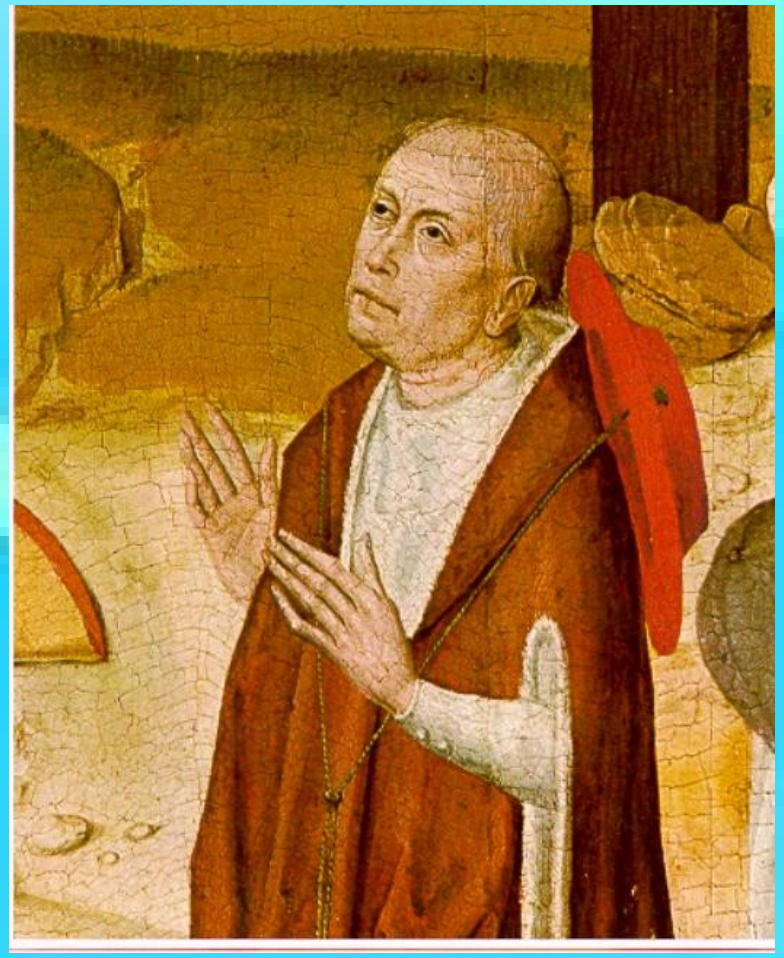
Истоки квантитативной лингвистики

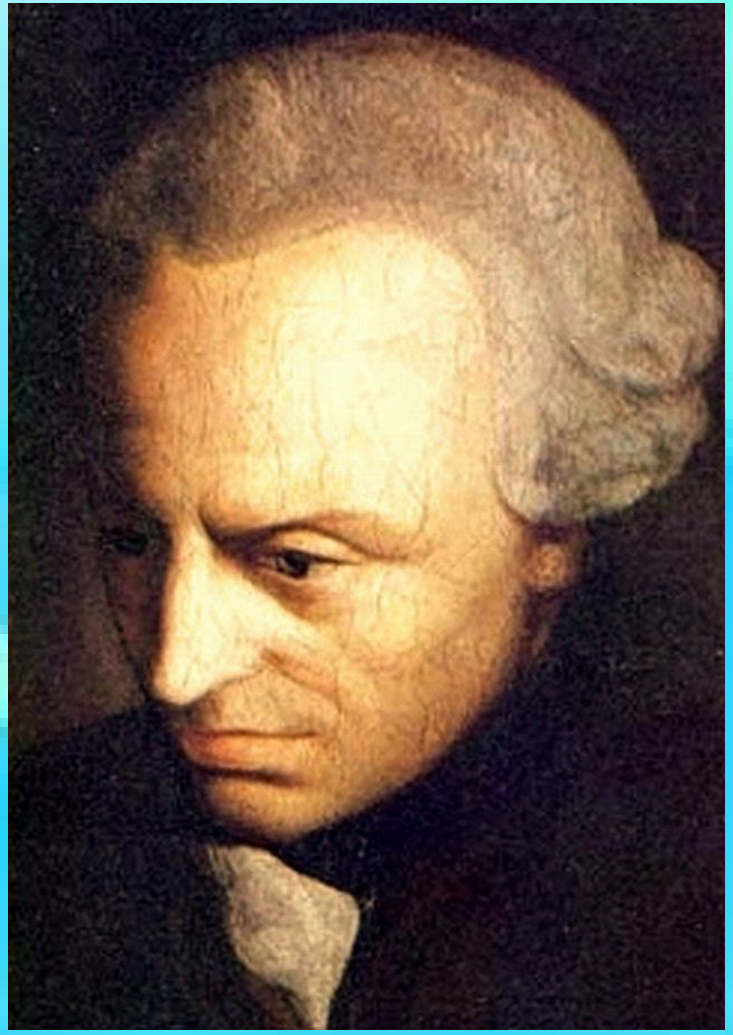
Математические методы в науке



Николай Кузанский

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

Николай Кузанский (1401-1464) – величайший из немецких гуманистов первого поколения, богослов, философ, математик и церковно-общественный деятель, родом из лотарингской деревни Кус (Cues), откуда и его прозвание.



