначить объект исследования, «компьютерная лингвистика – дисциплина, которая занимается проблемами автоматической обработки естественного языка» [Hess, 2005. S. 4].

В отечественной науке изучение проблем автоматической обработки единиц естественного языка соотносится с предметной областью прикладной лингвистики, причем значение термина «прикладная лингвистика» отличается от интерпретации термина-кальки *applied linguistics* в англо-американской литературе, да в западноевропейской науке в целом. Вплоть до последнего времени под прикладной лингвистикой понималась дисциплина, занимающаяся проблемами методики обучения языкам: «До настоящего времени основная часть разработок в прикладной лингвистике была посвящена обучению и изучению языков, в первую очередь английского как иностранного или второго языка» [Naves, 2002, p. 4]. В последнее время наметилась тенденция расширения предметной области прикладной лингвистики, в которую также включаются проблемы логопедии и перевода, ср. определение в Оксфордском словаре: «Отрасль языкознания, занимающаяся практическим применением исследований языка, например, в обучении языкам, переводе и логопедии» [Definition, 2012].

Такая интерпретация существенно отличается от понимания прикладной лингвистики в отечественной науке. Ю.В. Рождественский считает, что «задача прикладного языкознания состоит в том, чтобы ввести новые материалы речи, наметить наиболее эффективные пути речевой коммуникации на базе новой техники, утвердить норму языка путем обучения, распространить новые виды речевой коммуникации путем обучения новым видам текстов (создания этих текстов, их обучения и применения) [Рождественский, 1990, с. 215]. По его мнению, прикладная лингвистика включает три основных направления: лингводидактику, лингвосемиотику, информационное обслуживание [Там же. С. 299]. В

информационное обслуживание входят такие области деятельности, как библиотечное, архивное, канцелярское дело, информационный поиск, реферирование, составление информационных словарей, двуязычный перевод, автоматизированные системы управления. Для обозначения «языковедческой части» теории информационного обслуживания Ю.В. Рождественский предлагает термин «лингвистическая информатика» [Там же. С. 354].

Очевидно, что такое понимание прикладной лингвистики значительно шире того, которое принято за рубежом и охватывает ряд областей, которые в англо-американской традиции относятся к другим предметным областям. Так, проблемы информационного поиска обычно включаются в предметную область информационной науки (*information science*), а проблемы разработки автоматизированных систем управления относятся к компьютерной науке (*computer science*) [Information science, 2012]. Вместе с тем, логопедия, с точки зрения интерпретации, предложенной Ю.В. Рождественским, не должна входить в прикладное языкознание, а является приложением языкознания, поскольку в этом случае языковые данные используются «для решения практической задачи, за которую ответственна другая область науки или практики» [Рождественский, 1990, с. 298]. В работе А.Н. Баранова, напротив, указывается, что под прикладной лингвистикой следует понимать «деятельность по приложению научных знаний об устройстве и функционировании языка в нелингвистических научных дисциплинах…» [Баранов, 2001, с. 7].

Кроме наличия различных, в том числе и противоречивых интерпретаций, особенностью термина «прикладная лингвистика» также является то, что он обозначает целый ряд видов деятельности не связанных напрямую с автоматической обработкой текстов, в том числе лингвистическую семиотику, лингводидактику, терминологию, которые указаны в паспорте специальности 10.02.21 Прикладная и математическая лингвистика [Паспорта, 2002-2012].

В данной статье мы попытаемся рассмотреть основные понятия и структуру компьютерной лингвистики, раскрыть ее междисциплинарную сущность. Особое внимание будет уделено сопоставительному рассмотрению терминов компьютерной лингвистики и теоретической лингвистики, что позволит выявить специфику рассматриваемой предметной области.

В прототипическом представлении термин «компьютерная лингвистика» связан, с одной стороны, с наукой о естественном языке, а с другой – с семантическими компонентами «компьютеры», «программы», «интернет», что позволяет ограничить предметную область проблемами разработки лингвистического программного обеспечения, аппаратных средств и технологий. Под лингвистическими аппаратными средствами мы понимаем вычислительное оборудование, специально предназначенное для обработки текстов на естественном языке. Такое оборудование широко используется в системах распознавания артикуляционных и акустических параметров устной речи, а также в оптических системах распознавания символов. Под лингвистическим программным обеспечением мы понимаем программы, приложения и системы, на входе которых – текст на естественном языке, и которые функционируют на основе лингвистических алгоритмов – алгоритмов, применяющийся для обработки единиц естественного языка.

В основе функционирования многих лингвистических приложений лежит процесс лексической декомпозиции текста, в результате которой во входном тексте распознаются токены и на выходе генерируется их список. *Токен* можно определить как последовательность буквенных и/или цифровых символов, отделенную слева и справа знаками форматирования текста и/или препинания. Разбивка текста на токены называется *токенизацией*, а программы, выполняющие токенизацию, – *токенайзерами.* В тексте *«By-by, dearie», he smiled*» токен *By-by* распознается по кавычкам слева и запятой справа, токен *dearie* – по пробелу слева и запятой справа, токен *he* – по пробелам слева и справа, токен *smiled* – по пробелу слева и точке справа.

