ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Кафедра информационных систем в искусстве и гуманитарных науках

ДОПУСТИТЬ К ЗАЩИТЕ
Заведующий Кафедрой информационных систем в искусстве и гуманитарных науках
_____**(Борисов Н.В.)**
“23” мая 2016 г.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
Основная образовательная программа
«Прикладная информатика в области искусств и гуманитарных наук»
Направление 230700 «Прикладная информатика»
Уровень Бакалавриат

«Зависимость реализации процедур контекстной предсказуемости от жанровых и стилевых характеристик текста»

Студента *Крутченко Ольги Витальевны*
_____
(подпись студента)

Руководитель *профессор СПбГУ, доктор филол. наук,*
Ягунова Елена Викторовна
_____
(подпись руководителя)

Санкт-Петербург
2016

АННОТАЦИЯ

выпускной квалификационной работы

Крутченко Ольги Витальевны

название выпускной квалификационной работы

Зависимость реализации процедур контекстной предсказуемости от жанровых и стилевых характеристик текста

Пояснительная записка 65 стр., 5 ч., 5 рис., 13 табл., 39 источников, 4 прил.

КОНТЕКСТНАЯ ПРЕДСКАЗУЕМОСТЬ, БИГРАММА, КОРПУС ТЕКСТОВ, НАУЧНЫЕ ТЕКСТЫ, ХУДОЖЕСТВЕННЫЕ ТЕКСТЫ, DICE, SURPRISAL, УСЛОВНАЯ ВЕРОЯТНОСТЬ, ИНФОРМАЦИОННАЯ ЭНТРОПИЯ, ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ, CLOZE-ТЕСТ

Объектом исследования является контекстная предсказуемость в текстах художественного и научного функциональных стилей.

Цель данной работы – выявить зависимость реализации процедур контекстной предсказуемости и от жанровых и стилевых характеристик текста, простроить модель-прототип, позволяющую предсказывать элементы текста, обозначить ее особенности для текстов разных функциональных стилей.

Для достижения поставленной цели необходимо было решить следующие задачи:

- выбрать и обосновать методы исследования;
- написать модульную программу, реализующую вычисления контекстной информации и предсказательную силу для каждого элемента текста;
- построить модели для художественных и научных текстов;
- оценить эффективность статистически и с помощью эксперимента с информантами.

В процессе работы был произведен анализ различных методов исследования контекстной предсказуемости, проведен вычислительный эксперимент на основе корпусов научных и художественных текстов и эксперимент с информантами.

В результате были построены модели текстов, произведена оценка работоспособности модели по выбранным признакам контекстной предсказуемости.

Полученные результаты и написанная программа могут применяться для дальнейших исследований в области автоматической обработки текстов.

Работу планируется продолжать в магистратуре.

Содержание

Введение

Информационная избыточность – это неотъемлемое свойство любого текста, особенно с точки зрения теории информации. И именно благодаря этому свойству человек успешно воспринимает и понимает как устный, так и письменный тексту. Избыточность является неотъемлемым свойством любого языка и поэтому присуща всем текстам без исключения, но в разной степени, в зависимости от функционального стиля текста [1].

С вопросом об избыточности текста тесно связано понятие контекстной предсказуемости, т.е. предугадывания слова на основе контекста. Эффект контекстной предсказуемости по сути является противопоставлением информационной избыточности, демонстрируя, что для восприятия и понимания текста не все его слова являются равнозначными.

В данной работе производится анализ различных вычислительных методов исследования контекстной предсказуемости, выделяются наиболее адекватные метрики и признаки для дальнейшей проверки в ходе построения модели текста и оценки ее работы по каждому из признаков, производится составление корпусов текстов художественного и научного стилей. Исследование, проведенное в рамках выпускной квалификационной работы, предполагает проведение вычислительного анализа на основе корпусов научных и художественных текстов и эксперимента с информантами.

Основной целью проведения исследования является выявление зависимости реализации процедур контекстной предсказуемости от жанровых и стилевых характеристик текста.

Решаются следующие задачи:

- Анализ литературы, позволяющий выбрать наиболее адекватные методы исследования контекстной предсказуемости;
- Подбор и обоснование материала для будущего исследования;

- Формирование корпусов текстов различных стилей и жанров;
- Выбор и обоснование основных методик исследования;
- Написание программы, реализующей такие модули как :
  - препроцессинг (токенизация, ламатизация),
  - модули, реализующие признаки контекстной предсказуемости (энтропийные признаки, различные меры связности и т.д.)
- Проведение эксперимента с информантами
- Оценка эффективности модели по каждому из признаков.

Изучение контекстной предсказуемости предполагает учет многих аспектов, так как эта тема является междисциплинарной. Один из них – психологический аспект. Существует много различных исследований зависимости контекстной предсказуемости и скорости чтения человека, его движении глаз при чтении [2] и др.

С другой стоны, изучение контекстной предсказуемости необходимо непосредственно для лингвистики, психологии, восприятия и анализа текста. Такие методы исследования как проведение cloze-текстов, тестов направленных на восстановление недостающих элементов текста, позволяют оценить степень владения языком информантами, readability текста (например, решение вопроса о понятности текстов наподобие текстов инструкций) [3], а так же проанализировать особенности обучения данному языку [3, 4, 5, 6, 7].

Но особенно актуален вопрос контекстной предсказуемости в компьютерной лингвистике, при решении задач связанных с автоматической обработкой текстов [8].

В частности, для распознавания и исправления опечаток в тексте при решении различных задач, связанных с дальнейшей обработкой текста. Используя принципы контекстной предсказуемости, при невозможности распознать слово, можно предположить, что в нем допущена опечатка, и

далее – восстановить правильное слово. В таком случае, восстановить исходное слово, то которое подразумевалось, возможно с помощью контекста. И после сравнения наиболее вероятных вариантов в этом контексте со словом с возможной опечаткой, сделать выводы.

Также контекстная предсказуемость может помочь в выделении ключевых слов в тексте и коллокаций [9]. Словосочетание, являющееся коллокацией, имеет признаки целостной семантической и синтаксической единицы, для него показатели контекстной предсказуемости будут велики. Ключевые слова, напротив, являются основным источником новой и значимой информации в тексте, следовательно, их контекстная предсказуемость будет невелика, особенно при первых их появлениях.

Таким образом, актуальность и практическая значимость исследования контекстной предсказуемости очень высоки для разнообразных областей, связанных с автоматической обработкой текста.

1. Анализ литературы

1.1 Исследование контекстной предсказуемости с помощью cloze-теста

Начальной точкой изучения контекстной предсказуемости можно считать введение в лингвистические исследования такой формы тестирования как cloze-тест. Cloze-тест был разработан и предложен американским ученым В. Тейлором для определения readability текста (показателя, насколько текст труден для чтения и восприятия). Методика составления cloze-теста такова: выбирается отрывок прозы объемом 100 – 400 слов, в котором пропускается каждое n-ое слово. Испытуемому предлагается восстановить пропущенные слова. Успешность выполнения данного текста непосредственно зависит от времени, необходимому испытуемому для понимания всего текста и восстановления связи между событиями. Это в свою очередь определяется тем, насколько хорошо испытуемый владеет лексикой данного языка, в какой степени у него развита языковая догадка и как адекватно он понимает текст каждой конкретной ситуации [5].

Данный вид тестов может быть использован для контроля в процессе обучения иностранному языку, поскольку данный метод позволяет точно и объективно установить степень сформированности навыков чтения и уровень владения лексикой при чтении.

Однако определение степени владения различными навыками иностранного языка является не единственным применением cloze-тестов. С помощью данного вида тестов возможно также оценить языковую модель конкретного языка. Данные исследования демонстрируют сравнение результатов проведения cloze-тестов среди носителей языка и статистическую языковую модель. Данные эксперименты демонстрируют, что возможно получить подробную информацию о производительности языковой модели через cloze-тесты с информантами [3].

Метод cloze-тестов используется также и для оценки понимания речи на слух. Причем данный подход важен не только в целях контроля в обучении иностранному языку, но и для изучения механизмов восприятия звучащей речи, обладающей своими отличительными особенностями: эллипсис, нечеткое произнесение безударных слогов, объективные помехи канала связи и т.д. Данный вопрос подробно рассматривается в работах Ягуновой Е.В. «Вариативность стратегий восприятия звучащего текста» [10], «Исследование избыточности русского звучащего текста» [11] и др. Для данного типа исследований особенно часто применяются скрытые Марковские модели, позволяющие рассматривать текст как совокупность процессов перехода из одного состояния в другое [12].

С точки зрения компьютерной лингвистики предсказуемость слов в контексте исследована незначительно. Однако в последнее время появляется все больше исследований на эту тему.

Основными подходами в данном исследовании контекстной предсказуемости являются анализ статистических данных, основанных на корпусах текстов, и проведение cloze-тестов с информантами. Для проведения комплексного исследования необходимо использовать сочетание двух подходов и сопоставление результатов на каждом из этапов, причем для анализа корпусных ресурсов необходимо использовать различные методы. На первичном этапе анализа корпусных данных возникает два основных вопроса: как оценивать контекстную предсказуемость на основе статистических данных и на основе каких материалов (корпусов) проводить исследование.

Контекстную предсказуемость слова в тексте можно оценить различными способами. В первую очередь это статистические меры ассоциации, используемые в основном для выявления коллокаций. Это такие меры как MI, t-score, Dice [9, 13, 14] и др. Их значения для исследования

контекстной предсказуемости могут быть интересны как при вычислении на отдельном тексте, так и на корпусе сразу [15]. Другой возможный подход к контекстной предсказуемости это информационная энтропия и условная вероятность. Далее эти меры будут рассмотрены более подробно.

1.2. Основные математические модели контекстной предсказуемости

В последние годы появилось много новых исследований, посвященных такой проблеме как «сложность» языковых систем. Возникший нтерес к этой теме является относительно новым. Если рассматривать исследования, в которых предлагаются объективные критерии для определения сложности произвольного языка и ранжирования различных языков по сложности, то первой работой в этом направлении можно считать статью Джона Мак-Уортера (в 2001 году). В своей работе он критикует сложившееся мнение об одинаковой сложности всех языков и доказывает, что некоторые современные языки проще «старинных» [16]. В дальнейшем, идеи Джона Мак-Уортера были развиты в работах других исследователей, таких как Ваутера Кюстерса [17], Эстена Даля [18], Питера Традгила [19] и др.

Для данной дипломной работы это направление связано не с исследованием языков разной типологии, а квантитативной типологией стилей и жанров, активно развивающейся в наше время. Исходя из этого была предположена зависимость процедур контекстной предсказуемости и модели текста.