Иммануи́л Кант

Философ И. Кант был убежден, что точное естествознание простирается до тех границ, в пределах которых возможно применение математического метода.

Иммануил Кант (1724-1804) – немецкий философ, родоначальник немецкой классической философии, стоящий на грани эпох Просвещения и Романтизма.

Кант отвергал догматический способ познания и считал, что вместо него нужно взять за основу метод критического философствования, сущность которого заключается в исследовании самого разума, границ, которые может достичь разумом человек, и изучении отдельных способов человеческого познания.

Связь языкознания и математики

Н. Винер

Если науки естественного цикла сравнительно давно заговорили на языке математики, то гуманитарные науки обратились к нему только в XX веке. Первой среди них была лингвистика, занимающая особое, центральное положение среди всех областей человеческого познания. Системность языка, обобщенный характер его единиц — вот та благодатная почва, в которой стали плодотворно укореняться идеи и методы современной математики.

В лингвистике есть все условия, необходимые, с точки зрения известного кибернетика Н. Винера, для математического исследования. Во-первых, здесь влияние наблюдателя ничтожно мало, осознания явления наблюдателем недостаточно для того, чтобы его изменить. Во-вторых, язык обладает длинными статистическими рядами.

Норберт Винер (1894-1964) — американский математик. В своем фундаментальном труде «Кибернетика» (1948 год) сформулировал основные ее положения. Винер — автор трудов по математическому анализу, теории вероятностей, электрическим сетям и вычислительной технике. Его детище, кибернетика — наука об управлении и связях в машинах и живых организмах, родилось из сплава прежде не пересекавшихся математики, биологии, социологии и экономики.

«Высшее назначение математики состоит в том, чтобы находить скрытый порядок в хаосе, который нас окружает»

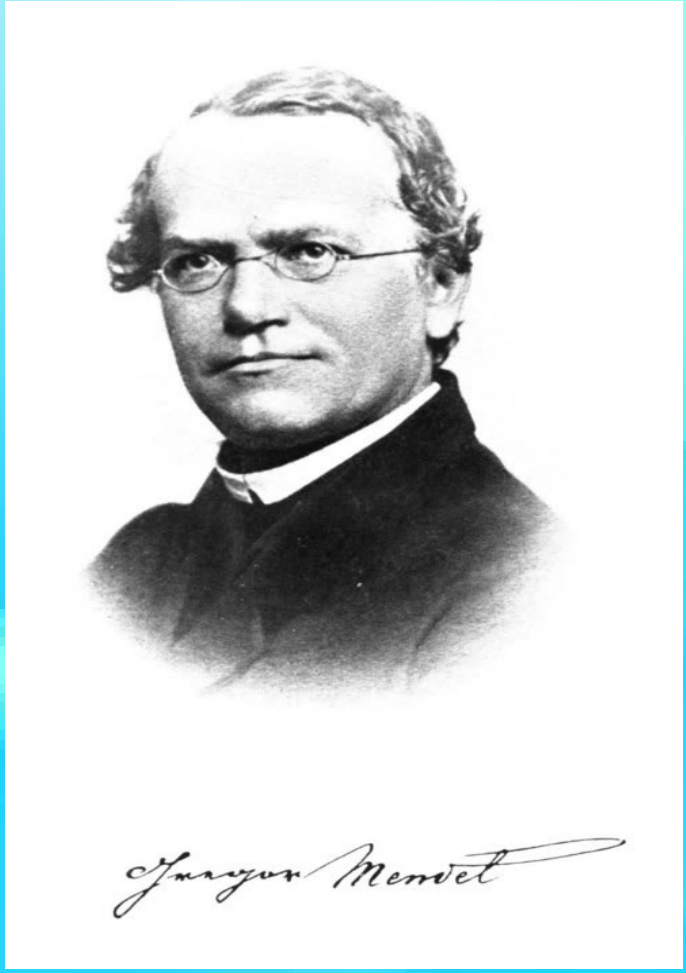
Н. Винер

В. Я. Буняковский

Языкознание первым из гуманитарных наук от установки на полное и исчерпывающее описание отдельных фактов перешло к установке на обобщение, на поиски единого закона, объясняющего необозримое множество отдельных фактов. Эта познавательная установка и определила интерес к математическим методам.

Известный русский математик В.Я. Буняковский писал о необходимости применения математики в области грамматических и этимологических разысканий. Связь языкознания с математикой наметилась уже давно.

Виктор Яковлевич Буняко́вский (1804-1889) – российский математик, вице-президент академии наук в 1864—1889 годах.



Г. И. Мендель

Грегор Мендель (Грегор Иоганн Мендель) (1822-84) — австрийский естествоиспытатель, ученый-ботаник и религиозный деятель, монах, основоположник учения о наследственности (менделизм).

Применив статистические методы для анализа результатов по гибридизации сортов гороха, сформулировал закономерности наследственности.

В 1868 Грегор Мендель был избран настоятелем монастыря и практически отошел от научных занятий. В его архиве сохранились заметки по метеорологии, пчеловодству, **лингвистике**. На месте монастыря в Брно ныне создан музей Менделя.

Г. И. Мендель пытался применять статистические методики не только в области биологии, но и в языкознании.