Как видно из приведенного примера, токены обычно совпадают со словами, поэтому термину *токен* соответствует термин *слово* в теоретической лингвистике. В инструкциях для пользователей и в интерфейсах лингвистического ПО достаточно часто используется термин *слово* (*word*), а не *токен* (*token*),

поскольку он более понятен и привычен. Однако, с точки зрения теоретической интерпретации, между двумя терминами имеются существенные различия. Интерпретация термина *слово* в языкознании обычно дается на основе соотношения между знаком, обозначаемым объектом и значением и представляется в виде известного семантического треугольника. В компьютерной лингвистике, как было показано выше, учитываются только знаки, выделяемые по формальным признакам. Соответственно, различаются и цели изучения этих единиц. В лингвистике проводятся исследования, направленные на толкование значений слов, разграничение окказиальных и узуальных словоупотреблений выявление новых значений и условий их актуализации. В компьютерной лингвистике рассматриваются статистические особенности распределения токенов в тексте, на основе которых разрабатываются формулы взвешивания, необходимые для выявления наиболее статистические значимых терминов (salient terms). В этой связи проводится разграничение между уникальными токенами и общими токенами. Термин *уникальный токен* обозначает токен без учета количества его повторов в тексте, а термин *общие токены* – количество токенов с учетом их частотностей. В Британском национальном корпусе [BYU-BNC, 2012] уникальный токен *the* повторяется 5 973 437 раз, т.е. дает 5 973 437 общих токенов, а в Корпусе современного американского английского [The Corpus, 2012] его частотность составляет 25 063 954. Таким образом, количество общих токенов, как правило, больше количества уникальных токенов. Это позволяет при взвешивании терминов использовать вероятностные величины и устранить зависимость весовых коэффициентов от размера текста. Вероятностный коэффициент для *the* в Британском национальном корпусе составляет 0,05973 (при размере корпуса 100 000 000 слов), а в Корпусе современного американского английского – 0,055 69 (размер корпуса – 450 000 000 слов). Разница в вероятностных величинах составляет около четырех тысячных, в то время как разница между сырыми частотностями – около девятнадцати миллионов. Процесс преобразования сырых частотностей с целью устранения зависимости от размера текста, а также приведения различных величин к единому виду в компьютерной лингвистике называется нормализацией. В качестве средства нормализации широко применяется логарифмирование по основанию 2. Двоичный логарифм от 5 973 437 равен 22,510 13, а от 25 063 954 – 24,579 11, что дает различие приблизительно в две целых, а не в девятнадцать миллионов.

Для систем, связанных с представлением смысла текста, в процессе лексической декомпозиции важно распознавать в качестве одного токена такие словосочетания, как географические названия, личные имена, сокращения, устойчивые сочетания. При разбивке на отдельные токены сочетаний *New York* или *N.A.T.O.* может не воспроизвестись или даже исказиться смысл текста. Потому в ряде лингвистических программ для распознавания сочетаний применяются специальные списки и дополнительные правила. При обработке текста [Missing, 2009] Essence [Sillanpää, 2009] приложение, предназначенное для автоматического реферирования, распознает в качестве отдельных токенов сочетания *Jamrul Hussain*, *Nilufa Begum*, *NEW YORK*. Программа статистического анализа AntConc [Laurence, 2011], обрабатывая тот же самый текст, разделяет все эти сочетания на отдельные токены. Такое различие объясняется разной функциональностью и пользовательской аудиторией. Программы статистического анализа выдают данные о частотностях единиц текста; их пользователями являются специалисты в области автоматической обработки текста, а также лингвисты, которые используют эти данные в своей профессиональной деятельности. Системы реферирования относятся к ПО общего назначения и предназначены для намного более широкого круга пользователей.

В языкознании словарный состав языка классифицируется по семантическим, синтаксическим, этимологическим, стилистическим критериям. В компьютерной лингвистике широко применяется классификация лексических единиц на знаменательные и служебные слова (стоп слова). В литературе [Tsz-Wai, 2005; Francis, 1982] в качестве основного признака стоп слов выделяется их равномерная распределенность по текстам, относящимся к разным жанрово-стилистическим группам. В любом достаточно большом тексте на английском языке наиболее частотными будут артик-

ли, местоимения, предлоги, союзы. Как отмечает У. Фрэнсис, 10 наиболее частотных слов английского языка дают от 20 до 30 % общих токенов. Удаление стоп слов позволяет существенно (в ряде случае почти на 40 %) уменьшить размеры лингвистических баз данных, повысить быстродействие и точность поиска. Вместе с тем, стоп слова используются в качестве одного из основных параметров в процессе автоматической классификации текстов: тот факт, что они встречаются в любых текстах, независимо от их жанрово-стилистических особенностей, позволяет провести сопоставительный анализ текстов и выявить особенности распределения стоп слов, присущие отдельным группам, категориям, жанрам текстов. Фильтрация стоп слов является важной процедурой обработки текста в информационно-поисковых системах и системах автоматической классификации текстов, которая выполняется на основе специальных списков стоп слов, либо алгоритмически. С целью фильтрации стоп слов часто применяют формулу TF*IDF, предложенную Дж. Солтоном и Ч. Янгом в 1973 г. [Salton, 1973], а также ее интерпретации [Yatsko, 2013]. В соответствии с формулой распределение терминов в анализируемом тексте сопоставляется с их распределением в коллекции текстовых документов; при этом наибольший вес получают термины, встречающиеся с наибольшей частотностью в данном документе, но редко встречающиеся в других текстовых документах коллекции, в то время как термины, встречающиеся в текущем документе и во всех текстах коллекции, получают нулевые коэффициенты. Таким образом, формула описывает определенную закономерность распределения текстовой информации. Основной проблемой, возникающей при использовании формулы TF*IDF, является неопределенность количественного и жанрово-стилистического состава коллекции текстов, с которой сопоставляется анализируемый текст. Одним из подходов к решению этой проблемы может быть использование зонального анализа текста на основе интерпретации закона Брэдфорда [Яцко, 2013].

