**Применение методов математической статистики и теории вероятностей в задачах теоретической лингвистики при анализе устной и звучащей речи на русском и английском языках**

2007 г.

**Оглавление**

1. Введение

2. Анализ стихотворений

2.1 Построение дискретного вариационного ряда

2.2 Непрерывные вариационные ряды

2.3 Графическое построение дискретных лингвистических вариационных рядов для рассматриваемых стихотворений

2.4 Ряды распределения дискретных случайных величин

2.5 Математическое ожидание дискретной случайной величины

2.6 Дисперсия дискретной случайной величины

2.7 Энтропия дискретной случайной величины

2.8 Вероятность появления гласных звуков в стихотворениях, сравнение

2.9 Коэффициент темпа речи

3. Объединённый коэффициент синтаксической и ритмомелодической сложности

4. Вывод

5. Список литературы

1. Введение

В эпоху научно-технической революции математизация охватывает все сферы человеческой деятельности, в том числе и языкознание. Проникновение математических методов в лингвистику обусловлено двумя причинами. Во-первых, развитие языковедческой теории и практики требует введения все более точных и объективных методов для анализа языка и текста. Одновременно использование математических приемов при систематизации, измерении и обобщении лингвистического материала в сочетании с качественной интерпретацией результатов позволяет языковедам глубже проникнуть в тайны построения языка и образования текста. Во-вторых, все расширяющиеся контакты языкознания с другими науками, например с акустикой, физиологией высшей нервной деятельности, кибернетикой и вычислительной техникой, могут осуществляться только при использовании математического языка, обладающего высокой степенью общности и универсальности для различных отраслей знаний.

Особенно настойчиво математизируется языкознание в связи с использованием естественного языка в информационных и управленческих системах человек–машина–человек. В действующих системах машинного перевода, автоматического аннотирования, человеко-машинного диалога всякое сообщение на естественном языке перекодируется в математическом языке компьютера. Примером того является голосовое управление в современных мобильных телефонах.

Говоря об особенностях взаимодействия языкознания и математики, следует иметь в виду, что как естественный язык, так и язык математики являются знаковыми (семиотическими) системами передачи информации.

Основные расхождения между этими языками связаны с различным построением языкового знака и знака математического.

лингвистический знак (слово, словосочетание, предложение) обычно включает в себя четыре компонента – имя (материальный носитель информации), денотат (отражение предмета из внешнего мира), десигнат (понятие о предмете) и коннотат (комплекс чувственно-оценочных оттенков, связанных с предметом и понятием о нем); знак математического языка включает только имя и десигнат - математическое понятие;

лингвистический знак многозначен - значения его представляют собой нечеткие множества с размытыми границами; математический знак имеет, как правило, одно концептуальное значение;

лингвистический знак потенциально метафоричен, у знака математического метафоричность полностью отсутствует.

Особенности построения лингвистического языка приводят к тому, что естественный язык представляет собой нежестко организованную диффузную систему, которая воспринимается и используется человеком в значительной мере интуитивно. Напротив, язык математики является хорошо организованной системой, существующей и функционирующей в виде логического построения, каждый элемент которого имеет осознанную значимость.

Конфронтация естественного языка и языка математики требует, чтобы каждому лингвистическому объекту был поставлен в соответствие некоторый математический объект. Лингвистический знак, например, словосочетание или слово и составляющие этот знак фигуры – фонемы, буквы, слоги – должны интерпретироваться с помощью знаков математических. Эта математическая интерпретация связана с расчленением лингвистического объекта и выделением в нем одного смыслового или сигнального компонента, который становится предметом дальнейшего исследования.

Применение математических методов в языкознании имеет своей целью заменить обычно диффузную, интуитивно сформулированную и не имеющую полного решения лингвистическую задачу одной или несколькими более простыми, логически сформулированными и имеющими алгоритмическое решение математическими задачами. Такое расчленение сложной лингвистической проблемы на более простые алгоритмизуемые задачи мы будем называть математической экспликацией лингвистического объекта или явления.

Выбор математического аппарата в лингвистических исследованиях – вопрос не простой. Его решение зависит в первую очередь от того, как определяется предмет и основные понятия языкознания и его теоретического ядра – структурно-математической лингвистики.

