

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ  
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ  
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

**К. К. Боярский**

**ВВЕДЕНИЕ В КОМПЬЮТЕРНУЮ  
ЛИНГВИСТИКУ**

Учебное пособие



Санкт-Петербург  
2013

Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. – СПб: НИУ ИТМО, 2013. – 72 с.

Рассматриваются основные принципы компьютерного анализа текстов на естественном языке. Приведены примеры анализа на трех уровнях — морфологическом, синтаксическом и семантическом с использованием соответствующего инструментария: словарей и корпусов текстов. Обсуждены возможные применения результатов анализа, в том числе в области классификации текстов и извлечения из них информации.

Для студентов специальности 036000 «Интеллектуальные системы в гуманитарной сфере».

Печатается по решению Ученого Совета ЕНФ.

Протокол № 7 от 22 октября 2013 г.



В 2009 году Университет стал победителем многоэтапного конкурса, в результате которого определены 12 ведущих университетов России, которым присвоена категория «Национальный исследовательский университет». Министерством образования и науки Российской Федерации была утверждена программа его развития на 2009–2018 годы. В 2011 году Университет получил наименование «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики».

© Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, 2011

© К. К. Боярский, 2013

## Оглавление

1	Предмет компьютерной лингвистики .....	4
1.1	Что такое компьютерная лингвистика?.....	4
1.2	Основные направления компьютерной лингвистики .....	5
1.3	Компьютерный анализ текста .....	9
1.4	Задачи лингвистических информационных технологий .....	14
2	Инструментарий компьютерной лингвистики .....	17
2.1	Словари .....	17
2.2	Корпуса текстов .....	26
2.3	Национальный корпус русского языка (НКРЯ) ( <a href="http://ruscorpora.ru">http://ruscorpora.ru</a> ) .....	27
3	Автоматический анализ текста .....	30
3.1	Морфологический уровень .....	30
3.2	Синтаксический уровень .....	34
3.3	Анафора и кореферентность.....	40
4	Классификация и кластеризация.....	45
4.1	Закон Ципфа .....	45
4.2	Модель $TF*IDF$ .....	48
4.3	Классификация документов .....	50
4.4	Классификация с обучением. Наивный байесовский классификатор .....	52
4.5	Классификация с обучением. Другие алгоритмы .....	57
4.6	Оценка результатов классификации. $F$ -мера .....	58
4.7	Кластеризация .....	62
4.8	Контент-анализ .....	66
5	Литература.....	71

# 1 Предмет компьютерной лингвистики

## 1.1 Что такое компьютерная лингвистика?

Создание ЭВМ в середине 20-го века и быстрое развитие кибернетических идей стимулировали появление новых наук, которые ранее просто невозможно было представить. Как правило, они возникали на стыке наук, часто не связанных друг с другом. Так, на стыке биологии и инженерных наук возникла бионика. На стыке вычислительной техники и лингвистики родилась наука, которая несколько раз меняла название: сначала она называлась математической лингвистикой, потом структурной лингвистикой и вычислительной лингвистикой. Наконец за ней прочно укрепилось ее современное название — **компьютерная лингвистика**.

Две причины обусловили появление новой науки. Во-первых, исследователи-лингвисты надеялись, что современные точные науки (и прежде всего математика) помогут лингвистике обрести недостающую ей точность. Появление ЭВМ укрепило эти надежды, так как многим языковедам с самого начала было ясно, что компьютеры — это не только «быстро работающие арифмометры», но и мощное средство для автоматизации работы с текстами. Появилась возможность автоматизировать многие трудоемкие процессы, например статистическую обработку текстов, ведение разнообразных словарных и лексических картотек. Во-вторых, с появлением компьютеров почти сразу же возникла проблема общения с ними неподготовленных пользователей. Бесспорно, наилучшей формой для таких пользователей мог быть привычный естественный язык. Но для организации такого взаимодействия надо прежде понять законы и особенности использования естественного языка в процессе общения людей между собой. А, как вскоре выяснилось, традиционная лингвистика изучением этих законов практически не занималась.

Таким образом:

«**компьютерная лингвистика**» = «лингвистика в изложении для компьютера» (компьютер выступает в роли субъекта восприятия лингвистики)

или:

«**компьютерная лингвистика**» = «лингвистика, которую делают на компьютере».

Важнейшими понятийными категориями компьютерной лингвистики являются такие структуры знаний, как «**фреймы**» (понятийные, или, как принято говорить, концептуальные структуры для декларативного представления знаний о типизированной тематически единой ситуации), «**сценарии**» (концептуальные структуры для процедурного представления знаний о стереотипной ситуации или стереотипном поведении), «**планы**» (структуры знаний, фиксирующие представления о возможных действиях,

ведущих к достижению определенной цели).

Эти структуры соответствуют двум основным типам знания.

«**Знание что**» — декларативные знания, представляются обычно в виде совокупности пропозиций, утверждений о чем-либо. Например, чашка — небольшой сосуд для питья округлой формы, обычно с ручкой, из фарфора, фаянса и т. п. Декларативные знания поддаются процедуре верификации в терминах «истина–ложь».

«**Знание как**» — процедурные знания, представляются как последовательность (список) операций, действий, которые следует выполнить. Характерный пример процедурных знаний — инструкции по пользованию бытовыми приборами. В отличие от декларативных знаний, процедурные знания невозможно верифицировать как истинные или ложные. Их можно оценивать только по успешности – неуспешности алгоритма.

Декларативное знание легче осознается человеком, чем процедурное. В то же время инженерная лингвистика представляет собой скорее знание «как», а не знание «что».

## 1.2 Основные направления компьютерной лингвистики

- ✓ Анализ текстов на естественном языке.

В январе 1954 г. в рамках так называемого Джорджтаунского проекта группа американских лингвистов выдвинула идею, продемонстрировала полностью автоматический перевод 60 предложений с русского языка на английский. Организаторы эксперимента уверяли, что в течение трёх—пяти лет проблема машинного перевода будет решена. Идея заинтересовала лингвистов многих стран и активизировала работы в области анализа текстов. В ходе этих работ надо было прежде всего ответить на вопрос: «Существуют ли строгие формальные правила, по которым строится структура предложения и структура текста?» Если о структуре предложения лингвисты накопили много материала, то структура текста ими не изучалась.

В результате проведенных исследований стало ясно, что за каждым текстом (в том числе и за отдельным предложением, являющимся своего рода мини-текстом) скрывается не одна, а несколько формальных структур, которые можно разделить на три уровня.

Первый уровень — это **поверхностная синтаксическая структура**. В этой структуре каждое предложение текста рассматривается изолированно от других и для каждого проводится что-то вроде разбора предложения по его членам, как все мы делали в школе. Выделяются подлежащее и сказуемое, определения, дополнения и обстоятельства разного вида. Но этой структуры для анализа оказывается мало.

Следующий шаг — построение **глубинной синтаксической структуры** (второй уровень). Идея существования глубинной синтаксической

структуры связана с пониманием того, что различные естественные языки, отличаясь друг от друга многими внешними синтаксическими особенностями, передают весь спектр взаимосвязей между объектами, явлениями, их свойствами и протекающими с их участием процессами, характерными для окружающего мира. И этот мир един, каким бы языком мы его ни описывали. Следовательно, в каждом тексте существуют не зависящие от особенностей языка некие глубинные структуры, которые определяют адекватное отображение той или иной ситуации в окружающем мире. С этой идеей тесно связано использование так называемых глубинных падежей, или падежей Филмора<sup>1</sup>, названных по имени американского исследователя, впервые введшего их в научный оборот. Рассмотрим как пример две фразы: «*Мальчик сорвал цветок*» и «*Цветок, сорванный мальчиком*». В первом предложении субъект действия *сорвал* — это *мальчик*. И это слово играет здесь роль подлежащего, о чем свидетельствует именительный падеж. Во втором же предложении роль подлежащего играет слово *цветок*. Но субъектом действия *сорвал* и здесь остается все тот же *мальчик*. А *цветок* в любом из двух приведенных предложений играет роль объекта действия. Понимание ситуации, описываемой любым из этих предложений, заключается, в частности, в том, что мы выделяем в тексте некоторое действие, а также его субъект и объект. Позиции субъекта и объекта служат примером тех самых глубинных падежей, которые ввел Филмор. Эти два падежа (субъектный и объектный) не единственные. Разные исследователи выделяют разное количество таких падежей (инструментальный, временной, пространственный и т. д.), но их общее количество не превосходит полутора десятка.

Синтаксическая структура, построенная на основе глубинных падежей, позволяет перейти от синтаксического уровня предложения к его **семантическому уровню** (третий уровень). На этом уровне для анализа привлекаются дополнительные данные, связанные с наличием у лексических единиц языка (в частности, слов) определенных значений. В семантических структурах также можно выделить поверхностный и глубинный уровни, в чем-то похожие на соответствующие уровни в синтаксических структурах. Поверхностный семантический уровень тесно связан с глубинной синтаксической структурой, а глубинный семантический уровень как бы отрывается от нее, передавая смысл для целого класса однотипных ситуаций. Структуры наиболее глубокого уровня, возникающие при анализе предложений, могут быть названы **прагматическими**. Из них следует понимание того, к чему обязывает или призывает данное предложение. Прагматические структуры устанавливают связь между предложениями в

---

<sup>1</sup> Чарльз Дж. Филмор – американский лингвист, в 1990 г. президент Американского лингвистического общества

текстах, связывают текст в единое целое, а также побуждают нас делать те или иные действия в реальном мире (как, например, надпись: «*Стой! Проход запрещен!*»).

В процессе анализа текстов, содержащих более одного предложения, возникают новые структуры, обеспечивающие сцепление этих предложений в рамках некоторой описываемой ситуации или последовательности ситуаций. Возникают межфразовые связи, позволяющие понять текст как единое целое. Эти структуры пока изучены значительно хуже, чем структуры, лежащие в основе одного предложения. Необходимость в исполнении тех или иных этапов при анализе конкретного текста зависит от тех целей, для которых тот анализ осуществляется.

✓ Синтез текстов на естественном языке.

Задача синтеза может рассматриваться как обратная по отношению к анализу. Если заданы некоторая тема и цель будущего текста, то можно считать заданной прагматическую структуру текста. Ее надо декомпозировать в прагматические структуры отдельных предложений и для каждого предложения пройти все этапы анализа в обратном направлении. На сегодняшний день здесь еще масса нерешенных проблем. Один из возможных путей генерации текста состоит в использовании **актантов** действий. С каждым действием связан некоторый набор сопутствующих ему объектов и характеристик. Они, как правило, совпадают с глубинными падежами Филмора. Если, например, мы имеем дело с действием «*идти*», то с ним тесно связаны субъект, совершающий это действие, пункты начала и конца движения, цель движения и т. п. Это позволяет связать с глаголом «*идти*» некоторую структуру с набором пустых пока мест

ИДТИ (Субъект: \_\_\_\_\_ Куда: \_\_\_\_\_ Откуда: \_\_\_\_\_ Цель: \_\_\_\_\_).

Заполненной структуре типа

ИДТИ (Субъект: ПЕТР Куда \_\_\_\_\_ Откуда \_\_\_\_\_ Цель: за молоком)

соответствует фраза: «*Петр идет за молоком*».

Таким образом, процесс синтеза текстов представляется в виде ряда следующих друг за другом шагов. На первом шаге генерируется нужная последовательность глаголов-действий. На следующем шаге заполняются их актантные структуры, что приводит к появлению глубинной семантической структуры отдельных предложений. Затем эти структуры связываются с учётом общих действующих субъектов и используемых объектов, а также иных связывающих параметров в единый текст.

Синтез текстов производится не только в системах автоматического перевода. Таким образом работает, например, автоматическая система создания текстов волшебных сказок, носящая название TALE (сказка), созданная в нашей стране в 70-х годов. На первом шаге она выдает тексты примерно такого вида: «*Жил-был X. Не было у X желаемого Y. Стал просить X Бога. Бог обещал. Появился Y. Вырос Y. Ушел, раз X и не велел Y де-*

*лать Z. Но Y сделал Z. Вернулся X. Y нет. Понял X, что Y сделал Z. Пошел X искать Y ...»* В памяти системы хранятся данные для заполнения актантов, а одинаковые переменные показывают, что на эти места всюду надо поставить одни и те же заполнители. Так возникает текст: *«Жил-был царь. И не было у царя желаемого наследника. Стал царь просить Бога. Бог обещал. Появился наследник. Вырос наследник...»*

Для случая текстов типа волшебных сказок используются формальные специальные грамматики, созданные впервые советским ученым В. Проппом<sup>1</sup> еще в конце 20-х гг. Эти грамматики позволяют строить последовательности действий, не нарушающие логического порядка повествования (например, действие «*ушел X*» относительно действия «*X отсутствует*» возникнет раньше).

✓ Понимание текстов.

Проблема понимания текстов на естественном языке включает не только лингвистические аспекты. С ней тесно связаны задачи, традиционно решаемые в рамках психологии, философии и семиотики. Анализ текста сам служит инструментом для понимания содержания текста.

✓ Оживление текста.

Это направление своим появлением обязано персональным компьютерам, которые впервые дали возможность организовать общение с пользователем не только путем обмена текстами, но и посредством зрительных образов на экране дисплея. Одной из особенностей мышления человека (едва ли не основной для возможности самого мышления) является его разномодальность. Психологи пользуются этим термином, чтобы подчеркнуть, что наши представления об окружающем мире и о нас самих могут иметь различную природу (различную модальность). Две модальности: символическая (текстовая) и зрительная — являются для человека основными. Легко проверить, что между этими модальностями имеется весьма тесная связь. Обычно название чего-то или текстовое описание некоторой ситуации тут же вызывает зрительные представления об этих объектах и ситуациях. И наоборот, стоит нам увидеть нечто, как мы тут же готовы описать увиденное с помощью нашего родного языка. Так текст и сопутствующая ему зрительная картина оказываются объединенными в нашем сознании и интегрированными в некоторое единство.

Изучение того, как происходит эта интеграция и как по одной составляющей представления появляется вторая, — одна из увлекательных задач, стоящих перед специалистами в области компьютерной лингвистики и их коллегами — создателями интеллектуальных систем. Уже найдены неко-

---

<sup>1</sup> Пропп, Владимир Яковлевич (1895–1970), русский фольклорист, один из основоположников современной теории текста. Его книга «Морфология сказки» стала мировым научным бестселлером

торые важные законы интеграции текстов и зрительных образов. Созданы первые экспериментальные модели этого процесса и первые интеллектуальные системы, способные описывать в виде текста предъявляемую им картинку (например, пейзаж), а также воссоздавать одну из возможных картин, соответствующих введенному в систему тексту.

✓ Модели коммуникации.

Появление искусственных систем, способных воспринимать и понимать человеческую речь (пока в весьма ограниченном объеме) и тексты на естественном языке, создало предпосылки для непосредственного общения человека и компьютера. Это, в свою очередь, повысило интерес лингвистов к процессам, сопутствующим организации и ведению диалога. Примерами могут служить: способ построения сценария диалога на основе тех целей, которые активная сторона в диалоге ставит перед собой; поддержка выбранного сценария с учетом интересов партнера и его возможного противодействия тому сценарию, который используется; нахождение средств маскировки истинных намерений говорящего; организация пассивной поддержки коммуникационного процесса и т. д.

Эти направления, которые активно развиваются в компьютерной лингвистике, естественно не исчерпывают всего содержания этой науки. Но сказанного вполне достаточно, чтобы оценить ее важность и значимость не только для самой лингвистики, но и для создания технических систем, по способностям к диалогу не уступающих человеку.

### 1.3 Компьютерный анализ текста

Текст или речь?

В соответствии с традицией инженерно-лингвистической и вопреки традиции классического общего языкознания (де Соссюра<sup>1</sup>, Щербы<sup>2</sup>):

Текст — письменная форма речи

Речь — звуковая форма текста

Отражение уровневой организации языка в архитектуре систем компьютерного анализа естественно-языкового (ЕЯ) текста:

- ✓ фонетический уровень;
- ✓ графематический уровень;
- ✓ морфологический уровень;

---

<sup>1</sup> Фердинанд де Соссюр (Ferdinand de Saussure, 1857 – 1913, – швейцарский лингвист, которого часто называют отцом лингвистики XX века

<sup>2</sup> Щерба Лев Владимирович, 1880 –1944, – языковед, академик АН СССР, внёсший большой вклад в развитие психолингвистики, лексикографии и фонологии. Автор фразы «Глокая куздра штеко будланула бокра и курдячит бокренка».

- ✓ синтаксический уровень;
- ✓ семантический уровень.

**Фонетика** (от греч. φωνή — «звук», φωνηεντικός — «звуковой») — раздел языкознания, изучающий речевые звуки и звуковое строение языка (слоги, звукосочетания, закономерности соединения звуков в речевую цепочку). Различает звуки гласные и согласные, звонкие и глухие и т. д. Изучает:

- звук речи с точки зрения его создания, какие органы речи участвуют в его произношении;
- звук как колебание воздуха и фиксирует его физические характеристики: частоту (высоту), силу (амплитуду), длительность
- функции звуков в языке, оперирует фонемами, т. е. минимальными языковыми единицами, обладающими смыслоразличительной функцией.

### Графематический анализ

Этап графематического анализа предназначен для выделения элементов структуры текста: параграфов, абзацев, предложений, отдельных слов и т. д. В задачу графематического анализа входят:

- Выделение абзацев, заголовков, примечаний;
- Выделение предложений из входного текста;
- Разделение входного текста на слова, цифровые комплексы, формулы и т. д. *Токенизация*;
- Сборка слов, написанных в разрядку;
- Выделение устойчивых оборотов, не имеющих словоизменительных вариантов;
- Выделение ФИО (фамилия, имя, отчество), когда имя и отчество написаны инициалами;
- Выделение иностранных лексем, записанных латиницей;
- Выделение электронных адресов и имен файлов;

Выделение предложений из сплошного текста — процедура необходимая для дальнейшего анализа текста в любой системе анализа естественных языков.

Что такое предложение? Первый ответ на этот вопрос — это что-то, заканчивающееся на символы «.», «!», или «?». Но если рассмотреть встречающиеся тексты более внимательно, то можно обнаружить, что «.» используется не только для определения конца предложения, но и для аббревиатур и сокращений, а иногда выполняет обе эти роли. Другие знаки пунктуации также могут использоваться для выделения фрагментов, которые мы могли бы идентифицировать как предложения. Иногда эти фрагменты выделяются такими знаками как «:», «;» и «—».

А что такое слово? Последовательность символов, ограниченных пробелами или знаками препинания? Обычно это так, но не всегда. Сложности представляют алфавитно-цифровые комплексы (3.142; 2/3; 15-летний; 31.12.2012), составные слова (*кресло-кровать*; *Тот-кого-нельзя-называть*), имена собственные (*Санкт-Петербург*), неразрывные неизменяемые словосочетания (и\_так\_далее; таким\_образом), интернет-адреса (<http://yandex.ru>) и т. д. Минимальные линейные компоненты текста, которые в дальнейшем рассматриваются как неделимые единицы, называются **токенами**.

## Морфологический анализ

**Морфология** — это та часть грамматического строя языка, которая объединяет грамматические классы слов (части речи), принадлежащие этим классам грамматические (морфологические) категории и формы слов. На этапе морфологического анализа обрабатываются отдельные слова, в них выделяются основы и флексии (изменяемые части слов) — приставки, суффиксы, окончания. В дальнейшем флексии используются для установления грамматических отношений между словами в рамках одного предложения.

Слова как единицы грамматические и лексические группируются в части речи, т. е. в грамматические классы слов, объединяющиеся на основании обобщенных значений. Обобщенное значение, характеризующее все слова той или иной части речи, есть абстрактное представление того общего, что присутствует в лексических и морфологических значениях конкретных слов данного класса. Наиболее обобщенными значениями для частей речи являются значения предмета (субстанции) и признака — процессуального (представляемого как действие или состояние) и непроцессуального (представляемого как качество или свойство).

Так, все слова, входящие в часть речи «имя существительное», обладают значением предметности: они называют субстанции — конкретные предметы или предметно представленные факты, события, явления, свойства, качества, понятия и действия. Все слова, входящие в часть речи «глагол», обладают значением процессуального признака; они называют признаки (действия или состояния) как процессы. Все слова, входящие в части речи «прилагательное» и «наречие», обладают значением непроцессуального признака: они называют признаки как свойства или качества предмета (прилагательные и отчасти наречия) или как качества другого признака — процессуального или непроцессуального (наречия).