Однако основным ориентиром в исследовании послужили разработки моделей контекстной предсказуемости в информатике и смежных дисциплинах. Чаще всего такого рода модели опираются на скрытые Марковские процессы. Скрытые Марковские модели позволяют рассматривать текст как совокупность процессов перехода из одного

состояния в другое. При этом, если проанализировать текст достаточно большого объема, то возможно использовать полученные частоты для получения вероятности перехода в отдельные состояния. Например, проанализировав сказку Льиюса Кэролла «Приключения Алисы в Стране чудес», получили, что состояние «л» (появление в тексте буквы «л») встречается в тексте 100 раз. Затем при использовании полученной модели получаем это состояние 33 раза из 100, и следующим состоянием с большой долей вероятности будет состояние «и», поскольку слово «Алиса» является достаточно частотным словом в тексте, выбранном для первоначального анализа [12].

Ряд вероятностных методов, таких как скрытые модели Маркова, Марковские случайные поля активно используются в последние годы для вероятностного анализа, для задачи извлечения текстовых данных. Некоторые примеры таких задач включают моделирование языка, классификации документов, кластеризацию и извлечение информации [20].

Следует так же заметить при этом, что работы, связанные с изучением избыточности, велись и в нашей стране уже в 60-е годы. Например, исследования Н. Н. Леонтьевой, Р. Г. Пиотровского, Т. Н. Никитиной, М. И. Откупщиковой, специально посвященные этой теме [21]. Особенно полно данный вопрос рассматривается в статьях Пиотровского Р. Г. «Лингвистический автомат (в исследовании и непрерывном обучении)» [22] и «Информационные измерения языка» [23].

В качестве предварительного этапа данного исследования скрытые модели Маркова были также рассмотрены, была написана программа на языке Python, получены предварительные результаты. Однако, эти результаты не всегда подлежали полноценной интерпретации, поэтому на этапе работы над дипломом основное внимание было уделено взаимодействию

вероятностных (энтропия, удивительность и т.д.) метрик и метрик связанности (Dice, MI, t-score). Учет контекста происходил следующим образом: с помощью построения связных цепочек на основании меры Dice (с максимальной длиной цепочки равной 7 токенам), которые позволяли рассматривать контекст, предположительно соотносимый с тем минимальным контекстом, который может воспринимать человек в ходе анализа текста. Таким образом рассматривалось большое количество параметров в совокупности и отдельно. В планах на будущее (на магистерскую диссертацию) вернуться к скрытым Марковским процессам и интерпретации результатов разных моделей и разных наборов признаков.

1.2.1. Информационная энтропия

На понятии информационная энтропия из теории информации основана такая метрика как энтропийная характеристика. Рассчитывается по формуле:

$$H(x) = -log_2 P(x)\,,$$

где $P(x)$ – вероятность появления в тексте слова x.

В теории информации информационная энтропия – это мера неопределенности или непредсказуемости [24], неопределенность появления какого-либо символа первичного алфавита. В условиях исследования контекстной предсказуемости, в качестве элементарного символа выступает единичное словоупотребление, рассматриваемое на основе первичного алфавита, состоящего из словаря всех возможных словоупотреблений корпуса (или текста). В данном вопросе интересна именно частная энтропия, характеризующая только появление конкретного словоупотребления.

1.2.2. Условная вероятность

Одним из наиболее очевидных способов оценить контекстную предсказуемость является условная вероятность. В теории вероятности условная вероятность – вероятность одного события при условии, что другое событие уже произошло [24]. Перенося данное определение на контекстную предсказуемость, возможно рассчитать вероятность встретить одно слово при условии, что оно идет в тексте после другого. В данной ситуации контекст выступает в роли события, которое уже произошло.

Условная вероятность для контекстной предсказуемости слова рассчитывается по формуле:

$$P(x|context) = \frac{f(x, context)}{f(context)} ,$$

где $f(x, context)$ – частота совместной встречаемости слова x после заданного контекста, $f(context)$ – частото встречи контекста.

Здесь преимуществом так же является то, что размер контекста никак не ограничен, он выбирается исходя из поставленных задач.

1.2.3. Метрика MI (Mutual Information)

В основе метрики MI лежит такое понятие как взаимная информация, взятое из теории информации. MI, или коэффициент взаимной информации, относится к точечным оценкам силы связи и позволяет оценить независимость появления двух слов в тексте. Этот коэффициент сравнивает зависимые контекстно-связанные частоты (pMI) с независимыми, считая появление слов в тексте случайным. Определяется по формуле:

$$pMI(x_1 x_2) = log_2 \frac{f(x_1 x_2) \times N}{f(x_1) \times f(x_2)} ,$$

где x_2 – исследуемое слово, x_1 – слово предшествующего контекста, $f(x_1x_2)$ – частота встречаемости слов x_1 и x_2 в паре, $f(x_1)$, $f(x_2)$ – частоты слов x_1 и x_2 в корпусе, N – размер корпуса (в количестве словоупотреблений) [9, 13]. В дальнейшем будем использовать номинацию MI (несмотря на то, что оцениваются частотные характеристики).

Мера MI зависит от размера корпуса, что дает более высокий средний показатель для корпусов большего объема. Это говорит о большей достоверности полученных данных на большом корпусе, но исключает возможность сравнения полученных значений в разны корпусах текстов.

Основной принцип работы данной меры – присвоить большее значение сочетаниям с редкими словами, в том числе сюда могут попасть слова с опечатками, сочетания с иностранными словами. Для этого необходимо вводить нижний порог по частотности в корпусе [9, 15].

MI также не учитывает порядок следования слов, что важно при исследовании контекстной предсказуемости. Одним из возможных вариантов ее использования является самостоятельный учет порядка слов в словосочетании и разделение сочетаний на пары вида (x_1x_2) и (x_2x_1). В этом случае необходимо считать MI для каждой пары.

Возможен так же вариант видоизменения меры MI с возведением значения в куб – метрика MI3 [13]. Она рассчитывается по формуле:

$$MI3(x_1, x_2) = log_2 \frac{f(x_1x_2)^3 \times N}{f(x_1) \times f(x_2)},$$

где используются те же обозначения, что и в MI. Данный вариант расчета этой меры считается одним из возможных вариантов нормализации [9].

1.2.4. Метрика t-score

T-score - мера ассоциации, которая относится к асимптотическим критериям для проверки гипотезы. Определяется по формуле:

$$t - score(x_1, x_2) = \frac{f(x_1 x_2) - \frac{f(x_1) \times f(x_2)}{N}}{\sqrt{f(x_1 x_2)}},$$

где x_2 – исследуемое слово, x_1 – слово предшествующего контекста, $f(x_1 x_2)$ – частота встречаемости слов x_1 и x_2 в паре, $f(x_1), f(x_2)$ – частоты слов x_1 и x_2 в корпусе, N – размер корпуса (в количестве словоупотреблений) [9, 13].

Мера t-score, так же как и MI, учитывает размер корпуса и не учитывает порядок слов n и c (учесть порядок слов возможно аналогично). По сути, данная мера является скорректированным ранжированием словосочетаний по частоте встречаемости. В отличие от MI, данная мера не завышает значение для редких словосочетаний, и следовательно для ее использования не нужен нижний порог по частоте в корпусе [9].

В силу своих особенностей t-score интересна в исследовании контекстной предсказуемости слов тем, что она позволяет лучше, по сравнению с MI, выявлять стилистические особенности и устойчивые конструкции, встречающиеся в корпусе.

1.2.5. Метрика Dice

Метрика Dice, как и MI, относится к точечным оценкам меры связи. Она вычисляется по формуле:

$$Dice(x_1, x_2) = \frac{2 * f(x_1 x_2)}{f(x_1) + f(x_2)},$$

где x_2 – исследуемое слово, x_1 – слово предшествующего контекста, $f(x_1x_2)$ – частота встречаемости слов x_1 и x_2 в паре, $f(x_1), f(x_2)$ – частоты слов x_1 и x_2 в корпусе.

Данная мера не зависит от размера корпуса, она учитывает только частоту совместной встречаемости и независимые частоты. Однако, как и MI, эта мера дает завышенную оценку низкочастотных словосочетаний [13, 14]. Хотя это завышение у меры Dice гораздо менее критично, чем у меры MI.

Для исследования контекстной предсказуемости может быть интересен следующий алгоритм для оценки n-словных сочетаний с использованием меры Dice: для всех пар слов в корпусе (или тексте) считается коэффициент Dice, далее осуществляется компоновка элементов по одному из двух принципов (т.н. cosegment процедура).

Первый вариант: в один элемент объединяются пары слов на основании значения коэффициентов у этой пары слов и ближайшего контекста. Слово не присоединяется к предыдущему, если значение коэффициента Dice для данной пары ниже порогового, или если оно ниже, чем среднее арифметическое того же коэффициента для левой и правой пары. Накладывается условие, что связанные цепочки не могут состоять более, чем из 7 слов [14]. Данный алгоритм был подробно описан в статьях V.Daudaravicius [25], реализован и программа доступна для скачивания с сайта разработчика.

Второй вариант: для каждого словосочетания формируется группа путем последовательного объединения со словосочетаниями контекста. Для каждой группы высчитывается коэффициент Dice с учетом пяти словосочетаний из левого контекста и двух словосочетаний из правого контекста (такие цифры учета «окна» контекста являются приближенными к возможности восприятия контекста человеком).

1.2.6. Метрика surprise

Метрика surprisal (иначе собственная информация) является мерой содержания информации, связанной с событием в вероятностном пространстве. Чем меньше вероятность события, тем больше коэффициент surprisal связан с информацией, что это событие произойдет [26].

Метрика surprisal является еще одним способом оценить контекстную предсказуемость с помощью условной вероятности. Предложенная Х.Левви в 2001 году [27], эта мера стала достаточно стандартной для задач, связанных с оценкой контекстной предсказуемости. Она рассчитывается по формуле:

$$I(x, context) = log_2 \frac{1}{P(x|context)} ,$$

где $P(x|context)$ – условная вероятность появления слова x в заданном контексте.

Данную метрику можно рассматривать как информационную энтропию исследуемого слова и контекста вместе взятых, поскольку для ее расчета используется именно условная вероятность (т.е. оценивается зависимость слова от контекста) [26].

Эта метрика универсальна тем, что нет никаких ограничений, накладываемых на контекст. Это может быть как одно слово, так и n-словное сочетание.