Некоторые математики и лингвисты считают, что предметом математической лингвистики должно быть изучение грамматики, порождающей текст. При этом грамматика понимается как конечное множество детерминированных правил, в том числе неграмматических, а язык рассматривается как бесконечное число регулярных цепочек слов, порождаемых этой грамматикой. При этом подходе экспликация лингвистических объектов должна опираться на теорию множеств, математическую логику, теорию алгоритмов.

На основе применения «неколичественного» математического аппарата в теоретическом языкознании сформировалось направление, условно называемое комбинаторной лингвистикой – в ней используются методы

математической статистики

теории вероятностей,

теории информации,

математического анализа

Современные инструментальные методы экспериментальной фонетики связаны с применением различных приборов, главным образом электроакустических (спектрографы, интонографы и тому подобное), а также регистрирующие движения органов речи (артикуляцию). Поэтому фонетика тесно связана с физикой, физиологией и математикой. Методы математической логики применяются для формального описания категорий естественных языков. Языкознание оказалось той гуманитарной наукой, которая, не порывая связей с другими науками о человеке и его культуре, первой решительно стала использовать не только инструментальные методы наблюдения (в фонетике) и экспериментальные приёмы (в психолингвистике), но и систематически применять математические способы (в том числе и ЭВМ) для получения и записи своих выводов.

Цель моего реферата – выявить и изучить статистические закономерности стиля двух равных текстов (по 105 слов в каждом) поэтессы Зинаиды Гиппиус (1869 - 1945) «Свободный стих» и английского поэта Вильяма Блейка «Колыбельная» (William Blake, 1757-1827, «A Cradle Song») согласно звуковым характеристикам языка – ударению, сложности восприятия, темпу речи и другим. При анализе я использую следующие термины:

вероятность события,

вариационные ряды,

математическое ожидание,

закон распределения вероятности,

дисперсия,

энтропия.

Также я приведу примеры использования методов математической статистики и теории вероятностей при анализе устной и звучащей речи.

**2. Анализ стихотворений**

**2.1 Построение дискретного вариационного ряда**

«Свободный стих»

Приманной легкостью играя,

Зовет, влечет свободный стих.

И соблазнил он, соблазняя,

Ленивых малых и простых.

Сулит он быстрые ответы

И достиженья без борьбы.

За мной! За мной! И вот, поэты –

Стиха свободного рабы.

Они следят его извивы,

Сухую ломкость, скрип углов,

Узор пятнисто-похотливый

Икающих и пьяных слов...

Немало слов с подолом грязным

Войти боялись... А теперь

Каким ручьем однообразным

Втекают в сломанную дверь!

Втекли, вшумели и впылились...

Гогочет уличная рать.

Что ж! Вы недаром покорились:

Рабы не смеют выбирать.

Без утра пробил час вечерний,

И гаснет серая заря...

Вы отданы на посмех черни

Коварной волею царя!

А мне лукавый стих угоден.

Мы с ним веселые друзья.

Вариационные ряды длин словоупотребления в фонемах:

7 9 6

5 6 9 4

1 9 2 10

7 5 1 7

5 2 7 6

1 10 3 5

2 4 2 4 1 3 5

5 10 4

3 6 3 6

6 7 5 5

4 18

8 1 6 4

6 4 1 7

5 7 1 5

5 6 12

8 1 9 4

6 7 1 8

7 8 3

3 1 2 7 9

4 2 6 7

3 4 6 3 8

1 6 6 4

2 5 2 6 5

8 6 4

1 3 7 4 6

2 1 3 7 6

(порядок следования чисел здесь повторяет порядок следования слов в стихотворении построчно)

Рассматривая приведенную здесь последовательность чисел нетрудно заметить, что величина длины словоформ варьирует от одной единицы совокупности к другой. Моя задача – определить и изучить вариацию признака в данной совокупности.

Возможные значения признака в статистике называют вариантами. Различия между вариантами могут быть как количественными (дискретными или непрерывными) и качественными.