Морфологический анализ обеспечивает определение нормальной формы, от которой была образована данная словоформа, и набора параметров, приписанных данной словоформе. Нормальная форма (именительный падеж для существительных, инфинитив для глаголов и т. д.) называется **леммой**, а сам процесс определения лемм — **лемматизацией**. Лемма-

тизация производится для того, чтобы ориентироваться в дальнейшем только на нормальную форму, а не на все словоформы, используя параметры, например, для проверки согласования слов.

Каждая часть речи имеет свой собственный комплекс параметров, или грамматических категорий, в которых представлено то обобщенное значение, которое свойственно всем словам этой части речи. Так, значение предметности, свойственное существительному, грамматически представляется морфологическими категориями рода, числа и падежа; значение процесса, свойственное глаголу, — категориями вида, залога, наклонения, времени и лица.

### Синтаксический анализ

**Синтаксис** — часть грамматики, которая имеет дело с единицами, более протяженными, чем слово, — словосочетаниями и предложениями.

При синтаксическом анализе во-первых, предложение надо разделить на составляющие части меньшей длины. Например, рассматривая предложение *Сырые дрова плохо горят*, следует убедиться в том, что это предложение состоит из четырех слов [*Сырые*], [*дрова*], [*плохо*] и [*горят*], и, кроме того, делится на [*Сырые дрова*] и [*плохо горят*]. Во-вторых, необходимо установить отношения между частями предложения. Например, надо понять, что *сырые* — определение при слове *дрова*, а *горят* — сказуемое.

В результате синтаксического анализа линейная последовательность токенов (слов) преобразуется в набор синтаксических отношений. Грамматично построенные предложения являются связными, т. е. лишенными разрывов в цепочке синтаксических отношений. Отношения являются бинарными. Синтаксическое отношение неравноправно: определяемое слово «главнее» своего определения. Важная особенность синтаксических зависимостей заключается в том, что они далеко не всегда связывают слова, находящиеся рядом в цепочке.



Выполнение задачи осложняется огромным количеством альтернативных вариантов, возникающих в ходе разбора, связанных как с многозначностью входных данных (одна и та же словоформа может быть получена от различных нормальных форм), так и неоднозначностью самих правил разбора.

## Семантический анализ

**Семантика** (от греч. *σημαντικός* — обозначающий) — часть анализа, направленная на решение задач, связанных с возможностью определения значения слова в зависимости от контекста и конкретной ситуации, понимания смысла фразы. Элемент значения языкового знака называется *семой* (используются также термины *семантический компонент*, *семантический множитель*, *семантический маркер*, *дифференциальный признак* и некоторые другие).

Например, *холостяк* = [взрослый] [неженатый] [мужчина]; *мужчина* = [человек] [мужского пола] и т. д. Однако представление значения слова в виде набора сем часто не позволяет объяснить его реальное употребление. Например, никому не придет в голову назвать *холостяком* папу римского. Одна из основных задач семантической декомпозиции состоит в том, чтобы объяснить и даже предсказать особенности сочетаемости толкуемого слова. Например, мы говорим *пить чай*, а не *есть чай* потому, что слова *чай* и *пить* содержат сему [жидкий], а слово *есть* — нет.

Во многих случаях смысловой элемент состоит из нескольких слов. Последовательность из двух или более слов, частотность совместного появления которых в тексте выше, чем ожидаемая вероятность их совместного появления, называется **коллокацией**. В отличие от свободного словосочетания (*красивый мальчик/хороший мальчик/красивый цветок*), коллокация определяет, какие слова могут быть использованы вместе, например, какими предлогами управляет тот, или иной глагол (*уйти от кого-то, чего-то, но прийти к кому-то/чему-то*), или какие глаголы и существительные обычно используются вместе. Например, можно сказать *мощный двигатель* и *крепкий чай*, но нельзя, не меняя значения, заменить эти коллокации на словосочетания *крепкий двигатель* и *мощный чай* соответственно. В коллокациях выбор одного из компонентов (ключевого слова) осуществляется по смыслу, а выбор второго (коллоканта) зависит от выбора первого (например, *ставить условия* — выбор глагола *ставить* определяется традицией и зависит от существительного *условия*, при слове *предложение* будет другой глагол — *вносить*).

Коллокации частично некомпозициональны, то есть значение целого не равно сумме значений частей. Полностью некомпозициональные словосочетания, у которых смысл никак не соотносится со смыслом отдельных слов, называются **идиомами** (*дать дуба, медведь на ухо наступил*).

К коллокациям также часто причисляют составные топонимы и другие совместно употребляемые наименования (*крейсер «Варяг», Кировский завод*). В некоторых случаях коллокации могут быть разрывными: «*жизнь кипит*» и «*жизнь его постоянно кипит*».

С помощью компьютерных технологий коллокации могут автоматически извлекаться из текстов. Для этого используются различные меры ас-

социативной связи, которые оценивают, является ли взаимное появление лексических единиц случайным, или оно статистически значимо. Однако часто статистически значимое совместное появление двух слов не образует коллокации, например, словосочетание *Гарри Поттер* в текстах про Гарри Поттера.

## 1.4 Задачи лингвистических информационных технологий

✓ **Распознавание** звучащей речи и **синтез речи** по тексту. Первое устройство для распознавания речи появилось в 1952 году, оно могло распознавать произнесённые человеком цифры. Существует несколько основных способов распознавания речи. Распознавание отдельных команд из небольшого заранее заданного словаря позволяет достичь самой высокой достоверности распознавания. Примером использования является голосовая навигация по сайтам. Распознавание фраз, соответствующих определенным заданным правилам (грамматике) широко применяется в системах голосового самообслуживания. Поиск ключевых слов в потоке слитной речи, в этом случае речь не полностью преобразуется в текст — в ней автоматически находятся лишь те участки, которые содержат заданные слова или словосочетания. Используется в поисковых системах, в системах мониторинга речи. Распознавание слитной речи на большом словаре — эта технология наиболее близка к мечте человека о взаимодействии человека и машины — все, что сказано, дословно преобразуется в текст. Поэтому иногда эта технология так и называется STT — *speech to text*. До конца эта задача не решена нигде в мире, однако, достоверность распознавания уже достаточно высока для использования технологии на практике.

✓ **Поддержка ввода текста** на электронные носители. Одним из первых приложений в этом направлении были программы *автоматического переноса* слов и программы орфографической проверки текста (*спеллеры*). Эти программы обеспечивают коррекцию на лексико-морфологическом и синтаксическом уровне, т. е. производят 1) сличение ввода со списком допустимых структур (распознавание с дискретным входом) и, в случае неудачи, 2) поиск ближайшего соответствия. Близки к этим задачам также распознавание печатного и рукописного текста и автозавершение.

✓ **Машинный перевод.** Первые программы перевода были построены более 50 лет назад и были основаны на простейшей стратегии пословного перевода. Однако довольно быстро было осознано, что машинный перевод требует полной лингвистической модели, учитывающей все уровни языка, вплоть до семантики и прагматики. В настоящее время существует целый спектр компьютерных систем перевода разного качества, но, несмотря на многие десятилетия развития всего этого направления, в целом задача машинного перевода еще весьма далека до полного решения.

✓ **Информационный поиск.** Создание поискового образа документа предполагает *индексирование* его текста, т.е. выделение в нем ключевых слов. Поскольку очень часто гораздо точнее тему и содержание документа отображают не отдельные слова, а словосочетания, в качестве ключевых слов стали рассматриваться словосочетания. Это существенно усложнило процедуру индексирования документов, так как для отбора значимых словосочетаний текста потребовалось использовать различные комбинации статистических и лингвистических критериев.

✓ **Компрессия** текста (реферирование и аннотирование). Решение этой задачи состоит из двух этапов: 1) сегментация на высказывания (части высказываний) и затем 2) выбор наиболее значимых (синтез). **Реферирование** текста — сокращение его объема и получение его краткого изложения — реферата. Общий реферат может составляться также для нескольких близких по теме документов. Основным методом автоматического реферирования является отбор наиболее значимых предложений реферируемого текста, для чего обычно сначала вычисляются ключевые слова текста, и рассчитывается коэффициент значимости предложений текста. Близкая к реферированию задача — **аннотирование** текста документа, т. е. составление его аннотации. В простейшей форме аннотация представляет собой перечень основных тем текста, для выделения которых могут использоваться процедуры индексирования.

✓ **Классификация** текстов. При создании больших коллекций документов актуальны задачи классификации и кластеризации текстов с целью создания классов близких по теме документов. **Классификация** означает отнесение каждого документа к определенному классу с заранее известными параметрами, а **кластеризация** — разбиение множества документов на кластеры, т. е. подмножества тематически близких документов. Очень близка к классификации задача **рубрицирования** текста — его отнесение к одной из заранее известных тематических рубрик (обычно рубрики образуют иерархическое дерево тематик). Задача классификации получает все большее распространение, она решается, например, при распознавании спама, а сравнительно новое приложение — классификация SMS-сообщений в мобильных устройствах.

✓ **Извлечение фактов и знаний** (Information Extraction). Извлечение информации из текстов часто требуется при решении задач экономической и производственной аналитики. Для этого осуществляется выделение в тексте ЕЯ определенных объектов — именованных сущностей (имен, персоналий, географических названий), их отношений и связанных с ними событий. Как правило, это реализуется на основе частичного синтаксического анализа текста, позволяющего выполнять, например, обработку потоков новостей от информационных агентств.

✓ **Анализ нормативных текстов.** Тексты законов, постановлений, планы работ анализируются на предмет выявления противоречий, логических пропусков и т. д. Так, например, анализ одного из региональных законов выявил, что в нем полностью описано (кто выполняет, что выполняет, форма отчетности и т. д.) только 30% необходимых действий. Естественно, такой закон нормально функционировать не может.

✓ **Анализ «под заказ»** — распознавание заранее заданных сюжетных схем.

✓ **Вопросно-ответные системы (Question Answering).** Эта задача решается путем определения типа вопроса, поиском текстов, потенциально содержащих ответ на этот вопрос, и извлечением ответа из этих текстов.

Задачи реферирования, выделения феноменов и понятий, классификации и кластеризации, ответов на запросы, тематического индексирования и поиска по ключевым словам принято относить к технологиям **Text Mining**, т. е. интеллектуального анализа текстов.

✓ **Диалог с компьютерными системами** на естественном языке. Наиболее часто эта задача решалась для специализированных баз данных — в этом случае язык запросов достаточно ограничен (лексически и грамматически), что позволяет использовать упрощенные модели языка. Запросы к базе, сформулированные на ЕЯ, переводятся на формальный язык, после чего выполняется поиск нужной информации и строится соответствующая фраза ответа.

## 2 Инструментарий компьютерной лингвистики

### 2.1 Словари

Тип отдельного словаря определяется основной информацией, которую он содержит, его общим назначением. Прежде всего, словари делятся на **энциклопедические** и **лингвистические**. В энциклопедических словарях содержится информация о предметах, понятиях, явлениях, в лингвистических — информация о словах, называющих предметы, понятия, явления и пр.

Лингвистические словари можно подразделить на многоязычные, двуязычные и одноязычные. Много- и двуязычные словари используются для перевода. В них значения слов одного языка объясняются через сопоставление со словами другого языка. В одноязычных словарях слова объясняются посредством слов того же языка.

По функциям и цели создания толковые словари разделяются на **дескриптивные** и **нормативные**. Дескриптивные словари предназначены для полного описания лексики определенной сферы и фиксации всех имеющихся там употреблений; в них фиксируются все имеющиеся релевантные случаи. Оценка качества дескриптивного словаря зависит от того, с какой степенью полноты словник словаря отражает проблемную область, насколько точно описаны значения лексем, представленных в материале. Типичным примером дескриптивного словаря является «Толковый словарь живого великорусского языка» В. И. Даля (первое издание в четырех томах выходило в 1863-1866 гг.). Цель создателя словаря заключалась не в нормировании языка, а в по возможности полном описании всего многообразия великорусской речи — в том числе ее диалектных форм, просторечия. Дескриптивными по определению являются словари сленгов и жаргонов, диалектные словари. Понятие нормы вполне может быть применено и к сленгу, и к жаргону, однако норма в этих сферах бытования языка, как правило, менее устойчива и — что наиболее существенно — не становится объектом языковой политики государства.

Цель нормативного словаря — дать норму употребления слова, исключив не только неправильные употребления слов, связанные с неправильным пониманием их значений, но и те употребления, которые не соответствуют коммуникативной ситуации (литературный язык vs<sup>1</sup> диалект vs жаргон), ср. *отчини* вместо *открой*, *стрелка* вместо *встреча* и пр. Иными словами, нормативные словари рекомендуют, предписывают стандарт употребления слова, задают литературную норму. Первым нормативным сло-

---

<sup>1</sup> Vs – сокращение от латинского *Versus* – против.

варем русского языка двадцатого века является четырехтомный «Толковый словарь русского языка» под редакцией Д. Н. Ушакова, вышедший с 1935 по 1940 гг. Авторский коллектив словаря видел свою задачу в «попытке отразить процесс переработки словарного материала в эпоху пролетарской революции, полагающей начало новому этапу в жизни русского языка, и вместе с тем указать нормы употребления слов».

Среди типов одноязычных словарей можно выделить:

✓ Толковые — содержат в себе слова и понятия языка с кратким описанием того, что эти слова означают, часто сопровождая толкование примерами использования слов. Толковый словарь изъясняет лексическое значение того или иного слова.

✓ Тезаурусы — охватывают понятия, определения и термины специальной области знаний или сферы деятельности. Тезаурусы способствуют пониманию в общении и взаимодействии лиц, связанных одной дисциплиной или профессией.

✓ Идеографические — словари, в которых статьи упорядочены не по алфавиту, как обычно, а по смыслу (лексическому значению заглавного слова или фразы). Если алфавитный словарь служит для того, чтобы узнать что-то о данном слове, то идеографический словарь служит для того, чтобы узнать что-то о данном смысле — например, какими словами можно выразить данное значение.

Кроме того, словари могут различаться по принципам отбора лексики, например:

- по сфере употребления (разговорные, диалектные, поэтической лексики, ...);
- историческому аспекту (архаизмов, неологизмов, ...)
- по раскрытию отдельных параметров слова (орфографические, этимологические, ...);
- по раскрытию системных отношений между словами (словообразовательные, омонимические, синонимические, ...);
- частотные;
- обратные (см. табл. 2.1);
- и т. д.

Рассмотрим на примерах структуру и содержание словарных статей в словарях некоторых типов.

### Орфографический словарь

До<sup>2</sup>, нескл., с. (нота).

До<sup>1</sup>, предлог.

Добавить(ся), -влю, -вит(ся)

<...>

Доблесть, -и.

## Толковый словарь

### ДО

1. предлог, кого-чего.

1. Указывает на пространственный, временной или количественный предел действия, движения, состояния, качества и т.п.

%% Добежать до леса.

%% До деревни три километра

<...>

2. Указывает на высшую (предельную) степень, которой достигает действие, состояние и т.п.

%% Промокнуть до нитки

<...>

3. Указывает на действие, которому предшествует другое действие.

%% Уехать до первых заморозков.

<...>

### ДО

2. неизм.; ср.

Начальный звук музыкальной гаммы; нота, обозначающая этот звук.

%% Нижнее до.

<...>

### ДОБАВИТЬ

&& -влю, -вишь; св. что (чего).

1. Прибавить (1 зн.).

%% Д. краску в раствор.

%% Д. сто рублей.

2. Сказать или написать в дополнение; дополнить.

%% Д. несколько слов к письму.

<...>

### ДОБЛЕСТЬ

&& -и; ж.

Высок. Отвага, мужество, геройство.

%% Воинская д.

%% Пример доблести русских моряков.

## Этимологический словарь

ДО (предлог). Общеслав. индо-европ. характера (ср. нем. zu "к", англ. to — тж., лат. do в endo "внутри" и т. д.). Значение "до" выводится из значения "к". Того же корня, что и да.

ДОБЛЕСТЬ. Заимств. из ст.-сл. яз., где является суф. производным от добль "храбрый", того же корня, что и добрый

### Словарь синонимов

ДОБАВИТЬ+> прибавить; добавить, присоединить, придать, подбавить; подбросить, подкинуть, поддать, подпустить (разг.); прибросить, подвалить (прост.); присовокупить (книжн.) / обычно о деньгах: набавить; надбавить, накинуть, прикинуть (разг.) / при подсчете: причислить; приплюсовать (разг.) / к сказанному, написанному: примолвить (устар.)

<...>

ДОБЛЕСТЬ+> смелость; отвага, храбрость, мужество, безбоязненность, бесстрашие, неустрашимость, доблесть, героизм, геройство, дерзость; кураж (разг. шутл.); дерзание, дерзновенность (высок.); дородство (устар.); дерзновение (устар. высок.)

### Обратный словарь

В обратном словаре словарные единицы расположены в алфавитном порядке «задом наперед». Может быть полезен, например, для нахождения рифм. Четыре фрагмента обратного словаря приведены в табл. 2.1.

**Таблица 2.1.** Фрагмент обратного словаря

жаба	приправа	юнга	дымка
раба	оправа	тревога	рюмка
полнеба	справа	синагога	ямка
треба	расправа	изжога	лямка

### Идеографический словарь

**Таблица 2.2.** Фрагмент идеографического словаря

ПРИРОДА	ЯВЛЕНИЕ	ЯВЛЕНИЕ	ЯВЛЕНИЕ (феномен)	явление – природный объект; форма проявления сущности феномен - необычное явление.
			СОВЕРШАТЬСЯ	происходить с кем-чем, изменяться во времени; существовать динамически.
			ПРОЦЕСС	динамическое состояние.
	ВРЕМЯ	ПЕРИОД ВРЕМЕНИ	промежуток времени; время (через какое-то #); дни, годы, эпоха.	

			ДЛИТЕЛЬНОСТЬ	протяженность существования.
			МОМЕНТ	временная точка; элемент времени, связанный с чем-л.; фаза существования.
			РАНЬШЕ (чего)	до какого-л. момента, прежде.
		ВЗАИМОДЕЙСТВИЕ	ВОЗДЕЙСТВИЕ	изменяющий фактор; активное непосредственное влияние на что-л.
			ЭФФЕКТ	результат воздействия (давать #. потрясающий #).
			ПЕРЕДАЧА (явление)	изменение принадлежности объекта.
	МАТЕРИЯ	МАТЕРИЯ	ПРИРОДА	весь материальный мир в многообразии его форм, проявлений.
			ЭФИР	первичная мировая несжимаемая среда; начало всех начал (разг.).
			ИЗЛУЧЕНИЕ	излучать – выделять лучистую энергию.
	ВЕЩЕСТВО	СОСТАВ ВЕЩЕСТВА	МОЛЕКУЛА	наименьшая частица химического вещества.
			ХИМИЧЕСКОЕ СОЕДИНЕНИЕ	соединение химических элементов.
			ХИМИЧЕСКАЯ РЕАКЦИЯ	изменение химического состава; взаимодействие атомов, сопровождающееся качественной перестройкой их электронных оболочек.

## Частотный словарь

Частотный словарь — вид словаря, в котором лексические единицы характеризуются с точки зрения степени их употребительности в совокупности текстов, представительных либо для языка в целом, либо для отдельного функционального стиля, либо для одного автора. Словарь может быть отсортирован по частоте, по алфавиту (тогда для каждого слова будет указана его частота), по группам слов (например, первая тысяча наиболее частотных слов, за ней вторая) и т. д. Частотные словари могут строиться на основе словоформ, лемм (нормальных форм слова) или словосочетаний.

В связи с тем, что размеры корпусов, на основании которых составляется словарь, различны, обычно производится приведение частот встречаемости к относительным единицам *ipm* (частота на миллион словоформ, *instances per million words*).

Естественно, что практически в любом наборе текстов на первых местах по встречаемости будут служебные слова — союзы, предлоги и т. д. Самое частотное слово русского языка — союз *и*, имеющее частоту около 30000 *ipm*. Для Национального корпуса русского языка (НКРЯ, см. ниже раздел 2.3) наиболее частотные существительные, глаголы и прилагательные приведены в табл. 2.3.