1.2.7. Метрика salience

Метрика salience для оценки сочетаемости слов встречается намного реже метрик MI и t-score. Однако ее можно рассматривать как один из нормализованных вариантов метрики Dice. Коэффициент salience рассчитывается по формуле:

$$salience(x_1, x_2) = 14 + log_2 \frac{2 \times f(x_1 x_2)}{f(x_1) + f(x_2)},$$

где x_2 – исследуемое слово, x_1 – слово предшествующего контекста, $f(x_1 x_2)$ – частота встречаемости слов x_1 и x_2 в паре, $f(x_1), f(x_2)$ – частоты слов x_1 и x_2 в корпусе [13].

2. Методика и материалы исследования

2.1. Выбор методик для дальнейшего исследования

Все рассмотренные метрики для выявления контекстной предсказуемости можно классифицировать следующим образом: вероятностные оценки (энтропийная характеристика, условная вероятность, surprisal), точечные (MI, Dice, salience) и асимптотические (t-score) оценки мер связи. Некоторые из них очень похожи между собой, они отличаются друг от друга только нормализацией.

Для практической части данного исследования имеют интерес следующие меры:

- Условная вероятность и энтропийная характеристика, поскольку они являются основными вероятностными метриками.
- Метрика surprisal, т.к данная метрика является стандартной для оценки контекстной предсказуемости.
- Метрика Dice, которая будет использоваться для реализации алгоритма объединения коллокаций в связанные сегменты.

Все отобранные метрики достаточно разнообразны в подсчете коэффициента, и как следствие интересны для сравнения их работоспособности. А так же все из них имеют свои особенности, преимущества и недостатки. В связи с этим они наиболее интересны для дальнейшей проверки на корпусах текстов и отдельных текстах в ходе построения модели и анализе ее работы. Сравнение различных методик

позволит наглядно выявить их различия и эффективность работы и проанализировать результаты в отдельности по каждой из метрик.

2.2. Обоснование материала. Формирование корпусов текстов

Как уже упоминалось ранее, избыточность является неотъемлемым свойством естественного языка и текста на естественном языке в частности, необходимым для восприятия и понимания. Избыточность присуща всем текстам без исключения, однако она не является постоянной величиной и зависит от многих параметров, одним из которых является функциональный стиль текста [1, 11].

Общее количество информации, содержащейся в тексте, называется информационной насыщенностью текста. Информационная насыщенность является абсолютным показателем качества текста (в отличие от информативности, которая зависит от степени новизны темы для читателя, и следовательно является относительным показателем качества). По степени информационной насыщенности пять основных функциональных стилей можно расположить следующим образом в порядке возрастания: разговорный, художественный, публицистический, научный, официально-деловой [1, 28].

В соответствии с этой классификацией наибольшей избыточностью обладает разговорный и художественный стили, в то время как научный и официально-деловой стремятся к повышению информационной насыщенности, т.е. к уменьшению избыточности.

Исходя из выше сказанного, для исследования контекстной предсказуемости были выбраны для сравнения два функциональных стиля: научный и художественный, существенно отличающихся избыточностью текстов. Причем научные тексты должны быть отобраны и разделены на

подкорпуса, каждый из которых принадлежит одной предметной области и однороден по жанру и теме.

На данном этапе подготовки к вычислительному анализу контекстной предсказуемости на основе двух корпусов текстов можно предположить, что значение контекстной предсказуемости для научных текстов будет намного выше по сравнению с художественными ввиду большей информационной насыщенности.

Для корпуса художественных текстов были отобраны тексты различающиеся по следующим параметрам: количество словоупотреблений в тексте, жанр, «узнаваемость» данного художественного произведения.

Корпус художественных текстов состоит из 6 текстов. По количеству словоупотреблений тексты варьируются от 9 500 до 363 500 словоупотреблений. Общее количество словоупотреблений в корпусе – 782 300.

Для корпуса научных текстов было сформировано 2 подкурпуса: научные статьи по корпусной лингвистике (объем – 15 093) и когнитивной психологии (объем – 22 703). Общее количество словоупотреблений в корпусе – 37 796.

Ввиду небольшого объема научных статей в ходе вычислительного эксперимента имеет смысл проводить анализ непосредственно на всем корпусе, в то время как художественные тексты можно рассматривать по отдельности. Результаты анализа корпуса научных текстов и отдельно взятых художественных текстов могут быть сравнимы по причине общности темы научных статей, принадлежности одной предметной области, присутствии схожих ключевых слов (корпус научных статей схожих по этим признакам можно воспринимать как единый текст).

Сформированные корпуса текстов послужат основой для исследования и получения предварительных результатов. Эти корпуса будут рассматриваться как «ядерные». В ходе исследования они могут пополняться для решения частных задач.

2.3. Выбор программных средств для построения модели

Существуют различные программные средства для создания и анализа собственных корпусов текстов. Самые известные из них: Intelli Text [29] и Sketch Engine [30] – программные интерфейсы для создания и анализа корпусов электронных текстов - программный интерфейс для работы с корпусами онлайн, и AntConc [31] - кроссплатформенная программа для проведения корпусных лингвистических исследований и управления данными. Очевидным плюсом данных программных средств является быстродействие при работе с большими объемами данных, широкий выбор доступных функций, однако они также обладают своими недостатками. Многие качественные ресурсы являются коммерческими. К тому же несмотря на большое количество доступных функций, ни один из ресурсов не может предоставить реализацию всех необходимых для данного исследования признаков, а именно подсчет всех выбранных ранее метрик. Поэтому используя готовые программные средства необходимо будет самостоятельно совмещать полученные с помощью разных ресурсов результаты, что является достаточно трудоемкой задачей при анализе больших корпусов текстов.

В связи с этим, в рамках данного вычислительного эксперимента гораздо удобнее написать собственную программу для построения модели текста. Несмотря на то, что собственный продукт, скорее всего, будет уступать в быстродействии крупным коммерческим ресурсам, он будет обладать существенным преимуществом: построенная модель будет удовлетворять всем требованиям поставленной задачи, реализуя необходимые признаки контекстной предсказуемости, и будет предоставлять результат в удобном для дальнейшего анализа виде.

Для написания собственной программы был выбран интерпретируемый, объектно-ориентированный высокоуровневый язык программирования

Python [32], поскольку он является достаточно производительным для решения задач связанных с обработкой текстов (анализ, преобразование, поиск, порождение текстовой информации). Данный язык программирования является расширяемым, имеет хорошую поддержку модульности, позволяя использовать помимо обширной стандартной библиотеки, как собственные библиотеки, так и библиотеки, созданные другими разработчиками. В языке имеется стандартный модуль для математических вычислений math [33], реализующий обширный функционал. Так же для Python созданы различные модули для работы с естественными языками, которые широко используются в различных лингвистических исследованиях.

Одной из отличительных черт языка Python является наличие таких встроенных стандартных типов данных как списки, словари и тьюплы [34], которые так же будут полезны для организации сложных структур хранения данных в лингвистических исследованиях.

В качестве среды разработки для написания программы был выбран продукт компании JetBrains – PyCharm [35]. Это интегрированная среда разработки для языка программирования Python, которая предоставляет следующие возможности:

- статический анализ кода, подсветку синтаксиса и ошибок;
- удобную навигацию по проекту и исходному коду: отображение файловой структуры проекта, быстрый переход между файлами, классами, методами и использованиями методов;
- рефакторинг (процесс изменения внутренней структуры программы, не затрагивающий её внешнего поведения): переименование, извлечение метода, введение переменной, введение константы, подъём и спуск метода и т. д.;
- встроенный отладчик для Python;

- встроенные инструменты для юнит-тестирования (процесса, позволяющего проверить на корректность отдельные модули исходного кода программы).

3. Написание модульной программы

3.1. Постановка задачи

Необходимо написать программу, результатом работы которой будет построенная на основе входного файла модель текста, результат работы модели будет представлен в виде таблицы в выходном файле.

Для представления признаков контекстной предсказуемости был выбран такой объект как направленный граф, вершинами которого являются множество слов текста, а дуги отображают связь слова с его возможными контекстами. Каждая дуга содержит максимально полные списки контекстной информации (см. рис. 1).

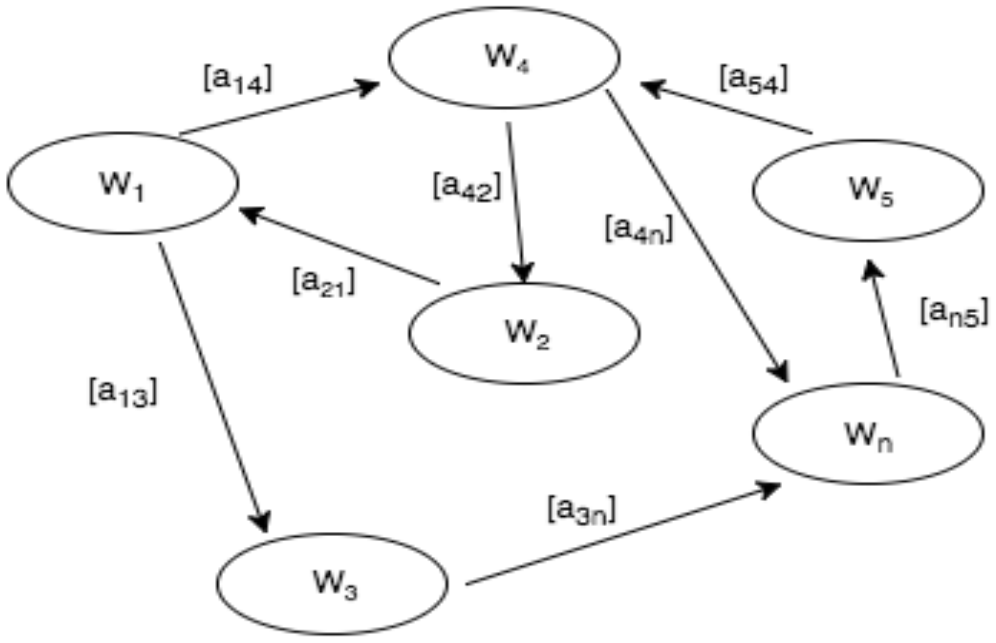

Рисунок 1 – Направленный граф из модели текста.

Программная реализация данного графа представляет собой сложную структуру данных на основе стандартных типов данных языка Python, содержащую всю вычисленную информацию по каждому из выделенных признаков контекстной предсказуемости.