Теперь я построю дискретный вариационный ряд длины словоформ в фонемах в данном стихотворении:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| N | 13 | 10 | 8 | 13 | 14 | 18 | 11 | 8 | 5 | 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

Где X – признак, N – сумма всех вариант, – варианты, – число повторений вариант

N = 105 (так как текст состоит из 105 слов, как было указано выше)

Теперь вместо абсолютных частот укажу относительные частоты (частости) в процентах:



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X |  |  |  |  |  |  |  |  |  |
| f\*100% | 12,38% | 9,52% | 7,62% | 12,38% | 13,33% | 17,14% | 10,47% | 7,62% | 4,76% |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X |  |  |  |  |  |  |  |  |  |
| f\*100% | 2,85% | 0% | 0,95% | 0% | 0% | 0% | 0% | 0% | 0,95% |

Самые распространённые слова в данном стихотворении имеют длину в 6 фонем (17,14%)

Проведу аналогичные действия со стихотворением «Колыбельная»:

A Cradle Song

Sweet dreams form a shade,

O'er my lovely infants head.

Sweet dreams of pleasant streams,

By happy silent moony beams

Sweet sleep with soft down,

Weave thy brows an infant crown.

Sweet sleep Angel mild,

Hover o'er happy child.

Sweet smiles in the night,

Hover over my delight.

Sweet smiles Mothers smiles

All the livelong night beguiles.

Sweet moans, dovelike sighs,

Chase not slumber from thy eyes,

Sweet moans, sweeter smiles,

All the dovelike moans beguiles.

Sleep sleep happy child.

All creation slept and smil'd.

Sleep sleep, happy sleep, 1

While o'er thee thy mother weep

Sweet babe in thy face,

Holy image I can trace.

Вариационные ряды длин словоупотребления в фонемах:

4 5 4 1 5

4 3 5 7 3

4 5 4 7 6

3 4 7 4 4

4 4 4 4

3 3 5 2 6 5

4 4 6 5

6 2 3 4 5

4 6 2 2 4

6 4 3 6

4 6 5 6

2 2 7 4 7

4 5 7 4

4 3 6 4 2 3

4 5 4 6

2 2 7 5 7

4 4 4 4

2 7 5 3 6

4 4 4 4

4 4 2 2 4 3

4 4 2 2

4 3 4 2 3 5

Дискретный вариационный ряд длины словоформ в фонемах в данном стихотворении будет таков:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| N | 1 | 15 | 13 | 41 | 14 | 12 | 9 |

Где так же, как и в предыдущем примере, X – признак (количество фонем в слове), N – сумма всех вариант, – варианты, – число повторений вариант.

N = 105

Очевидным является то, что дискретные вариантные ряды двух стихотворений сильно отличаются друг от друга, это можно представить нагляднее, если вместо абсолютных частот указать относительные частоты в процентах:



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X |  |  |  |  |  |  |  |
| f\*100% | 0,95% | 14,28% | 12,38% | 39,04% | 13,33% | 11,42% | 8,57% |

Различие между длинами словоформ в рассматриваемых стихотворениях состоит в том, что у английского автора преобладают слова в четыре фонемы (39,04%), в то время как у Зинаиды Гиппиус – в шесть. Так же несложно заметить, что количество вариантов в стихотворении «The Cradle Song» значительно меньше, чем в «Свободный стих».

**2.2 Непрерывные вариационные ряды**

Непрерывные вариационные ряды, как и дискретные, широко распространены в анализе устной и звучащей речи, так как здесь значения признака:

длина

частота

интенсивность звука

могут отличаться друг от друга на как угодно малую величину. Поскольку отличия между вариантами имеют непрерывный характер, используется только интервальное построение вариационного ряда. Для исследования данных фонетических аспектов нужны специальные измерительные приборы для замеров звучания слогов. Несмотря на невозможность проведения данного анализа, я расскажу о его основном принципе.

При наличии результатов эмпирических исследований, создаются непрерывные интервальные ряды, где - длина слогов в мс, а интервалы вариант выглядят следующим образом – (, (), () и так далее.

Ширина интервала определяется по формуле Стерджесса:

.

При этом интервальная разность *k* округляется до ближайшего целого числа, число интервалов *l* определяется из выражения

.

**2.3 Графическое построение дискретных лингвистических вариационных рядов для рассматриваемых стихотворений**

Несмотря на его простоту, слабой стороной табличного описания колебания признака является недостаточная наглядность. Поэтому для достижения большей наглядности я использую графическое изображение интересующего меня распределения (длин словоформ по фонемам) – многоугольник распределения признака (полигон).