**Таблица 2.3.** Частотные слова русского языка по НКРЯ

существительные		глаголы		прилагательные	
частота	слово	частота	слово	частота	слово
2369	человек	8900	быть	876	новый
1529	время	2398	мочь	554	последний
1490	год	2053	сказать	473	русский
1195	дело	1492	говорить	456	хороший
1119	жизнь	1427	знать	429	большой
1024	рука	1291	есть	373	высокий
1005	день	1186	стать	362	российский
839	слово	849	хотеть	339	молодой
835	раз	793	иметь	339	великий
747	глаз	758	видеть	326	старый

Частотные словари широко применяются как в компьютерной лингвистике (например, для классификации текстов), так и в лингвистике традиционной (например, для сравнения лексики разных авторов, анализа изменения лексики с течением времени и т. п.). Так, например, в частотном словаре лексики Лермонтова среди существительных лидируют слова *рука*, *душа*, *день*; среди глаголов на пятом месте идет *любить* (в НКРЯ — 21-е место).

Сравнивая лексику Лермонтова и Пушкина, исследователи отмечают, что у Пушкина больше слов разговорных, народных. Например (Лермонтов : Пушкин): *девица* — 38:108, *баба* — 4:44, *батька* — 0:5, *попадья* — 1:28, *печка* — 1:14.

Проблемы при создании частотных словарей заключаются в:

- воспроизводимости (будут ли результаты идентичны на другом аналогичном корпусе);
- всплесках частоты отдельных слов (частота слова в одном тексте может повлиять на его позицию в частотном списке);
- сложности определения позиции менее частотных слов, что не дает возможности ранжировать их рационально; например, слово *белиберда* входит в 20000 наиболее частотных слов, в то время как слово *хрюкнуть* находится за пределами списка первых 40 тысяч;
- омонимичности многих словоформ (см. раздел 3.1) (*стали* => *сталь* или *стать*, *банку* => *банк* или *банка*, *вера* => *вера* или *Вера*).

В словарях предшествующих поколений, составлявшихся в конце XX века, омонимия разрешалась вручную, так как объем обрабатываемого корпуса был незначителен. Очевидно, что для 100-миллионных корпусов такое решение не подходит. Поэтому задействуются системы компьютерного анализа текстов. Однако это порождает новую проблему: слова, отсутствующие в словаре анализатора. При составлении частотного словаря НКРЯ доля несловарных слов составляла 3% всех словоупотреблений и 45% списка словоформ корпуса. Автоматическая интерпретация несловарных форм в свою очередь может приводить к ошибкам, вызывая появление таких лемм как *Янсный* (от *Янсен*), *Барклаивать* (от *Барклай*).

Вопрос о способе лемматизации и ее необходимости вообще при автоматическом составлении частотных словарей нетривиален, и его решение зависит от целей работы. Рассмотрим для примера четыре фразы, представляющие собой ответы на один из вопросов социологической анкеты:

*Демократы не варвары, они защищают себя.*

*Защитные свойства демократии.*

*Защищались они варварски.*

*Российская демократия не встала на защиту населения.*

В таблице 2.4 приведены два варианта частотного словаря. В левой части — словарь с лемматизацией. В правой части словарь составлен без лемматизации, почти все словоформы одного слова, а также близких к ним слов, объединены в одну словарную единицу под названием той словоформы, которая встретилась в тексте первой. Близость слов определялась по специальным правилам сравнения слов.

**Таблица 2.4.** Два типа частотного словаря

С лемматизацией		Без лемматизации	
2	ДЕМОКРАТИЯ	3	ДЕМОКРАТЫ
2	НЕ	2	ВАРВАРЫ
2	ОНИ	2	ЗАЩИТУ
1	ВАРВАР	2	НЕ
1	ВАРВАРСКИ	2	ОНИ
1	ВСТАТЬ	1	ВСТАЛА
1	ДЕМОКРАТ	1	ЗАЩИЩАЛИСЬ
1	ЗАЩИТА	1	ЗАЩИЩАЮТ
1	ЗАЩИТНЫЙ	1	НА
1	ЗАЩИЩАТЬ	1	НАСЕЛЕНИЯ
1	ЗАЩИЩАТЬСЯ	1	РОССИЙСКАЯ
1	НА	1	СВОЙСТВА
1	НАСЕЛЕНИЕ	1	СЕБЯ
1	РОССИЙСКИЙ		
1	СВОЙСТВО		
1	СЕБЯ		

Безусловно, левый столбец выглядит более приятным. Однако легко заметить, что справа слова, близкие по значению, объединяются в одну словарную единицу, что может оказаться полезным при классификации. Кроме того, при составлении словаря без лемматизации снимаются проблемы омонимии и несловарных слов, а программа работает во много раз быстрее.

### Способы сравнения слов

Частотные словари, представленные в таблице 2.4 были получены с помощью диалоговой система классификации и анализа текстов Вега, разработанной в ИТМО и Экономико-математическом институте РАН. В ней используются два способа сравнения слов. Первый способ «**по основе**» заключается в том, что вначале определялся размер «основы» слова (здесь под основой понимается не грамматическая основа, а начальная часть слова). Основа определяет то количество символов, которое учитывается при сравнении двух слов между собой. Для подавляющего большинства слов размер основы коррелирует с размером самого слова (табл. 2.5).

**Таблица 2.5.** Размер основы слова

Длина слова	1	2	3	4	5	6	7-10	11	12-14	15-17	18-20	21-23	>23
Длина основы	1	2	3	4	5	6	7	8	9	12	15	18	21

Алгоритм определения длины основы заключается в следующем. Во-первых, отбрасываются окончания *-ся* или *-сь*. Во-вторых, если словоформа оканчивается на одно из стандартных сочетаний букв, соответствующее окончание нужно «вычесть». Ниже приведен фрагмент таких окончаний:

End1 = «АЕИЙОУЫЬЮЯ»

End2 = «АМ АХ АЯ ЕВ ЕЕ ЕЙ ЕМ ЕТ ИЕ ИИ ИЙ ИМ ИТ ИХ ИЮ ИЯ ОВ...»

End3 = «АМИ ЕГО ЕМУ ИЕЙ ИЕМ ИМИ ИМИ ИТЬ ИЯМ ИЯХ ЛСЯ ОГО...»

End4 = «АЯСЯ ЕЕСЯ ЕМСЯ ЕНИЕ ЕНИИ ЕНИЙ ЕНИЮ ЕНИЯ ЕТСЯ ИЕСЯ...»

End5 = «ЕГОСЯ ЕНИЕМ ЕНИЯМ ЕНИЯХ ЕШЬСЯ ИТЬСЯ ОСТЕЙ ОСТЬЮ ...»

Если длина основы по таблице равна  $n$ , а после вычитания окончаний —  $m$ , то окончательно за основу берется начальный фрагмент слова, имеющий минимальную длину из этих двух вариантов.

Однако можно выделить достаточно большое количество слов, которые являются исключением из этого правила. Для таких «нестандартных» слов существует специальный вспомогательный словарь. Например, размер основы равен 3 для следующих слов: *дочерью, дочка, дочкой*. Основа для всех этих слов будет «ДОЧ». Таким образом, при сравнении сравнивая любое из этих слов со словом *дочь*, у которого длина основы тоже 3 (4 – 1 по списку окончаний), получается совпадение.

Второй способ сравнения слов «по лемме» основан на применении морфологического анализатора. Однако результаты морфологического анализа показывают, что около 18 процентов словоформ произведены не от одной леммы, а от нескольких (см. раздел 3.1). В случае такого неоднозначного разбора слова возникает вопрос: какую лемму выбрать? Известно, что полностью снять неоднозначность можно только в том случае, если произвести синтаксический и семантический разбор предложения. Разбор фрагмента предложения (окно на 9 слов) позволяет снизить неоднозначность до 1.5%. Однако и при разборе одного слова можно предпринять некоторые меры по снижению неоднозначности.

Так, например, повелительные формы глаголов и деепричастия — *благодаря, для, зря, мая, моря, нашей, некая, почти, при, секретарь, три, хотя* — встречаются крайне редко и при анализе их можно удалять.

Если слово начинается с прописной буквы, то предпочтение отдается именам собственным. Так, в предложении *Я увидел Варвару* последнее слов получит лемму *ВАРВАРА*, а в предложении *Я подошел к варвару* последнее слово получит лемму *ВАРВАР*.

Леммы выбираются в зависимости от грамматических характеристик в определенном порядке, причем приоритет отдается знаменательным частям речи: наречиям, существительным, прилагательным и глаголам. Например, при разборе слова *совести* получают две леммы: *СОВЕСТИТЬ* и *СОВЕСТЬ*. Поскольку в списке частей речи глагол размещен после существительного, первая лемма удаляется, остается существительное.

## 2.2 Корпуса текстов

Корпус текстов — это некоторое собрание текстов, в основе которого лежит логический замысел, логическая идея, объединяющая эти тексты. Воплощение этой логической идеи: правила организации текстов в корпус алгоритмы и программы анализа корпуса текстов сопряжённая с этим идеология и методология

Первый корпус текстов (так называемый Брауновский корпус) был создан в США в 60-е годы и был предназначен для отражения лингвистических особенностей американской печатной прозы.

Лингвистический, или языковой, корпус текстов — большой, представленный в электронном виде, унифицированный, структурированный, размеченный, филологически компетентный массив языковых данных, предназначенный для решения конкретных лингвистических задач.

**Таблица 2.6.** Отличия компьютерных корпусов текстов

	Корпус текстов Пушкина	Компьютерный корпус текстов
Машинный носитель	– (+)	+
Разметка	– (+)	+
Способ отбора	–	+
Репрезентативность	–	+

**Репрезентативность** — важнейшее свойство корпуса. Корпус должен с максимальной объективностью представить разнообразие изучаемого явления, и дать в то же время объективную картину бытования этого явления в речевой практике носителей данного языка.

**Назначение языкового корпуса** — показать функционирование лингвистических единиц в их естественной контекстной среде.

На основе корпуса можно получить данные:

- ✓ о частоте словоформ, лексем, грамматических категорий,
- ✓ об изменениях частот
- ✓ об изменениях контекстов в различные периоды времени
- ✓ о поведении языковых единиц разных авторов
- ✓ о совместной встречаемости лексических единиц
- ✓ об особенностях их сочетаемости, управления
- ✓ и т. д.

## Классификация корпусов текстов

### **По степени организации и структурированности:**

- ✓ Электронный архив — это тексты на электронном носителе, но их форма, представленная на машинном носителе, не стандартизирована и не унифицирована.
- ✓ Электронная библиотека — тексты здесь представлены однородным и стандартизированным образом.
- ✓ Корпус текстов — форма стандартизирована и унифицирована, тексты предназначены для отражения части лингвистической реальности.

### **По индексации:**

- ✓ Простой.
- ✓ Аннотированный.

### **По языку:**

- ✓ Одноязычный;
- ✓ Двуязычный;
- ✓ Многоязычный.

### **По способу применения и использования корпуса:**

- ✓ Исследовательский;
- ✓ Иллюстративный;
- ✓ Параллельный.

## 2.3 Национальный корпус русского языка (НКРЯ) (<http://ruscorpora.ru>)

**Национальный корпус** представляет данный язык на определенном этапе (или этапах) его существования и во всём многообразии жанров, стилей, территориальных и социальных вариантов и т. п. Национальный корпус имеет две важные особенности. Во-первых, он характеризуется представительностью, или сбалансированным составом текстов. Это означает, что корпус содержит по возможности все типы письменных и устных текстов, представленные в данном языке (художественные разных жанров, публицистические, учебные, научные, деловые, разговорные, диалектные и т. п.), и что все эти тексты входят в корпус по возможности пропорционально их доле в языке соответствующего периода. Следует иметь в виду, что хорошая представительность достигается только при значительном объеме корпуса (десятки и сотни миллионов словоупотреблений).

Во-вторых, корпус содержит особую дополнительную информацию о свойствах входящих в него текстов (так называемую разметку, или аннотацию). Разметка — главная характеристика корпуса, она отличает корпус от простых коллекций (или «библиотек») текстов, в изобилии представ-

ленных в современном интернете, в том числе и на русском языке (таких, как, по-видимому, наиболее известная «Библиотека Мошкова» или, например, «Русская виртуальная библиотека»). Однако такие библиотеки в необработанном виде для научных исследований языка пригодны очень ограниченно. Не следует забывать также, что библиотеки создаются теми, кому интересно в большей степени содержание текстов, чем их языковые качества. Национальный корпус, в отличие от электронной библиотеки, — это не собрание «интересных» или «полезных» текстов; это собрание текстов, интересных или полезных для изучения языка. А такими могут оказаться и роман второстепенного писателя, и запись обычного телефонного разговора, и типовый договор аренды и т.п. — наряду, конечно, с классическими произведениями художественной литературы.

Чем богаче и разнообразнее разметка, тем выше научная и учебная ценность корпуса. В Национальном корпусе русского языка в настоящее время используется пять типов разметки: метатекстовая, морфологическая (словоизменяемая), синтаксическая, акцентная и семантическая.

## Разметка

*Англ.:* tagging, annotation.

Разметка — приписывание текстам и их компонентам специальных меток.

Виды разметки:

- ✓ **экстралингвистическая** (*метаразметка*) — сведения об авторе и сведения о тексте: автор, название, год и место издания, жанр, тематика;
- ✓ **структурная** — глава, абзац, предложение, словоформа
- ✓ собственно **лингвистическая**, в том числе:
  - **Морфологическая**, part-of-speech tagging (POS-tagging)
  - **Синтаксическая** — связи между словами (предикативные, предложные, союзные,...)
  - **Семантическая** — предметные/непредметные имена, части чего-либо, типы действий,...
  - **Анафорическая** — смысл одного элемента текста определяется смыслом другого (*Космонавт вернулся на борт станции. Он сообщил, что чувствует себя нормально*)
- ✓ **Просодическая** — ударения, ритмика речи, логические ударения,...
- ✓ и т. д.

## Структура морфологической информации в НКРЯ

Морфологическая информация, приписываемая произвольному слову в тексте, состоит из четырех «полей», или групп помет:

- ✓ Лексема, которой принадлежит словоформа (указывается «словарная

запись» данной лексемы и ее принадлежность к той или иной части речи).

✓ Множество грамматических признаков данной лексемы, или словоклассифицирующие характеристики (например, род для существительного, переходность для глагола).

✓ Множество грамматических признаков данной словоформы, или словоизменительные характеристики (например, падеж для существительного, число для глагола).

✓ Информация о нестандартности грамматической формы, орфографических искажениях и т. п.

## Семантическая разметка в Национальном корпусе русского языка

Семантическая информация в НКРЯ содержит три группы помет:

✓ Разряд: имя собственное, возвратное местоимение и т.д.

✓ Лексико-семантические характеристики: **таксономия** (тематический класс лексемы) — для имен существительных, прилагательных, глаголов и наречий; **мереология** (указание на отношения «часть — целое», «элемент — множество») — для предметных и непредметных имен; **топология** (топологический статус обозначаемого объекта) — для предметных имен; **каузация** — для глаголов; **служебный статус** — для глаголов; **оценка** — для предметных и непредметных имен, прилагательных и наречий.

Словообразовательные характеристики: **морфо-семантические** (например, «диминутив» — уменьшительное, «семельфактив» — однократное действие); **разряд производящего слова** (например, отглагольное существительное); **лексико-семантический** (таксономический) **тип производящего слова** (например, наречие, образованное от прилагательного размера); **морфологический тип словообразования** (сложное слово).

## Другие корпуса

Хельсинкский аннотированный корпус русских текстов ХАНКО  
(<http://www.ling.helsinki.fi/projects/hanco/>)

Машинный фонд русского языка (<http://cfri.ru/>)

Корпус русского литературного языка (<http://www.narusco.ru/>)

## 3 Автоматический анализ текста

### 3.1 Морфологический уровень

#### Представление морфологической информации

Будем считать, что словоформа характеризуется пятеркой — строкой словоформы; частью речи; нормальной формой (**леммой**), от которой была образована данная словоформа; частью речи нормальной формы; набором морфологических параметров, приписываемых к данной словоформе. Часть речи нормальной формы необходима, так как, например, деепричастие удобно считать формой глагола, а не выводить в отдельное слово.

Полный набор словоформ, образованных от одной леммы, называется **парадигмой**.

Морфологический параметр — это пара <имя параметра, значение параметра>. Именем параметра может служить род, число, время, склонение, краткость формы прилагательного и другие признаки слов, принятые в данном языке. Значение параметра — это конкретное значение, которое может принимать данный признак. Так, например, падеж может быть именительным, родительным, ...; род может быть мужским, женским, средним; число — единственным, множественным и т. д. Параметры равны между собой, если равны их имена и значения.

В ряде случаев значение параметра определить невозможно или в этом нет необходимости. Например, в русском языке существительным во множественном числе не приписывают род. Также существуют слова, которые имеют только форму множественного числа. Если словам, обладающим единственным числом значение рода может быть приписано из единственного числа, то слова, не обладающие единственным числом (очки, часы), такой информации лишены полностью. В этом случае можно считать, что значение параметра нулевое

Среди параметров слова выделяют словообразовательные и формообразовательные. Словообразовательные параметры не изменяются при изменении слова по формам. Так, например, слово «мама» остается женского рода в любой своей форме. Формообразовательные параметры изменяются при изменении слова по формам. Для существительных падеж будет формообразовательным параметром. Словообразовательные параметры для одних частей речи могут являться формообразовательными для других. Например, параметр рода не меняется у существительных, однако будет образовывать формы у прилагательных и глаголов.

В русском языке количество словоформ для одной леммы может быть очень большим. Так если наречия и предлоги имеют только одну форму, то существительные — 12, прилагательные — 24 (без учета краткости и степеней сравнения), а у глаголов число словоформ может превышать 300

(если считать деепричастия и причастия формами глагола).

## Словарь Зализняка

Основой большинства современных реализаций компьютерной морфологии русского языка является грамматический словарь Андрея Анатольевича Зализняка (академик РАН, лауреат Государственной премии России). Впервые издан в 1977 г. Содержит около 100 тыс. слов.

В этом словаре каждому слову сопоставлен словоизменительный класс. Примеры словарных статей:

ГАЗАНУТЬ гсНЗв [г – глагол; с – совершенного вида; Н – непереходный; Зв – словоизменительный класс]

ГАЗЕЛЬ жо8а (антилопа) [ж – женский род (только для существительных); о – одушевленный; 8а – словоизменительный класс]

ГАЗЕЛЬ ж8а (стихотворная форма)

Каждому классу словоизменения соответствуют несколько наборов окончаний

Например, схожие слова класса п1 [п – прилагательное] имеют отличающиеся наборы окончаний:

МАГНИТНЫЙ ая ое ого ой ому ую ом ою ые ых ым ыми ее ей 1ЕН ь  
2Н а о ы

Понимать эту запись следует так. Во-первых, отбрасываем окончание «ый» и вместо него подставляем окончания из списка. Получаем *магнитная*, *магнитное* и т. д. Обозначение «1ЕН» говорит, что нужно отбросить справа еще одну букву («н») и вместо нее подставить «ен». Твердый знак — признак нулевого окончания. Получаем *магнитен*. Далее отбрасываем две буквы («ен»), вместо них подставляем «н» и окончание, получаем *магнитна* и т. д.

МОБИЛЬНЫЙ ая ое ого ой ому ую ом ою ые ых ым ыми ее ей 2ЕН ь  
2ЬН а о ы

Глаголы класса г4сН:

СПЯТИТЬ л ло в вши вший 2Ч у 1Т ишь ит им ите ят ь ьте

Отбрасываем окончание «ть», подставляем окончания из списка: *спятил*, *спятило*... Убираем две буквы («ти»), заменяем на «ч» + окончание, получаем *спячу*. Убираем одну букву («ч»), заменяем на «т» + окончание: *спятишь* и т. д.

СХУЛИГАНИТЬ л ло в вши вший 1 ю ишь ит им ите ят ь ьте

Глаголы класса г1н:

ЧИТАТЬ — есть форма страдательного причастия прошедшего времени *читанный*

ПОЧИТАТЬ — нет страдательного причастия прошедшего времени

По совершенно особому типу склоняются русские фамилии типа *Иванов*, *Никитин* (сравните склонение слов БОЯРИН, БОЛГАРИН, БОРОДИН или ФИЛИН и ФОМИН).

Всего имеется около 1040 наборов окончаний.

### Омонимия в русском языке

Во многих случаях одной словоформе можно приписать несколько наборов параметров. Такая ситуация называется **омонимией**. Омонимы (от греч. *ομοσ* — одинаковый и *ονομα* — имя) — разные по значению, но одинаковые по звучанию и написанию единицы языка. Понятие омонима близко к понятию многозначности, но не совпадает с ним. Различают несколько типов омонимии.