Для построения модели текста на основе исходных текстовых данных на естественном языке необходима последовательная обработка текста на различных уровнях. Для анализа контекстной предсказуемости каждого слова текста необходимо из входного текст составить множество пар, состоящих из исследуемого слова и контекста. В рамках данной задачи в качестве контекста принято считать одно предыдущее слово в тексте. Следовательно, весь исходный текст разбивается на двухсловные сочетания – биграммы. Заметим, что биграммы и коллокации в данной работе являются различными терминами. В качестве биграмм рассматриваются все пары слов, встречаемые в тексте, в то время как под коллокациями понимается устойчивое сочетание двух и более слов, имеющее признаки синтаксически и семантически целостной единицы.

Множество уникальных биграмм текста является основой модели текста, для которой будут производиться дальнейшие вычисления выбранных метрик и признаков. Результатом работы программы – моделью текста – является структура данных, содержащая уникальное множество биграмм и результат вычисленных признаков и метрик по каждой из биграмм. Данный результат выводится в файл в виде таблицы, содержащей построчно результаты для каждой биграммы.

Для построения такой модели текстовые данные, который поступают на вход программе, последовательно обрабатываются на нескольких последовательных этапах, таких как:

- препроцессинг, включающий в себя токенизацию (разбиение входного текста на простые элементы – токены, в данной задаче слова) и лемматизацию (процесс приведения словоформы к лемме и получения набора грамматических характеристик);

- создание частотного словаря токенов с полным набором грамматических характеристик;
- генерация множества биграмм текста, выделение уникального множества биграмм и создание их частотного словаря;
- вычисление необходимых признаков и метрик для каждой из биграмм;
- генерация на основе полученных структур итоговой модели текста;
- структурированный вывод в файл полученной модели.

Далее реализация каждого этапа будет рассмотрена более подробно.

3.2. Токенизация

Первым этапом обработки поученного на вход программы текста является токенизация – выделение из входного текста элементарных значимых единиц, токенов. Также необходима дальнейшая обработка токенов – привидение к единообразному виду, удаление лишних символов (например, специфических знаков, кавычек) и др. В рамках данной задачи в качестве токенов выступают словоформы, приведенные к нижнему регистру, и знаки препинания. Причем отдельно выделяются такие знаки как запятая, точка с запятой, тире, двоеточие, группа знаков конца предложения. Группа знаков конца предложения включает в себя следующие символы: точка, восклицательный знак, вопросительный знак, троеточие, различные комбинации вопросительного и восклицательного знаков в конце предложения. Все эти знаки объединены в одну группу, поскольку имеют общую функцию – обозначение конца предложения. Рассмотрение каждого из этих знаков в отдельности не несет дополнительной информации на данном этапе исследования.

Процесс токенизации осуществляется с помощью библиотеки NLTK (Natural Language Toolkit) [36]. NLTK представляет собой пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python. Позволяет решать такие задачи как классификация, токенизация, стемминг, разметка, парсинг, семантические рассуждения.

Процесс токенизации в программе является отдельным модулем, состоящим из функции разбиения на токены. Для токенизации используется импортированный из библиотеки NLTK модуль tokenize. В этом модуле реализованы различные возможные варианты токенизации (по предложениям, словам, знакам препинания и др.). Для токенизации по словам используется функция word_tokenize, на вход которой подается текст для токенизации. Результатом работы функции является возвращаемый список (стандартная структура данных языка Python) токенов. Стоит отметить, что в функцию word_tokenize изначально передаются текстовые данные, приведенные к нижнему регистру с помощью стандартной функции обработки строк lower. Реализуется это следующим образом:

```
tokens = nltk.word_tokenize(file.read().lower())
```

Здесь tokens является полученным списком токенов, готовых для дальнейшей обработки.

На данном этапе исследования в качестве исследуемой единицы выступают отдельные словоформы и конечный список знаков препинания, поэтому необходимо удалить все ненужные символы из токенов. Это могут быть различные специфические комбинации знаков препинания, выделенные в отдельные токены, или различные символы, такие как кавычки, являющиеся частью токена, содержащего словоформу. В первом

случае необходимо удалить токен из общего списка, во втором удалить только специальный символ, оставив в токене словоформу.

Здесь последовательно просматривается список токоенов и в списке остаются только те токены, которые являются словоформами или принадлежат заранее обговоренному списку символов пунктуации. Далее удаляются все возможные варианты кавычек из токенов, путем замены символа кавычек на пустой символ.

Для дальнейших вычислений в ходе программы так же потребуется такая характеристика как относительная частота каждого токена в тексте. Для этого необходимо заранее подготовить частотный словарь токенов.

Для этого необходимо создать пустой словарь, который при поиске по ключу будет выдавать значение равное нулю, если данный ключ отсутствует в словаре. При просмотре списка токенов производится заполнение словаря и подсчет количества каждого токена. Относительная частота высчитывается при делении абсолютного количества каждого токена на общее число токенов в списке.

В результате работы данного модуля получаем список токенов и вспомогательную структуру данных – частотный словарь токенов.

3.3. Лемматизация

Следующий этап обработки входных данных – лемматизация, т.е процесс приведения словоформы к лемме, её нормальной (словарной) форме. Он осуществляется с помощью морфологического анализатора `pymorphy2` [37], подключаемого как внешний модуль. При работе используется словарь OpenCorpora, а для слов, отсутствующих в словаре строятся гипотезы. Основные функциональные возможности данного морфологического анализатора:

- приводить слово к нормальной форме;
- ставить слово в нужную форму;
- возвращать грамматическую информацию о слове (число, род, падеж, часть речи и т.д.).

В pymorphy2 для морфологического анализа слов русского языка есть класс `MorphAnalyzer`. С помощью метода `MorphAnalyzer.parse()` осуществляется разбор отдельного слова. Метод `MorphAnalyzer.parse()` возвращает один или несколько объектов типа `Parse` с информацией о том, как слово может быть разобрано.

Учитывая неоднозначность, анализатор может возвращать несколько вариантов разбора для одного слова. Каждый вариант разбора имеет параметр `score` – вероятностную оценку правильности данного варианта разбора. Выбор окончательного правильного варианта разбора должен проводиться вручную, с учетом контекста слова, однако для задач обработки большого объема данных допускается автоматический выбор наиболее вероятного варианта разбора с учетом характеристики `score`. Точность морфологического анализа, заявленная разработчиками pymorphy2, при таком подходе составляет около 79%, что является допустимым в рамках решаемой задачи.

Каждый разбор состоит из нормальной формы и набора тегов - набора граммем, характеризующих данное слово. Их можно получить, обратившись к атрибутам `normal_form` и `tag` соответственно.

В данном модуле создается словарь `res_table`, в котором элементом является пара ключ-значение, составленная из токена и его наиболее вероятного морфологического разбора. Тип данных «словарь» используется для того, чтобы не дублировать информацию по одинаковым токенам (т.к. в словаре все ключи являются уникальными). При просмотре списка токенов происходи проверка наличия данного токена в словаре, и

если он отсутствует, в словарь добавляется новая запись о данном токене и его морфологическом разборе.

Результатом работы данного модуля является словарь `res_table` – частотный словарь всех словоформ исследуемого корпуса текстов (или одиночного текста) с полным набором грамматических характеристик.

3.4. Генерация множества биграмм

Основной единицей анализа в данной задаче является пара, состоящая из контекста и исследуемого слова. Так как в качестве контекста рассматривается тоже единичная словоформа, то основным элементом модели является биграмм, состоящий из исследуемого слова и слова-контекста (предшествующего исследуемому слову). Для получения упорядоченного списка всех биграмм входного текста воспользуемся созданной ранее структурой – списком токенов, который также является упорядоченным (токены представлены в порядке следования в тексте).

Здесь возникает вопрос о том, как учитывать первое слово текста. Для данной задачи было решено рассматривать первое слово текста как значимое, путем составления биграмма из добавленного специального символа, указывающего на начало текста, и первого слова текста. В качестве специального символа был выбран знак «$», поскольку он реже остальных встречается в текстах, и в случае ошибки токенизации, вероятнее всего не встретится как отдельный токен. А следовательно, при появлении элемента «$» в биграмме, биграмма будет правильно растолкована.

Генерация всех остальных токенов происходит путем перебора последовательных пар из списка токенов:

```
t0 = '$'
for t1 in tokens:
```

```
yield t0, t1
t0 = t1
```

Далее аналогично частотному словарю токенов, необходимо создать частотный словарь биграмм. Заметим, что общее количество биграмм будет равно общему количеству токенов, так как в качестве первой биграммы берется специальный символ и первое слово, второй – первое и второе слово и т.д., а последней биграммой является пара слов (n-1, n).

Результатом работы данного модуля является частотный словарь биграмм анализируемого корпуса текстов.

3.5. Вычисление признаков и метрик

3.5.1. Энтропийная характеристика

Ранее в данной работе были рассмотрены различные методы анализа контекстной предсказуемости и выбраны те, которые имеют интерес для практического исследования. Одной из них является энтропийная характеристика. Энтропийная характеристика основана на понятии информационная энтропия из теории информации.

Энтропийная характеристика [24] основана на понятии информационная энтропия из теории информации. Рассчитывается по формуле:

$$H(x) = -log_2 P(x)\,,$$

где $P(x)$ – вероятность появления в тексте слова x.

В программе, основным элементом модели текста является биграмм. Энтропийная характеристика зависит только от одного слова, в качестве этого слова берется исследуемое слово из биграмма, т.е. при анализе биграммы (x_1, x_2) расчет энтропийной характеристики производится для слова x_2.

Так как энтропийная характеристика не зависит от контекста, наиболее удобным способом для ее расчета является составление словаря токенов, в котором каждому токену сопоставляется значение энтропии. В дальнейшем при генерации модели этот признак будет добавляться из словаря для исследуемого слова в биграмме.

В результате работы модуля получаем словарь со значением энтропийной характеристики для каждого токена.

3.5.2. Условная вероятность

Следующим рассматриваемым признаком контекстной предсказуемости является условная вероятность.

Условная вероятность [24] для контекстной предсказуемости слова рассчитывается по формуле:

$$P(x|context) = \frac{f(x, context)}{f(context)},$$

где $f(x, context)$ – частота совместной встречаемости слова x после заданного контекста, $f(context)$ – частото встречи контекста.

Для каждой биграммы из списка уникальных биграмм (таковым является множество ключей частотного словаря биграмм) рассчитывается условная вероятность. В качестве частоты совместной встречаемости слова x после заданного контекста берется значение из частотного словаря биграмм для данной биграммы. Частота встречи контекста берется из частотного словаря токенов, где в качестве ключа выступает слово-контекст.

Биграмма, включающая первое слово текста и специальный символ рассчитывается отдельно – для значения условной вероятности такой биграммы берется значение для первого слова из частотного словаря токенов.

Изначально вероятности присваивается нулевое значение.