Мы подробно остановились на процессе обработки лексических единиц текста, поскольку он наглядно демонстрирует междисциплинарные особенности предметной области, а лексическая декомпозиция является фундаментальным алгоритмом, который лежит в основе ряда алгоритмов, выполняемых на различных уровнях системы языка. На основе токенизации проводится морфологический анализ, аннотирование, фразовая декомпозиция, разбивка на н-грамы, клаузальная декомпозиция, разрешение анафоры.

С помощью алгоритмов морфологического анализа распознаются элементы морфологической структуры слова – корни, основа, суффиксы, окончания. К алгоритмам, широко применяемым на морфологическом уровне, относятся стемминг и лемматизация.

Цель стемминга – отождествить основы различных словоформ, имеющих одно значение. На входе стеммера – список токенов, на выходе – список их основ (стемм). Стемминг позволяет существенно повысить показатели точности и полноты поиска и широко используется в информационно-поисковых системах различных типов. В теоретической лингвистике под основой слова понимается его неизменяемая часть, выражающая лексическое значение. Термин *стемма* обозначает последовательность символов, остающуюся после удаления строк, содержащихся в определенных файлах данных, и выполняющую функцию отождествления токенов. Ланкастерский стеммер [Paice, 1990] в токене *daughter* удаляет *er*, так как такая строка есть в файле данных, на основе которого он работает. С точки зрения теоретической лингвистики в данном случае происходит ошибка, поскольку *er* входит в основу слова. С точки зрения компьютерной лингвистики ошибки не происходит, потому что с помощью стеммы *daught* можно отождествить токены *daughter* и *daughters*. Вместе с тем, при отделении *er* от *cater* по стемме *cat* отождествятся не только токены *catered, caters, catering* но также и *cat, cats, cat's*. Возникает ошибка избыточного стеммирования, поскольку по одной стемме отождествляются токены с разным значением. Лемматизация также предусматривает отождествление основ слов, однако проводится с учетом частей речи, к которым относятся словоформы. Например, стеммер отождествит *read*, *reads, reader, readers* с одной основой *read*, в то время как лемматайзер отождествит глагольные формы *read*, *reads* с основой (леммой) *read*, а именные формы *reader, readers* – с леммой *reader*. Задача лемматизации – отождествить

словоформы, соотносящиеся с одной лексемой. Словари лемм широко используются в корпусной лингвистике в целях поддержки лингвистических исследований.

Аннотирование проводится теггерами, на входе у которых – список токенов, на выходе – список, в котором каждому токену приписывается определенный тег - условное обозначение, указывающее на его лингвистические характеристики. Наиболее распространенным видом теггеров являются теггеры частей речи (POS taggers), которые распознают часть речи токена и приписывают ему соответствующий тег. Помимо информации о части речи обычно указывается и информация о лексико-грамматических и семантических характеристиках слова, например, *NN* – нарицательное существительное в единственном числе, *NNS* – нарицательное существительное во множественном числе, *AJC* – прилагательное в сравнительной степени и т.д.

Термин *тег* был введен в научный оборот в связи с разработкой электронных текстовых корпусов и не имеет аналогов в теоретической лингвистике, хотя широко используется в информатике, в частности для обозначения дескрипторов языков гипертекстовой разметки. В настоящее время аннотирование широко применяется в системах автоматической классификации текстов, а теги частей речи и их сочетания выступают в качестве параметров такой классификации. К другим видам тегов относятся семантические теги и теги когнитивных ролей (knowledge roles). В [Mustafaraj, 2007] проводилось аннотирование текстов диагностических отчетов о состоянии электроизоляции высоковольтных ротационных устройств когнитивными ролями *Observed Object*, *Symptom*, *Cause*. В результате была создана система, с помощью которой инженер мог получать информацию о признаках неполадки конкретного объекта, причинах и способах ее устранения.

Аннотирование семантическими и когнитивными ролями предусматривает распознавание как отдельных слов, так и словосочетаний. Такое аннотирование требует предварительной разработки и применения специальных грамматик фразовой структуры на синтаксическом уровне языковой системы.