Омонимия **частеречная** — словоформы относятся к различным частям речи: «*стали*» (*сталь*) и «*стали*» (*стать*).

Омонимия **лексическая** возникает вследствие звукового совпадения различных по происхождению слов, например *рысь* (бег / животное); в результате полного расхождения значений многозначного слова, например *мир* (вселенная / отсутствие войны, вражды); при параллельном словообразовании от той же основы, например *тройка* (лошадей / отметка).

Внутри одной части речи омонимия может быть **полная**, когда совпадают парадигмы обоих слов или **частичная**, когда в парадигмах имеются различающиеся формы, например *ласка* (животное / проявление нежности) расходятся в форме родительного падежа множественного числа (*ласок* — *ласк*).

**Грамматические омонимы**, или **омоформы** — слова, совпадающие лишь в отдельных формах. Грамматическая омонимия характерна для слов, относящихся к разным частям речи, однако возможна и внутри одной части речи, например *пришли* (прислать / прийти), *лечу* (лечить / лететь).

**Морфологическая** омонимия — одной словоформе, образованной от одной и той же леммы, может быть приписано несколько наборов параметров. Например, слово *мамы* образуется от леммы МАМА, возможные варианты: ед. ч., род. пад. или мн. ч., им. пад.

Омонимы русского языка собраны в словаре омонимов [3]. В [12] выделено 58 типов частеречной омонимии, наиболее распространенные из них приведены в табл. 3.1.

**Таблица 3.1.** Наиболее распространенные типы частеречной омонимии

Тип омонимии	Кол-во омонимов	Примеры
Нареч. / Крат. прил.(прич.)	922	совершенно, адекватно, безумно...
Глаг. / Сущ.	878	берег, вызову, души...
Сущ. / Прил.	379	больной, дорогой, это...
Сущ. / Крат. прил.(прич.)	263	гол, долги, знаком ...
Сущ. / Деепр.	167	пролив, буря, нагоняя...
Сущ. / Нареч.	107	часами, летом, бегом...
Глаг. / Сравн. степ.	80	темней, умней, красней...
Глаг / Крат. прил.(прич.)	77	допустим, одержим, сравним...
Предикат / Нареч. /Крат. прил.	66	важно, понятно, тревожно...
Сущ. / Сравн. степ.	26	суше, круче, чаще...
Инф. / Сущ.	15	вести, мести, сволочь...
Глаг. / Нареч. /Крат. прил.	14	вяло, пошло, убито...
Глаг. / Прил.	14	мой, синим, целую...
Деепр. / Прил.	11	скупая, строгая, заезжая...
Нареч. / Сравн. степ.	10	меньше, севернее, дольше...

Компьютерные морфологические анализаторы бывают двух основных типов. Первый использует словарь всех существующих в языке словоформ, при каждой из которых указаны лемма и набор параметров. Анализаторы второго типа подбирают допустимые леммы и параметры, учитывая словоизменительные классы и наборы окончаний. Анализаторы первого типа проще и работают быстрее, но требуют огромных словарных баз на несколько миллионов словоформ.

Практика показывает, что примерно в 50% случаев имеет какая-либо форма омонимии, и набор морфологических признаков оказывается неоднозначным. Уменьшить неоднозначность можно с помощью синтаксического и семантического анализа, а также используя статистические методы. Например, отбрасываются как крайне маловероятные (хотя грамматически возможные) такие варианты как *для* (убираем деепричастие от глагола *длить*, остается предлог), *пять* (убираем повелительное наклонение от глагола *пятить*, остается числительное), *соков* (убираем краткую форму прилагательного *соковый*, остается род. пад. от существительного *сок*).

Другой проблемой, с которой приходится сталкиваться уже на этапе морфологического анализа является разбор слов, отсутствующих в словаре. Казалось бы, число слов, используемых в речи или на письме не так уж велико. Так, например, в художественных и публицистических произведениях А. С. Пушкина, а также в его письмах и деловых бумагах встречается около 20 тысяч слов. Как указывалось выше, словарь Зализняка содержит

свыше 100 тысяч слов. Словарь, используемый в семантико-синтаксическом анализаторе SemSin, разработанном в ИТМО и Экономико-математическом институте РАН [8], насчитывает примерно 177 тысяч лексем. Тем не менее, при анализе текстов трех романов Гончарова общим объемом 467 тысяч словоформ было выявлено около 1300 новых слов. Среди них встречаются имена собственные, прилагательные, существительные, в меньшем количестве глаголы и наречия.

Еще хуже обстоит дело с текстами из интернета. Огромное количество искаженных слов, опечаток, неологизмов крайне затрудняет их анализ. Поэтому современные системы обязательно имеют в своем составе модули, облегчающие пополнение словаря, и модули обработки некорректно написанных слов. В некоторых системах новая лексема и тип словоизменения выводятся из состава слова (приставки, суффиксы, окончания). Позволяя резко уменьшить количество ручного труда, такие системы, тем не менее, зачастую допускают грубые ошибки.

Процесс снятия омонимии часто называется **дизамбигуацией** (от disambiguation — устранение конфликтов, неоднозначностей). Естественно, для проведения дизамбигуации необходим анализ не только неоднозначного слова, но и окружающего контекста. В той или иной степени соответствующие методы опираются на представления о частоте встречаемости двух-, трех- и многословных сочетаний. Например, в предложении *Лошадь перешла на рысь* сочетание *перейти на рысь* еще не гарантирует того, что *рысь* — это тип бега (сравните: *Болезнь перешла на рысь*).

### 3.2 Синтаксический уровень

Задачей синтаксического анализа является построение синтаксического представления текста, т. е. синтаксической структуры. Сфера действия синтаксического анализа ограничена предложением. На входе анализатора — цепочка словоформ с приписанными им грамматическими характеристиками (в том объеме, в котором это позволяет сделать лексикоморфологический анализ и снятие грамматических неоднозначностей). На выходе полного синтаксического анализа — иерархическая структура (обычно дерево).

#### Деревья составляющих

Рассмотрим предложение *Мама мыла раму*.

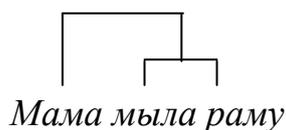
1	Мама	подлежащее
2	мыла	сказуемое
3	раму	Прямое дополнение

Необходимо с помощью формальных средств выразить отношения между словами. Первый способ заключается в объединении в группы

наиболее тесно связанных друг с другом слов.

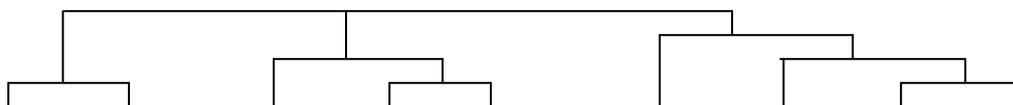
	Слово	Объединено в группу с...
1	Мама	(мыла+раму)
2	мыла	раму
3	раму	мыла

Такое объединение можно изобразить как (*Мама (мыла раму)*), или графически



Получившаяся структура носит название структуры составляющих.

**Составляющие** — это структурные единицы или отрезки предложения, которые целиком состоят из более тесно связанных друг с другом единиц меньшего размера. Формулировка «более тесно связанные» предполагает наличие смысловой связи между словами, которые входят в данную составляющую. О наличии такой связи решать лингвисту с учетом знания данного языка, а в формальном аспекте можно определить только метаязык, на котором можно записать все возможные в принципе конфигурации (структуры) составляющих для любого предложения.



*(Эти школьники) (скоро (будут писать)) (диктант (по (русскому языку)))*

Таким образом, в методе составляющих предложение рассматривается как конечное множество (элемент множества - словоупотребление) с определенным на нем отношением линейного порядка (следование слева направо). Благодаря линейному порядку может быть введено понятие отрезка.

Составляющие определяются в виде системы, т. е. в виде их (составляющих — одиночных слов и групп слов в виде отрезков) множества, на элементы которого накладываются некоторые, чисто формальные, ограничения. Система составляющих для конечного линейно упорядоченного множества  $S$  — это такое множество  $C$  отрезков этого множества, которое удовлетворяет следующим условиям:

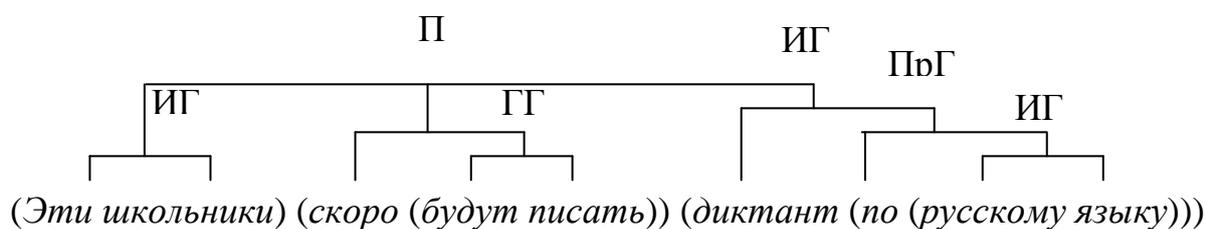
$S \in C$  — (само предложение целиком является элементом системы своих составляющих)

$\forall w \in S \ w \in C$  — (каждое отдельно взятое словоупотребление в предложении является элементом системы составляющих этого предложения)

$\forall \alpha, \beta$ , являющихся отрезками предложения  $S$  и входящих в  $C$ , либо

$\alpha \cap \beta = \emptyset$ , либо  $\alpha \subset \beta$ , либо  $\beta \subset \alpha$  — (любые две составляющие некоторого предложения или не пересекаются, или содержатся друг в друге).

Традиционный способ описания сходств и различий между синтаксическими свойствами слов состоит в том, что различные слова причисляются к различным грамматическим классам, или частям речи. Подобно тому, как отдельные слова-составляющие подразделяются на части речи, составляющие-группы также образуют небольшое множество грамматических классов. Желательно, чтобы номенклатура синтаксических групп, так же как и номенклатура частей речи, представляла собой классификацию, т. е. разбиение объектов на непересекающиеся множества, так, чтобы одна и та же группа не относилась более чем к одному множеству. Основанием для такой классификации может служить часть речи, к которой принадлежит вершина группы, т. е. слово, соответствующее корневому узлу в том фрагменте структуры зависимостей, который характеризует группу. Например, в группах {эти школьники}, {диктант по русскому языку} вершиной является существительное, такие группы называются **именными** группами (ИГ). Аналогично сочетание предлога с существительным называется **предложной** группой (ПрГ), главного глагола со вспомогательным — **глагольной** группой (ГГ) и т. д. Корневой узел дерева соответствует всему предложению (П).

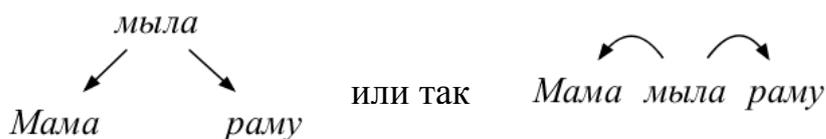


### Деревья подчинения

Вернемся к предложению *Мама мыла раму* и рассмотрим его под другим углом.

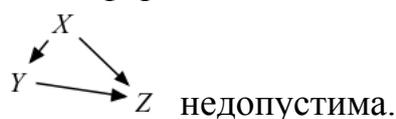
	Слово	Зависит от...
1	Мама	мыла
2	мыла	-
3	раму	мыла

Структуру зависимостей или **дерево подчинения** можно изобразить так:



В методе деревьев подчинения предложение рассматривается как конечное множество (элемент множества — словоупотребление). Всякое дерево, для которого данное предложение служит множеством узлов, называется **деревом подчинения** для данного предложения. Заметим, что из этого формального определения следует, что любое слово может быть объявлено корнем дерева, а остальные слова ему непосредственно или опосредованно подчинены.

Синтаксическая зависимость всегда является антисимметричной, т. е. если словоформа  $X$  — вершина словоформы  $Y$ , то отсюда следует, что словоформа  $Y$  не является вершиной словоформы  $X$ . Синтаксическая зависимость, кроме того, является антитранзитивным отношением. Это означает, что никакая словоформа  $X$  не может быть вершиной одновременно двух словоформ  $Y$  и  $Z$ , таких, что  $Y$  — вершина  $Z$ , следовательно структура



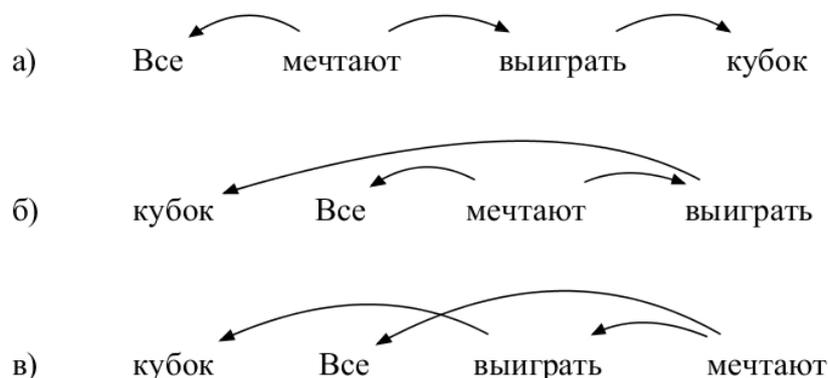
Линейный порядок слов в дереве не отражается, и одно и то же дерево может соответствовать нескольким порядкам. Однако соотношение синтаксических зависимостей и порядка слов не произвольно. Важным свойством, которым обладают деревья подчинения, является их **проективность**. Дерево проективно, если:

а) ни одна из стрелок не пересекает другую стрелку

и

б) никакая стрелка не накрывает вершину (корень)

Чтобы установить проективность или непроективность дерева синтаксического подчинения, нужно расположить все стрелки зависимостей по одну сторону от прямой, на которой записано предложение (рис. 3.1).



**Рис. 3.1.** Проективность связей: а) полная проективность, б) слабая непроективность, в) непроективность

При слабой непроективности стрелки не пересекаются, однако могут накрывать корневую вершину. Непроективные предложения обычно вы-

глядят коряво, «неграмматично», но, вместе с тем, распространены в поэзии:



В дереве подчинения синтаксические связи имеют различную природу. Классификация связей используется в описании правил, по которым слова (словоформы) языка соединяются в правильно построенные предложения. К сожалению, в настоящее время не существует общепринятого перечня типов синтаксических связей, поскольку в основу классификации разные авторы кладут разные принципы. Если воспользоваться типизацией синтаксических отношений (СинтО) по НКРЯ, то дерево зависимостей предложения «Даже маленькие дети быстро приучаются играть на компьютере» будет выглядеть следующим образом (рис. 3.2):

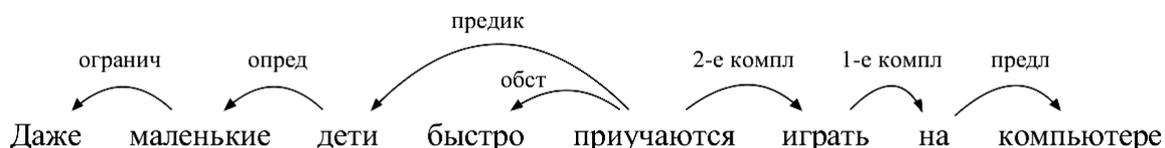


Рис. 3.2. Размеченное дерево зависимостей

Здесь:

*огранич* — ограничительное СинтО, связывает слово любой части речи с частицей или ограничительным наречием;

*опред* — определительное СинтО, связывает существительное или прилагательное с прилагательным или причастием;

*предик* — предикативное СинтО, связывает сказуемое в качестве хозяина с подлежащим в качестве слуги;

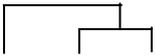
*обст* — обстоятельственное СинтО, связывает глагол или слово другой части речи, являющееся вершиной предложения с обстоятельством;

*1-компл*, *2-компл* — связывают предикатное слово (глагол, существительное, прилагательное или наречие) с его (не первыми) синтаксическими актантами (см. ниже);

*предл* — предложное СинтО, связывает предлог, с именной группой, зависящей от предлога.

Таким образом, имеется два варианта формального метаязыка для записи знаний о синтаксической структуре. Соответствия могут быть установлены между любыми структурами составляющих и проективными (только!) деревьями подчинения. Однако без дополнительной информации однозначные соответствия от одного типа структур к другому установить

невозможно (всегда есть более одного варианта соответствия). Например, структуре

 X Y Z могут соответствовать деревья  или  и т. д.

Каждый вариант представления синтаксиса имеет свои достоинства и недостатки. Так структуры составляющих не позволяют изображать непроективные структуры и разрывные словосочетания. С другой стороны деревья подчинения не могут адекватно выразить сочинительные отношения или двойное подчинение (например, придаточные, подсоединяемые словом *который*). Также невозможно ввести структурные единицы большие, чем слово.

### Синтаксическая омонимия

В русском языке часто встречаются неоднозначные предложения вроде *Он взял лук* или *Они видели в лесу норку*. Однако эти предложения имеют по два смысла только из-за лексических омонимов, входящих в их состав. Синтаксически же предложения устроены совершенно одинаково. Таким образом, здесь представлена лексическая, а не синтаксическая омонимия.

В случае же синтаксической омонимии разница в смыслах словосочетаний или предложений не связана с лексической омонимией слов, в них входящих. Предложение или словосочетание является синтаксически омонимичным, если ему можно приписать не менее двух синтаксических структур. Т. е. в одном словосочетании (предложении) можно по-разному выделить или грамматически по-разному проинтерпретировать члены предложения и/или по-разному установить или проинтерпретировать синтаксические связи между ними.

Очень часто можно по-разному установить синтаксические связи между словами, т. е. для зависимого слова можно найти разных «хозяев»: *Он умеет заставить себя слушать* (*заставить себя vs себя слушать*), или: *Приехал оркестр из Москвы* (*приехал из Москвы vs оркестр из Москвы*).

Другой вариант: пара «хозяин — слуга» выделяется единственным способом, но проинтерпретировать связь между этими словами можно по-разному: *Преследование тигра закончилось неудачей* (тигр преследует кого-то vs кто-то преследует тигра); *перевод Пушкина* (выполненный кем-то перевод произведений Пушкина vs выполненный Пушкиным перевод).

Можно по-разному определить форму слова: *Он не выучил правила* — или родительный падеж единственного числа, или винительный падеж множественного числа. Добавление определения сразу снимает омонимию

(Он не выучил этого правила vs Он не выучил эти правила). Моря<sub>им.п.</sub> окружают материки<sub>вин.п.</sub> vs Моря<sub>вин.п.</sub> окружают материки<sub>им.п.</sub>.

Рядом расположенные слова могут по-разному объединяться в группы. В концерте приняли участие известные ансамбли и самодеятельные коллективы — непонятно, известные только ансамбли, или самодеятельные коллективы тоже известные.

В языке встречаются и такие типы неоднозначных конструкций, которым приписывается одна и та же синтаксическая структура (т. е. и формы слов, и связи между ними совпадают), однако смысл может быть проинтерпретирован по-разному. Например: Коллега предложил закончить работу (коллега сам закончит работу vs он предложил, чтобы другие закончили работу). Мы узнали об открытии выставки — Мы узнали, что выставка 1) открылась, 2) открывается, 3) откроется, 4) будет открыта. Он два раза обвязывал шарф вокруг шеи (он два раза выполнял действие обвязывания vs он обвязывал шарф вокруг шеи двумя витками).

Зачастую человек часто не видит разных смыслов потому, что выбирает из них один, руководствуясь контекстом или внеязыковой ситуацией. При автоматическом же анализе фраза может быть проинтерпретирована совсем не в том смысле, который вложил в неё автор. Например: При заводе имеются курсы по подготовке в институт, детские сады и ясли. Безусловно, здесь в качестве однородных членов выступают курсы, детские сады и ясли. Однако если отвлечься от вневингвистических знаний, а ориентироваться только на строение предложения, в качестве однородных членов вполне можно выделить институт, детские сады и ясли.

### 3.3 Анафора и кореферентность

**Анафора** представляет собой явление, при котором смысл одного элемента текста (линейно вторичного, **анафора**) определяется смыслом другого элемента того же текста (линейно первичного, **антецедента**). Анафорические связи являются обязательным условием связности текста. Например, начало повести М. Ю. Лермонтова «Тамань»:

*Тамань — самый скверный городишко из всех приморских городов России. Я там чуть-чуть не умер с голода, да еще вдобавок меня хотели утопить.*

Смысл слова *там* во втором предложении нельзя понять, не прочитав предыдущей фразы. Таким образом, *Тамань* — антецедент для анафора *там*.