Расчет условной вероятности и метрик, которые будут описаны далее не выделяется в отдельный модуль. Все вычисления ведутся в момент генерации модели, при ее заполнении конкретными значениями.

3.5.3. Метрика Dice

Метрика Dice [13], как и MI, относится к точечным оценкам меры связи. Она вычисляется по формуле:

$$Dice(x_1, x_2) = \frac{2 * f(x_1 x_2)}{f(x_1) + f(x_2)} ,$$

где x_2 – исследуемое слово, x_1 – слово предшествующего контекста, $f(x_1 x_2)$ – частота встречаемости слов x_1 и x_2 в паре, $f(x_1), f(x_2)$ – частоты слов x_1 и x_2 в корпусе.

Здесь аналогично используются значения из частотного словаря биграмм для частоты совместной встречаемости и значения из частотного словаря токенов для одиночных частот.

В программе вычисление производится так же с помощью стандартной библиотеки math и не выносится в отдельный модуль:

3.5.4. Метрика surprisal

Метрика surprisal является еще одним способом оценить контекстную предсказуемость с помощью условной вероятности. Она рассчитывается по формуле:

$$I(x, context) = log_2 \frac{1}{P(x|context)} ,$$

где $P(x|context)$ – условная вероятность появления слова x в заданном контексте [27].

В рамках данной задачи в качестве контекста берем одиночное словоупотребление, причем условная вероятность встречаемости слова в зависимости от предыдущего слова-контекста была рассчитана ранее. Для случаев, в которых условная вероятность не рассчитана (равна нулю по умолчанию), значение данной меры принимается равным бесконечности, поскольку значение данной меры тем выше, чем меньше контекстная предсказуемость исследуемого слова в зависимости от предшествующего контекста.

Вычисление так же производится непосредственно для каждой биграммы при добавлении ее в модель.

3.6. Генерация модели текста

Все предыдущие этапы были вспомогательными для построения основной значимой структуры программы отражающей модель текста. Модель текста – такой объект, который отражает все исследуемые признаки. Как говорилось ранее, для отображения модели используется направленный граф (см. рис. 1), однако для более удобного представления информации в программной реализации для него была сконструирована специфическая сложная структура данных, которая обладает свойством моментального доступа к необходимой информации для ее извлечения.

Одной из отличительных черт языка Python является наличие таких встроенных стандартных типов данных как списки, словари и тьюплы. Эти типы данных возможно комбинировать, создавая структуры различной сложности (например, список списков, словарь, значениями которого являются сложные составные элементы и т,п.). Именно их комбинация будет использоваться для создания структуры данных - модели текста.

Для представления вычисленной информации модели текста была выбрана следующая структура: данная структура является словарем, ключи

которого - значения из уникального списка словоупотреблений. Значение, соответствующее ключу этого словаря, – список, который в свою очередь состоит из вложенных списков. Вложенный список является элементом, характеризующим один из возможных контекстов для данного слова. Каждый вложенный список содержит следующие элементы в определенной заранее последовательности:

- одно из слов возможного контекста, встречающегося для исследуемого слова в тексте;
- элементы результата вычисленных метрик, перечисленных в определенном заранее порядке одинаковом для всех сложенных списков этого словаря.

Схематично структуру словаря можно представить следующим образом (см. рис. 2):

model = { "$word_1$" : [[$context_1$, probability, entropy, MI, t-score, Dice, lemma, tags],
[$context_2$, probability, entropy, MI, t-score, Dice, lemma, tags],
...
[$context_m$, probability, entropy, MI, t-score, Dice, lemma, tags]]
...
"$word_n$" : [[$context_1$, probability, entropy, MI, t-score, Dice, lemma, tags],
...
[$context_k$, probability, entropy, MI, t-score, Dice, lemma, tags]] }

Рисунок 2 - Структура данных для хранения информации о контекстной предсказуемости.

Программная реализация генерации данной структуры имеет следующий вид:

```
model = {}
for (t0, t1), freq in bi.items():
     if t1 in model:
```

```
            model[t1].append([t0, res, mi, t_score, dice,
            log_dice, surprisal])
        else:
            model[t1] = [[t0, res, mi, t_score, dice, log_dice,
            surprisal]]
```

Процесс генерации структуры данных для хранения информации о контекстной предсказуемости и непосредственно вычисления всех метрик выделено в отдельный обособленный модуль, результатом работы которого является построенная структура. Главным преимуществом данной структуры является эффективно организованное хранение информации,

3.7. Структурированный вывод

Для анализа полученных результатов необходимо вывести в удобочитаемом виде полученные результаты. Наиболее удобная форма для представления такого рода информации – таблица. Каждая строчка будущей таблицы будет иметь следующий вид: биграмма и все рассчитанные для нее признаки. Для этого в текстовый файл выводятся построчно все уникальные биграммы и ее признаками с использованием разделителя в виде знака табуляции между элементами. Разделитель необходим для дальнейшего переноса данных в таблицы Microsoft Excel.

Реализуется такой вывод с помощью выделения из сложной структуры данных, созданной для хранения контекстной информации, всех возможных контекстов для исследуемого слова путем перебора всех возможных словоформ и их контекстов:

```
for word, inf in model.items():
    entropy                                                    =
    str('%.10f'%(token_dict_entropy[word])).replace('.', ',')
    normal_word = str(res_table[word].normal_form)
```

```
tag_word = str(res_table[word].tag)
for item in inf:
     prev = str(item[0])
     prob = str('%.10f'%(item[1])).replace('.', ',')
     mi = str('%.10f'%(item[2])).replace('.', ',')
     dice = str('%.10f'%(item[4])).replace('.', ',')
     surp = str('%.10f'%(item[6])).replace('.', ',')

     file.write(prev+'\t'+str(word)+'\t'+prob+'\t'+entropy
     +'\t'+mi+'\t'+dice+'\t'+log_dice+'\t'+surp+'\t'+norma
     l_word+'\t'+tag_word+'\n')
```

Реализуется вывод в файл так же в отдельном модуле. В результате работы модуля получаем выходной текстовый файл. Для дальнейшего анализа и сортировки полученных данных, результаты из текстового файла переносятся в Microsoft Excel.

3.8 Выделение сильно связанных сегментов текста

Как уже упоминалось в пункте 1.2.5 для исследования контекстной предсказуемости интересен алгоритм для оценки n-словных сочетаний с использованием меры Dice для выделения из текста сильно связанных, неделимых сегментов - chunks. Данный термин был введен в качестве когнитивного термина психологом Дж. А. Миллером в статье «Магическое число семь, плюс или минус два: некоторые ограничения на нашу способность для обработки информации"» в 1956 году для обозначения фрагмента (иными словами куска) текста, состоящего из нескольких слов, которые обычно используются вместе в фиксированном выражении. Пример таких фраз: «на мой взгляд», «Знаете, что я имею

ввиду?» и др. Выделение данных фраз (chunks) производилось в рамках исследования усвоения иностранного языка [38].

В данной работе этот термин будет в дальнейшем использоваться в кириллическом написании - чанк, для упрощения понимания восприятия текста.

Объединение слов в чанки происходит на основании заранее обговоренного признака связности двух элементов текста (слов).

При вычислительном эксперименте ранее был рассчитан коэффициент Dice для каждой биграммы (см. рис. 3).

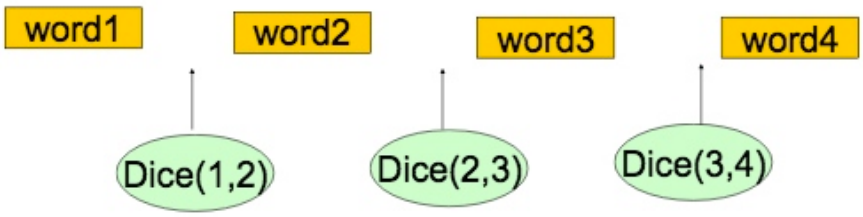

Рисунок 3 – Обозначения имеющихся коэффициентов Dice.

В данном исследовании в качестве признака для объединения двух элементов word2 и word3 (см. Рис.3) в чанк выбрано условие, основанное на среднем арифметическом двух значений коэффициента Dice справа и слева от исследуемых слов.

Первое анализируемое слово всегда является чанком. Для присоединения каждого последующего слова к чанку необходимо выполнение следующего условия:

$$Dice(2,3) > \frac{Dict(1,2) + Dice(3,4)}{2}$$

Слово не присоединяется к предыдущему, если значение коэффициента Dice для данной пары ниже порогового, т.е. чем среднее арифметическое того же коэффициента для левой и правой пары. Накладывается

дополнительное ограничение на длину чанков: количество элементов (слов) не более 7 [14, 25].

Объединение слов в чанки было выделено в отдельный модуль программы.

4. Анализ результатов вычислительного эксперимента

4.1. Сравнение значений признаков контекстной предсказуемости

В результате вычислительного эксперимента были получены модели текстов, вычисленная информация о контекстной предсказуемости представлена в виде результирующих таблиц. Для проведения дальнейшего анализа и сравнения результатов, полученных для разных текстов, необходимо учесть различный объем исследуемых текстов. В связи с этим была составлена таблица отражающая лексическое разнообразие текстов (см. табл. 1).

Таблица 1.

Текст	Объем текста (кол-во словоупотр.)	Процент уникальных словоформ в тексте	Процент уникальных лексем в тексте
Корпус научных текстов по когнитивной психологии	13434	30.4	23.3
Корпус научных текстов по компьютерной лингвистике	13434	39.6	25.6
В.Астафьев «Ловля пескарей в Грузии»	9624	50.2	36.6
В.Пелевин «Проблема верволка…»	10472	38.0	25.3
Э.М.Ремарк «Станция на горизонте»	49568	28.3	16.0
В.Скотт «Айвенго»	148466	19.7	8.3
К.Маккалоу «Поющие в терновнике»	200852	17.3	7.8
А.Дюма «Граф Монте-Кристо»	363554	12.1	4.7

Лексическое разнообразие исследуемых текстов.

Для анализа полученных результатов для каждого текста было посчитано среднее арифметическое значение по каждому из исследуемых признаков и метрик (см. прил. А). Для примера приведен фрагмент данной таблицы (см. табл. 2).

Таблица 2.

Текст	Энтропия (ср.ар.)	Dice	Surprisal
Корпус научных текстов по компьютерной лингвистике	11.4	0.27	2.35
В.Пелевин «Проблема верволка…»	10.3	0.22	2.68
А.Дюма «Граф Монте-Кристо»	12.5	0.06	5.3

Фрагмент таблицы, содержащей среднее значение признаков по каждому из исследуемых текстов.

