Компьютерная лингвистика: синтез математики, лингвистики и информатики

Международный Институт Менеджмента ЛИНК
ФАКУЛЬТЕТ лингвистики
К.п.н., доцент Матвеева Н.В.

Москва-Жуковский, 2017

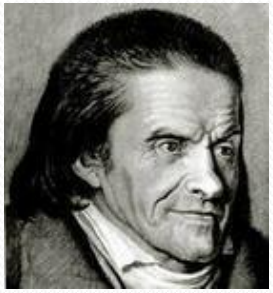
Наша цель

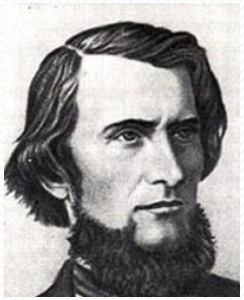
Показать, что **лингвистика, математика и информатика** давно и неразрывно связаны между собой неким общим содержанием, которое реализовано в научном направлении «Компьютерная лингвистика».

Познакомить с задачами, которые решает «Компьютерная лингвистика»

Осветить состояние на сегодняшний день, роль и значение учебного курса «Компьютерная лингвистика» для современного студента как будущего специалиста-лингвиста.

Компьютерная лингвистика в системе знаний





Все связано со всем, все науки связаны между собой! Об этом писали в своих трудах великие педагоги Я.А. Коменский, И.Г. Песталоцци, К.Д. Ушинский и другие.

Однако, структура школьных знаний и существующие формы обучения, до сих пор формирует «лоскутное» мировоззрение.

Компьютерная лингвистика аккумулирует в себе **математику, логику, лингвистику и информатику**, формируя целостное представление о реальной действительности.

Роль математики в науках

«Использование науками отвлеченных понятий и методов **математики** расширяет их возможности, способствует открытию новых, более глубоких закономерностей».

А.Т.Хроленко

Лингвистика – пионер использования математики в гуманитарных науках

Науки естественного цикла – физика и информатика – давно заговорили на языке математики.

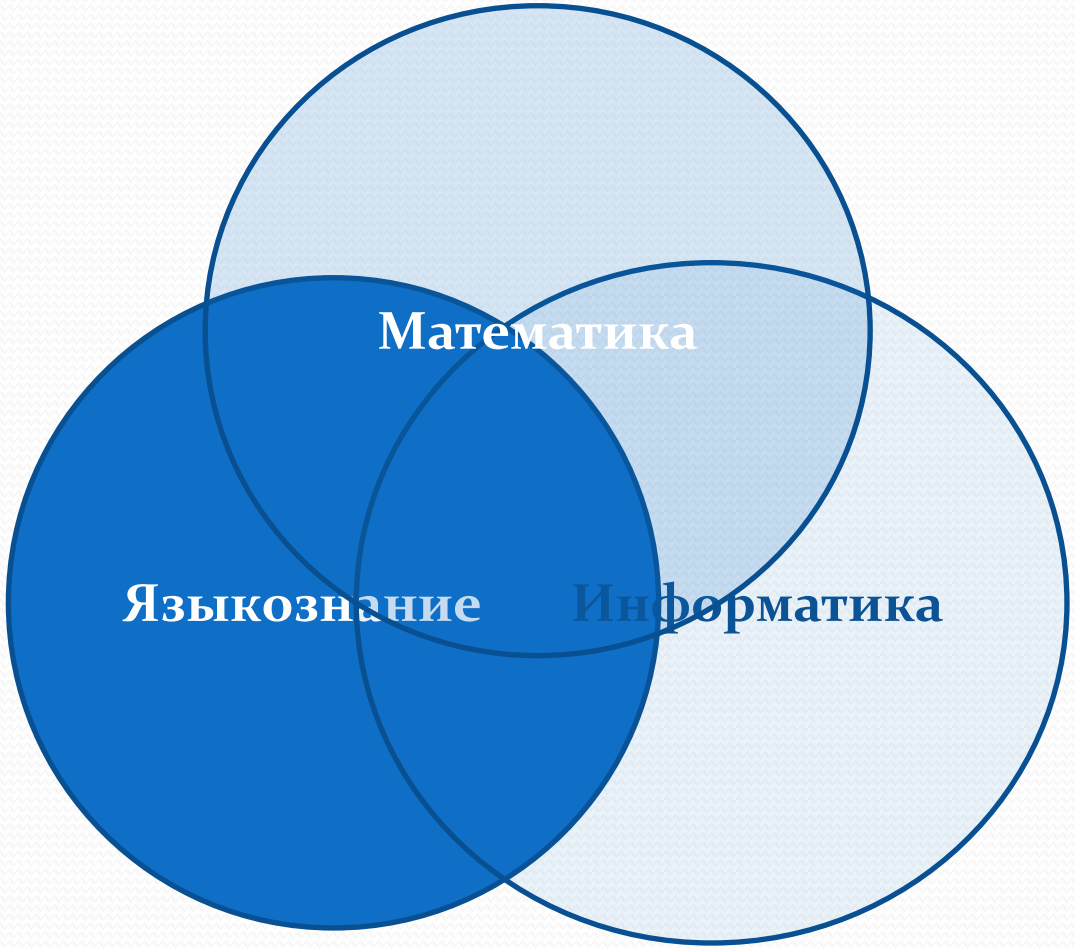
Гуманитарные науки обратились к математике только в **XX** веке.