К анафорической связи близко понятие **кореференции**.

В письменной или устной речи слова служат для обозначения объектов, сущностей, действий и т. п. При этом один объект может именоваться совершенно по-разному. Так в примере из Лермонтова одно и то же место именуется *Тамань* и *городишко*. Таким образом, слово в тексте всегда что-

то обозначает, на что-то указывает. Отношение между словом в тексте и обозначаемой этим словом сущностью некоего мира, в контексте которого текст порожден, называют референциальным отношением. Одно и то же слово может ссылаться на разные сущности; различные слова или словосочетания могут обозначать одну и ту же сущность.

**Кореференцией** называется отношение между такими словами или словосочетаниями, которые обозначают один и тот же объект, то есть имеют один и тот же референт. Явления кореференции обусловлены фундаментальными закономерностями организации текста. Поскольку текст линеен, а описываемая им ситуация, как правило, нелинейна, в тексте почти неизбежно должны содержаться повторные упоминания элементов описываемой ситуации. При каждом новом упоминании того же объекта производится новая номинация этого объекта, которая базируется на том, что уже было сказано об этом объекте, и на тех знаниях, которые в тексте не вербализованы (экстралингвистические знания говорящего о контексте предметной области). Несмотря на то, что в тексте возникают цепочки кореферентных имен, анафорическую связь будем представлять как бинарную, то есть вторичные номинации считать связанными анафорической связью только с первичной.

Хотя проблема анафоры в лингвистике достаточно подробно разработана, эти теоретические знания почти не находят воплощения в практике: на сегодняшний день не существует известных развитых разработок систем автоматического разрешения анафоры. Тем не менее, разработки в этой области ведутся. Например, в анализаторе SemSin [8] производится установление кореференции, для местоимений личных (*он, она, оно, они*), притяжательных (*его, ее, их*), возвратных (*себя, свой*), относительных (*который*). Выявление кореференции производится на основе продукционных правил, т. е. правил преобразования построенного ранее дерева разбора.

Для формулировки условий правил используется следующая информация:

- позиция местоимения и его предполагаемого антецедента в цепочке предложения;
- позиция местоимения и его предполагаемого антецедента в дереве зависимостей;
- принадлежность сегменту того или иного типа (причастный или деепричастный оборот, придаточное предложение);
- тип входной и выходной связи;
- наличие у предполагаемого антецедента определенных зависимых слов;
- род и число местоимения и его предполагаемого антецедента.

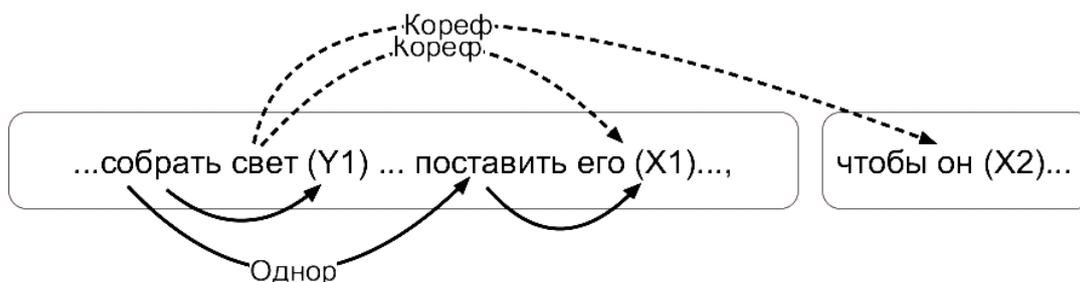
Следует отметить, что правила поиска антецедента намного сложнее простого поиска предшествующего местоимению **конгруэнтного** слова,

т. е. слова, согласованного с местоимением по роду и числу. Приведем несколько примеров, в которых переменная  $X$  отмечает местоимения, а переменные  $Y, Z$ , и т. д. отмечают анализируемые слова.

Пример 1. Предложение:

*Надо собрать лунный свет<sub>Y1</sub> в чашечку, поставить его<sub>X1</sub> в холодное место, чтобы он<sub>X2</sub> загустел, а потом принимать по две капли три раза в день.*

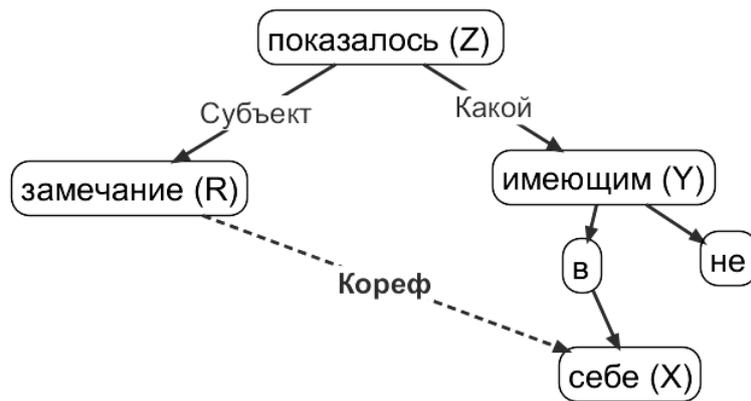
Правило: Если личное местоимение  $X$  находится в придаточном, вводимом союзом *чтобы*, ищем ближайшее конгруэнтное местоимению слово  $Z$  в предшествующем сегменте. Если  $X$  стоит в именительном падеже,  $Z$  не должен иметь входную связь «*Субъект*», если  $X$  стоит не в именительном падеже, этого запрета нет. Если  $Z$  отвечает этим условиям, и у  $Z$  нет антецедента,  $Z$  считается антецедентом  $X$ , и они соединяются связью «*Кореферент*». Если у  $Z$  есть антецедент  $Q$ ,  $Q$  считается антецедентом  $X$ , и они соединяются связью «*Кореферент*».



Пример 2. Предложение:

*Это замечание<sub>R</sub> показалось<sub>Z</sub> Алисе не имеющим<sub>Y</sub> в себе<sub>X</sub> никакого смысла.*

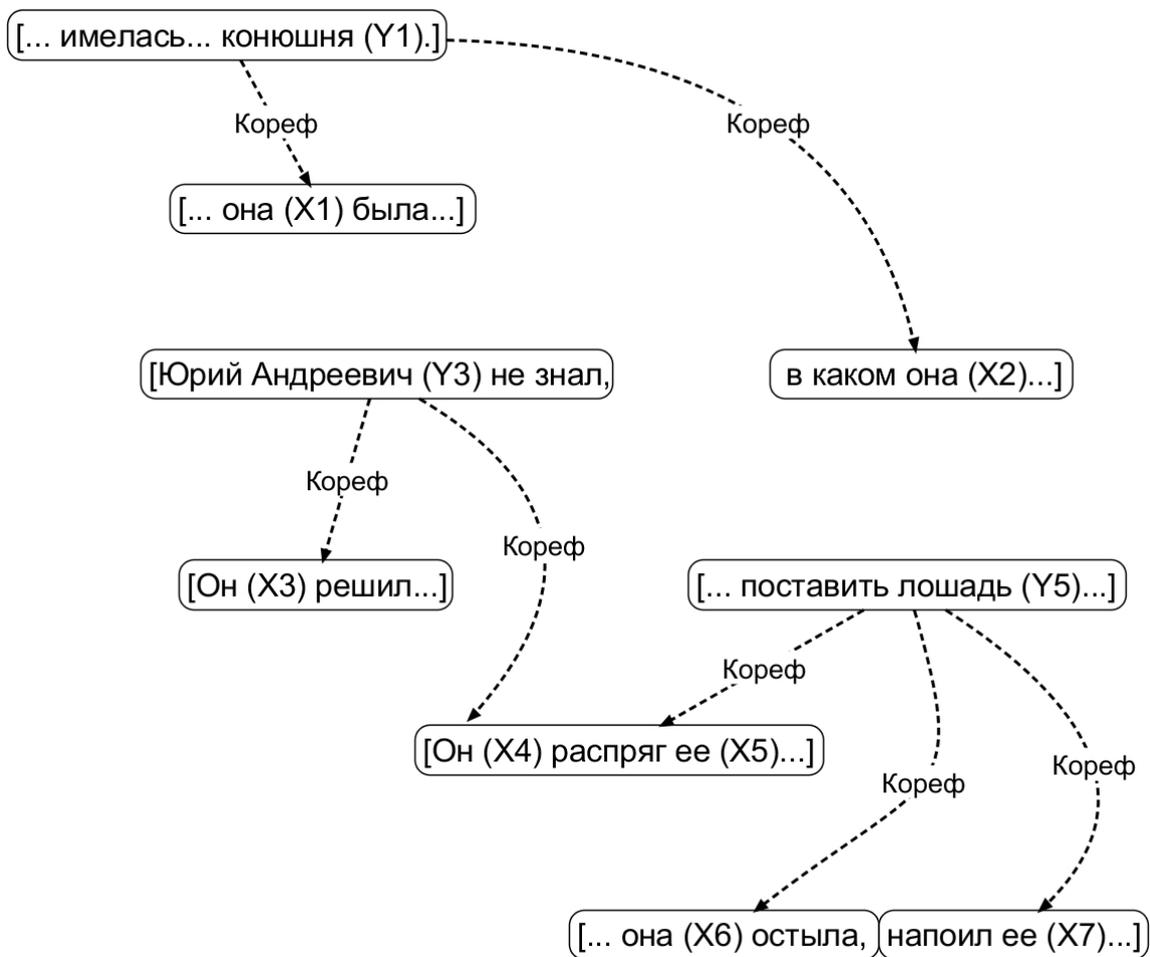
Правило: Если лемма местоимения  $(X)=\{\text{свой, себя}\}$  и  $X$  находится в необособленном причастном обороте, поднимаемся вверх по дереву от  $X$  до причастия  $Y$ . Если  $Z$  источник связи на  $Y$  и  $Z$  — существительное или местоимение, не имеющее антецедента, оно считается антецедентом  $X$ , и они соединяются связью «*Кореферент*». Если источник связи  $Z$  — глагол, антецедентом  $X$  считается слово  $R$  с входной связью «*Субъект*», не имеющее антецедента, и они соединяются связью «*Кореферент*».



Зачастую поиск антецедента не локализован в одном предложении. Тогда возникает значительно более сложная задача — анализ связного текста.

Пример 3. Текст:

*Среди служб во дворе у Микулицыных имелась вплотную к сараю пристроенная конюшня<sub>Y1</sub>. Но она<sub>X1</sub> была на запоре, Юрий Андреевич<sub>Y3</sub> не знал, в каком она<sub>X2</sub> состоянии. Он<sub>X3</sub> решил на первую ночь поставить лошадь<sub>Y5</sub> в легко отворившийся незапертый сарай. Он<sub>X4</sub> распряг ее<sub>X5</sub>, и когда она<sub>X6</sub> остыла, напоил ее<sub>X7</sub> принесенною из колодца водою.*



## 4 Классификация и кластеризация

### 4.1 Закон Ципфа

Классификатор — это алгоритм, соотносящий некие входные данные с одним или несколькими классами. В отличие от алгоритмов кластеризации эти классы должны быть определены заранее. Один из самых ярких примеров автоматической классификации — это фильтрация спама. Классификация используется также как инструмент для решения множества других задач:

- ✓ снятие омонимии при обработке натуральных языков;
- ✓ в поисковых системах — для ограничения области поиска в целях повышения точности (вертикальный поиск);
- ✓ автоматическое определение языка, на котором написан текст;
- ✓ анализ тональности (определение эмоциональной окраски текста).

Некоторые методы классификации будут обсуждаться ниже, пока же заметим, что для решения задач классификации текстов, документы в непосредственном виде не подходят для интерпретации классификатором. Поэтому необходимо применение процедуры индексации, которая переводит текст в удобное представление. Обычно документ представляется в виде вектора признаков или терминов, такое представление называется *векторной моделью* документа.

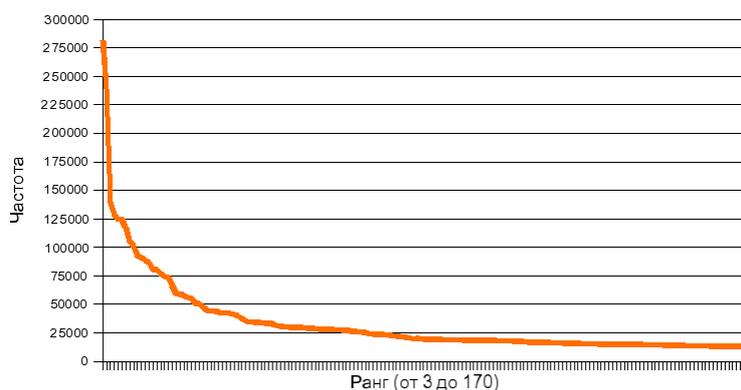
Различия в подходах заключаются как в понимании того, что такое термин, так и в способах определения веса термина.

Одним из самых распространенных методов перехода к математической модели документа, является «метод ключевых слов». *Ключевое слово* — слово в тексте, способное в совокупности с другими ключевыми словами представлять текст. Суть метода в следующем. Для каждого класса текстов создается список характерных для него слов, тогда каждый текст можно представить в виде вектора частот появления в нём слов из данного списка. Возникает проблема поиска и выделения из текста слов, которые будут для него ключевыми. Огромный объем информации, который подлежит обработке, делают особенно актуальной задачу автоматического выделения ключевых слов. Причем от чистоты этого выделения напрямую зависит точность классификации.

Во всех текстовых документах, созданных человеком, можно выделить статистические закономерности. В любом языке есть слова, которые встречаются чаще, чем остальные, но не имеют значения. Есть слова, которые встречаются реже, но имеют намного большее смысловое значение.

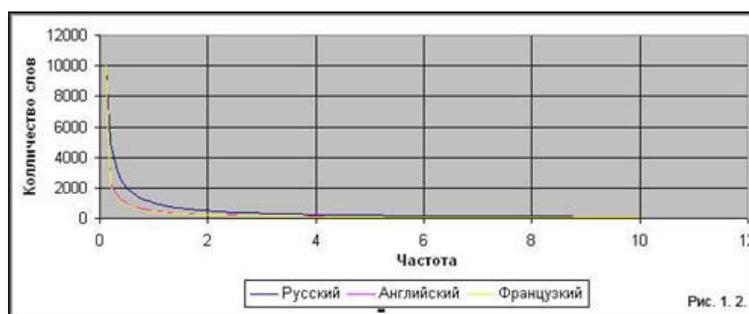
В 1949 году Джордж Ципф<sup>1</sup> сформулировал несколько закономерностей. Данные законы получены не на основе математических выводов, а на основе анализа статистики частоты слов текста на многих языках, то есть эмпирически.

Если все слова достаточно длинного текста упорядочить по убыванию частоты их использования, то частота  $n$ -го слова в таком списке окажется приблизительно обратно пропорциональна его порядковому номеру  $n$  (так называемому *рангу* этого слова). Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и т. д. (рис. 4.1).



**Рис. 4.1.** Закон Ципфа для русской Википедии

Это утверждение верно в пределах одного языка. Однако и межъязыковые различия невелики. На каком бы языке текст ни был написан, вид кривой Ципфа останется неизменной. Может немного отличаться лишь коэффициент гиперболы (рис. 4.2).



**Рис. 4.2.** Закон Ципфа для разных языков

Законы Ципфа позволяют находить ключевые слова.

Исследования показывают, что наиболее значимые для текста слова лежат в средней части графика (рис. 4.3). Этот факт имеет простое обоснование. Слова, которые попадают слишком часто, в основном оказываются

<sup>1</sup> George Kingsley Zipf, гарвардский профессор-лингвист и филолог

ся предлогами, местоимениями. Редко встречающиеся слова тоже, в большинстве случаев, не имеют решающего смыслового значения.

От установки ширины зависит качество отделения значимых слов. Если установить большую ширину диапазона, то в ключевые слова будут попадать вспомогательные слова; если установить узкий диапазон — можно потерять смысловые термины. Поэтому в каждом отдельном случае необходимо использовать ряд эвристик для определения ширины диапазона, а также пользоваться специальными методиками, уменьшающими влияние этой ширины.

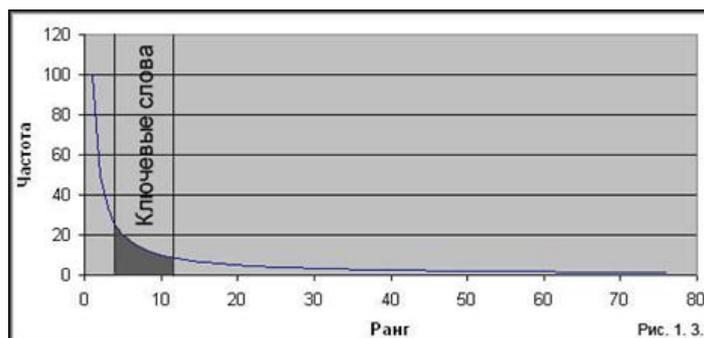
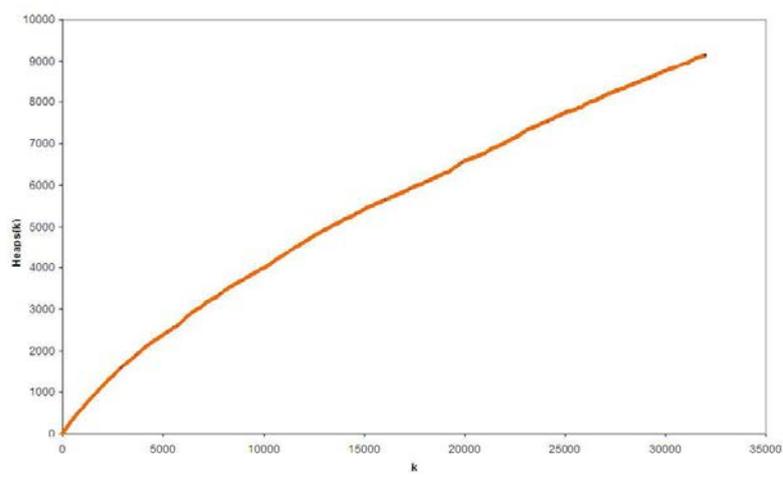


Рис. 4.3. Выбор ключевых слов

Одним из способов, например, является предварительное исключение из исследуемого текста слов, которые изначально не могут являться значимыми и, поэтому, являющиеся «шумом». Такие слова называются нейтральными или стоповыми (стоп-словами). Для русского текста стоповыми словами могли бы являться все предлоги, частицы, личные местоимения.

### Закон Хипса

В компьютерной лингвистике эмпирический закон Г. С. Хипса (H. S. Hears) связывает объем документа с объемом словаря уникальных слов, которые входят в этот документ. Казалось бы, словарь уникальных слов должен насыщаться, а его объем стабилизироваться при увеличении объемов текста. Оказывается, это не так! Для всех известных сегодня текстов в соответствии с законом Хипса, эти значения связаны соотношением  $v(n) = \alpha n^\beta$  где  $v$  — это объем словаря уникальных слов, составленный из текста, который состоит из  $n$  уникальных слов,  $\alpha$  и  $\beta$  — определенные эмпирически параметры (рис. 4.4). Для европейских языков  $\alpha$  принимает значение от 10 до 100, а  $\beta$  — от 0.4 до 0.6.



**Рис. 4.4.** Закон Хипса: по оси абсцисс – количество слов в тексте, по оси ординат – объем словаря

При больших объемах текста становится видно, что график закона Хипса идет не плавно, а ступенчато. Такая особенность связана с тем, что с некоторого момента тексты, относящиеся к какой-то узкой предметной области, заканчиваются, следующий текст относится уже к другой предметной области, а для нее характерна другая лексика. Если при добавлении в коллекцию нового текста наблюдаются отклонения от закона Хипса, например, число слов в словаре не возрастает, это может свидетельствовать о наличии плагиата.

## 4.2 Модель $TF*IDF$

Некоторые слова могут встречаться почти во всех документах некоторой коллекции и, соответственно, оказывать малое влияние на принадлежность документа к той или иной категории, а значит не быть ключевыми для этого документа. Для понижения значимости слов, которые встречаются почти во всех документах, вводят инверсную частоту термина  $IDF$  (inverse document frequency) — это логарифм отношения числа всех документов  $D$  к числу документов  $d$ , содержащих некоторое слово.