Стоит отметить, что значение среднего арифметического энтропии и метрики surprisal не являются взаимообратными, поскольку для вычисления энтропии использовалась вероятность появления в тексте одиночного словоупотребления, в то время как расчет значения метрики surprisal производился с учетом контекста (на основе условной вероятности).

На данном этапе исследования можно сделать вывод о том, что значения выбранных признаков непосредственно зависят не только от свойств контекстной предсказуемости текста, но так же и от объема исследуемого текста. Это говорит о необходимости пополнения корпусов текстов для проведения дальнейшего исследования. Не смотря на это, на основе полученных результатов уже можно сделать вывод, что гипотеза об ожидаемом более высоком показателе контекстной предсказуемости для корпуса научных текстов по сравнению с художественными подтверждается.

4.2. Практическое применение построенной модели. Исправление опечаток и снятие неоднозначности

Построенные модели текстов уже на данном этапе могут использоваться для решения практических задач, например связанных со снятием неоднозначности и исправлением опечаток.

Для распознавания опечаток необходимо проверить наличие в результирующей таблице слов, различающихся одной буквой, причем значение леммы для одного из слов не известно (см. табл. 3). В этом случае необходимо сравнить значения энтропии для этих слов. В случае, если энтропия одного из слов будет намного выше значения общей энтропии текста, на основе которого производилось построение модели, высока вероятность, что в данном слове допущена опечатка и необходимо проверить значения других признаков при одинаковом контексте для этих слов (если таковой контекст имеется).

Таблица 3.

Контекст	Слово	Энтропия	Dice	Surprisal	Lemma
сегодня	вечером	12	0,079	4,2	вечер
сегодня	вчером	18	0,026	6,2	UNKN

Пример исправления опечаток.

Другой случай применения полученной результирующей таблицы на практике – снятие неоднозначности. Так как морфологический разбор осуществляется автоматически, он несовершенен. Поэтому при использовании автоматического морфологического разбора возможны ошибки.

Для проверки необходимо сравнить значение некоторых признаков для одинаковой словоформы с различным контекстом. Необходимо в первую

очередь обратить внимание на значении метрики surprisal, заметно превышающее среднее значение по тексту (см. табл. 4). Данный показатель говорит о том, что данная биграмма является редко встречаемой, именно поэтому при выборе морфологического разбора возможно была допущена ошибка, поскольку при автоматическом морфологическом разборе анализатор выбирает наиболее вероятностный вариант.

Таблица 4.

Контекст	Слово	Энтропия	Dice	Surprisal	Lemma
моя	вина	16	0,026	6,0	вино
глоток	вина	13	0,045	1,6	вино

Пример снятия неоднозначности.

4.3 Анализ выделенных цепочек слов

В результате вычислительного эксперимента была получена таблица всех выделенных сильно связанных цепочек – чанков, длины которых составляют от 1 до 7 элементов.

Сравнивая полученные результаты для художественного и научного функциональных стилей можно выделить следующие закономерности:

- Средняя длина чанков в художественных текстах составляет 5 элементов, в научных – 3.
- В научном стиле предложение практически полностью разбивается на чанки, в то время как в художественном стиле в более длинных предложениях чаще встречается 1-2 объединенных блока.
- Чанки в текстах научного стиля это фразы клише, вводные конструкции и обороты. В художественном – устойчивые сочетания, коллокации.

Данные выводы можно проиллюстрировать на конкретных примерах с помощью графиков (см. рис. 4 и рис. 5), которые более наглядно отражают отличия художественного и научного функциональных стилей.

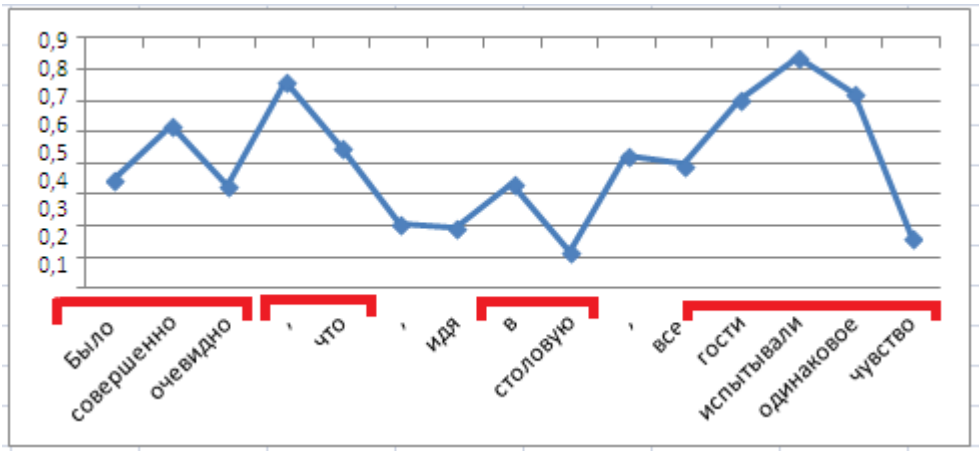

Рисунок 4 – Выделение чанков в предложении художественного стиля.

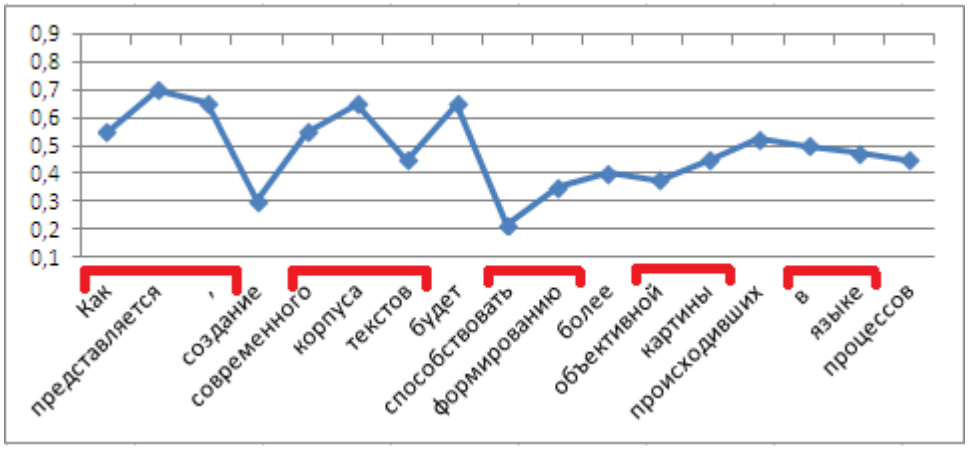

Рисунок 5 – Выделение чанков в предложении научного стиля.

5. Оценка выбранных признаков с помощью эксперимента с информантами

5.1. Подготовка и проведение эксперимента

Неотъемлемой частью вычислительного эксперимента является проверка его результатов. Ранее в пункте 1.1 подробно рассматривалась возможность изучения вопроса контекстной предсказуемости с помощью проведения cloze-теста. В рамках данного исследования такой вид теста был выбран для оценки работоспособности модели текста, построенной в ходе вычислительного эксперимента.

Процедура проведения теста стандартного варианта теста была подробно описана ранее (см. п. 1.1).

При разработке конкретного варианта теста использовались следующие параметры. Для проведения теста были выбраны 4 фрагмента текстов (2 фрагмента художественного стиля, 2 - научного), принадлежащих различным произведениям. Каждый из фрагментов по объему составляет от 100 до 120 слов. В каждом фрагменте пропущено по 10 слов, которые предлагается восстановить информантам.

Выбор того, какие слова фрагмента будут пропущены, осуществлялся на основе метрики surprisal (в последнее время она является наиболее основной и часто используемой при изучении контекстной предсказуемости). В рамках проведения эксперимента с информантами предполагалось проверить работоспособность модели текстов, построенной на основе вычислительного эксперимента. Поэтому в каждом фрагменте текста в качестве пропущенных слов были выбраны слова с высоким (8 - 11), средним (4 - 8), низким (0 - 3) значением метрики surprisal. Предполагается, что более высокое значение меры говорит о том, что данное слово хуже восстанавливается из контекста.

Мера surprisal выбрана в качестве основной для составления теста, однако результаты эксперимента будут сопоставляться и со значением энтропийной характеристики.

Списки исключенных слов по каждому из фрагментов с соответствующими значениями метрик представлены в таблицах 5 - 8.

Таблица 5.

№	Контекст	Слово	Surprisal	Энтропия
1	и	увидела	10,7	13
2	красок	небе	2	14
3	жемчужно-розовое	сиянье	0	18
4	взошло	солнце	1,6	12
5	чистейшей	воде	1,6	14
6	за	бортом	9,7	17
7	густо	синело	3,8	18
8	дно	покрывали	2	18
9	сами	собой	5,6	12
10	над	водой	7	14

Значения метрики surprisal и энтропийной характеристики для исключенных слов из 1-го фрагмента

Таблица 6.

№	Контекст	Слово	Surprisal	Энтропия
1	сильный	ветер	3,8	14
2	все	время	6	10
3	она	освещала	10,3	17
4	мог	спрятаться	8,3	17
5	спрятался	там	2,6	11
6	свист	ветра	0	15
7	,	ваше	9	11
8	лучше	сказать	5,1	11
9	это	время	6,4	10
10	раз	слышались	9,1	17

Значения метрики surprisal и энтропийной характеристики для исключенных слов из 2-го фрагмента

Таблица 7.

№	Контекст	Слово	Surprisal	Энтропия
1	-	несмотря	10,6	13
2	и	многообразие	9,6	15
3	как	правило	4,8	11
4	Вундта	рассматривают	1,6	14
5	.	Иначе	8	11
6	как	теоретическое	6,8	12
7	методологических	проблем	1	13
8	которыми	сталкиваются	2,3	14
9	непосредственный	опыт	0	11
10	сопровождается	чувством	1,6	14

Значения метрики surprisal и энтропийной характеристики для исключенных слов из 3-го фрагмента

Таблица 8.

№	Контекст	Слово	Surprisal	Энтропия
1	с	современными	7	13
2	ее	назначение	3,7	14
3	,	чтобы	6,9	10
4	нормативным	вариантом	0	13
5	грамматических	признаков	1	11
6	вариантах	написания	1	11
7	операция	требует	0	11
8	сохраняется	оригинальная	1,6	14
9	,	во-вторых	10,4	12
10	устаревший	вариант	0	12

Значения метрики surprisal и энтропийной характеристики для исключенных слов из 4-го фрагмента

Была составлена инструкция по прохождению теста для информантов (см. прил. Б) и сам бланк теста (см. прил. В). Тест состоит из двух частей – основной и дополнительной. Первая содержит непосредственно

содержательную часть – фрагменты текстов с пропущенными словами, вторая – вопросы для информантов, на которые необходимо было ответить после прохождения теста. Информантам предлагалась инструкция по прохождению теста и бланк теста, который необходимо было пройти за неограниченное время.