И первой из них была лингвистика!

Пересечение поля понятий

Понятия «язык», «слово», «знак» и «модель» являются общими понятиями для информатики, математики и филологии – то есть содержание этих предметов взаимно пересекается и дополняет друг друга.

Пересечение поля понятий



Внутренняя связь математики, информатики и филологии

Связь математики, информатики и филологии находит свое отражение не только в **общих понятиях**, но и в **способе мыслительной деятельности** – и математика, и информатика, и филология учат **моделированию явлений окружающей действительности**.

То есть в **методологической основе** всех трех дисциплин лежат одни и те же методы: **формализация и моделирование явлений окружающей действительности**.

Язык как общее понятие и универсальное средство моделирования

Язык – это не только общее понятие математики, информатики и лингвистики, но **универсальное средство моделирования**. Слово естественного языка есть его минимальная смысловая единица. Оно же – форма представления информации в сознании человека в виде понятия.

Язык подчиняется **определенным закономерностям**, что ввело лингвистику в **цикл дисциплин**, к которому принадлежат математика и классическая аристотелевская логика

Язык как системный механизм

НО ГЛАВНОЕ, что выявил Фердинант де Соссюр:

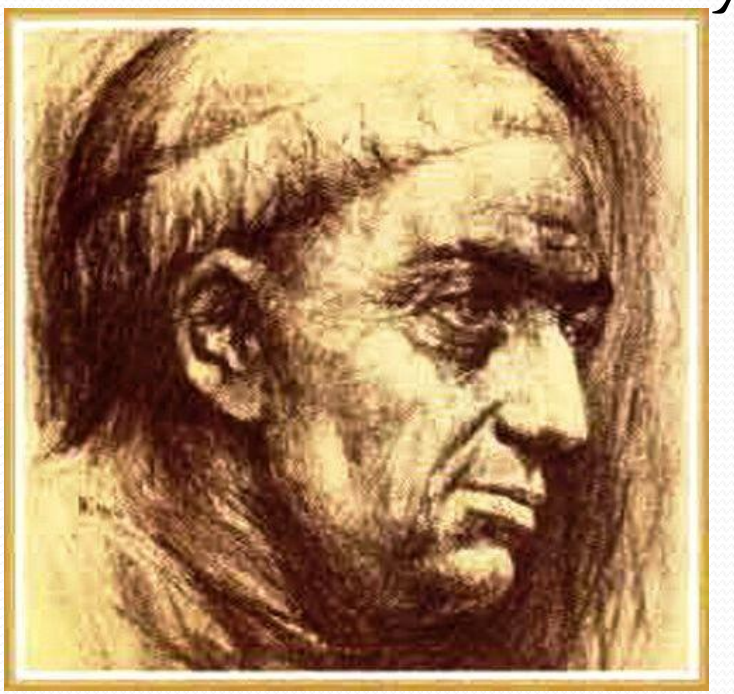
язык – это **системный механизм**, функционирование которого проявляется в речевой деятельности его носителей, который подчиняются ряду закономерностей (вероятностным, статистическим и пр.).