И.А. Бодуэн де Куртенэ

И.А. Бодуэн де Куртенэ, набрасывая контуры будущего языкознания, непременным условием его считал тесную и органическую связь с математикой: «Нужно чаще применять в языкознании количественное, математическое мышление и таким образом приблизить его все более к наукам точным».

Перспективные мысли высказаны Бодуэном в статье «Количественность в языковом мышлении». Выдающийся лингвист практически использовал квантитативную методику в исследованиях по фонетике (исчисление альтернаций) и по грамматике (описание типов склонения).

(СМ. ПРИЛОЖЕНИЕ)

Бодуэн де Куртенэ (Baudouin de Courtenav) Иван Александрович(1845-1929) – русско-польский языковед, член-корреспондент Петербургской АН (1897). Один из виднейших представителей общего и славянского историко-сравнительного языкознания, родоначальник т. н. казанской, позже петербургской лингвистических школ.

http://danefae.org/lib/baudouin/selecta2.htm

Е. Д. Поливанов

Крупнейший теоретик языка Е.Д. Поливанов, говоря о точках соприкосновения между математикой и лингвистикой, особо выделял следующие сферы:

а) анализ кимографических кривых;

б) диалектологическая статистика;

в) приложение теории вероятностей к определению относительной вероятности этимологий — как достоверных, так и гипотетических и, наконец, фантастических.

Евгений Дмитриевич Поливанов (1891–1938), ученик И.А. Бодуэна де Куртенэ, принадлежит к блестящей плеяде русских советских лингвистов. Он сформировался как ученый в первой четверти XX в., сочетал в себе яркий талант исследователя, необыкновенные способности полиглота и глубокие познания в области теории языка. Ему принадлежат работы, не потерявшие своей актуальности и в наши дни, по теории языка и его эволюции, сравнительной индоевропеистике, фонетике, фонологии, акцентологии, графике и орфографии, грамматике и фонетики японского, китайского, дунганского, грузинского, узбекского, турецкого и других языков.

Связь языкознания с математикой не была односторонней

А. А. Марков

Используя методы математики, лингвистика в свою очередь питала математику плодотворными идеями.

Наблюдения известного математика А.А. Маркова (1856—1922) над текстом «Евгения Онегина» (распределение доли гласных и согласных среди первых 20 тыс. букв –«испытания, связанные в цепь») привели к открытию знаменитых «марковских цепей»

Андре́й Андре́евич Ма́рков (1856-1922) — русский математик, академик, внёсший большой вклад в теорию вероятностей, математический анализ и теорию чисел.

Для описания и исследования лингвистических фактов привлекаются различные разделы математики:

- алгебра;
- теория множеств;
- математическая логика;
- теория информации;
- теория вероятностей;
- математическая статистика.

Среди математических наук рождается новая дисциплина:

Математическая лингвистика – математическая дисциплина, предметом которой является разработка и изучение понятий, образующих основу формального аппарата для описания строения естественных языков (т. е. метаязыка лингвистики). *(Математическая энциклопедия)*

http://enc-dic.com/enc_math/Matematicheskaja-lingvistika-2132/

Квантитативная лингвистика в системе математической лингвистики

Математическая лингвистика развивается в нескольких направлениях:

1) **алгебраическая лингвистика**;

2) **комбинаторная лингвистика**, которая опирается на разделы «неколичественной» математики (теория множеств, математическая логика, теория алгоритмов);

3) **квантитативная лингвистика**, которая изучает лингвистические явления с помощью «количественной» математики (математическая статистика, теория вероятностей, теория информации и др.).

Понятие *Алгебраическая лингвистика*

В современной науке это направление математической лингвистики ещё не достаточно изучено, поэтому в словарях нет точного определения этого понятия.

В словарной статье Языкознание (БСЭ) мы можем прочитать следующее:

Новое направление — **структурная лингвистика** – возникло в начале 20 в. с появлением «Курса общей лингвистики» Ф. де Соссюра. Это направление начало складываться в общем русле структурализма, развивавшегося аналогично и почти одновременно в разных областях — в общем изучении систем, в психологии (Гештальтпсихология), в теории литературы и искусства («Формальный метод» в литературоведении) и др.

Основные принципы нового направления:

1) подлинной и основной реальностью является не отдельный факт какого-либо языка, а язык как система; каждый элемент языка существует лишь в силу его отношений к другим элементам в составе системы; система не суммируется из элементов, а, напротив, определяет их;

2) костяк, структуру системы создают вневременные отношения; отношения в рамках системы доминируют над элементами;

3) поэтому возможно вневременное **«алгебраическое» изучение системы языка, основанного на отношениях, а не на индивидуальности элементов или их материальности; возможно применение строгих, математических методов в языкознании;**

4) язык есть система особого рода — знаковая система, существующая, с одной стороны, объективно, вне психики человека, в межличностном общении людей, с другой стороны — эта система существует и в психике людей;

5) подобно языку организованы некоторые другие системы, действующие в человеческих обществах, — фольклор, обычаи и ритуалы, отношения родства и др.; все они могут изучаться, подобно языку, лингвистически, в частности формализоваться «алгебраически» или иными способами (Семиотика).

http://enc-dic.com/enc_sovet/Jazkoznanie-104478.html

Проанализировав данные словарей, мы пришли к выводу, что **алгебраическая лингвистика связана с такими понятиями: язык как система, семиотика, структурная лингвистика.**

Понятие *Комбинаторная лингвистика*

Комбинаторная лингвистика — применяет для изучения лингвистических явлений "неколичественный" или "качественный" математический аппарат (теория множеств, математическая логика, теория алгоритмов, теория порождающих грамматик и др.).

http://enc-dic.com/word/k/Kombinatornaja-lingvistika-57139.html

Квантитативная (или количественная) лингвистика противопоставляется комбинаторной лингвистике.