5.2. Анализ полученных результатов

В данном эксперименте участвовали 10 информантов различных возрастных категорий.

По результатам ответов информантов была составлена сводная результирующая таблица (см. прил. Г), в которой обобщены ответы информантов, приводятся подсчитанное количество правильно вставленных в пропуски словоформ, их процент об общего числа, процент совпадения части речи вставленного на место пропуска слова с частью речи исходного. При анализе результатов было отмечено, что процент лексем исходного слова, вставленных в пропуски, совпадает с процентом словоформ.

Основные предположения эксперимента подтвердились - слова с низким значением меры surprisal восстанавливаются информантами однозначно или с использованием синонимов в 85-100% случаев.

По результатам эксперимента была выделена группа слов (10% от общего количества пропущенных), которые однозначно восстанавливаются информантами, однако имеют высокое значение меры surprisal (от 8 до 11). Изначально предполагалось, что данная группа слов будет менее восстанавливаема. Значения же энтропийной характеристики для данной группы слов незначительно превышают среднюю энтропию текста (составляют 13 – 14, при средней энтропии текста 11).

Данный результат можно объяснить тем, что для расчета меры surprisal использовалось в качестве контекста одно предшествующее слово, а информант, заполняя пропуски, ориентировался на контекст большей длины. Факт использования более широко контекста подтверждается тем, что слова этой группы входят в выделенные чанки, образованные на основании метрики Dice с учетом более широкого контекста (пять слов из левого контекста и два слова из правого).

Рассмотрим подобный случай на примере. В предложении «Она вышла на палубу и увидела новую, незнакомую Австралию» было пропущено слово «увидела». По результатам эксперимента 100% информантов правильно восстановили в данном случае словоформу. Значение метрики surprisal для пропущенного слова составляет 10,7. Однако несмотря на высокое значение метрики surprisal, при выделении чанков на основании метрики Dice была выделена цепочка «вышла и увидела». Этот пример в частности и вся группа данных слов в целом подтверждают необходимость использовать при исследовании контекстной предсказуемости более широкий контекст, приближенный к рамкам восприятия информации человеком.

Так же было отмечено, что слов данной группы в фрагментах из художественных текстов более чем в два раза больше, по сравнению с научными.

В научных текстах практически безошибочно (в 95% случаев) восстанавливаются слова, являющиеся клише. Например, такие как «иначе говоря», «несмотря на», «во-первых, …, во-вторых,…» и др. Но несмотря на низкое значение меры surprisal (а следовательно предположительно лучшую восстанавливаемость из контекста), термины и научная лексика в фрагментах научных текстов в 75% случаях не восстанавливаются информантами.

В общем, результаты проведенного эксперимента подтверждают работоспособность модели текстов, построенной при вычислительном эксперименте. Однако в дополнении к ней выделяют особые случаи, в которых проявляются характеристики человеческого восприятия (анализ более широкого контекста, определенные знания и опыт конкретного информанта), которые не учитывались при вычислительном эксперименте.

Заключение

Вопрос исследования контекстной предсказуемости в современной компьютерной лингвистике является актуальным и практически значимым для решения различных задач связанных с автоматической обработкой текста. Данная тема является интересной в силу своей междисциплинарности и многоплановости возможных исследований. Научная новизна исследования заключается в сопоставлении признаков контекстной предсказуемости для различных функциональных стилей.

В рамках данного исследования были проанализированы различные методы изучения контекстной предсказуемости и выбраны наиболее адекватные метрики и признаки для дальнейшей проверки в ходе вычислительного эксперимента.

На втором этапе были сформированы корпуса разных стилей и жанров в соответствии с составленным списком признаков, выбранных методов и информационных технологий для дальнейшего исследования.

Подготовленные материалы – сформированные корпуса текстов художественного и научного стилей и список отобранных метрик – послужили основой для дальнейшей практической части исследования, в рамках которой был проведен вычислительный эксперимент.

Была написана модульная программа на языке Python, позволяющая построить модель текста, реализующую различные признаки контекстной предсказуемости и проведена оценка ее работы, в том числе с помощью проведения cloze-теста.

Эксперимент с информантами подтвердил эффективность и работоспособность построенной модели текстов, а также наметил возможные пути для дальнейшего развития исследования и совершенствования написанной программы.

По результатам построения моделей текстов подтвердилась гипотеза об ожидаемом более высоком показателе контекстной предсказуемости для корпуса научных текстов по сравнению с художественными.

Полученные результаты могут использоваться для решения различных практических задач, связанных с автоматической обработкой текстов.

Апробация данной работы прошла на заседании секции "Компьютерная лингвистика" Ежегодного научно-практического семинара "Новые информационные технологии в автоматизированных системах" (Институт прикладной математики им. М.В. Келдыша РАН, МИЭМ ВШЭ, Московский государственный технический университет им. Н.Э. Баумана). По материалам проведенного семинара была опубликована статья Крутченко О.В. Классификация ключевых слов для описания новостных кластеров (в соавторстве с Мартина В.С., Соколова Д.Ю., Флуд Д.В.), статья индексируема в базе РИНЦ [39].

Список использованных источников

1. Ягунова Е.В. Основы теоретической, вычислительной и экспериментальной лингвистики, или Размышления о месте лингвиста в компьютерной лингвистике // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011.

2. Biemann Ch., Remus St. and Hofmann M. J. Predicting word 'predictability' in cloze completion, electroencephalographic and eye movement data // Natural Language Processing and Cognitive Science / Bernadette Sharp, Wiesław Lubaszewski and Rodolfo Delmonte (eds). Libreria Editrice Cafoscarina, Venezia. P.83-95.

3. Owens M., O'Boyle P., Mcmahon J., Ming J., Smith Fj. A comparison of human and statistical language model performance using missing-word tests // Language and speech, 1997, vol. 40, №4. – P. 377-389.

4. Richard D. Robinson. The Cloze Procedure: a New Tool for Adult Education // Adult Education Qarterly. 1973 – P. 23, 97-98.

5. Taylor W. L. Cloze procedure: a new tool for measuring readability // Journalism Quarterly, 1953. – P. 415-433.

6. Oller J. W., Jr., Grover Kh Yii, Greenberg L.A., Hurtado R. The learning effect from textual coherence measured with cloze // Cloze and coherence / J. W. Oller, Jr., J. Jonz (Eds). – Cranbury, NJ, 1994. – P. 247-268.

7. Nusbaum H. C. et al. Why cloze procedure? // Cloze and coherence / J.W. Oller, Jr., J. Jonz (Eds) – Cranbury, NJ, 1994. – P. 1-20.

8. Ягунова Е.В. Исследование контекстной предсказуемости единиц текста с помощью корпусных ресурсов // Труды международной

конференции "Корпусная лингвистика– 2008". – СПб. : СПбГУ, 2008б. – С. 396-403

9. Ягунова Е.В., Пивоварова Л.М. Природа коллокаций в русском языке. Опыт автоматического извлечения и классификации на материале новостных текстов // Сб. НТИ, Сер.2, №6. М., 2010.

10. Ягунова Е.В. Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей) / СПбГУ – Пермь, 2008.

11. Ягунова Е.В. Исследование избыточности русского звучащего текста // Избыточность в грамматическом строе языка / Отв. ред. М. Д. Воейкова. СПб.: Наука, 2010. — 462 с.

12. Markov Models for Text Analysis [Электронный ресурс] // Purdue University, Department of Statistics. 2009. Режим доступа: http://www.stat.purdue.edu/~mdw/CSOI/MarkovLab.html (дата обращения: 15.04.2016).

13. Хохлова М.В. Исследование лексико-семантической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов). // Санкт-Петербург, 2010.

14. Ягунова Е.В., Пивоварова Л.М. Исследование структуры новостного текста как последовательности связных сегментов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 25-29 мая 2011 г.). Вып. 10 (17).- М.: Изд-во РГГУ, 2011.

15. Ягунова Е.В. Исследование контекстной предсказуемости единиц текста с помощью корпусных ресурсов // Труды международной конференции "Корпусная лингвистика– 2008". – СПб. : СПбГУ, 2008. – С. 396-403

16. J. McWhorter. The world’s simplest grammars are creole grammars // Linguistic typology. 2001. 5(2–3).

17. W. Kusters. Linguistic complexity: the infl uence of social change on verbal infl ection. Utrecht, 2003.

18. Ö. Dahl. The growth and maintenance of linguistic complexity. Amsterdam, 2004

19. P. Trudgill. Sociolinguistic typology: social determinants of linguistic complexity. Oxford, 2011.

20. Y. Sun, H. Deng, J. Han. Probabilistic Models for Text Mining // Mining Text Data. 2012. - P 259-295.

21. Бердичевский А. Языковая сложность (Language complexity) // Вопросы языкознания. 2012. №5.

22. Пиотровский Р. Г. Лингвистический автомат (в исследовании и непрерывном обучении). СПб., 1999.

23. Пиотровский Р. Г. Информационные измерения языка. Л., 1968.

24. D. MacKay. Information Theory, Inference, and Learning Algorithms. // Cambridge University Press, 2003.

25. V. Daudaravicius. Automatic Identification of Lexical Units. // Computational Linguistics and Intelligent text processing CICling, 2009.

26. Decision Trees: Entropy, Information Gain, Gain Ratio. Marina Santini [Электронный ресурс] / Режим доступа: http://www.slideshare.net/marinasantini1/lecture-4-decision-trees-2-entropy-information-gain-gain-ratio-55241087?related=1 (дата обращения: 18.03.2016).

27. Myslín, Mark, & Roger Levy. Codeswitching and predictability of meaning in discourse. // Language 91(4), 2015.

28. Бабайлова А.Э. Текст как продукт, средство и объект коммуникации при обучении неродному языку. // Изд. Саратовского университета, 1987.

29. IntelliText [Электронный ресурс] / Режим доступа: http://corpus.leeds.ac.uk/ (дата обращения: 19.04.2016).

30. Sketch Engine [Электронный ресурс] / Режим доступа: https://sketchengine.co.uk/ (дата обращения: 19.04.2016).

31. Laurence Anthony's Website. Sowfware [Электронный ресурс] / Режим доступа: http://laurenceanthony.net/software.html (дата обращения: 19.04.2016).