Языкознание и математика

Взаимодействие **языкознания** и **математики** имеет богатую историческую традицию.

Например, задача построения формальной математической модели языка и речи на протяжении десятилетий занимала и занимает умы многих исследователей.

Все надо записывать в цифрах



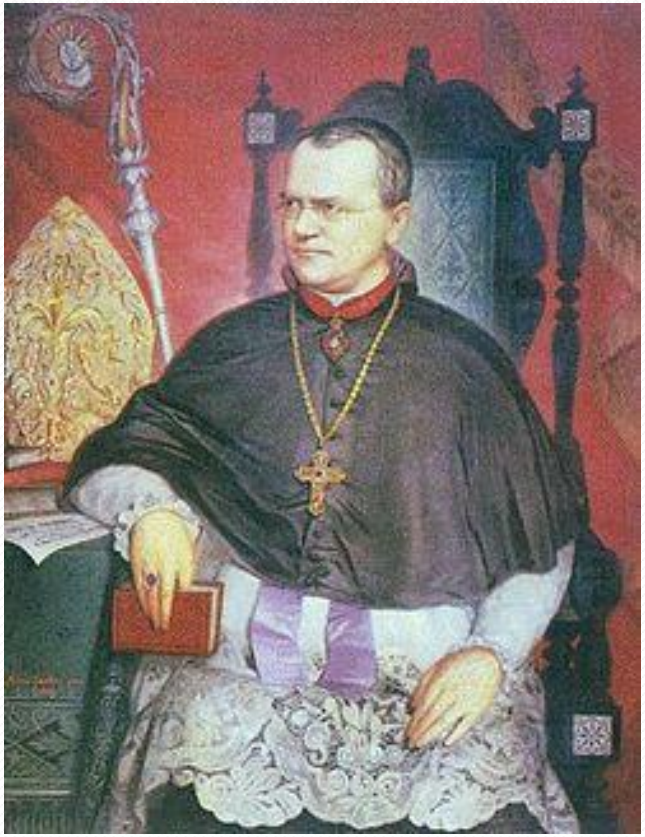
Уже в X веке ученый и философ эпохи возрождения Николай Кузанский в своем трактате «Об ученом познании» утверждал, что все познания о природе необходимо записывать в цифрах, а все опыты над нею производить с весами в руках.

Математику **нужно** применять в лингвистических исследованиях

В начале XIX века, известный русский математик Виктор Яковлевич Буняковский говорил о **необходимости применения математики в лингвистических исследованиях**.

Универсальные статистические методы исследования

Тогда же, в начале XIX века, основоположник генетики австриец Грегор Иоганн **Мендель** пробовал применить статистические методы исследования не только **в биологии, и метеорологии**, но и **лингвистике**.



Тесная связь языкознания и математики

В начале **XX** века российский и польский языковед И.А.Бодуэ́н де Куртене́ указывал на тесную связь языкознания и математики.

Он говорил о том, что применение **количественных методов** в языкознании **приблизит его к точным наукам**.

Марков АА – марковские цепи и лингвистика

Тогда же, в начале XX века, известный математик Андрей Андреевич Марков, ученик П.Л.Чебышева, применяя **методы теории вероятностей и математической статистики** пришел к открытию «марковских цепей».

Он исследовал доли гласных и согласных в тексте А.С.Пушкина «Евгений Онегин».

Норберт Винер – сын лингвиста Лео Винера

Математик и основоположник кибернетики Норберт Винер – сын известного американского лингвиста Лео Винера, который перевел с русского на английский 24-томное собрание сочинений Л. Н. Толстого – не без оснований считал, что в науке **лингвистике есть все условия, необходимые для математического исследования.**

С чего же все началось?

В ходе Нюбрнбергского процесса в связи с необходимостью синхронного перевода был задуман м**ашинный перевод. Это было в 1944-45** годах.