Одним из фундаментальных алгоритмов, применяемых на синтаксическом уровне, является синтаксическая декомпозиция, которая выполняется синтаксическими сплиттерами. На входе у сплиттера – текст, на выходе – список предложений текста. Алгоритмы синтаксической декомпозиции предусматривают распознавание предложений на основе символов форматирования текста: пробелов, знаков пунктуации, знаков конца строк. Таким образом, термин *предложение* в компьютерной лингвистике обозначает последовательность строк, отделенную справа и слева символами форматирования текста и знаками пунктуации. Распознавание предложений осложняется отсутствием стандартного форматирования текста; точки, восклицательные, вопросительные знаки, которые обычно применяются в качестве разделителей, могут использоваться не только в конце, но и в середине предложения. Целый ряд единиц текста, которые форматируются как предложения, на самом деле предложениями не являются. К ним относятся такие элементы, как оглавление, заглавия отдельных разделов, названия рисунков, таблиц, текст, использующийся внутри самих таблиц и рисунков, колонтитулы. Между тем именно предложения являются основной единицей анализа во многих системах, а в системах автоматического реферирования и выходной текст состоит из предложений. Ошибки в распознавании предложений существенно снижают эффективность таких систем в целом. Нами была предложена дедукционно-инверсионная архитектура декомпозиции текста, в соответствии с которой вначале текст разбивается на абзацы, затем – на слова, затем из слов генерируются предложения. Таким образом, декомпозиция начинается с большей единицы (абзаца), затем осуществляется переход к меньшей единице (слову), затем – снова к большей (предложению). Дедукционно-инверсионная архитектура декомпозиции позволяет игнорировать такие компоненты текста, как заголовки, подзаголовки, оглавления, поскольку они не входят в состав абзацев [Яцко, 2009].

Синтаксическая декомпозиция является основой для выполнения целого ряда алгоритмов распознавания фразовой структуры предложения. Широко распространены алгоритмы выделения *n-gram* – словосочетаний, состоящих из двух (биграмы), трех (триграмы) и более (тетраграмы, пентаграмы, гексагра-

мы, гептаграмы, октограмы) токенов [Bickel, 2005]. Разбивка на словосочетания в данном случае проводится с учетом позиции токена в предложении. Например, предложение *John has a dog* включает 4 юниграма, 3 диграма (*John has*, *has a*, *a dog*), 2 триграма (*John has a*, *has a dog*), 1 тетраграм – все предложение. Количество биграм для каждого предложения ($ng_{(s)}$) будет составлять n-1; тригарм – n-2, где n-количество токенов в предложении, т.е. $ng_{(s)} = w_{i-(n-1)}, w_{i-(n-2)}\ldots w_{i-(n-n)}$, где w_i – порядковый уровень *n-gram*, начиная с биграм. Распознавание *n-gram* проводится на основе соответствующих правил.

Анализ распределения *n-gram* позволяет выявить статистически значимые словосочетания и часто применяется в стохастических алгоритмах аннотирования тегами частей речи. Распределения *n-gram* используются с целью автоматической классификации и категоризации, поскольку выступают в качестве важного параметра, позволяющего определить принадлежность текста к определенной категории, типу, группе, жанру. При анализе на синтаксическом уровне в качестве основной единицы выступают биграммы и диграммы, поскольку рекуррентность словосочетаний с большим количеством токенов маловероятна. Анализ *n-gram* большего порядка применяется в системах автоматической коррекции орфографии, а также в системах оптического распознавания символов (*optical character recognition*), где основной единицей выступают символы в токенах.

Для анализа морфологически значимых словосочетаний применяются программы фразовой декомпозиции-чанкеры, которые на выходе выдают списки фраз определенного типа (именных, глагольных, предложных, адъективных, адвербиальных). Наиболее распространены именные (noun phrase) чанкеры, распознающие словосочетания с управляющим существительным. Именно этим типом словосочетаний обозначаются объекты, описываемые в тексте, а их ранжирование по весовым коэффициентам позволяет получить список ключевых слов, отражающих основное содержание текста. Распознавание словосочетаний этого типа выполняется на основе предварительного аннотирования тегами частей и объединения отдельных частей речи во фразы на основе правил грамматики.

Правила фразовой структуры были разработаны для английского языка в рамках концепции генеративной грамматики, предложенной Н. Хомским. Грамматические правила записываются в виде:

NP→NN; NP→DetNN; NP→ DetANN,

где указывается состав словосочетания, в данном случае именного, а также порядок слов. В первом случае показано, что именное словосочетание может состоять только из одного существительного (*NN*); во втором случае оно состоит из детерминанта (*Det*) и существительного, причем детерминант занимает позицию перед существительным, а обратный порядок слов неправилен; в третьем случае словосочетание состоит из детерминанта, прилагательного (*A*), существительного, причем другие варианты словопорядка неправильны.

К настоящему времени на основе концепции Н. Хомского создан целый ряд грамматик, которые делятся на два основных вида – деривационные и недеривационные. В деривационных грамматиках проводится разграничение между поверхностной и глубиной структурой словосочетания и предложения, и формулируются дополнительные правила вывода (деривации) поверхностных структур их глубинных. Синтаксическая структура представляется в виде иерархического дерева зависимости. Недеривационные грамматики описывают поверхностные, как правило, линейные синтаксические структуры. Выбор того или иного типа грамматики обуславливается задачами конкретного исследовательского проекта. Деривационные грамматики лежат в основе функционирования синтаксических парсеров, которые выдают на выходе граф синтаксической структуры предложения. Так же, как и теггеры частей речи, синтаксические парсеры обучаются на предложениях с размеченной вручную синтаксической структурой; в них применяются правила для определения наиболее вероятного варианта на основе скрытых моделей Маркова. В качестве примера можно привести *Lexparser*, разработанный в Стэнфордском университете США [The Stanford parser, 2014].

Иерархические синтаксические структуры применяются в системах машинного перевода для установления эквивалентности синтаксических структур в двух языках.