Комбинаторная лингвистика – направление в языкознании, которое занимается изучением синтагматических связей языковых единиц и их комбинаторных свойств. **Комбинаторная лингвистика представляет собой синтез двух областей**:

1) **синтагматика**, являющая собой аспект языка, содержащий языковые правила сочетаемости одноуровневых единиц;

2) **комбинаторика**, содержанием которой является составление и изучение комбинаций слов, подчиненных определенным коммуникативным задачам при данных условиях их реализации, и которые можно образовать из заданного количества слов.

http://journals.tsu.ru/uploads/import/855/files/342-007.pdf

Понятие ***Квантитативная лингвистика*** в словарях и энциклопедиях

Квантитативное от лат. quantitas — количество *(Словарь-справочник лингвистических терминов)* *http://enc-dic.com/linguistics/Kvantitativnoe-1390.html*

Квантитативная лингвистика *(Толковый переводоведческий словарь)* :

1. Изучает и эксплицирует лингвистические явления с помощью методов "количественной" математики (теория вероятностей, математическая статистика, теория информации, математический анализ и др.). Это направление противопоставляется комбинаторной лингвистике.

2. См. статистическая лингвистика.

3. В целом может рассматриваться как:

- техника лингвистического наблюдения и описания, обработки данных наблюдения;
- метод исследования языка и речи, не обязательно противополагаясь сопоставленному, сравнительно-историческому и другим методам языкознания;
- концепция, как система количественных идей и представлений об объекте лингвистической науки.
  http://enc-dic.com/word/k/Kvantitativnaja-lingvistika-57121.html

Квантитативная лингвистика в системе прикладной лингвистики

Прикладная лингвистика

Лингвистику принято подразделять на *теоретическую*, иногда её ещё называют «научная лингвистика», или «теория языкознания» (в рамках этого направления рассматриваются различные научные концепции, лингвистические теории, лингвистические школы, язык с точки зрения его структуры и систем, **и *прикладную*** (практическую) лингвистику.

Прикладна́я лингви́стика (прикладное языкознание) — наряду с теоретической лингвистикой является частью науки, занимающейся языком. Специализируется на решении практических задач, связанных с изучением языка, а также на практическом использовании лингвистической теории в других областях.

Основные направления прикладной лингвистики

- Компьютерная лингвистика и её инструментарий;
- Гипертекстовые технологии представления текста;
- Корпусная лингвистика;
- Машинный перевод;
- Оптимизация общения с ЭВМ: системы обработки естественного языка
- Теория и практика информационно-поисковых систем;
- Лексикография как дисциплина прикладной лингвистики;
- Теория и методика преподавания языка;
- Политическая лингвистика;
- Квантитативная лингвистика;
- Терминоведение — наука об упорядочении и стандартизации научно-технической терминологии;
- Лингвистическая экспертиза.

Практическое применение квантитативной лингвистики

КВАНТИТАТИВНАЯ ЛИНГВИСТИКА: ИССЛЕДОВАНИЯ И МОДЕЛИ

Выделяя основные направления квантитативной лингвистики, мы ориентировались на монографию Арапова М.В. «Квантитативная лингвистика».

- Значение количественных данных для изучения языка.
- Частота как характеристика употребительности слова в тексте.
- Изменчивость употребительности слова в синхронии.
- Историческая изменчивость употребительности слова (употребительность и возраст слова).
- Длина слова и его употребительность.
- Полисемия слова и его употребительность.
- Продуктивность классов слов.
- Однородность и регулярность отношений между единицами словаря.

Содержание раздела Квантитативная лингвистика курса «Прикладная и математическая лингвистика» в МГУ:

1. Применение различных статистических методов анализа данных в лингвистике. Описательная статистика в лингвистических исследованиях.
2. Основные статистические критерии проверки зависимости / независимости признаков и однородности выборок, применяемые в лингвистических исследованиях.
3. Закон Ципфа — Мандельброта и его следствия.
4. Квантитативные методы автоматического выделения ключевых слов и терминов.
5. Квантитативные методы, применяемые в лексикографии.
6. Квантитативные методы, применяемые в корпусной лингвистике.
7. Задачи атрибуции текстов и стилеметрия.

Частотные словари

Практическим результатом статистического изучения лексики являются **частотные словари**, отличающиеся от обычных лингвистических (толковых, орфографических и др.) тем, что словарные единицы располагаются **не только в алфавитном порядке**, но и **в порядке убывающей частотности**. **В первом случае это будет алфавитный частотный словарь, а во втором — ранговый частотный словарь**.

Частотные словари характеризуются следующими параметрами:

- **объем текста** (число словоупотреблений);
- **объем словаря словоформ**;
- **объем словаря лексем**.

Первым частотным словарем был словарь Кединга (1898).

Затем в течение девяноста лет было составлено несколько сот частотных словарей и частотных списков для нескольких десятков языков.