Личный переводчик Эйзенхауэра и его личный знакомый – сейчас он руководитель одного из отделов фирмы IBM – затеяли так называемый «Джорджтаунский эксперимент»: **перевод с русского на английский с помощью ЭВМ**.

Была создана программа. Словарь содержал **250** слов.

При переводе использовалось 6 правил. Данные хранились на перфокартах и вводились латиницей.

Начало машинного перевода в СССР и России

Именно в нашей стране в конце **1955** г. были проведены первые опыты **перевода научно-технического текста с английского языка** на русский при помощи электронной счетной машины (ЭВМ) БЭСМ Академии наук СССР ...

Начало машинного перевода в СССР и России

Именно в нашей стране в **1958** году вышла работа **Панова** Дмитрия Юрьевича «Автоматический перевод».

В книге рассказывается о проблеме **автоматического перевода с одного языка на другой**.

Занимался этой проблемой Институт точной механики и вычислительной техники и Институт научной информации Академии наук СССР.

А.П.Ершов и компьютерная лингвистика

Во второй половине XX века, **основоположник школьной информатики** А.П.Ершов, один из пионеров **теоретического и системного программирования**, создатель Сибирской школы информатики, **становится одним из основателей российской корпусной и компьютерной лингвистики**.

Ершов - лингвист, математик и информатик в одном лице

Ершов — один из **пионеров российской корпусной лингвистики.**

По его инициативе начал создаваться Машинный фонд русского языка при Институте русского языка АН СССР.

Лингвист Мельчук И.А.

В 60-тые годы А.П. Ершов уже работает над проблемой **общения человека с ЭВМ на естественном языке**.

К решению этой проблемы он привлекает **лингвиста** Мельчука Игоря Александровича (р. 1932) – в настоящее время канадского лингвиста российского происхождения, создателем лингвистической теории «Смысл — Текст».

Лингвистика и информатика давно в тесном содружестве

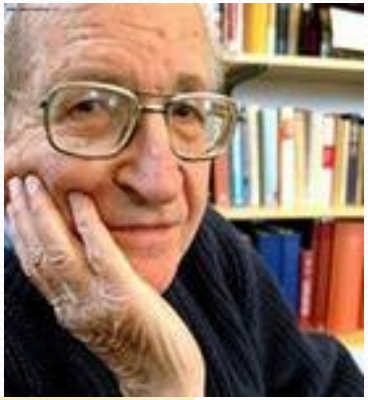
Мельчук окончил испанское отделение филологического факультета МГУ.

В **1956** работал в **Институте языкознания** АН СССР, где занимался проблемой машинного перевода и к началу 70-х годов он является лидером в области **структурной прикладной лингвистики**.

В 1974 начинает работу над интегральной моделью языка «Смысл — Текст», которая в то время значительно опережала аналогичную теорию **Н. Хомского**

В настоящее время Мельчук профессор университета в Монреале.

Ноам Хомский и искусственный интеллект …

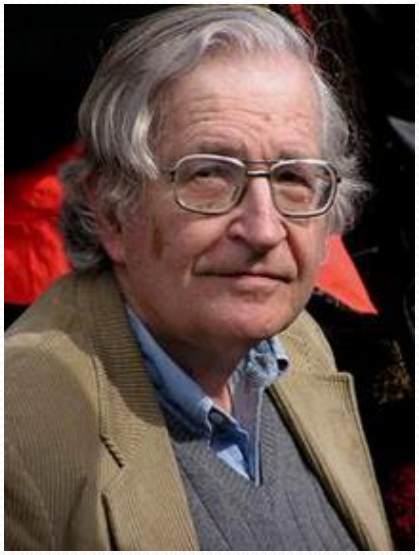


Ноам Хомский – лингвист, основатель когнитивной психологии, политик и публицист., автор классификации формальных языков. **Но важно, что …**

В **60-ые** годы он принимает участие в конференциях по **искусственному интеллекту**, где обсуждались вопросы **моделирования** процессов **порождения и восприятия речи.**

Сейчас проблема искусственного интеллекта – это один из важнейших разделов науки **информатики и компьютерной лингвистики**.