**2.4 Ряды распределения дискретных случайных величин**

Так как дискретная случайная величина может принимать возможные значения с различными вероятностями, чтобы охарактеризовать её в статистическом смысле, необходимо указать вероятности всех её значений.

Законом распределения вероятностей дискретной случайной величины называется таблица соответствия между возможными значениями этой величины и их вероятностями. Эта таблица – ряд распределения дискретной случайной величины.

Для первого стихотворения:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *X* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|  | 0.1238 | 0.0952 | 0.0762 | 0.1238 | 0.1333 | 0.1714 | 0.1047 | 0.0762 | 0.0476 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *X* | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|  | 0.0285 | 0 | 0.0095 | 0 | 0 | 0 | 0 | 0 | 0.0095 |

Для второго стихотворения:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|  | 0.095 | 0.1428 | 0.1238 | 0.3904 | 0.1333 | 0.1142 | 0.0857 |

По определению, сумма вероятностей событий в каждом из стихотворений должна быть равна 1

Сделаю проверку результатов. Для первого стихотворения:

0.1238 + 0.0952 + 0.0762 + 0.1238 + 0.1333 + 0.1714 + 0.1047 + 0.0762 + 0.0476 + 0.0285 + 0.0095 + 0.0095 = 0.9997 -

подсчёты произведены с небольшой погрешностью

Для второго стихотворения:

0.095 + 0.1428 + 0.1238 + 0.3904 + 0.1333 + 0.1142 + 0.0857 = 0.9971

Из данных результатов следует, что предыдущие исследования сделаны без ошибок.

**2.5 Математическое ожидание дискретной случайной величины**

Математическим ожиданием дискретной случайной величины *Х* называется сумма произведений её всех возможных значений на соответствующие вероятности, обозначается через *М(Х).*

Если случайная величина принимает значения , соответственно с вероятностями , … , то

Стоит заметить, что математическое ожидание является величиной постоянной, его часто называют статистическим значением случайной величины, а также центром распределения, так как около него группируются отдельные значения случайной величины.

Для «Свободного стихотворения»:

M(X) = 10.1238 + 20.0952 + 30.0762 + 40.1238 + 50.1333 +60.1714 + 70.1047 + 80.0762 + 90.0476 + 100.0285 + 120.0095 + 180.0095 = 5.0738

Для «The Cradle Song»:

M(X) = 10.095+ 20.1428+ 30.1238+ 40.3904+ 50.1333 +60.1142+ 70.0857 = 4.1797

Соответственно, M(X)> M(X), исходя из данного результата можно утверждать, что первое стихотворение сложнее для восприятия на слух, чем второе, что немаловажно для анализа звучащей речи.

**2.6 Дисперсия дискретной случайной величины**

Дисперсией дискретной случайной величины *Х* называется математическое ожидание квадрата её отклонения от среднего статистического значения и обозначается через *D(X).*

Для первого стихотворения:

*D(X)=* 0.1238(1 - 5.0738 )+ 0.0952(2 - 5.0738)+ 0.0762(3 - 5.0738)+ 0.1238(4 - 5.0738) 0.1333(5 - 5.0738 )+ 0.1714(6 - 5.0738 )+ 0.1047(7 - 5.0738)+ 0.0762(8 - 5.0738)+ 0.0476(9 - 5.0738)+ 0.0285(10 - 5.0738 )+ 0.0095(12 - 5.0738 )+ 0.0095(18 - 5.0738 )= 8.0928

Для второго стихотворения:

*D(X)=* 0.095(1 - 4.1797)+ 0.1428(2 - 4.1797)+ 0.1238(3 - 4.1797)+ 0.3904(4 - 4.1797)+ 0.1333 (5 - 4.1797)+ 0.1142(6 - 4.1797)+ 0.0857(7 - 4.1797) = 2.9732

**2.7 Энтропия дискретной случайной величины**

Теория энтропии – основа современной теории информации, которая является актуальным направлением исследований в области теории вероятностей и высшей математики в целом. Энтропия является информационной характеристикой дискретной случайной величины. Вычисляется она по формуле К. Шеннона:

Для первого стихотворения H(X) = 3,282844098 бит

Для второго стихотворения H(X) = 2,675265 бит

Энтропия в лингвистике – это одна из наиболее универсальных теоретико-информационных характеристик текста. Это показатель сложности текста в теоретико-информационном смысле.