Первым частотным словарем русского языка был словарь Г. Йоссельсона (США, Детройт, 1953).

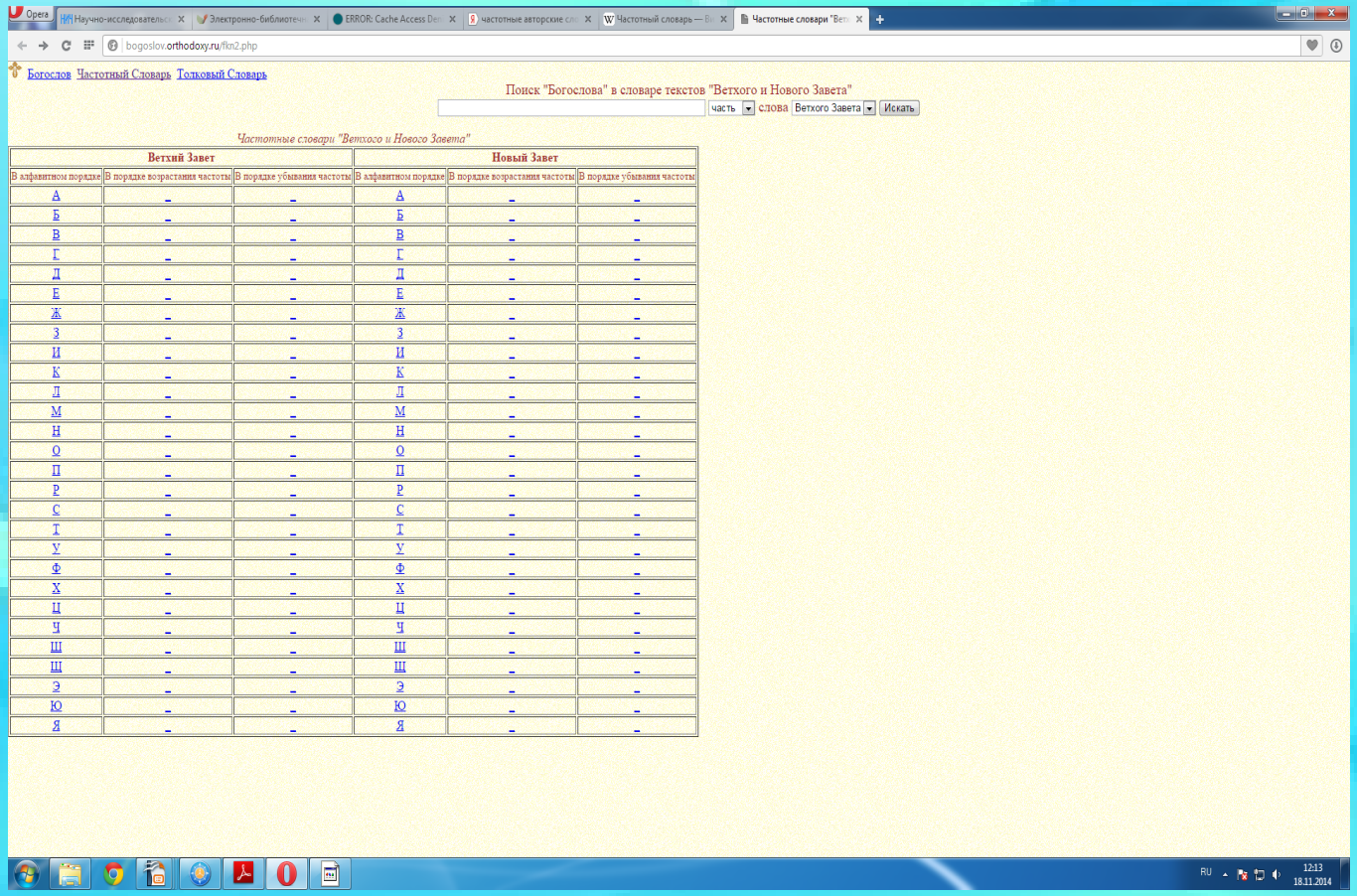
В нашей стране первый частотный словарь русского языка был составлен Э. Штейнфельд (1963).

Интересны материалы к частотному словарю языка Пушкина (1963).

В 1977 г. вышел в свет «Частотный словарь русского языка» под редакцией Л.Н. Засориной. Создавался он на основе выборки в один миллион словоупотребленийиз четырех жанров (художественная проза, драматургия, научная публицистика, газетно журнальные материалы). В нем около 40 тыс. слов. Самое частотное слово — предлог в (во), далее идут служебные слова и местоимения (и, не, на, я, быть, что, он, с, а, как, это). Самое частотное существительное — год.

В 90 х годах XX в. в Швеции вышел в свет «Частотный словарь современного русского языка» (Уппсала, 1993).