Универсальные правила языка



Воспринимая речь на конкретном языке, **естественный механизм овладения языком** извлекает из нее **универсальные правила**, этого языка

– это то, что компьютер сделать не может!

Но с помощью компьютера можно выявить **количественные соотношения** лингвистических единиц, на основе которых выявить различные языковые закономерности!

Интересный факт

Хомский показал, что у **детей** быстро развиваются языковые навыки, что **позволяет им извлекать одинаковый смысл (глубинную структуру) из различного порядка слов (поверхностной структуры)**

Например, в предложениях: «собака погналась за мальчиком» и «мальчика преследовала собака» **поверхностная структура различается, но глубинная структура одинакова**

Плунгян Владимир

Что есть язык для компьютерной лингвистики

В компьютерной лингвистике **язык** рассматривается как одна из ментальных или интеллектуальных способностей человека наряду с памятью, вниманием, интуицией и пр.

Компьютер – это электронное устройство с программным обеспечением, который не понимает человеческого языка. Между собой компьютеры «разговаривают» на компьютерном языке – языке **двоичных кодов**.

Внутренний язык компьютера

Внутренний язык компьютера – язык двоичных кодов – по своей природе не совместим с **языком человеческим,**

и поэтому – пользоваться людям языком компьютера в коммуникационном процессе общения

абсолютно невозможно!

Компьютеры становятся все более совершенными, но «собственный язык» компьютеров за все эти годы не изменился: кроме сложения нулей и единиц компьютер ничего делать не умеет.

Машинный перевод: успехи, неудачи, надежды

История машинного перевода текстов с одного языка на другой с помощью компьютера уже насчитывает без малого **шестьдесят** лет.

За это время сменилось несколько поколений систем машинного перевода: от почти игрушечных моделей, переводивших текст слово за словом без учета контекста, ученые перешли к сложным системам, создавая правила, учитывающие тонкие смысловые оттенки переводимого текста.

Машинный перевод: успехи, неудачи, надежды

В истории машинного перевода были свои взлеты и падения: энтузиазм первопроходцев сменялся глубоким пессимизмом, когда видные специалисты приходили к убеждению, что задача машинного перевода не может быть решена в обозримом будущем.

Сейчас машинный перевод **переживает второе рождение**: благодаря сочетанию различных методов и подходов качество перевода заметно улучшается и в эту область вовлекаются все новые языки, становится возможным перевод речи и синтез устнойречи.

Лингвистика, информатика, компьютер и информационные технологии

Суть нашего сообщения в том, чтобы показать, что лингвистика, математика, информатика и информационные технологии давно и неразрывно связаны между собой определенным общим содержанием (знаниями и способами деятельности – по Ледневу В.С.).

Это общее содержание концентрируется в дисциплине «Компьютерная лингвистика», которая только сейчас вводится в высших учебных заведениях, в то время как **лингвистика, математика и информатика пошли навстречу друг другу** (А.Т.Хроленко, А.В.Денисов) тогда, **когда**:

Лингвистика, математика и информатика пошли навстречу друг другу, когда:

- была осознана возможность машинного перевода (**1945,** Нюрнбергский процесс);
- были открыты закономерности частотного распределения слов в тексте на любом естественном языке (**1949,** законы Зипфа);
- появились теоретические работы о связи речи и статистики (1970, Головин Б.Н.);
- компьютеры стали персональными (**1981, IBM**),
- начали интенсивно развиваться **текстовые процессоры**;
- появились средства **проверки** орфографии и грамматики;
- появились программы **оптического распознавания текстов и речи** …

Язык и статистика

Головин
Борис
Николаевич
(1916 – 1984)
Советский лингвист
Доктор
филологических наук