$IDF = \lg \frac{D}{d}$  Значение этого параметра тем меньше, чем чаще слово встречается в документах коллекции. Таким образом, для слов, которые встречаются в большом числе документов,  $IDF$  будет близок к нулю (если слово встречается во всех документах  $IDF$  равен нулю), что помогает выделить важные слова.

Параметр  $TF$  (term frequency) — это отношение числа раз  $k_i$ , которое некоторое слово встретилось в документе, к общему числу слов в документе  $n$ . Нормализация длиной документа нужна для того, чтобы уравнивать в правах короткие и длинные документы.

$$TF = \frac{k_i}{n}.$$

Коэффициент  $TF*IDF$  равен произведению  $TF$  и  $IDF$ , при этом  $TF$  играет роль повышающего множителя,  $IDF$  — понижающего. Тогда весовыми параметрами векторной модели некоторого документа можно принять коэффициенты  $TF*IDF$  входящих в него слов. Для того чтобы веса находились в интервале  $(0, 1)$ , а векторы документов имели равную длину, значения  $TF*IDF$  обычно нормализуются.

Отметим, что эта формула оценивает значимость термина только с точки зрения частоты вхождения в документ, тем самым не учитывая порядок следования терминов в документе и их синтаксическую роль; другими словами, семантика документа сводится к лексической семантике входящих в него терминов, а композиционная семантика не рассматривается.

Ключевыми в данном случае будут являться слова набравший наибольший вес. Слова с малым весом, вообще можно не учитывать при классификации.

Проиллюстрируем на простом примере.

Пусть коллекция состоит из 3 документов.

1. Мама мыла мылом Машу.
2. Мама мыла, мыла раму.
3. В магазине купила мама мыло.

Вид словаря тогда будет следующим:

Слово	Всего	Встретилось в документах ( $d$ )	$IDF$
Мама	3	3	0
мыть	3	2	0,18
мыло	2	2	0,18
Маша	1	1	0,47
рама	1	1	0,47
магазин	1	1	0,47
купить	1	1	0,47

Вид векторов:

1 документ			2 документ			3 документ		
Слово	TF	TF*IDF	Слово	TF	TF*IDF	Слово	TF	TF*IDF
Маша	0,25	0,12	рама	0,25	0,11	магазин	0,25	0,12
мыло	0,25	0,05	мыть	0,5	0,09	купить	0,25	0,12
мыть	0,25	0,05	мама	0,25	0	мыло	0,25	0,05
мама	0,25	0				мама	0,25	0

Влияние  $TF$  видно во втором векторе. Так как слово «мыть» встреча-

ется 2 раза, он выше, чем у остальных слов. Однако из-за того, что это слово встречается и в других документах, у него ниже параметр  $IDF$ , поэтому его общий вес в векторе будет ниже, чем у слова «рама». Так влияет параметр  $IDF$ .

Слово «мама» же вообще можно не учитывать в векторном представлении. Так как оно встречается во всех предложениях коллекции, его значение  $TF*IDF$  всегда будет равно нулю.

Заметим, что все слова примера мы приводим к нормальной форме (лемматизируем).

Однако у метода  $TF*IDF$  есть существенный недостаток: при построении вектора не учитывается порядок слов, контекст, то есть важная семантическая составляющая текста.

Таким образом, данную коллекцию документов можно описать матрицей частот

Номер доку-мента	Номер слова	1 (мыть)	2 (мыло)	3 (Маша)	4 (рама)	5 (мага-зин)	6 (купить)
1		0,05	0,05	0,12			
2		0,09			0,11		
3			0,05			0,12	0,12

Дальше обычно частоты нормируются так, чтобы сумма квадратов по строке (т. е. по документу) равнялась единице. Тогда окончательно получаем представление документов в виде векторов:

$$\mathbf{d}_1 = (0,36; 0,36; 0,86; 0; 0; 0);$$

$$\mathbf{d}_2 = (0,63; 0; 0; 0,77; 0; 0);$$

$$\mathbf{d}_3 = (0; 0,28; 0; 0; 0,68; 0,68).$$

### 4.3 Классификация документов

Процесс классификации документов как векторов основан на гипотезе о том, что тематически близкие документы окажутся в пространстве терминов геометрически близко расположенными. Поэтому в основе алгоритмов классификации лежит понятие сходства или расстояния между документами в пространстве терминов. В данном случае понятия расстояния и сходства являются взаимно обратными, расстояние можно было бы называть различием. Выбор способа вычисления расстояния влияет на результат классификации. Часто расстояние между документами, представленными векторами  $\mathbf{d}_i$  и  $\mathbf{d}_j$  определяют как

$$dist(\mathbf{d}_i, \mathbf{d}_j) = \left( \sum_k |d_{ik} - d_{jk}|^r \right)^{\frac{1}{r}} \quad (4.1)$$

где  $r$  — это параметр, заданный пользователем. При  $r = 1$  получаем так называемое манхэттенское расстояние, или расстояние городских кварталов; при  $r = 2$  — евклидово расстояние; при  $r \rightarrow \infty$  — расстояние Чебышева, которое вычисляется как максимум модуля разности компонент этих векторов.

Другой часто используемой на практике мерой сходства является косинусная мера, которая представляет собой скалярное произведение векторов.

$$\text{sim}(\mathbf{d}_i, \mathbf{d}_j) = \cos(\angle(\mathbf{d}_i, \mathbf{d}_j)) = \frac{\sum_k d_{ik}d_{jk}}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}} \quad (4.2)$$

Если векторы ортогональны, то их мера близости равна 0, если совпадают, то 1.

Одним из возможных решений задачи классификации является классификация основная на правилах. Например, формулируются правила определения класса документа по его тексту типа

```
If (text.contains("выиграть") || text.contains("миллион") && ||
text.contains("доллар")) then "СПАМ" else "НЕ СПАМ"
```

Этот подход может быть хорошим вариантом при работе с небольшой коллекцией документов, которую можно тщательно проанализировать. Но есть у этого подхода и очевидные минусы. Для того чтобы выбрать значимые для классификации слова необходимо обладать экспертными знаниями в предметной области. Кроме того, отнюдь не всегда факт наличия или отсутствия какого-либо одного слова является решающим фактором для принятия решения. Усложнение правил с добавлением вложенных if'ов достаточно бесперспективно потому, что возможности человека в формулировании таких правил очень ограничены, а сложность правил катастрофически растет с количеством выбранных для классификации слов.

Можно пойти другим путем. Для каждого слова выберем некий условный вес (табл. 4.1), который будет означать, насколько вероятно, что сообщение с этим словом является спамом (0 — никогда не является спамом, 1 — всегда спам).

**Таблица 4.1.** Условные веса слов для выявления спама

Слово	СПАМ	НЕ_СПАМ
выиграть	0,99	0,01
миллион	0,95	0,05
доллар	0,95	0,05
приглашение	0,50	0,50
конференция	0,01	0,99

Суммарный вес документа определяется как произведение весов всех известных слов документа отдельно для класса «СПАМ» и класса «НЕ\_СПАМ». Слова, для которых у нас нет веса при классификации, пропускаются. Какой суммарный вес оказался больше тот класс и побеждает.

Это более разумный подход, так как он более гибок и принимает решение на основании всех известных слов в тексте. Если у нас будет некий способ автоматически подобрать оптимальные веса слов, то данный подход можно считать методом машинного обучения.

#### 4.4 Классификация с обучением. Наивный байесовский классификатор

Самый простой, но вместе с тем один из самых часто используемых при обработке натуральных языков алгоритм классификации — наивный байесовский классификатор (*Naive Bayes Classifier, NBC*).

В основе *NBC* лежит теорема Байеса:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (4.3),$$

где  $P(c|d)$  — вероятность, что документ  $d$  принадлежит классу  $c$ , именно её надо рассчитать;

$P(d|c)$  — вероятность встретить документ  $d$  среди всех документов класса  $c$ ;

$P(c)$  — безусловная вероятность встретить документ класса  $c$  в корпусе документов;

$P(d)$  — безусловная вероятность документа  $d$  в корпусе документов.

Теорема Байеса позволяет как бы переставить местами причину и следствие. Зная, с какой вероятностью причина приводит к некоему событию, эта теорема позволяет рассчитать вероятность того, что именно эта причина привела к наблюдаемому событию.

Цель классификации состоит в том, чтобы понять к какому классу принадлежит документ, поэтому нужна не сама вероятность, а наиболее вероятный класс. Байесовский классификатор использует оценку апостериорного<sup>1</sup> максимума (*Maximum a posteriori estimation*) для определения наиболее вероятного класса. То есть рассчитывается вероятность для всех классов и выбирается тот класс, который обладает максимальной вероятностью. Поскольку знаменатель (вероятность документа) является константой и никак не может повлиять на ранжирование классов, в данной задаче его можно игнорировать и полагать, что

---

<sup>1</sup> Апостериорный – основанный на опыте, в отличие от априорный – не основанный на опыте, предшествующий ему.

$$C_{map} = \max_{c \in C} [P(d|c) P(c)]. \quad (4.4)$$

В естественных языках вероятность появления слова сильно зависит от контекста. Байесовский же классификатор называется «наивным», поскольку представляет документ как набор слов, вероятности которых не зависят друг от друга. Исходя из этого предположения условная вероятность документа аппроксимируется произведением условных вероятностей всех слов (точнее говоря, терминов, отобранных, например, по критерию  $TF*IDF$ ), входящих в документ:

$$P(d|c) \approx P(w_1|c) P(w_2|c) \dots P(w_n|c) = \prod_{i=1}^n P(w_i|c). \quad (4.5)$$

Подставляя в (4.4), получаем

$$C_{map} = \max_{c \in C} \left[ P(c) \prod_{i=1}^n P(w_i|c) \right]. \quad (4.6)$$

При достаточно большой длине документа придется перемножать большое количество очень маленьких чисел. Для того чтобы избежать возникающих при этом проблем, связанных с арифметическим переполнением или потерей точности, зачастую пользуются свойством логарифма произведения  $\log(ab) = \log a + \log b$ . Так как логарифм функция монотонная, ее применение к обеим частям выражения изменит только его численное значение, но не параметры при которых достигается максимум. При этом логарифм от числа близкого к нулю будет числом отрицательным, но в абсолютном значении существенно большим, чем исходное число, что делает логарифмические значения вероятностей более удобными для анализа. Формула (4.6) приобретает вид

$$C_{map} = \max_{c \in C} \left[ \log(P(c)) + \sum_{i=1}^n \log(P(w_i|c)) \right] \quad (4.7)$$

Основание логарифма в данном случае не имеет значения.

Оценка вероятностей  $P(c)$  и  $P(w_i|c)$  осуществляется на обучающей выборке. Вероятность класса мы можем оценить как

$$P(c) = \frac{D_c}{D},$$

где  $D_c$  — количество документов, принадлежащих классу  $c$ , а  $D$  — общее количество документов в обучающей выборке.

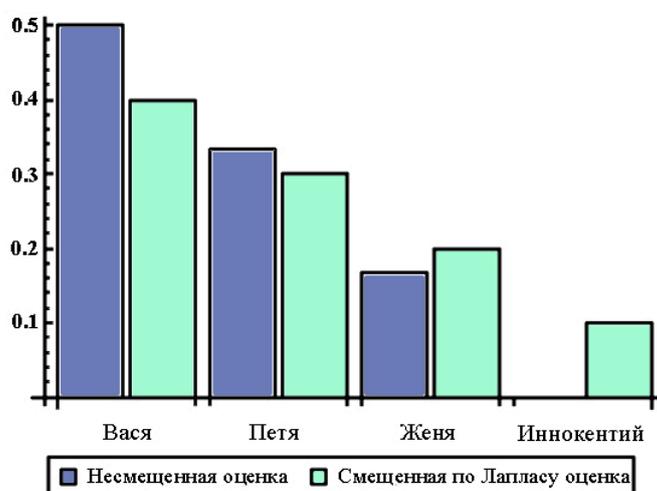
Оценка вероятности слова в классе может делаться несколькими путями. Например, в *multinomial bayes model* это

$$P(w_i | c) = \frac{W_{ic}}{\sum_{k \in V} W_{kc}}. \quad (4.8)$$

Здесь  $W_{ic}$  — количество раз, которое  $i$ -е слово встречается в документах класса  $c$ ,  $V$  — словарь корпуса документов (список всех уникальных слов). Другими словами, числитель описывает, сколько раз слово встречается в документах класса (включая повторы), а знаменатель — это суммарное количество слов во всех документах этого класса.

С формулой (4.8) есть одна небольшая проблема. Если на этапе классификации встретится слово, которого не было на этапе обучения, то значения  $W_{ic}$ , а следовательно и  $P(w_i|c)$  будут равны нулю. Это приведет к тому, что документ с этим словом нельзя будет классифицировать, так как он будет иметь нулевую вероятность по всем классам. Избавиться от этой проблемы путем анализа большего количества документов не получится. Невозможно составить обучающую выборку, содержащую все возможные слова, включая неологизмы, опечатки, синонимы и т.д. Типичным решением проблемы неизвестных слов является аддитивное сглаживание (сглаживание Лапласа). Идея заключается в том, что к частоте каждого слова прибавляется единица. Добавление единицы к каждой частоте встречаемости термина можно интерпретировать как априорное равномерное распределение (каждый термин встречается в каждом классе по одному разу), которое затем на обучающем множестве уточняется. Логически данный подход смещает оценку вероятностей в сторону менее вероятных исходов. Таким образом, слова, которые отсутствовали на этапе обучения модели, получают пусть маленькую, но все же не нулевую вероятность.

Допустим, на этапе обучения мы видели три имени собственных указанное количество раз: *Вася* – 3; *Петя* – 2; *Женя* – 1. На этапе классификации появляется имя *Иннокентий*. Тогда оригинальная и смещенная по Лапласу оценка вероятностей будут выглядеть следующим образом (рис. 4.5).



**Рис. 4.5.** Несмещённая и смещённая и оценка вероятности

Смещённая оценка никогда не бывает нулевой, что защищает от проблемы неизвестных слов

Подставив выбранные нами оценки в (4.7), получаем окончательную формулу, по которой происходит байесовская классификация:

$$c_{map} = \max_{c \in C} \left[ \log \left( \frac{D_c}{D} \right) + \sum_{i=1}^n \log \left( \frac{W_{ic} + 1}{|V| + \sum_{k \in V} W_{kc}} \right) \right]. \quad (4.9)$$

Таким образом, для реализации Байесовского классификатора необходима обучающая выборка, в которой проставлены соответствия между текстовыми документами и их классами. Затем нужно собрать следующую статистику из выборки, которая будет использоваться на этапе классификации:

- относительные частоты классов в корпусе документов. То есть, как часто встречаются документы того или иного класса;
- суммарное количество слов в документах каждого класса;
- относительные частоты слов в пределах каждого класса;
- размер словаря выборки. Количество уникальных слов в выборке.

Совокупность этой информации будем называть моделью классификатора. Затем на этапе классификации необходимо для каждого класса рассчитать значение следующего выражения и выбрать класс с максимальным значением (упрощённая запись формулы (4.9))

$$\log \left( \frac{D_c}{D} \right) + \sum_{i \in Q} \log \left( \frac{W_{ic} + 1}{|V| + L_c} \right).$$

В этой формуле:

$D_c$  — количество документов в обучающей выборке принадлежащих классу  $c$ ;

$D$  — общее количество документов в обучающей выборке;

$|V|$  — количество уникальных слов во всех документах обучающей выборки;

$L_c$  — суммарное количество слов в документах класса  $c$  в обучающей выборке;

$W_{ic}$  — сколько раз  $i$ -ое слово встречалось в документах класса  $c$  в обучающей выборке;

$Q$  — множество слов классифицируемого документа (включая повторы).

Допустим, у нас есть три документа, для которых известны их классы:  
[Спам] *вы выиграли миллион долларов*;

[Спам] *приглашаем играть в лотерею*;

[Не\_спам] приглашаем на конференцию в Лондон.

Модель классификатора будет выглядеть следующим образом:

	Частоты классов	Суммарное кол-во слов
СПАМ	2	6
НЕ_СПАМ	1	3

	выиг- рать	мил- лион	дол- лар	при- глашать	играть	лоте- рея	конфе- ренция	Лон- дон
СПАМ	1	1	1	1	1	1		
НЕ_СПАМ				1			1	1

Теперь классифицируем фразу «Приглашаем на конференцию в Одессу». Рассчитаем значение выражения для класса СПАМ (используем натуральные логарифмы):

$$\log(2/3) + \log(2/(8+6)) + \log(1/(8+6)) + \log(1/(8+6)) \approx -7,629.$$

Для класса НЕ\_СПАМ:

$$\log(1/3) + \log(2/(8+3)) + \log(2/(8+3)) + \log(1/(8+3)) \approx -6,906.$$

В данном случае выиграл класс НЕ\_СПАМ.

В простейшем случае выбирается класс, который получил максимальную оценку. Но если нужно, например, пометить сообщение как спам только если соответствующая вероятность больше 80%, то сравнение логарифмических оценок ничего не даст. Оценки, которые выдает алгоритм, не удовлетворяют двум формальным свойствам, которым должны удовлетворять все вероятностные оценки: они лежат в интервале от нуля до единицы и их сумма должна быть равна единице. Для того чтобы решить эту задачу, необходимо из логарифмических оценок сформировать вероятностное пространство. А именно: избавиться от логарифмов и нормировать сумму по единице.

$$P(c|d) = \frac{e^{q_c}}{\sum_{c \in C} e^{q_c}}$$

Здесь  $q_c$  — это логарифмическая оценка алгоритма для класса  $c$ , а возведение  $e$  (основания натурального логарифма) в степень оценки используется для того чтобы избавиться от логарифма. Таким образом, если в расчетах использовались не натуральные логарифмы, а десятичные, необходимо использовать не  $e$ , а 10.

Для вышеприведенного примера вероятность, что сообщение спам равно:

$$\frac{e^{-7,629}}{e^{-7,629} + e^{-6,906}} = 0,327.$$

то есть сообщение является спамом с вероятностью 32.7%.

## 4.5 Классификация с обучением. Другие алгоритмы

Рассмотрим кратко принципы работы некоторых других алгоритмов классификации текстов.

### Алгоритм Роккио

Алгоритм Роккио рассматривает документы в векторном пространстве терминов и ищет границы между классами как множества точек, равноудалённых от центроидов этих классов. Центроидом класса называется усреднённый вектор, или центр масс членов класса:

$$\bar{\mu}_{c_j} = \frac{1}{D_{c_j}} \sum_{d_i \in c_j} \bar{d}_i$$

Граница между двумя классами в многомерном пространстве терминов имеет вид гиперплоскости. Правило классификации заключается в определении области, в которую попадает новый документ, то есть в поиске центроида, к которому образ нового документа ближе, чем к остальным центроидам.



Рис. 4.6. Иллюстрация работы алгоритма Роккио

На рис. 4.6 к документу «звездочка» ближе всех центроид класса «кружков». Алгоритм Роккио предполагает, что классы имеют форму сфер с примерно одинаковыми радиусами. Если это предположение не выполняется, то алгоритм может привести к неудовлетворительным результатам. Например, на рис. 4.6 документ «квадрат» больше подходит классу «крестиков», а алгоритм отнесёт его к классу «треугольников».

### Алгоритм $k$ -ближайших соседей

Алгоритм  $k$ -ближайших соседей использует гипотезу компактности векторного пространства, которая заключается в том, что документы одного класса образуют в пространстве терминов компактную область, причём области разных классов не пересекаются. Тогда можно ожидать, что

тестовый документ будет иметь такую же метку класса, как и окружающие его документы из обучающего множества. Алгоритм  $k$ -ближайшего соседа относит тестовый документ к преобладающему классу его  $k$  соседей. При  $k = 1$  алгоритм относит документ к классу, самого ближайшего ему документа.

Данный алгоритм лучше справляется с несферическими или несвязными классами, чем алгоритм Роккио, поскольку определяет границы между классами локально. Для всех документов обучающего множества пространство терминов представляется разделенным на ячейки (выпуклые многогранники), состоящие из точек, которые ближе к данному объекту, чем к другим (рис. 4.7).