На синтаксическом уровне может проводиться декомпозиция не только на словосочетания и предложения, но и на клаузы – элементарные предикативные структуры, выражающие суждение. Понятие клаузы в какой-степени соответствует понятию пропозиции в лингвистике, однако клаузы выделяются по формальным признакам, к которым может относиться, например, наличие именной группы и следующей за ней глагольной группы в утвердительном предложении. Разбивка на клаузы применяется в системах интеллектуального анализа для более адекватной передачи содержания текста.

Наиболее распространенными алгоритмом, применяемыми на дискурсивном уровне являются алгоритмы разрешения анафоры, которые предусматривают замену анафорических местоимений предшествующими референтными именами объектов. Под дискурсом в компьютерной лингвистике понимается текст, связи между компонентами которого (клаузами, предложениями) манифестируются повторами лексических и/или синтаксических единиц. Как мы полагаем, исследование логико-сематических связей между единицами текста и проблема моделирование его логико-семантической структуры выходят за рамки предметной области компьютерной лингвистики и являются частью проблем исследования искусственного интеллекта.

В таблице представлены алгоритмы и программы, характеризующие специфику предметной области компьютерной лингвистики, сгруппированные по уровням системы языка.

Алгоритмы и программы автоматической обработки текста

Алгоритмы	Программы	Распознаваемая/обрабатываемая единица	Лингвистический термин	Уровни языка
Распознавание символов	OCR	Символ	Графема	Графемный
Стемминг	Stemmers	Стемма	Основа слова	Морфологический
Лемматизация	Lemmatizers	Лемма	Лексема	
Токенизация	Tokenizers	Токен	Слово	Лексический
Аннотирование	Taggers	Тэг	–	
Взвешивание терминов	Weighting filters	Весовой коэффициент	–	

Представленные в таблице алгоритмы и программы лежат в основе лингвистического программного обеспечения, которое можно классифицировать по целому ряду критериев. По материальной форме входного текста выделяются системы обработки устных текстов и письменных текстов. В первом случае обычно говорят об обработке речи (speech processing), а во втором – об обработке текста (text processing).

Начало развития компьютерной лингвистики связано с проблемами обработки письменных текстов, создания ИПС, систем реферирования и машинного перевода в конце 1950-х-1960-х гг. Системы обработки устной речи стали интенсивно разрабатываться в 1990-х гг., когда появились бытовые системы распознавания речи. В настоящее время они широко применяются в автоответчиках; в таких системах распознавания индивидуальных характеристик личности, как возраст, пол и даже уровень алкогольного опьянения [Levit, 2001]; в системах голосового управления техническими объектами, в том числе и наносистемами [Потапова, 2007]. По форме речевой деятельности можно выделить алгоритмы, предназначенные для обработки монологической речи и диалогической речи. Долгое время объектом автоматического анализа текста были монологические тексты, в основном тексты научных работ. Развитие Интернета обусловило появление жанров диалогической письменной речи: чатов, блогов, форумов.

Обработка таких текстов имеет свою специфику и требует применения специальных алгоритмов, учитывающих их паралингвистические особенности. Интенсивно развиваются и системы обработки диалогической устной речи: вопросно-ответные системы, системы машинного перевода. По степени интеллектуальности получаемых пользователями результатов можно выделить в отдельную группу алгоритмы, с помощью которых выдается информация, содержащаяся в тексте имплицитно, либо новая информация, которой нет в обрабатываемом тексте. Такие алгоритмы разрабатываются в процессе интеллектуального анализа текста (*text mining*) и существенно отличаются от традиционных алгоритмов информационного поиска и реферирования, в результате применения которых выявляется наиболее значимая информация, содержащаяся в тексте. Интеллектуальный анализ текста широко применяется в технике и медицине как средство обмена опытом. В медицине перевод историй болезней пациентов в электронную форму и их аннотирование тегами когнитивных ролей позволяет врачу с помощью поисковых систем находить диагнозы, соответствующие определенным симптомам, методики лечения, применявшиеся другими врачами, назначавшиеся медикаменты и препараты, результаты лечения [Li, 2013]. Успешно развивается интеллектуальный анализ мнений пользователей о коммерческих продуктах [Яцко, 2011], позволяющий фирмам-производителям выявлять достоинства и недостатки продукции и проводить эффективную маркетинговую политику.

По целевым группам пользователей можно выделить универсальное, специальное и профессиональное лингвистическое ПО. Системы универсального типа предназначены для любых групп пользователей, независимо от их профессии, возраста, социального положения. Типичный пример – ИПС интернета, которыми каждый день пользуются миллиарды людей в мире. Специальное лингвистическое ПО предназначено для определенных групп пользователей. Системы интеллектуального анализа текста обычно позиционируются как системы, предназначенные для поддержки принятия решений представителями определенной профессиональной группы. Профессиональное лингвистическое ПО предназначено для специалистов в области компьютерной лингвистики и поддержки исследований в области автоматического анализа текста. Существует ряд программ статистического анализа, предоставляющих информацию о количестве уникальных и общих токенов, количестве н-грам, контексте использования лексических единиц, вероятностные и статистические показатели их совместной встречаемости [Laurence, 2011; Scott, 2012]. К профессиональным также относятся инструментальные программные средства, предназначенные для оценки эффективности и качества лингвистических программ, приложений, систем.