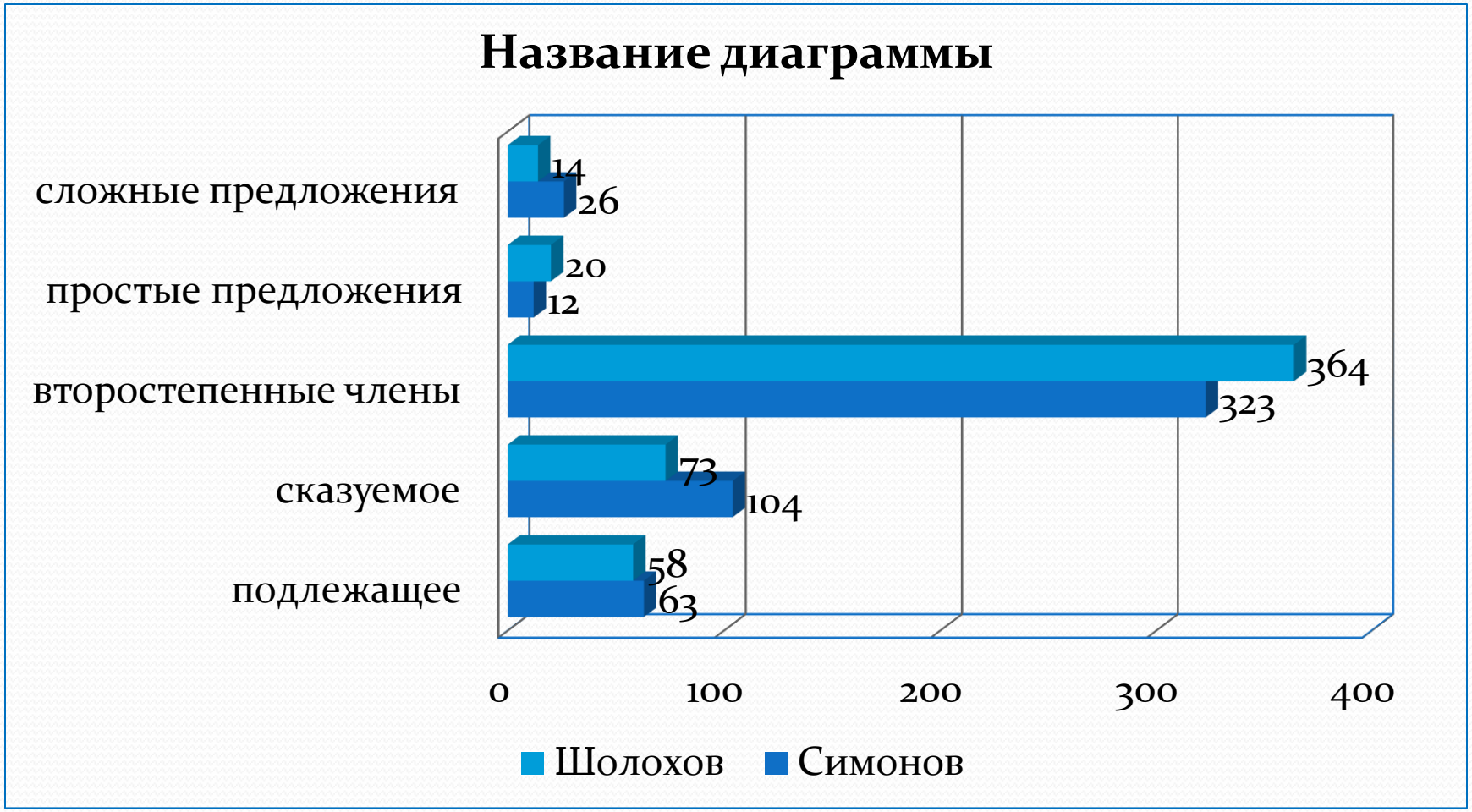
Анализ стилей

Симонов
Константин
Михайлович
(1915-1979)

Шолохов
Михаил
Александрович
(1905-1984)

Сравнение использования различных частей речи

Анализ структуры предложений



Законы Зипфа

Законы Зипфа описывают закономерности частотного распределения слов в тексте на любом естественном языке. Их опубликовал в 1949 году американский лингвист Джордж Зипф (*George Kingsley Zipf* — из-за его происхождения фамилия часто встречается в немецком прочтении «Ципф»).

Законы **эмпирические** — они не имеют строгого математического доказательства и основаны **на статистическом анализе распределения слов** в больших массивах текстов на разных языках.