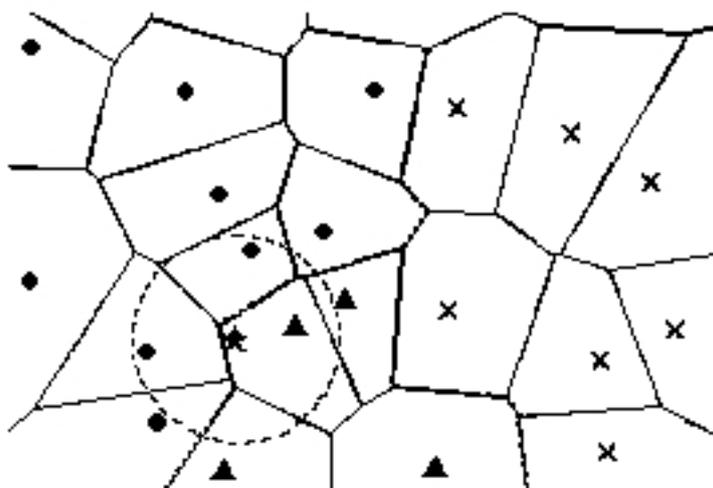


Рис. 4.6. Иллюстрация работы алгоритма  $k$ -ближайших соседей

Это в случае  $k = 1$ . В случае  $k > 1$  внутри ячеек также множество  $k$ -ближайших соседей остаётся инвариантным. На рис. 4.7 видно, что новый документ «звёздочка» попадает в ячейку объекта класса «треугольников», и при  $k = 1$  будет отнесён к этому же классу. Однако при  $k = 3$  «звёздочка» будет отнесён к классу «кружочков». При  $k = 1$  алгоритм неустойчив, так как классификация зависит всего от одного обучающего документа, а он может быть нетипичным или иметь неверную метку класса. На практике значение  $k$  выбирают на основе опыта эксперта и имеющихся знаний о решаемой задаче. Кроме того, число соседей можно подобрать на обучающем множестве так, чтобы максимизировать качество классификации.

#### 4.6 Оценка результатов классификации. $F$ -мера.

Как оценивать качество алгоритма? Допустим, вы хотите внести изменения в алгоритм. Откуда вы знаете, что эти изменения сделают алгоритм лучше? Конечно же, надо проверять алгоритм на реальных данных. Качество построенного классификатора оценивается по его ошибке на тестовом подмножестве обучающего множества документов. Ошибка — это доля неправильных решений классификатора. Решения классификатора

сравнивают с решениями экспертов, формирующих обучающее множество.

Для вычисления мер качества в задачах анализа информации необходимо для каждого класса  $c_j$  составить таблицу категорий принятых решений по множеству документов  $d_i$  (табл. 4.2):

**Таблица 4.2.** Таблица категорий

Эксперт решил Классификатор решил	$d_i \in c_j$	$d_i \notin c_j$
$d_i \in c_j$	$T_P$	$F_P$
$d_i \notin c_j$	$F_N$	$T_N$

Здесь  $T_P$  — количество истинно-положительных решений, т. е. и эксперт и классификатор отнесли эти документы к классу  $c_j$ ;  $T_N$  — количество истинно-отрицательных решений;  $F_P$  — количество ложно-положительных решений;  $F_N$  — количество ложно-отрицательных решений.

В простейшем случае параметром, характеризующим классификатор, может быть доля документов, по которым классификатор принял правильное решение (*Аккуратность* или Accuracy):

$$A = \frac{T_P + T_N}{T_P + T_N + F_P + F_N},$$

т. е. отношение числа документов, по которым классификатор принял правильное решение, к размеру обучающей выборки.

Однако эта мера присваивает всем документам одинаковый вес, что может быть не корректно в случае, если распределение документов в обучающей выборке сильно смещено в сторону какого-то одного или нескольких классов. В этом случае у классификатора есть больше информации по этим классам и соответственно в рамках этих классов он будет принимать более адекватные решения. На практике это приводит к тому, можно иметь  $A = 80\%$ , но при этом в рамках какого-то конкретного класса классификатор работает из рук вон плохо, не определяя правильно даже треть документов. При наличии небольших классов, то есть классов, доля документов которых меньше 10%, высокой правильности можно достичь, всегда отвечая «не принадлежит». Например, если относительная частота класса коллекции составляет 1%, то классификатор по принципу «всегда не принадлежит» даст правильность 99%.

Для более точных оценок используются такие параметры как *точность* (*precision*) — доля документов, действительно принадлежащих данному классу, относительно всех документов, которые система отнесла к этому классу

$$P = \frac{T_P}{T_P + F_P},$$

и *полнота* (*recall*) — доля найденных классификатором документов, при-

надлежащих классу, относительно всех документов этого класса в тестовой выборке

$$R = \frac{T_P}{T_P + F_N}$$

Рассмотрим пример. Допустим, имеется тестовая выборка, в которой 10 сообщений, из них 4 — спам. Обработав все сообщения, классификатор пометил 2 сообщения как спам, причем одно действительно является спамом, а второе было помечено в тестовой выборке как нормальное. Мы имеем одно истинно-положительное решение, три ложно-отрицательных и одно ложно-положительное. Тогда для класса «СПАМ» точность классификатора составляет 1/2 (50% положительных решений правильные), а полнота — 1/4 (классификатор нашел 25% всех спам-сообщений).

На практике значения точности и полноты гораздо удобней рассчитывать с использованием *матрицы неточностей (confusion matrix)*. Матрица неточностей — это матрица размера  $N$  на  $N$ , где  $N$  — это количество классов. Столбцы этой матрицы резервируются за экспертными решениями, а строки за решениями классификатора. При классификации очередного документа из тестовой выборки увеличивается на единицу число, стоящее на пересечении строки класса, который вернул классификатор, и столбца класса, к которому действительно относится документ.

На рис. 4.8. показан пример матрицы неточностей для 24 классов. Как видно, большинство документов классификатор определяет верно. Диагональные элементы матрицы явно выражены. Тем не менее, в рамках некоторых классов (3, 5, 8, 22) классификатор показывает низкую точность.

	0.91	0.96	0.94	0.75	1.00	0.83	0.85	0.97	1.00	0.86	1.00	0.79	1.00	0.75	1.00	1.00	0.96	0.90	0.81	0.89	0.94	0.98	0.86	0.89	0.94	
0.80		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	
0.95	1	94	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
1.00	2	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.29	3	0	0	6	0	0	3	2	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	3	0	0
1.00	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0
0.92	6	1	0	0	0	0	152	0	0	1	0	0	0	0	0	0	0	1	4	2	3	0	0	0	0	0
0.97	7	1	0	1	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0
0.33	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0.97	9	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.82	10	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
0.87	11	0	0	0	0	0	0	0	0	0	0	34	0	4	0	0	0	0	0	0	0	0	0	0	1	0
1.00	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0
0.57	13	0	0	0	0	0	0	0	0	0	0	9	0	12	0	0	0	0	0	0	0	0	0	0	0	0
0.63	14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	3	0	0	0	0	0	0	0	0
0.50	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0
0.77	16	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	47	0	1	3	4	0	0	2	0	0
0.87	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	69	1	2	5	0	0	0	0	0
0.97	18	0	0	0	0	1	4	0	0	1	0	0	0	0	0	0	0	0	197	1	0	0	0	0	0	0
0.78	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	35	183	13	0	0	0	2	0	
0.97	20	0	0	0	0	0	10	3	0	1	0	0	0	0	0	0	0	0	0	4	702	0	0	0	0	0
0.93	21	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54	0	2	0	0
0.29	22	0	0	1	0	0	2	0	0	6	0	0	0	0	0	0	0	0	1	1	1	0	6	2	0	0
0.91	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	6	0	0	115	0	
1.00	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	

**Рис. 4.8.** Матрица неточностей

Имея такую матрицу, просто рассчитать точность и полноту для каждого класса. Точность равняется отношению соответствующего диагонального элемента матрицы и суммы всей строки класса. Полнота — отношению диагонального элемента матрицы и суммы всего столбца класса. Формально:

$$P_c = \frac{A_{c,c}}{\sum_{i=1}^n A_{c,i}}$$

$$R_c = \frac{A_{c,c}}{\sum_{i=1}^n A_{i,c}}$$

Полнота и точность — меры, противоречащие друг другу в том смысле, что 100%-ую полноту легко достичь, просто поместив все документы в класс  $c_j$  (точность будет мала), и наоборот 100%-ую точность можно достичь, строго отбрасывая документы, помещая в класс  $c_j$  малое число документов (полнота будет мала).

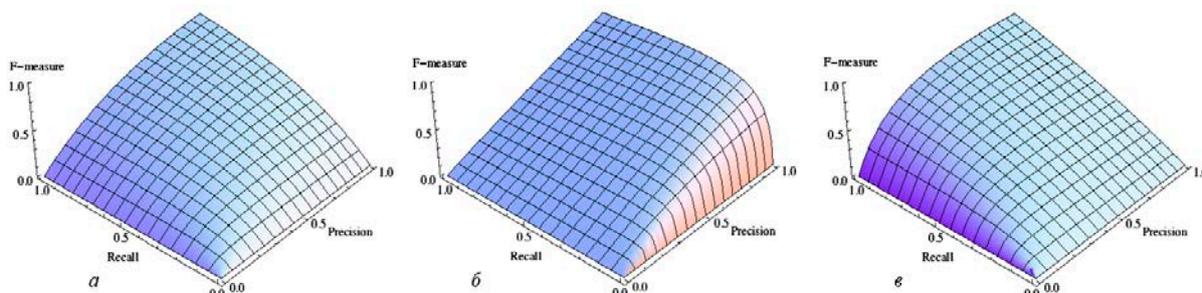
Показатель, позволяющий найти баланс между полнотой и точность, называют  $F_\beta$ -мерой, которая вычисляется как взвешенное среднее гармоническое:

$$F_\beta = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1) PR}{\beta^2 P + R}; \quad \beta^2 = \frac{1-\alpha}{\alpha}, \quad (4.10)$$

где  $\alpha \in [0;1]$ ,  $\beta \in [0;\infty]$ . При  $\beta > 1$  предпочтение отдаётся полноте, при  $\beta < 1$  — точности. На практике чаще применяют сбалансированный вариант  $F_\beta$ -меры —  $F_1$ -меру, т. е.  $\beta = 1$ ,  $\alpha = 0,5$ :

$$F_1 = \frac{2PR}{P+R} \quad (4.11)$$

Сравнение  $F_\beta$ -мер с разными  $\beta$  показано на рис. 4.9.



**Рис. 4.9.**  $F_\beta$ -меры: сбалансированная  $\beta = 1$  (а), с приоритетом точности  $\beta^2 = 1/4$  (б), с

приоритетом полноты  $\beta^2 = 4$  (в)

Для обобщения мер качества для всех классов применяют следующие подходы к усреднению:

- а) макроусреднение — обобщение на уровне классов;
- б) микроусреднение — обобщение на уровне документов.

Макроусреднение выполняется путём составления отдельных таблиц принятия решений для каждого класса, вычисления мер по каждой таблице и затем обычного усреднения значений мер по всем классам. Микроусреднение выполняется путём составления единой таблицы для всех классов, в которую сразу записываются решения по всем документам, затем по этой таблице вычисляют меры качества.

Макроусреднение приписывает равные веса решениям классификатора для каждого класса, а микроусреднение — равные веса решениям классификатора для каждого документа. Классы с большим числом документов (и решений по ним) вносят больший вклад в микроусреднение. Таким образом, результат микроусреднения в большей степени оценивают качество классификатора для крупных классов коллекции документов. Чтобы оценить его на малых классах следует применять макроусреднение.

## 4.7 Кластеризация

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов  $X = \{x_1, x_2, \dots, x_n\}$  на непересекающиеся подмножества, называемые **кластерами** (cluster). Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Целями кластеризации могут являться:

- ✓ Понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятия решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»).
- ✓ Сжатие данных. Если исходная выборка избыточно большая, то можно сократить её, оставив по одному наиболее типичному представителю от каждого кластера.
- ✓ Обнаружение новизны (novelty detection). Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров.

Кластеризация относится к технологиям Data Mining (добыча знаний).

Алгоритмы кластеризации:

**Иерархические:** кластеры последовательно строятся из уже найденных кластеров

- Агломеративные (объединительные, восходящие) — начинаем с индивидуальных элементов, затем объединяем
- Дивизимные (разделительные, нисходящие) — начинаем с одного кластера, потом делим.

Нисходящие алгоритмы концептуально более сложные. Нисходящая иерархическая кластеризация может оказаться весьма эффективной, если, например, нет необходимости генерировать полное дерево вплоть до отдельных документов, а ограничиться только верхними уровнями.

**Неиерархические:** оптимизируется некая целевая функция.

### Восходящая кластеризация

Рассмотрим подробнее работу агломеративного алгоритма на простом примере.

Пусть у нас имеется шесть документов, для которых определены ключевые термины (табл. 4.3).

**Таблица 4.3.** Документы, подлежащие кластеризации

docId	Слова в документе
1	китайский Пекин китайский
2	китайский китайский Шанхай
3	китайский Харбин
4	Токио Япония китайский
5	китайский китайский китайский Токио Япония
6	Токио Пекин

Определим для этих слов их веса по методу  $TF*IDF$ , отнормируем их и представим документы в виде векторов (табл. 4.4.):

**Таблица 4.4.** Нормированные вектора документов в пространстве терминов

Слово docId	китай- ский	Пекин	Шанхай	Харбин	Япония	Токио
1	0,31	0,94	0	0	0	0
2	0,20	0	0,98	0	0	0
3	0,10	0	0	0,99	0	0
4	0,25	0	0	0	0,93	0,32
5	0,39	0	0	0	0,78	0,49
6	0	0,84	0	0	0	0,57

Теперь вычислим матрицу расстояний (евклидовых) между векторами (табл. 4.5).

**Таблица 4.5.** Матрица расстояний

	d1	d2	d3	d4	d5	d6
d1	0					
d2	1,36	0				
d3	1,37	1,39	0			
d4	1,36	1,39	1,40	0		
d5	1,32	1,36	1,38	0,27	0	
d6	0,66	1,43	1,41	1,30	1,21	0

Дальнейшие действия определяются выбором критерия, используемого для принятия решения о том, какие кластеры следует объединить на текущем шаге алгоритма. Большое распространение получили три следующих критерия:

1. одиночная связь (минимальное расстояние, или максимальное сходство): сходство двух кластеров есть сходство между их наиболее похожими документами;
2. полная связь (максимальное расстояние, или минимальное сходство): сходство двух кластеров есть сходство между их наиболее непохожими документами;
3. групповое усреднение (усреднение всех показателей сходства): сходство двух кластеров есть среднее сходство всех пар документов, включая пары документов из одного кластера, исключая близость документа самому себе.

Кластеризация с одиночной связью создаёт протяженные («цепочные») кластеры, «сцепленные вместе» элементами, возможно, случайно оказавшимися ближе остальных друг к другу (рис. 4.10, *а*). Этот критерий носит локальный характер, так как не учитывает всю структуру кластера, например, его другие более удалённые части. Кластеризация с полной связью создаёт компактные кластеры (рис. 4.10, *б*) и носит глобальный характер, так как на решение об объединении кластеров влияет вся структура кластера, однако это одновременно повышает чувствительность к выбросам.

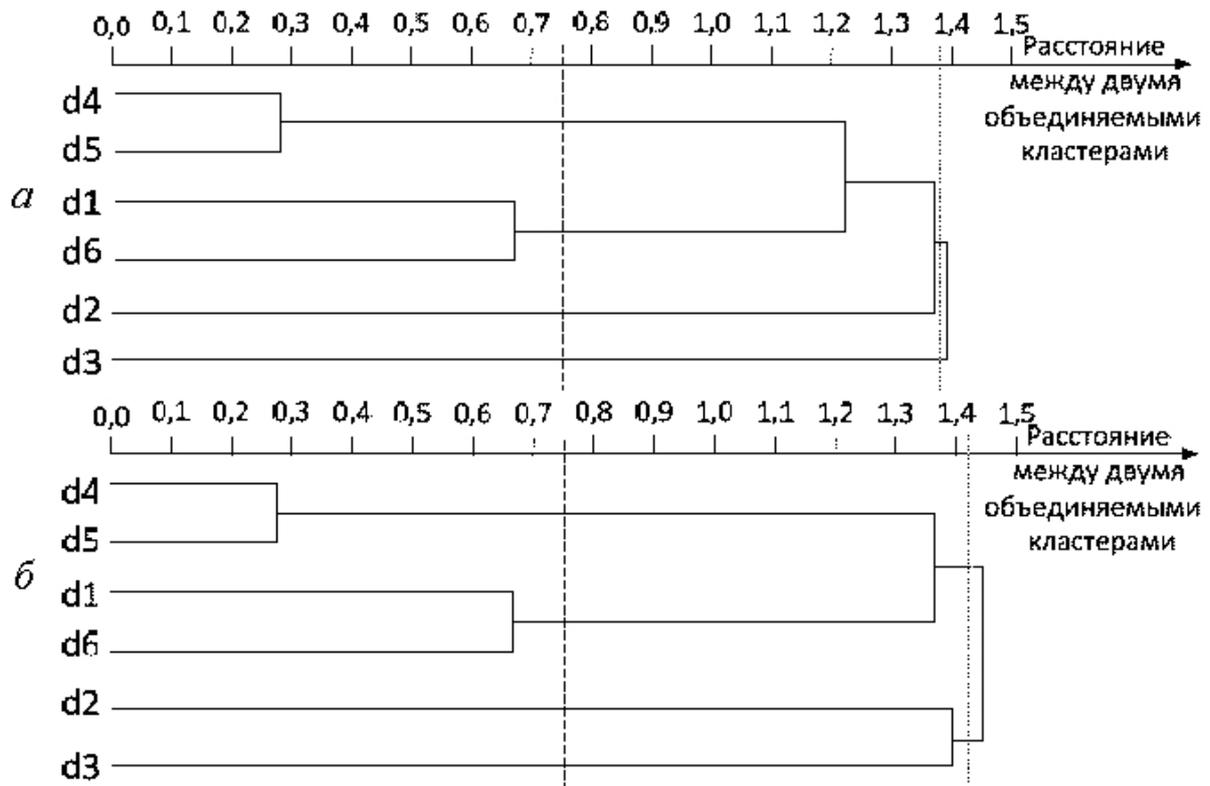


Рис. 4.10. Результат агломеративной кластеризации с одиночной связью (а) и с полной связью (б)

При объединении кластеров надо когда-то остановиться. Можно, например, указать точное число кластеров  $k$ . Разбиение при  $k = 2$  показано пунктирной линией:  $C1 = \{d1, d2, d4, d5, d6\}$ ,  $C2 = \{d3\}$  и  $C1 = \{d1, d4, d5, d6\}$ ,  $C2 = \{d2, d3\}$  соответственно.

Другим критерием прекращения процесса может являться максимальная разница между двумя последовательными мерами сходства (различия) между кластерами, показана штриховой линией:  $C1 = \{d4, d5\}$ ,  $C2 = \{d1, d6\}$ ,  $C3 = \{d2\}$ ,  $C4 = \{d3\}$  в обоих случаях.

### Неиерархическая кластеризация

Среди неиерархических алгоритмов кратко охарактеризуем алгоритм

$k$ -means. Минимизируется мера ошибки  $E(X, C) = \sum_{i=1}^n \|x_i - \mu_i\|^2$ , где  $\mu_i$  —

ближайший к  $x_i$  центроид кластера. В этом алгоритме не точки приписываются к кластерам, а двигается центроид кластера, а принадлежность точек определяется автоматически. Этапы работы: проинициализировать; классифицировать объекты по ближайшему к ним центроиду кластера; перевычислить каждый из центроидов; если ничего не изменилось — остановиться, если изменилось — повторить.

Идеальным кластером алгоритм  $k$ -means считает сферу с центроидом

в центре сферы.

В этом алгоритме необходимо знать число кластеров заранее.

Действие алгоритма начинается с выбора  $k$  начальных центров кластеров. Обычно исходные центры кластеров выбираются случайным образом (рис. 4.11,*а*). Затем каждый объект присваивается тому кластеру, чей центр является наиболее близким документу, и выполняется повторное вычисление центра каждого кластера как центроида, или среднего своих членов (рис. 4.11,*б*). Производится новое распределение объектов по кластерам (рис. 4.11,*в*).