В зависимости от режима функционирования лингвистические приложения и системы можно разделить на автоматические и автоматизированные. Автоматизированные системы работают в дискретном режиме; к этому виду относится большинство разрабатываемого в настоящее время ПО. В качестве примера можно привести информационно-поисковые системы, функционирование которых начинается с запроса пользователя и заканчивается выдачей результата. Автоматические системы функционируют в непрерывном режиме, как, например, системы реферирования устной речи, позволяющие отслеживать новостные события. Заметим, что в названиях конкретных видов лингвистического ПО нет строго разграничения между терминами «автоматический» и «автоматизированный». Информационно-поисковые системы совершенно верно характеризуются как автоматизированные, в то время как системы реферирования по традиции называются автоматическими (ср. название известного сборника Advances in automatic text summarization), хотя, на самом деле, имеются ввиду системы, работающие в дискретном режиме. Наряду с автоматическими и автоматизированными системами и приложениями разрабатывается также ПО для компьютерно-опосредованной (computer-assisted/aided) обработки текстов. Системы этого типа наиболее широко используются в практике двуязычного перевода и в обучении иностранным языкам, повышая эффективность деятельности преподавателя и переводчика. Системы типа переводческой памяти (translation memory) содержат базы данных, включающие ранее переведенные тексты, словари, корпуса. Выполняя пе-

ревод, его автор может вставлять в текст слова фразы, предложения из переводов, сделанных ранее, а также с помощью словарей и корпусов проверять контекст использования лексических единиц. Как мы полагаем, неверно использовать термин «автоматизированный перевод» для обозначения систем подобного типа, поскольку в них не применяются алгоритмы, указанные в таблице, а проводится построчное сопоставление текстов. К автоматизированным относятся системы машинного перевода, работающие в дискретном режиме.

Функционирование лингвистического ПО поддерживается лексикографическими ресурсами, к которым относятся списки терминов, терминологические словари, терминологостатистические словари, тезаурусы, онтологии [Яцко, 2013].

Вышеизложенное позволяет определить компьютерную лингвистику как дисциплину, изучающую закономерности распределения текстовой информации, проблемы, принципы, методы и алгоритмы разработки лингвистического программного и аппаратного обеспечения. На входе лингвистического программного и аппаратного обеспечения – текст на естественном языке. На выходе пользователю может предоставляться также текст на естественном языке, представляющий содержание входного текста; некоторая модель входного текста; статистические данные о распределении единиц входного текста. К предметной области компьютерной лингвистики не относятся системы, на входе которых – текст на искусственном языке, например, криптографические системы.

Компьютерная лингвистика представляет собой междисциплинарную науку, развитие которой детерминируется математическими, техническими, лингвистическими основами. Математика и языкознание выполняют методологическую роль, которая проявляется разнопланово. Методологическая роль математики возрастает по мере повышения уровня разработок и исследований. Разработка прикладного лингвистического ПО требует знания общих для всех видов программирования элементов булевой алгебры и исчисления высказываний, в то время как выполнение теоретических и фундаментальных исследований невозможно без знания соответствующих разделов математики, например, теории множеств, теории графов, теории вероятностей, статистического анализа, а также закономерностей и законов распределения текстовой информации. В настоящее время одной из фундаментальных проблем предметной области, решение которой невозможно без применения математического аппарата, является разработка критериев репрезентативности текстовых корпусов. Методологическая роль лингвистики возрастает по мере возрастания сложности обрабатываемых и распознаваемых лингвистических единиц. Если токенизация может проводиться по символам форматирования текста, то стемминг требует знания морфологической структуры слова, чанкинг и парсинг – знания фразовой структуры предложения, клауз-сплитинг – знания структуры предикативных конструкций, разрешение анафоры – знания межфразовых связей между предложениями. Актуальной теоретической задачей, требующей серьезного лингвистического анализа, является разработка ролевых грамматик для поддержки систем интеллектуального анализа текста.

Одной из основных проблем, влияющих на развитие предметной области, является сложность подготовки специалистов, которые должны владеть сочетанием гуманитарных, технических и математических знаний. Подготовка таких специалистов в зарубежных университетах проводится в рамках бакалаврских и магистерских программ, содержание которых включает технические, математические и лингвистические компоненты. В качестве примера можно привести магистерскую программу лингвистического факультета (Department of linguistics) Вашингтонского университета в Сиэтле [UW, 2014]. Технический компонент предусматривает хорошие навыки программирования на C++ и Java (рекомендуется также знание Perl и/или Python); знание структур баз данных и алгоритмов, конечных автоматов и измерительных преобразователей; умение использовать серверные кластеры на платформе UNIX. Математический компонент включает теорию вероятностей и статистический анализ. Лингвистический компонент включает введение в фонетику и синтаксис с акцентом на изучение артикуляционных и акустических коррелятов фонологических единиц и разработку формальных грамматик, необходимых для создания

прикладных программ; изучение методов поверхностной (shallow) обработки единиц естественного языка, включая токенизацию, аннотирование, морфологический анализ, парсинг; изучение методов глубокой обработки единиц естественного языка, включая алгоритмы и грамматики, необходимые для соотнесения глубинных структур с поверхностными синтаксическими структурами. На заключительном этапе обучения изучаются методы разработки информационно-поисковых и вопросно-ответных систем, систем машинного перевода, а также приложений и программ обучения языкам, проверки орфографии и грамматики, распознавания рукописного ввода и оптического распознавания символов, кластеризации документов, распознавания и синтеза речи. Во время обучения студенты походят практику в таких крупнейших компаниях, занимающихся разработкой лингвистического программного обучения, как Microsoft, Google, InXight.