Частотный словарь «Ветхого и Нового Завета»



Богослов Частотный Словарь Толковый Словарь

Поиск "Богослова" в словаре текстов "Ветхого и Нового Завета"

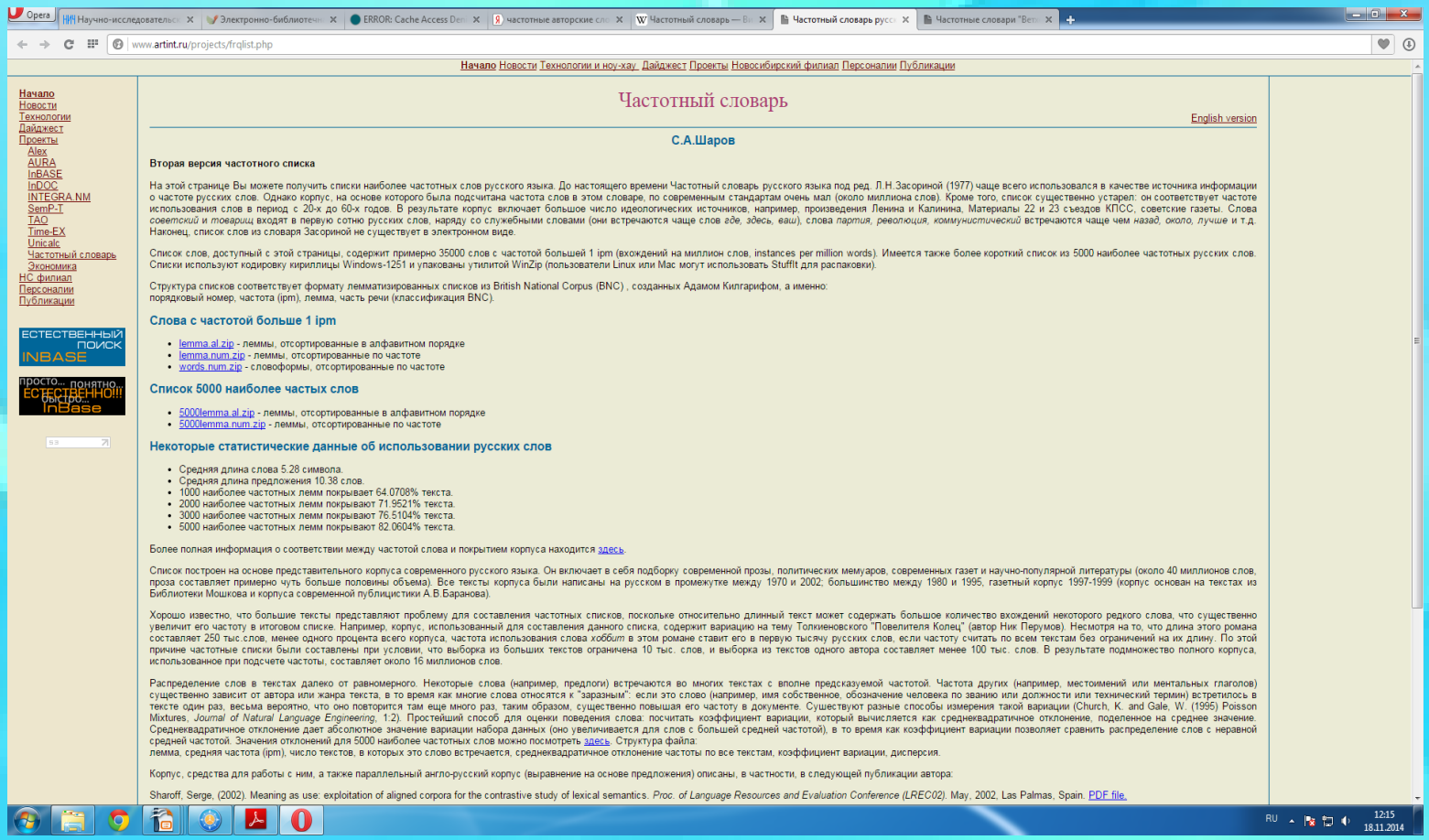
часть слова Ветхого Завета Искать

Частотные словари "Ветхого и Нового Завета"

Ветхий Завет			Новый Завет		
В алфавитном порядке	В порядке возрастания частоты	В порядке убывания частоты	В алфавитном порядке	В порядке возрастания частоты	В порядке убывания частоты
А	_	_	А	_	_
Б	_	_	Б	_	_
В	_	_	В	_	_
Г	_	_	Г	_	_
Д	_	_	Д	_	_
Е	_	_	Е	_	_
Ж	_	_	Ж	_	_
З	_	_	З	_	_
И	_	_	И	_	_
К	_	_	К	_	_
Л	_	_	Л	_	_
М	_	_	М	_	_
Н	_	_	Н	_	_
О	_	_	О	_	_
П	_	_	П	_	_
Р	_	_	Р	_	_
С	_	_	С	_	_
Т	_	_	Т	_	_
У	_	_	У	_	_
Ф	_	_	Ф	_	_
Х	_	_	Х	_	_
Ц	_	_	Ц	_	_
Ч	_	_	Ч	_	_
Ш	_	_	Ш	_	_
Щ	_	_	Щ	_	_
Э	_	_	Э	_	_
Ю	_	_	Ю	_	_
Я	_	_	Я	_	_