Первый закон Зипфа: «ранг – частота»

Вероятность обнаружения любого слова, умноженная на его ранг — постоянная величина.

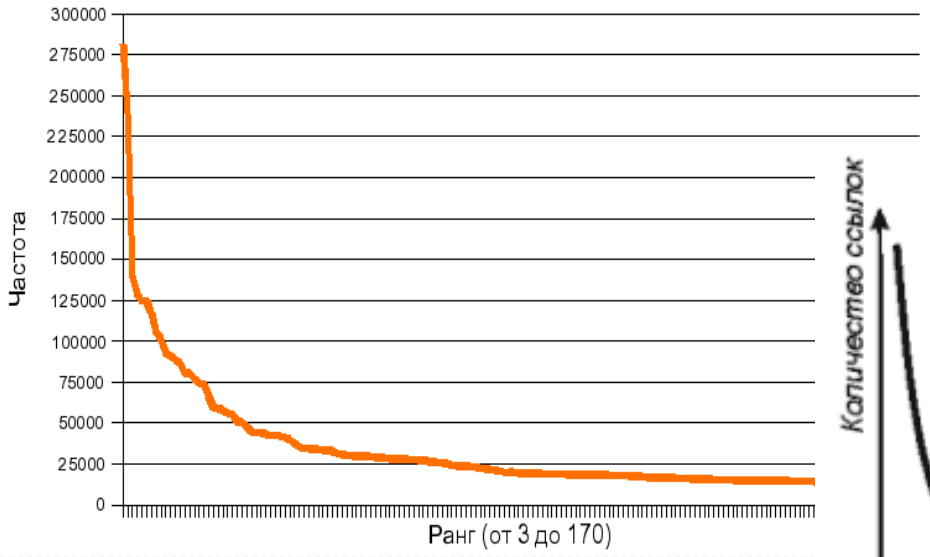
В любом тексте, написанном человеком, этот закон статистически верен.

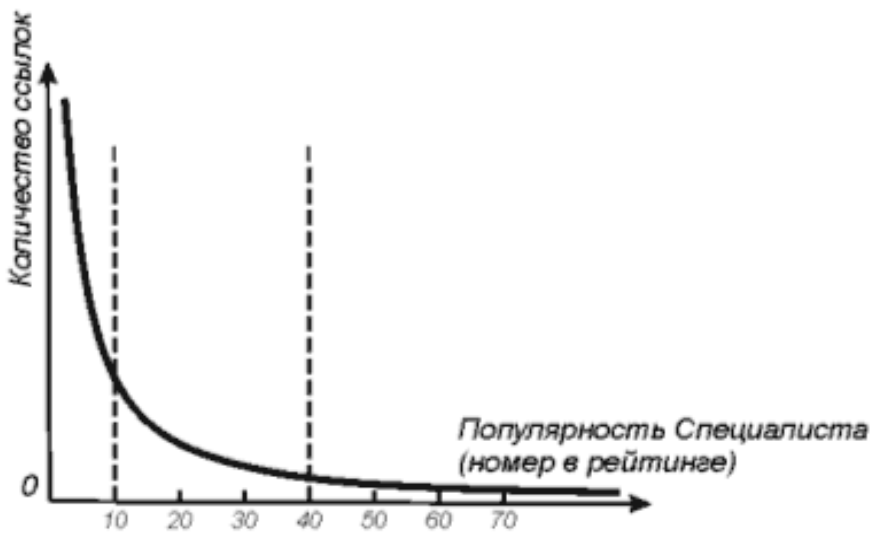
Второй закон Зипфа: «количество – частота»

Разные слова в большом массиве текста могут иметь одинаковое количество вхождений. Если построить график, где ось X отображает частоту слова, а ось Y — количество слов, входящих в текст с такой частотой, то **для любого массива текста этот график будет одинаковым**.

В логарифмическом масштабе этот график близок к прямой линии. В линейном масштабе график напоминает гиперболу, наклон кривой на начальном участке различается для разных языков, но для всех текстов на одном языке кривая распределения одинакова.

Количество ссылок и популярность специалиста





Лингвистика, математика и информатика всегда были рядом и вместе,

а не когда информатика достигла чего-то!