Данная магистерская программа представляет интерес, так как дает представление о междисциплинарной природе и структуре предметной области. Из трех компонентов наиболее объемным является лингвистический, что вполне естественно, и не случаен тот факт, что такие программы обучения обычно реализуются именно на лингвистических факультетах. Вместе с тем ведущим компонентом образовательной программы является технический, а хорошее владение навыками программирования – обязательным условием для поступающих. По результатам обучения и защиты магистерской диссертации выпускникам присваивается степень магистра естественных наук (*Master of Science*), а не гуманитарных наук (*Master of Arts*), которая присваивается по лингвистическим дисциплинам. Оформилась профессиональная специализация для обозначения которой общепринятым является термин *computational linguist*. Как утверждается на одном из сайтов «компьютерные лингвисты разрабатывают системы обработки человеческого языка. Им необходимо хорошее знание как программирования, так и лингвистики. Это – технически сложная предметная область, однако, квалифицированные компьютерные лингвисты востребованы и высоко оплачиваемы» [Uszkoreit, 2012].

Предлагаемая магистерская программа, как утверждается на сайте Вашингтонского университета, относится к числу немногих лучших программ международного уровня, по которым готовятся специалисты в области компьютерной лингвистики.

Очевидно, что развитие лингвистических технологий и создание аналогичных программ обучения является актуальным и для России.

Библиографический список:

1. *Баранов, А.Н.* Введение в прикладную лингвистику [Текст]: учеб. пособие / А.Н. Баранов. – М.: Эдиториал УРСС, 2001. – 360 с.

2. *Паспорта* Номенклатуры специальностей научных работников [Электронный ресурс] / Информика. – 2002-2012. – Режим доступа: http://www.edu.ru/db/portal/spec_pass/vuz_ds_pasport.php?spec=10.02.21 (дата обращения: 01.03.2014).

3. *Потапова, Р.К.* Нанотехнологии и лингвистика: прогнозы и перспективы взаимодействия [Текст] / Р.К. Потапова // Нанотехнологии в лингвистике и лингводидактике: миф или реальность? Опыт создания общего образовательного пространства стран СНГ. Тезисы Международной научно-практической конференции. – М., 2007. – С. 9-11.

4. *Рождественский, Ю.В.* Лекции по общему языкознанию [Текст] / Ю.В. Рождественский. – М.: Высш. шк., 1990. – 381 с.

5. *Алгоритмы* предварительной обработки текста: декомпозиция, аннотирование, морфологический анализ [Текст] / В.А. Яцко, М.С. Стариков, Е.В. Ларченко [и др.] // Научно-техническая информация. Сер. 2. – 2009. – № 11. – С. 8-18.

6. *Яцко, В.А.* Лексикографические ресурсы для автоматического анализа текста [Текст] / В.А. Яцко // Вестник Иркутского государственного лингвистического университета. – 2013. – №2. – С. 19-24.

7. *Яцко, В.А.* Метод зонального анализа данных [Текст] / В.А. Яцко // В мире научных открытий. – 2013. – № 6.1. – С. 166-182.

8. *Яцко, В.А.* Опыт разработки онтологии для автоматического анализа мнений пользователей о коммерческих продуктах [Текст] / В.А. Яцко, М.С. Стариков // Научно- техническая информация. – 2011. – № 7. – С. 9-14.

9. *Advanced* natural language processing. MIT OpenCourseWare [Electronic resource]. – 2001-2014. – URL: http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-864-advanced-natural-language-processing-fall-2005 (дата обращения: 01.03.2014).

10. *Association* for Computational Linguistics. ACL Homepage [Electronic resource]. – 2014. – URL: https://www.aclweb.org (дата обращения: 01.03.2014).

11. *Bickel, S.* Predicting sentences using n-gram language models [Electronic resource] / S. Bickel, P. Haider, T. Scheffer. – 2005. – URL: http://www.mpi-inf.

mpg.de/~bickel/publications/bickel_emnlp_2005.pdf (дата обращения: 01.03.2014).

12. *BYU-BNC*: British National Corpus [Electronic resource] / Brigham Young University. – 2012. – URL: http://corpus.byu.edu/bnc (дата обращения: 01.03.2014).

13. *Chair* of Foundations of Computational Linguistics [Electronic resource]. – 2012. – URL: http://www3.ims.uni-stuttgart.de/gcl/index.html.en (дата обращения: 01.03.2014).

14. *Computational* linguistics [Electronic resource] / Stanford linguistics. – 2014. – URL: http://linguistics.stanford.edu/research/computational-linguistics (дата обращения: 01.03.2014).

15. *Definition* of applied linguistics [Electronic resource] // Oxford dictionaries. – 2012. – URL: http://oxforddictionaries.com/definition/english/applied%2Blinguistics (дата обращения: 01.03.2014).

16. *Foundations* of Computational Linguistics. Institute for natural language processing. University of Stuttgart [Electronic resource]. – 2013. – URL: http://www.ims.uni-stuttgart.de/institut/arbeitsgruppen/gcl/index.en.html (дата обращения: 01.03.2014).