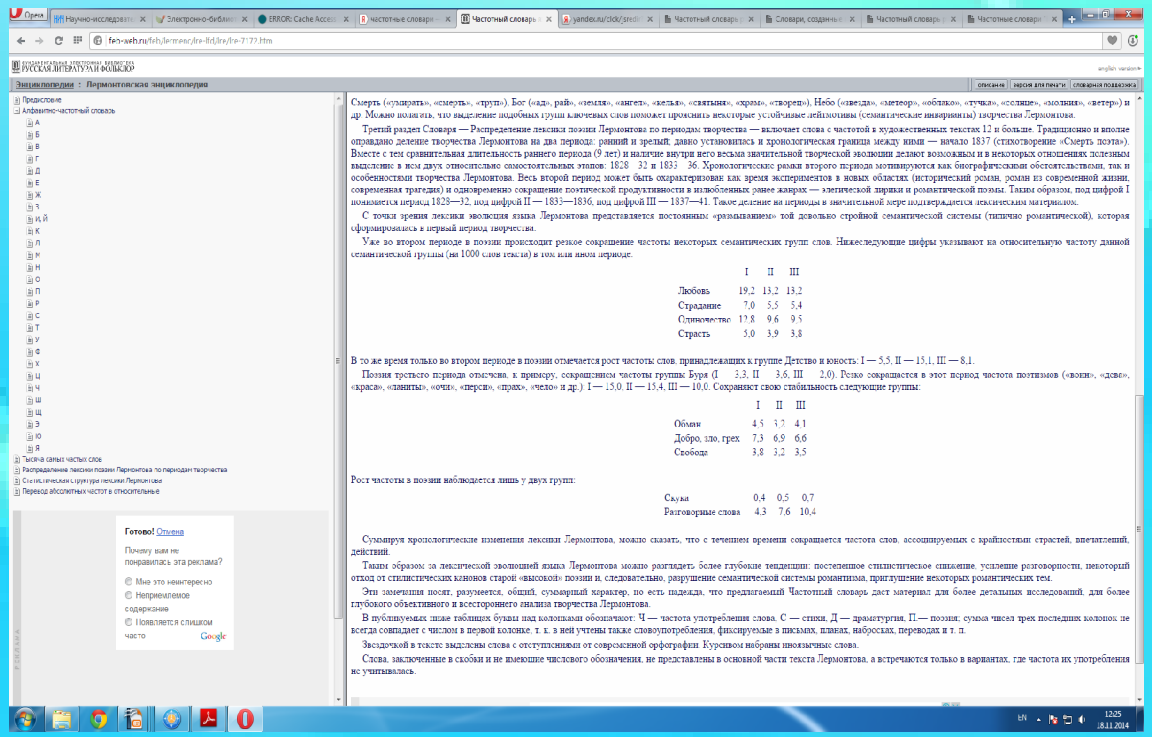
http://bogoslov.orthodoxy.ru/fkn2.php

Частотный словарь С. А. Шарова



http://www.artint.ru/projects/frqlist.php

Авторский идеостиль



Частотный словарь языка М. Ю. Лермонтова

Сопоставление частотного словаря языка писателя с частотным словарем национального языка — один из путей выявления особенностей языка и стиля писателя. С помощью формально количественных методов изучается авторский идиостиль, под которым понимают взаимосвязь между языковыми средствами и особенностями творческой позиции писателя, его взгляда на мир, на окружающую действительность.

http://feb-web.ru/feb/lermenc/lre-lfd/lre/lre-7172.htm

Определение авторства

Определение авторства с помощью формально количественных и статистических методов стимулировало поиск и выявление характерных структур авторского языка. На этом строятся многообразные методики, представленные в книге «От Нестора до Фонвизина. Новые методы определения авторства» (М., 1994).

Специалисты исследовали несколько простых параметров авторского стиля и на базе большого количества произведений писателей XVIII—XX вв. статистически доказали, что доля всех служебных слов в данном прозаическом произведении является авторским инвариантом. Один из исследователей, опираясь на модель цепей А.А. Маркова, предложил методику определения авторства, основанную на том, что по произведениям автора, которые достоверно им созданы, вычисляется матрица переходных частот употреблений пар букв. Затем такие матрицы строятся для каждого из авторов, «подозреваемых» в написании анонимного текста, и для каждого автора оценивается вероятность того, что именно он написал анонимный фрагмент текста. В результате автором анонимного текста полагается тот, у которого вычисленная оценка вероятности больше.

Знаменитый шедевр древнерусской словесности XII в. «Слово о полку Игореве», уникальность которого вот уже около двух столетий ставится скептиками под сомнение, был подвергнут жесткой формально количественной ревизии. Применение анализа частот парной встречаемости грамматических классов слов позволило убедительно доказать, что глубинная структура «Слова» — структура языка XI столетия. Этот формально количественный анализ не отверг гипотезы историка Б.А. Рыбакова о боярине Петре Бориславиче как авторе «Слова о полку Игореве». Возможно, отчасти она и подтверждена. Однако, полагают исследователи, необходимо еще более детальное исследование текстов, которое будет проведено в ближайшее время.