Лингвистика и информатика **изначально, уже больше 60 лет идут рука об руку**, взаимно поддерживая и «толкая» вперед друг друга!

Компьютерная лингвистика

Компьютерная лингвистика смотрит на компьютер как на эффективное и удобное **средство** максимально полного и всестороннего изучения языка во всех его ипостасях и формах.

Растет производительность компьютеров и тогда …

… возникает заманчивая идея использовать компьютер для решения таких лингвистических задач, как: **индексирование, реферирование, аннотирование, перевод, обработка текста, установление авторства, формирование репрезентативных корпусов и пр.**

Производительность компьютера – это хорошо

Однако, без должного **лингвистического обеспечения** современные информационные системы и технологии зайдут в тупик и не смогут помогать человеку справляться с колоссальным потоком информации.

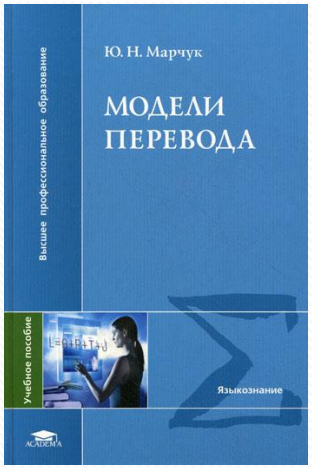
Что такое компьютерная лингвистика

Компьютерная лингвистика – одна из наиболее актуальных современных **лингвистических дисциплин**.

«По мере того, как расширяется информатизация современного общества, возрастает значение компьютерной лингвистики, науки, находящейся на стыке глубоко человечной, гуманитарной науки лингвистики (языкознания), изучающей законы развития и пользования могучим средством мышления и коммуникации – языком, – и компьютерного знания, с помощью которого машине передастся все большая часть интеллектуального труда человека»

(Марчук Ю.Н.)

Марчук Юрий Николаевич



Профессор МГУ, доктор филологических наук, специалист по прикладной и **компьютерной лингвистике**, машинному переводу, автоматическому анализу и синтезу текстов, терминологии и терминоведению, лексикологии и лексикографии, общему языкознанию.

Компьютерная лингвистика

Компьютерная лингвистика – это актуальная область современного языкознания и информатики.

Компьютерная лингвистика занимается разными проблемами **компьютерной обработки** текстов на естественных языках, в том числе проблемами компьютерного перевода и др.

Компьютерная лингвистика

Компьютерная лингвистика нужна специалистам, перед которыми стоят задачи построения **систем искусственного интеллекта**

и

всем изучающим и интересующимся лингвистикой

и ее приложениями.

Задачи компьютерной лингвистики

Компьютерная лингвистика решает задачи **автоматического анализа текста**, такие как:

- машинный **перевод**;
- **распознавание** текста и речи;
- создание **количественной характеристики** текста;
- **порождение** текстов (стихов, народных сказок и пр.);
- **информационный поиск**;

и многое др.

Дисциплина «Компьютерная лингвистика» по Марчуку Ю.Н. (МГУ)

Как учебная дисциплина, «компьютерная лингвистика находится **в процессе становления и пока** нет единой точки зрения на ее содержание, задачи и методы обучения»

(Марчук Ю.Н.)

Это должны знать современные лингвисты

Владимир Селегей:

За полвека существования компьютерная лингвистика пережила периоды больших надежд и таких же больших разочарований.

Как будет развиваться компьютерная лингвистика, на что она способна сегодня, и что сможет завтра – все это должны знать современные специалисты в области лингвистики

Заключение

Мы показали, что **лингвистика, математика и информатика** давно и неразрывно связаны между собой неким общим содержанием, которое реализовано в научном направлении «Компьютерная лингвистика».

Компьютерная лингвистика решает насущные **потребности** современного информационного общества: помогает выявлять закономерности развития естественных языков, создавать словари и справочники, осуществлять машинный перевод и анализ текстов, получать количественные характеристики текстов, определять авторство и многое другое.

Учебный курс «Компьютерная лингвистика» заслуживает особого внимания!

СПАСИБО ЗА ВНИМАНИЕ!