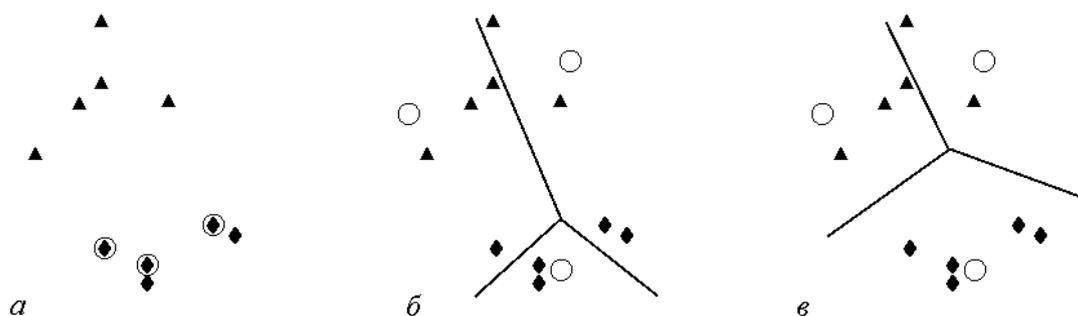


Рис. 4.11. Алгоритм *k-means*

Начальный выбор центров (*а*), первое приближение и новые центры (*б*), второе приближение (*в*)

Видно, что на рис. 4.11,*в* нижний кластер уже полностью совпадает с классом «ромбиков». Такое перемещение объектов и повторное вычисление центроидов кластеров продолжается до тех пор, пока не будет достигнуто условие остановки. Таким условием остановки может служить либо достижение порогового числа итераций, либо неизменность положения центроидов кластеров при очередной итерации, либо достижение порогового значения ошибки кластеризации. На практике используют комбинацию критериев остановки, чтобы одновременно ограничить время работы алгоритма и получить приемлемое качество.

Результат работы алгоритма может зависеть от начального положения центроидов.

Отметим еще существование **нечетких кластеров**: один объект может принадлежать нескольким кластерам с какой-то вероятностью. Точки на краю кластера «меньше принадлежат» кластеру, чем в центре.

## 4.8 Контент-анализ

**Контент-анализ** (англ. *content analysis*; от *content* — содержание) — формализованный метод изучения текстовой и графической информации, заключающийся в переводе изучаемой информации в количественные показатели и ее статистической обработке.

Контент-анализ (КА) стал использоваться в социальных науках, начиная с 30-х гг. XX в. в США. Впервые этот метод был применен в журнали-

стике и литературоведении. Основные процедуры контент-анализа были разработаны американскими социологами Х. Лассуэллом и Б. Берелсоном.

КА связывают с применением процедур подсчета человеком-кодировщиком различных элементов содержания в процессе анализа однотипных текстов. Анализ полученного распределения этих элементов позволяет сделать те или иные выводы о мнении граждан, внимании прессы и т. п.

С помощью КА исследуют различные фрагменты текста. Это может быть совокупность статей, биографий, глав или абзацев учебника, или, наконец, совокупность анкет с текстовыми ответами респондентов на открытые вопросы. Сущность КА заключается в том, что все многообразие текстов по интересующей исследователя тематике сводится к набору определенных элементов, которые затем подвергаются подсчету и анализу. При этом фиксируется как сама частота выделенных элементов, так и частота связанности одних групп элементов с другими, а в ряде случаев и отношение к этим элементам со стороны источника информации (автора текста).

Таким образом, именно в преобразовании качественных данных в количественные и заключается отличие КА от других видов научных исследований.

В отличие от лингвистического анализа при КА подсчитывают не лингвистические единицы, а элементы содержания, которые можно определять по-разному, чем и вызвана некая субъективность результатов. Отдельное слово характеризуется лишь номинативной, назывной функцией, а его появление в тексте отражает лишь факт обращения мысли автора к данному предмету. Единицей выражения мысли, безусловно, является предложение, поэтому для КА часто используется такая единица анализа как «элементарное высказывание» или **фраза**. Под фразой будем понимать часть текста, посвященная одному высказыванию, одному суждению. Такая фраза может состоять из одного или нескольких предложений, или даже из одного слова

Рассмотрим некоторые варианты применения КА с использованием компьютерной системы ВЕГА, разработанной в Экономико-математическом институте РАН совместно с ИТМО.

ВЕГА — это диалоговая система классификации и анализа текстов, использующая принципы КА текстов, словари и классификаторы. Система в основном предназначена для обработки структурированной и, прежде всего, анкетной информации, представляющей собой ответы респондентов на открытые и полузакрытые вопросы социологических анкет. Она позволяет выполнять также некоторые элементы анализа текста: составление словарей, подсчет встречаемости слов, поиск слов по словарю и по тексту и т. д. Система обеспечивает статистический анализ ответов на закрытые и полузакрытые вопросы. Кроме того, существует огромное количество «текстовых документов», для которых можно подготовить описание по

стандартной форме. Таким типом документов может быть подборка публикаций в журналах или газетах, совокупность социологических учебников, сочинения школьников, приговоры суда и т. п. В этом случае можно просто разработать своего рода стандартную анкету и заполнить ее на каждый документ.

Процесс работы состоит из следующих основных этапов.

1. Создается первичный вариант классификатора. ВЕГА позволяет работать с «двухъярусным» классификатором, в котором каждый класс еще подразделяется на группы.

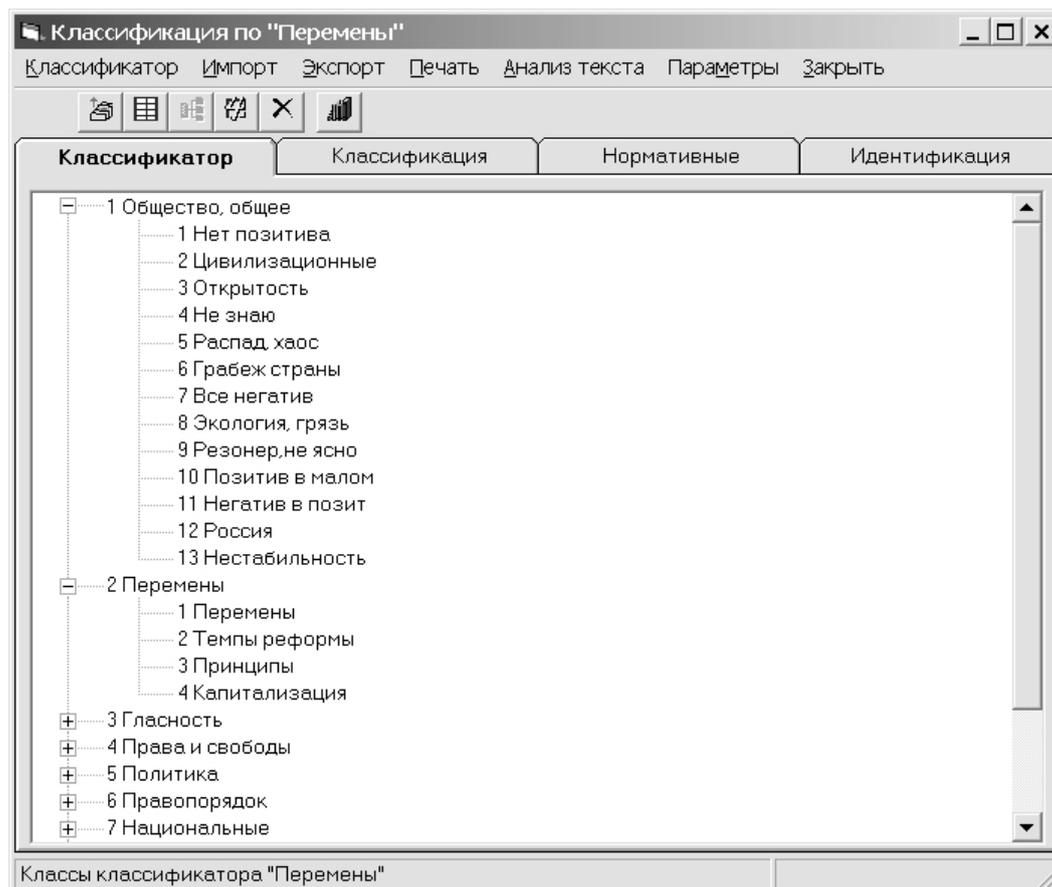


Рис. 4.11. Представление классификатора в виде дерева

Пример такого классификатора, созданного для анализа открытых ответов на вопросы социологической анкеты о реформах 90-х гг. показан на рис. 4.11

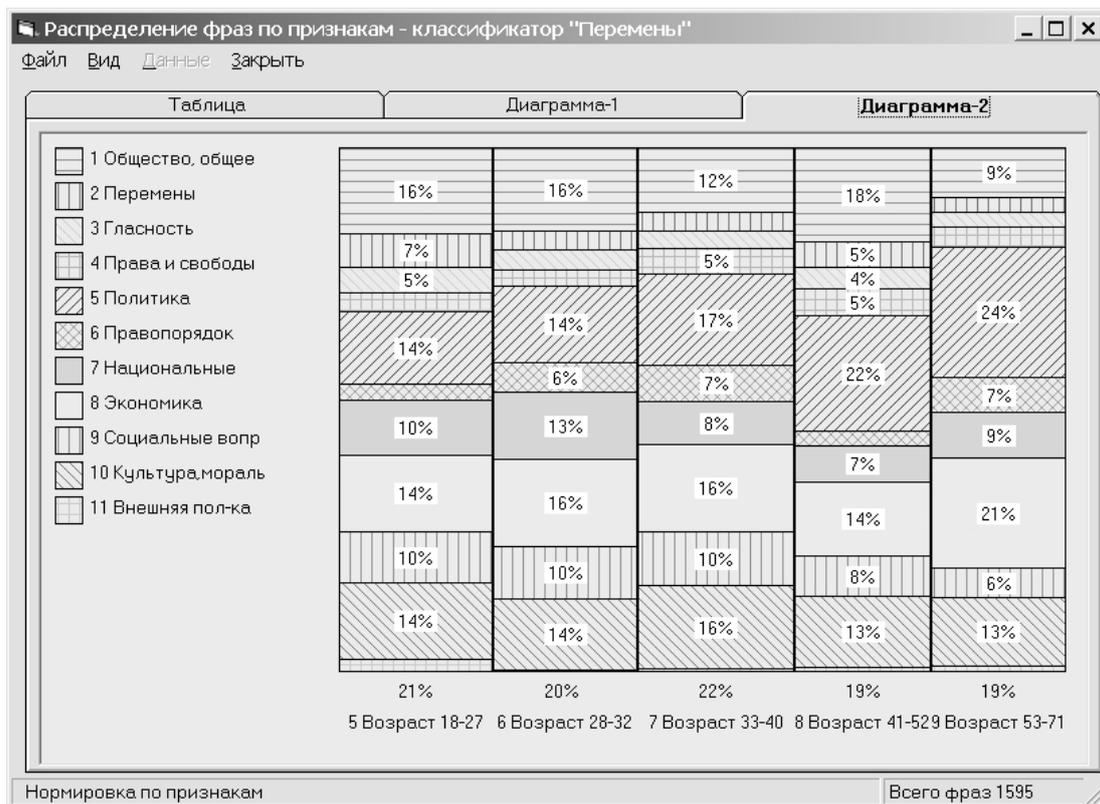
2. Отбираются наиболее характерные фразы. Каждой фразе присваивается соответствующие ее смыслу класс и группа. После этого фраза становится своеобразным эталоном для последующего анализа — **нормативной фразой**.
3. Далее переходят к **идентификации** фраз из текста посредством их сравнения с нормативными фразами. Идентификация может осуществляться под контролем исследователя, а может проводиться автоматически по методикам, аналогичным описанным выше. Если очередная фра-

за совпадают по смыслу с одной из нормативных, она получает свой **идентификатор** — номер нормативной фразы.

4. Если фраза из текста не совпадает по смыслу ни с одной нормативной, классификатор уточняется, после чего процесс идентификации повторяется.
5. Этот процесс повторяют несколько раз до тех пор, пока все фразы из текста не будут отождествлены с нормативными. После этого процедура КА закончена, создан окончательный вариант классификатора, и каждая исходная фраза получила свой идентификатор. А поскольку каждая нормативная фраза имеет свой класс и группу, это позволяет в дальнейшем использовать для обработки обычные статистические методы, имея под рукой исходный текст.

Таким образом, проводится классификация только части фраз, а все остальные фразы отождествляются с заранее отобранными. Классификатор создается непосредственно в процессе классификации фраз. Это дает двойное преимущество. Во-первых, сокращается объем работы по классификации фраз. Во-вторых, при любом изменении классификатора — а это не исключение, а правило при анализе текстов — достаточно изменить класс и группу у ряда нормативных фраз, чтобы автоматически произошли соответствующие изменения и у всех фраз, совпадающих с ними.

Исследователя часто интересует зависимость полученного распределения по классам от пола, возраста, зарплаты, национальности и других характеристик респондентов. Такие характеристики чаще всего представлены в числовой форме. Для такого рода исследований можно использовать анализ по признакам. Каждый признак представляет собой условие отбора записей для анализа. Так на рис. 4.12 представлено распределение количества фраз в зависимости от возраста респондентов.



**Рис. 4.12.** Распределение фраз по возрасту

Из диаграммы ясно видна, прежде всего, неравномерность распределения фраз по классам — больше всего респондентов волновали состояние общества (класс 1), политика (класс 5), экономика (класс 8) и культура, мораль (класс 10). Интересно, что самые молодые респонденты наиболее заинтересованы именно теми темами, которые перечислены выше. Респонденты в возрасте 28 – 32 года дополнительно интересуются национальными вопросами (класс 7). Почему-то больше всего внимания к теме культуры и морали обращают респонденты в средней возрастной группе (от 33 до 40 лет), хотя еще большее внимание они уделяют политике. Наиболее обеспокоены политикой респонденты в возрасте от 41 до 52 лет, они же больше всех интересуются состоянием общества (класс 1). Самых пожилых респондентов политика и экономика волнует больше, чем всех других.

## 5 Литература

1. Автоматическая обработка текста (материалы сайта) [Электронный ресурс]. – Режим доступа: <http://www.aot.ru>, свободный.
2. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика // Большакова Е.И. и др. – М.: МИЭМ, 2011. – 272 с.
3. Аношкина Ж.Г. Словарь омонимичных словоформ русского языка. М.: Машинный фонд русского языка Института русского языка РАН, 2001. ([Электронный ресурс]. – Режим доступа: <http://irlras-cfrl.rema.ru:8100/homofoms/index.htm>, свободный.).
4. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003—2005. М.: Индрик, 2005. (см. [Электронный ресурс]. – Режим доступа: <http://ruscorpora.ru/sbornik2005/12apresyan.pdf>, свободный.).
5. Баженов Д. О задачах классификации [Электронный ресурс]. – Режим доступа: // <http://bazhenov.me/blog/>, свободный.
6. Баранов А. Н. Введение в прикладную лингвистику. М., 2003.
7. Боярский К. К., Каневский Е. А. Вега — компьютерная система классификации и анализа текстов. Lambert Academic Publishing, 2011.
8. Боярский К. К., Каневский Е. А. Семантико-синтаксический анализатор SemSin Международная конференция по компьютерной лингвистике «Диалог-2012», [Электронный ресурс]. – Режим доступа: <http://www.dialog-21.ru/digest/2012/?type=doc>, свободный.
9. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения. М.: «Наука», 1985. — 143 с.
10. Зализняк А. А. Грамматический словарь русского языка, М., 2003.
11. Золотова Г. А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. – М., 2006. 2-е изд.
12. Кобзарева Т. Ю., Афанасьев Р. Ю. Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций // «Компьютерная лингвистика и интеллектуальные технологии». Труды Международного семинара «Диалог'2002» (Протвино, 6 – 11 июня 2002 г.). В двух томах. Под редакцией А. С. Нариньяни. Т. 2. Прикладные проблемы. С. 258 – 268. (см. [Электронный ресурс]. – Режим доступа: <http://www.dialog-21.ru/materials/archive.asp?id=7569&y=2002&vol=6078>, свободный.).
13. Кобозева И. М. Лингвистическая семантика. – М., 2000.
14. Коваль С. А. Лингвистические проблемы компьютерной морфологии. СПб., 2005.

15. Леонтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы. М., 2006.
16. Маннинг К. Д., Рагхаван П., Шютце Х. Введение в информационный поиск.: Пер. с англ. – М.: ООО «Вильямс», 2011. – 528 с.
17. Мельчук И. А. Опыт теории лингвистических моделей «Смысл <=> Текст». М., 1974. М., 1999.
18. Муравенко Е. В. Что такое синтаксическая омонимия? Лингвистика для всех. Летние лингвистические школы 2005 и 2006 / Ред.–сост. Е. В. Муравенко, О. Ю. Шеманаева. – М.: МЦНМО, 2008. – 440 с.
19. Национальный корпус русского языка // [Электронный ресурс]. – Режим доступа: <http://www.ruscorgora.ru>, свободный.
20. Рогожникова Р. П. Толковый словарь сочетаний, эквивалентных слову. М., Астрель–АСТ, 2003. – 416 с.
21. Рубашкин В. Ш. Прикладная лингвистика и языковая инженерия // Труды международной конференции «MegaLing'2005. Прикладная лингвистика в поиске новых путей». СПб., 2005.
22. Рыков В. В. Лекции и статьи по корпусной лингвистике // [Электронный ресурс]. – Режим доступа: <http://rykov-cl.narod.ru>, свободный.
23. Тестелец Я. Г. Введение в общий синтаксис. М., 2001 – 800 с.
24. Тузов В.А. Компьютерная семантика русского языка. – СПб: Изд-во СПбГУ, 2004. – 400 с.
25. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – 2-е изд., перераб. и доп.– СПб.: БХВ-Петербург, 2007.



В 2009 году Университет стал победителем многоэтапного конкурса, в результате которого определены 12 ведущих университетов России, которым присвоена категория «Национальный исследовательский университет». Министерством образования и науки Российской Федерации была утверждена программа его развития на 2009–2018 годы. В 2011 году Университет получил наименование «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

## КАФЕДРА ФИЗИКИ

Кафедра физики как одна из общеобразовательных кафедр существует с момента организации Ленинградского института точной механики и оптики. В довоенные и послевоенные годы кафедру возглавляли А.П. Ющенко, затем профессора В.Ф. Трояновский, Л.С. Поллак, И.В. Поройков, К.К. Аглинцев, Д.Б. Гогоберидзе, Н.А. Толстой, С.В. Андреев, А.Я. Вятский, основоположник теплофизической школы ЛИТМО профессор Г.М. Кондратьев и его ученики доцент А.Ф. Бегункова и профессор Н.А. Ярышев, профессор С.К. Стафеев. В настоящее время коллектив кафедры составляют выпускники нашего университета, Ленинградского Политехнического института, физического факультета Ленинградского государственного университета. На кафедре успешно работают свыше 20 профессоров и доцентов.

С момента образования кафедры ее сотрудники уделяют большое внимание совершенствованию методики преподавания физики, как одной из базовых дисциплин подготовки будущих инженеров и формирующему интеллект предмету. Сотрудниками кафедры написано более сорока учебных пособий для студентов по различным разделам инженерного курса физики. В настоящее время проводится комплексная работа по совершенствованию всего учебного процесса, включая создание фронтальных компьютеризированных учебных лабораторий, банков контроля и проверки усвоения знаний, подготовку программно-методического обеспечения по дистанционному обучению студентов через компьютерные сети RUNNET и INTERNET. В тесном сотрудничестве с объединением «Росучприбор» ведется разработка лабораторных учебных стендов и практикумов.

Коллектив кафедры ведет активную научную работу. В 1957-1973 годах было сформировано научное направление по исследованию физики взаимодействия электронных пучков с веществом. С 1973 года получили развитие научные исследования в области нестационарной теплопроводности и теплотрии. С 1979 года стали проводиться научные разработки в области спектроскопии разупорядоченных конденсированных систем, с 1987 года по физике волновых процессов, нелинейной оптике и радиофизике анизотропных сред, с 1994 года – по оптическому и рентгеновскому рассеянию надмолекулярными, в частности, фрактальными структурами, с 1999 года – по фотонным кристаллам. Научные разработки кафедры неоднократно удостоивались грантов Министерства образования, Российских и Международных научных фондов. Десятки студентов и аспирантов, руководимые преподавателями кафедры, удостоивались стипендий Президента России, Правительства России, Администрации Санкт-Петербурга, Международных организаций (ISSEP, SPIE и др.).