Квантитативные методы, применяемые в корпусной лингвистике

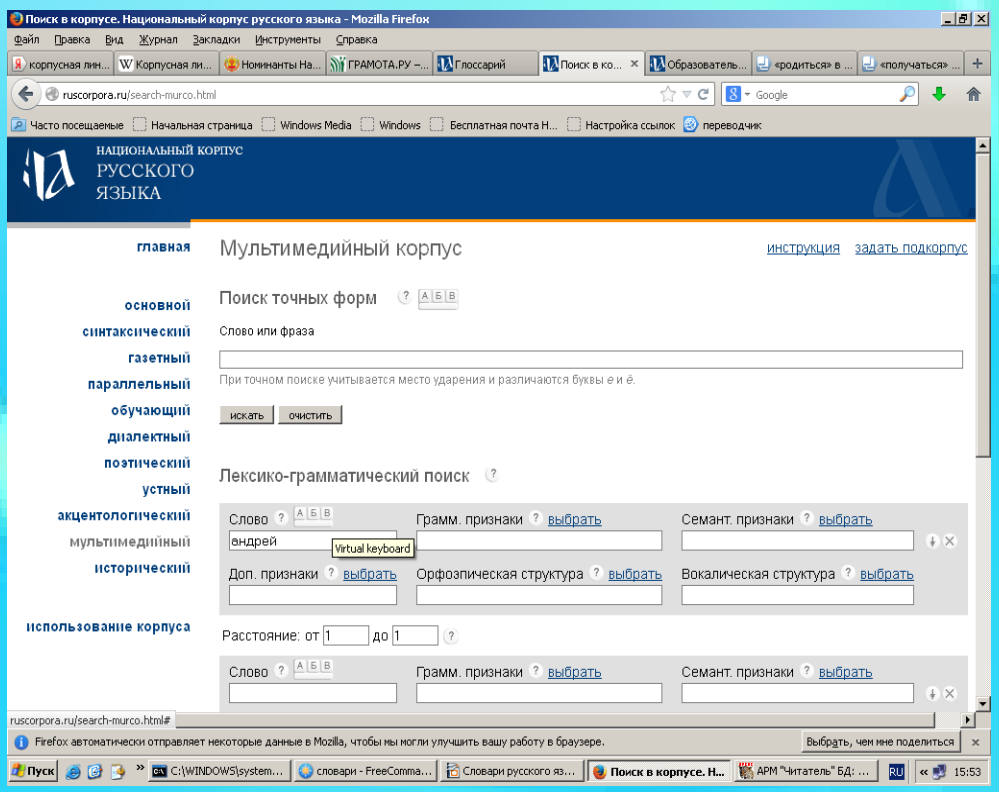
- Статистические методы оценки продуктивности аффиксов;
- Корпусные статистические методы анализа лексики. Статистические методы выделения терминов, устойчивых словосочетаний, синонимических групп, семантических полей.
- Статистические методы машинного перевода.
- Методы самообучения в применении к частеречной разметке корпуса (автоматический тэггинг). Применение методов скрытых марковских моделей при частеречной разметке корпуса текстов.
- Статистические методы синтаксической разметки корпуса. Стохастические грамматики

КОРПУСНАЯ ЛИНГВИСТИКА

Корпусная лингвистика — раздел языкознания, занимающийся разработкой, созданием и использованием текстовых (лингвистических) корпусов. Термин введён в употребление в 60-х годах XX века в связи с развитием практики создания корпусов, которому начиная с 80-х способствовало развитие вычислительной техники.

Лингвистическим корпусом называют совокупность текстов, собранных в соответствии с определёнными принципами, размеченных по определённому стандарту и обеспеченных специализированной поисковой системой. Иногда корпусом («корпус первого порядка») называют просто любое собрание текстов, объединённых каким-то общим признаком (языком, жанром, автором, периодом создания текстов).

Национальный корпус русского языка



http://ruscorpora.ru/search-murco.html

На этом сайте помещен корпус современного русского языка общим объемом более 500 млн слов. Корпус русского языка — это информационно-справочная система, основанная на собрании русских текстов в электронной форме.

Национальный корпус представляет язык во всём многообразии жанров, стилей, территориальных и социальных вариантов. Национальный корпус создается лингвистами для научных исследований и для обучения языку, он охватывает период с середины 18 века до сегодняшнего дня.

http://ruscorpora.ru

Национальный корпус создается лингвистами (специалистами по так называемой *корпусной лингвистике*, быстро развивающейся современной области языкознания) для научных исследований и обучения языку. Большинство крупных языков мира уже имеет свои национальные корпуса (различающиеся по полноте и уровню научной обработки текстов). Общепризнанным образцом является, в частности, **Британский национальный корпус (BNC http://www.natcorp.ox.ac.uk/)**: на него ориентированы многие другие современные корпуса. Среди корпусов славянских языков выделяется **Чешский национальный корпус (http://ucnk.ff.cuni.cz/)**, созданный в Карловом университете Праги.

Корпус позволяет работать с современной русской речью, представленной в самых разных ситуациях общения, в разных жанрах и социальных вариантах

- Корпус позволяет отобрать тексты того типа и того периода, с которым Вы хотите работать
- С помощью корпуса можно получить примеры определенного грамматического явления, определенной синтаксической конструкции или семантической категории
- Корпус дает возможность получить статистические данные об интересующих Вас явлениях.

Используя материалы Портала в ваших публикациях, не забывайте ссылаться на автора и адрес портала!

Структура национального корпуса русского языка:

Национальный корпус русского языка состоит из следующих корпусов:

- **основной**
- **синтаксический**
- **газетный**
- **параллельный**
- **обучающий**
- **диалектный**
- **поэтический**
- **устный**
- **акцентологический**
- **мультимедийный**
- **исторический**