32. Python [Электронный ресурс] / Режим доступа: https://python.org (дата обращения: 25.01.2016).

33. Python. Documentation. The Python Standard Library. [Электронный ресурс] / Режим доступа: https://docs.python.org/2/library/math.html (дата обращения: 25.01.2016).

34. М.Лутц. Программирование на Python. 4-е издание. // Изд. O'Reilly, 2011.

35. PyCharm [Электронный ресурс] / Режим доступа: https://www.jetbrains.com/pycharm/ (дата обращения: 25.10.2015).

36. NLTK 3.0 documentation. [Электронный ресурс] / Режим доступа: http://www.nltk.org/ (дата обращения: 19.04.2016).

37. Морфологический анализатор PyMorphy2. [Электронный ресурс] / Режим доступа: https://pymorphy2.readthedocs.org (дата обращения: 19.04.2016).

38. Миллер Дж. А. Магическое число семь плюс или минус два. О некоторых пределах нашей способности перерабатывать информацию / Ред. Ю.Б. Гиппенрейтер, В.Я. Романов. – Москва : ЧеРо, 1998. – С. 564-582.

39. Мартина В.С., Соколова Д.Ю., Флуд Д.В, Крутченко О.В. Классификация ключевых слов для описания новостных кластеров // Новые информационные технологии в автоматизированных системах: материалы девятнадцатого научно-практического семинара. – М.: ИПМ им. М.В. Келдыша, 2016. – С. 94 – 100.

Приложение А.

Среднее значение признаков контекстной предсказуемости по каждому из исследуемых текстов

Таблица 1.

Текст	Усл.вер.	Энтропия (ср.ар.)	Средняя энтропия	MI	Dice	Surprisal
Корпус научных текстов по компьютерной лингвистике	0.45	11.4	11.1	23	0.27	2.35
Корпус научных текстов по когнитивной психологии	0.39	11.2	11.0	22.6	0.29	2.43
В.Астафьев «Ловля пескарей в Грузии»	0.53	10.7	10.71	21.7	0.32	2.26
В.Пелевин «Проблема верволка…»	0.42	10.3	10.27	21.2	0.22	2.68
Э.М.Ремарк «Станция на горизонте»	0.35	11.8	11.2	23.6	0.14	3.64
В.Скотт «Айвенго»	0.27	12.4	11.7	25.2	0.09	4.38
К.Маккалоу «Поющие в терновнике»	0.25	12.5	11.5	25.3	0.09	4.72
А.Дюма «Граф Монте-Кристо»	0.2	12.5	11.44	25.8	0.06	5.3

Среднее значение признаков по каждому из исследуемых текстов.

Приложение Б.

Инструкция по прохождению теста

Данный тест состоит из двух частей.

Основная часть содержит 4 фрагмента текстов с пропущенными словами. Первые два фрагмента – отрывки из художественных произведений, вторые два – отрывки из научных статей.

Вам предлагается предварительно ознакомиться с фрагментом текста. Затем, читая его повторно, заполнить пропуски. Обратите внимание, что в каждый пропуск необходимо заполнить только одним словом. Не рекомендуется возвращаться к фрагменту после двух прочтений.

Переходите к следующему фрагменту текста, только после завершения работы с предыдущим.

Дополнительная часть теста содержит несколько вопросов, на которые вам предложено ответить после прохождения основной части теста.

Приложение В.

Бланк теста, предложенный информантам

Основная часть теста.

I) Она вышла на палубу и 1) _____ новую, незнакомую Австралию. В прозрачном, нежном, лишенном красок 2) _____ медленно разливалось, поднималось все выше жемчужно-розовое 3) _____, и вот уже на востоке, на краю океана взошло 4) _____, новорожденный алый свет превратился в белый день. Пароход неслышно скользил по чистейшей 5)_____ , такой прозрачной, что за 6) _____, глубоко внизу, можно было разглядеть сумрачные лиловые пещеры и проносящихся мимо ярких рыб. Вдали море густо 7) _____, порой отсвечивало зеленым, а местами, там, где дно 8) _____ водоросли или кораллы, темнели пятна цвета густого вина — и повсюду, словно сами 9) _____, как кристаллы в кварце, возникали острова, то ослепительно белые песчаные, поросшие пальмами, то гористые, сплошь покрытые джунглями, то плоские, в зелени кустарника, едва приподнятые над 10) _____.

II) Был конец сентября, дул 1) _____ ветер; бледную луну все 2)_____ закрывали несущиеся по небу черные тучи; она 3) _____ только песок на аллеях, ведущих к дому, а под деревьями тень была такая густая, что там вполне мог 4) _____ человек, не опасаясь, что его заметят.

Я спрятался 5)_____, где он ближе всего должен был пройти; едва я успел скрыться, как сквозь свист 6) _____, гнущего деревья, мне послышались стоны. Но вы знаете, 7) _____сиятельство, или лучше 8)_____, вы не знаете, что тому, кто готовится совершить убийство, всегда чудятся глухие крики.

Прошло два часа, и за это 9)_____ мне несколько раз 10)_____ те же стоны.

III) 1) _____ на отсутствие единой теории и 2) _____ интерпретаций понятия сознания, в психологической науке, как 3) _____, факты сознания ассоциируются с осознаваемыми переживаниями. Сознание со времен В. Вундта 4) _____ в основном как синоним осознания. 5)_____ говоря, «сознание» понимается как эмпирический термин, а не как 6) _____ понятие. (В этом, возможно, кроется одна из основных методологических 7) _____, с которыми 8) _____ психологи при построении теории). Осознание – это феномен, доступный

интроспективному анализу, это тот непосредственный 9) _____, который открывается носителю сознания и сопровождается 10)_____ субъективной очевидности ощущаемого, воспринимаемого, представляемого, мыслимого и т.д.

IV) Нормализация орфографии не означает ее унификацию в текстах в соответствии с 1)_____ правилами. Ее 2)_____ состоит не в том, чтобы исправить в тексте все отклонения от современных норм, а в том, 3)_____ снабдить все вариативные написания соответствующим нормативным 4)_____. В процессе морфологической разметки разбирается нормативная форма, а набор грамматических 5)_____приписывается всему комплексу, так что на поисковый запрос выдаются контексты, содержащие запрашиваемое слово во всех вариантах 6)_____, при этом оно отображено на экране в том реальном виде, в каком представлено в тексте.

Хотя эта операция 7)_____ дополнительных затрат труда лингвиста-эксперта, они оправданы тем, что во-первых, на выходе сохраняется 8)_____орфография текста, 9)_____, обеспечивается поиск всех орфографических вариантов слова по морфологическим признакам (без этой операции найти в корпусе устаревший 10)_____ написания можно только при точном поиске), в-третьих, происходит пополнение словаря корпуса.

Дополнительная часть теста.

1) К какой возрастной категории Вы относитесь?
   - 18 - 25 лет
   - 25 – 35 лет
   - 35 – 45 лет
   - больше 45 лет.

2) Укажите свой пол?
   - мужской
   - женский

3) Узнали ли Вы художественные произведения, из которых были выбраны первые два фрагмента? Если да, укажите из каких.

Приложение Г.

Сводные таблицы ответов информантов

Таблица 1.

№	Исх. слово	Инф.1	Инф.2	…	Кол-во прав. сл.	общий % прав. сл.	Кол-во прав. ч.р.	общий % прав. ч.р.
1	увидела	увидела	увидела		10	100	10	100
2	небе	небе	небе		9	90	10	100
3	сиянье	облако	солнце		1	10	10	100
4	солнце	солнце	светило		9	90	10	100
5	воде	воде	воде		10	100	10	100
6	бортом	бортом	ней		4	40	10	100
7	синело	синело	виднелись		3	30	10	100
8	покрывали	заполнили	заполнили		3	30	10	100
9	собой	собой	бриллианты		5	50	6	60
10	водой	землей	небом		4	40	10	100

Обобщенные ответы информантов по 1-му фрагменту текста

Таблица 2.

№	Исх. слово	Инф.1	Инф.2	…	Кол-во прав. сл.	общий % прав. сл.	Кол-во прав. ч.р.	общий % прав. ч.р.
1	сильный	сильный	сильный		8	80	10	100
2	время	сильнее	время		9	90	9	90
3	освещала	освещала	освещала		10	100	10	100
4	спрятаться	спрятаться	спрятаться		8	80	10	100
5	там	там	там		10	100	10	100
6	ветра	ветра	ветра		8	80	10	100
7	ваше	ваше	ваше		10	100	10	100
8	сказать	благородие	величество		2	20	2	20
9	время	время	время		10	100	10	100
10	слышались	слышались	слышались		10	100	10	100

Обобщенные ответы информантов по 2-му фрагменту текста

Таблица 3.

№	Исх. слово	Инф.1	Инф.2	…	Кол-во прав. сл.	общий % прав. сл.	Кол-во прав. ч.р.	общий % прав. ч.р.
1	иначе	иначе	иначе		8	80	8	80
2	многообразие	обширность	разнообразие		2	20	10	100
3	несмотря	несмотря	взгляды		7	70	8	80
4	правило	правило	чувство		3	30	8	80
5	рассматривают	считают	определяют		1	10	10	100
6	теоретическое	теоретическое	теоретическое		7	70	10	100
7	проблем	проблем	понятий		5	50	10	100
8	сталкиваются	работают	исследуют		5	50	10	100
9	опыт	опыт	опыт		9	90	10	100
10	чувством	оценкой	чувством		1	10	8	80

Обобщенные ответы информантов по 3-му фрагменту текста

Таблица 4.

№	Исх. слово	Инф.1	Инф.2	…	Кол-во прав. сл.	общий % прав. сл.	Кол-во прав. ч.р.	общий % прав. ч.р.
1	современными	существующими	отдельными		2	20	10	100
2	назначение	цель	цель		2	20	0	0
3	чтобы	чтобы	чтобы		10	100	10	100
4	вариантом	формами	формам		2	20	10	100
5	признаков	ошибок	характеристик		0	0	10	100
6	написания	предоставления	использования		2	20	10	100
7	требует	требует	требует		9	90	10	100
8	оригинальная	исходная	изначальная		0	0	10	100
9	во-вторых	во-вторых	во-вторых		4	40	8	80
10	вариант	вариант	вариант		10	100	10	100

Обобщенные ответы информантов по 4-му фрагменту текста

Выпускная квалификационная работа выполнена мною самостоятельно. Использованные в работе материалы из опубликованной научной, учебной литературы и Интернет имеют ссылки на них.

Отпечатано в 1 экземпляре.

Библиография 39 наименований.

Один экземпляр сдан на кафедру.

Крутченко Ольга Витальевна 

23.05.2016