17. *Francis, W.N*. Frequency analysis of English usage lexicon and grammar [Text] / W.N. Francis. – Boston: Houghton Mifflin, 1982. – 561 p.

18. *Hess, M.* Einführung in die Computerlinguistik [Electronic resource] / M. Hess. – 2005. – URL: https://files.ifi.uzh.ch/cl/hess/classes/ecl1/ecl1.0.l.pdf (дата обращения: 01.03.2014).

19. *Information* science – definition and more [Electronic resource] // Merriam-Webster dictionary. – 2012. – URL: http://www.merriam-webster.com/dictionary/information%20science (дата обращения: 01.03.2014).

20. *Institut* für Computerlinguistik [Electronic resource] / Universität Heidelberg. – 2014. – URL: http://www.cl.uni-heidelberg.de/ (дата обращения: 01.03.2014).

21. *Laurence* Anthony's software [Electronic resource] – 2011. – URL: http://www.antlab.sci.waseda.ac.jp/software.html (дата обращения: 01.03.2014).

22. *Levit, M.* Use of prosodic speech characteristics for automated detection of alcohol intoxication [Text] / M. Levit, R. Huber, A. Batliner, E. Nöth // Proceedings of the workshop on prosody and speech recognition. – Red Bank, NJ, 2001. – P. 103-106.

23. *Li, Q.* A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction [Text] / Q. Li, H. Zhai, L. Deleger // Journal of the American Medical Informatics Association. – 2013. – Vol. 20. – Issue 5. – P. 915-921.

24. *Missing* tot's trail goes cold after three month [Electronic resource]. – 2009. – January 19. – URL: http://edition.cnn.com/2009/CRIME/01/13/grace.coldcase.hussain/index.html?eref=rss_crime (дата обращения: 01.03.2014).

25. *Mustafaraj, E.* Mining diagnostic text reports by learning to annotate knowledge roles [Текст] / E. Mustafaraj, V. Hoof, D. Freisleben // Natural language processing and text mining / Ed-s Kao A., S. Poteet. – London, 2007. – P. 45-68.

26. *Naves, T.* Applied linguistics what it is and the history of the discipline [Electronic resource] / T. Naves. – 2002. – URL: http://diposit.ub.edu/dspace/bitstream/2445/4701/1/Naves2008ALDiscipline PartIonGrabe2002.pdf (дата обращения: 01.03.2014).

27. *Paice, C.D.* Another stemmer [Text] / C.D. Paice // SIGIR forum. – 1990. – Vol. 24, № 3. – P. 56-61.

28. *Philosophische* fakultät der HHUD: computerlinguistik [Electronic resource]. – 2014. – URL: http://www.phil-fak.uni-duesseldorf.de/cl (дата обращения: 01.03.2014).

29. *Salton, G.* On the specification of term values in automatic indexing [Текст] / G. Salton, C.S. Yang // Journal of documentation. – 1973. – Vol. 29, Issue 4. – P. 351-372.

30. *Scott, M.* WordSmith Tools version 6 [Electronic resource] / M. Scott. – Liverpool: Lexical Analysis Software, 2012. – URL: http://www.lexically.net/wordsmith/index.html (дата обращения: 01.03.2014).

31. *Sillanpää, M.* Lost knowledge – DM Partner's Essence [Electronic resource] / M. Sillanpää. – 2009. – June 4. – URL: http://bigmenoncontent.com/2009/06/04/lost-knowledge-—-dm-partner's-essence (дата обращения: 01.03.2014).

32. *TCL* – Thai computational linguistics laboratory [Electronic resource]. – 2005. – URL: http://www.tcllab.org/ (дата обращения: 01.03.2014).

33. *The corpus* of contemporary American English (COCA) [Electronic resource] / Brigham Young University. – 2012. – URL: http://corpus.byu.edu/coca/ (дата обращения: 01.03.2014).

34. *The Stanford* natural language processing group [Electronic resource]. – 2012. – URL: http://www-nlp.stanford.edu/ (дата обращения: 01.03.2014).

35. *The Stanford* parser: a statistical parser [Electronic resource] // The Stanford natural language processing group. – 2014. – URL: http://nlp.stanford.edu/software/lex-parser.shtml (дата обращения: 01.03.2014).

36. *Tsz-Wai, L.R.* Automatically building a stopword list for an information retrieval system [Text] / L.R. Tsz-Wai, B. He, I. Ounis // Journal on digital information management: special issue on the 5th Dutch-Belgian information retrieval workshop (DIR'05). – 2005. – V. 3, № 1. – P. 3-8.

37. *Uszkoreit, H.* Linguistics jobs: computational linguist [Electronic resource] / H. Uszkoreit. – 2013. – URL: http://allthingslinguistic.com/post/60695678526/linguistics-jobs-computational-linguist (дата обращения: 01.03.2014).

38. *UW* professional masters in computational linguistics [Electronic resource] / University of Washington. – 2014. – URL: http://www.compling.uw.edu/about (дата обращения: 01.03.2014).

39. *Yatsko, V.A.* TF*IDF Revisited [Electronic resource] / V.A. Yatsko // International journal of computational linguistics and natural language engineering. – 2013. – Vol. 2, Issue 6. – P. 385-387. – URL: http://www.ijclnlp.org/vol2issue6/paper60.pdf (дата обращения: 01.03.2014).