Из данных результатов несложно сделать вывод, что стиль и звучание «Свободного стихотворения» Зинаиды Гиппиус намного сложнее стихотворения «The Cradle Song». Оно более вариативно и несколько труднее воспринимается на слух.

**2.8 Вероятность появления гласных звуков в стихотворениях, сравнение**

Рассчитать вероятность гласных звуков в стихотворении будет не сложно, тем не менее, результаты данного исследования дадут нам возможность сравнить стихотворения по уровню их певучести, плавности.

Обозначим количество гласных звуков в первом стихотворении , во втором -

По результатам подсчётов =216, = 205, но это ещё не значит, что русское стихотворение певучей английского, для этого следует рассчитать вероятности и по общей формуле

533 – для первого стихотворения, соответственно 439

Очевидно то, что в стихотворении Вильяма Блейка вероятность появления гласных звуков превышает соответствующую вероятность в стихотворении Зинаиды Гиппиус, поэтому можно с полной уверенностью утверждать, что произведение «The Cradle Song» названо автором как нельзя кстати – «колыбельная» - певучая, плавная, спокойная.

**2.9 Коэффициент темпа речи**

*T=*

Где *n –* количество знаменательных слов,

*Р* – количество подлежащих,

*S* – количество сказуемых,

*N –* количество простых предложений,

*N –* количество двусоставных предложений.

Для «Свободного стихотворения»:

*T==*9.33

Для «Колыбельной»:

*T=*1.23

Темп речи первого стихотворения значительно превышает соответствующий показатель во втором, отсюда следует, что второе стихотворение более спокойное, плавное, что снова подтверждает, что Вильям Блейк отлично подобрал название для своего творения.

**3. Объединённый коэффициент синтаксической и ритмомелодической сложности**

Так как предметами моего исследования являются два стихотворения, то формула для расчета объединённого коэффициента синтаксической и ритмомелодической сложности может сыграть огромную роль в изучении данных текстов с точки зрения устной и звучащей речи и выявлении разницы между ними.

Общая формула выглядит следующим образом:

*Сл =*

Где *n –* количество знаменательных слов,

T- количество безударных слогов,

*l* – количество строк,

*N* =105, как нам известно из предыдущих исследований.

Сл1.0762

Сл0.314

Из данных подсчётов можно сделать вывод, что стихотворение воспринимается на слух сложнее, чем стихотворение Вильяма Блейка.

**4. Вывод**

Рассмотренное исследование наглядно иллюстрирует возможности методов математической статистики и теории вероятностей в задачах математической лингвистики. Там, где одной только интуиции читателя недостаточно, так как она всегда субъективна и недостаточно достоверна, применяется математический подход – строгий, объективный, основывающийся на математической модели стиля определенного вида. В моём примере была рассмотрена вероятностная модель текста – наиболее распространённая для решения сложных задач лингвистического анализа, но отнюдь не единственная.

Сравнив «Свободное стихотворение» Зинаиды Гиппиус и «The Cradle Song» Вильяма Блейка с помощью методов математической статистики и теории вероятностей, я пришла к выводу, что стиль Зинаиды Гиппиус более разнообразный, сложный, непредсказуемый, а стихотворение английского поэта – более плавное, спокойное, певучее, мелодичное, легко воспринимающееся на слух.

**5. Список литературы**

1. Р.Г. Пиотровский, К.Б. Бектаев, А.А. Пиотровская, Математическая Лингвистика, - М.: Высшая школа, 1977
2. В.В. Савченко, Теория вероятностей и математическая статистика: Конспект лекций, - Н. Новгород: НГЛУ, 2003
3. В.В. Савченко, В. В. Ретивина, Математика и информатика для лингвистов, краткий конспект лекций, - Н. Новгород: НГЛУ, 2006
4. ВВ Власов, Конспект лекций по высшей математике, - М.: Айрис, 1997
5. Р.Г. Пиотровский, К.Б. Бектаев, Математические методы в языкознании. Часть 2. математическая статистика и моделирование текста, Алма-ата: КазГУ, 1973