ФЕДЕРАЛЬНОЕ АГЕНТСВО ПО ОБРАЗОВАНИЮ

Федеральное государственное бюджетное образовательное учреждение

высшего образования

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГУМАНИТАРНЫЙ УНИВЕРСИТЕТ»

ИНСТИТУТ ЛИНГВИСТИКИ

Лапошина Антонина Николаевна

ОТБОР РЕЛЕВАНТНЫХ ПРИЗНАКОВ ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ СЛОЖНОСТИ РУССКОГО ТЕКСТА КАК ИНОСТРАННОГО

Магистерская диссертация   
по направлению подготовки 45.04.03

«Фундаментальная и прикладная лингвистика»

|  |  |
| --- | --- |
| Допущена к защите на ГАК  Директор УНЦ компьютерной лингвистики Института лингвистики РГГУ  \_\_\_\_\_\_\_\_\_\_\_\_   Селегей В.П.    «\_\_\_»\_\_\_\_\_\_\_\_\_2017г. | Научный руководитель  Кандидат филологических  Наук, доцент УНЦ КЛ  \_\_\_\_\_\_\_\_\_\_\_\_   Пиперски А.Ч. «\_\_\_»\_\_\_\_\_\_\_\_\_2017г. |

Москва, 2017

Оглавление

[Введение 3](#_Toc484684828)

[Глава 1. О сложности текста вообще и определения сложности иностранного текста в частности 4](#_Toc484684829)

[1.1. О проблеме сложности текста вообще 4](#_Toc484684830)

[1.2. Особенности определения сложности текста как иностранного 13](#_Toc484684831)

[Глава 2. Описание работы по созданию и обучению предсказательной модели 24](#_Toc484684832)

[2. 1. Сбор и описание корпуса 24](#_Toc484684833)

[2.2. Сбор признаков 29](#_Toc484684834)

[2.3. Отбор релевантных признаков 38](#_Toc484684835)

[2.4. Построение предсказательной модели. 42](#_Toc484684836)

[2.5. Примеры работы предсказательной модели 47](#_Toc484684837)

[2.6. Анализ ошибок 56](#_Toc484684838)

[Заключение и направления дальнейшей работы 58](#_Toc484684839)

[Приложение 1. Программный код для анализа текста и сбора   
признаков. 60](#_Toc484684840)

[Приложение 2. Программный код для настройки работы предсказательной модели. 66](#_Toc484684841)

[Список литературы 69](#_Toc484684842)

# Введение

Чтение считается одним из важнейших аспектов методики обучения иностранному языку: раздел, посвященный чтению, входит во все сертификационные тесты, одной из необходимых составляющих понятия владения иностранным языком является способность читать и понимать неадаптированные тексты на этом языке. Но для достижения хорошего результата обучения текстовый материал должен подходить учащимся по уровню сложности грамматических конструкций, объему, лексическому наполнению и т.д. Подходить - значить иметь оптимальное соотношение знакомой и новой информации. Подбор текстов для учебных пособий, поиск интересных неадаптированных текстов для занятий или самостоятельного чтения - всё это делает исследования в области автоматического определения сложности текстов актуальными и практически применимыми.

Целью данного исследования является поиск и анализ релевантных признаков для автоматического определения сложности текста в практике преподавания русского языка как иностранного. В первой главе, коротко затронув историю изучения проблемы сложности текста и читабельности, мы остановимся подробнее на поиске возможных особенностей определения сложности русского текста именно с позиции иностранных учащихся. Далее, во второй главе, с целью проверить наши предположения на практике мы проведем эксперимент по созданию предсказательной модели для автоматического определения уровня текста. Для этих целей будет собран специальный корпус текстов из различных пособий по русскому языку как иностранному, проведен их анализ и поиск признаков с наилучшей корреляцией со сложностью, а также построена модель регрессионного анализа. После описания эксперимента будут представлены примеры работы нашей модели и анализ ошибок.

# Глава 1. О сложности текста вообще и определения сложности иностранного текста в частности

### 1.1. О проблеме сложности текста вообще

Формально сложность текста можно обозначить как сумму всех элементов текста, влияющих на понимание темы, скорость чтения и уровень интереса к прочитанному. В широком понимании этого термина на сложность текста могут влиять такие факторы как шрифт, сопровождение иллюстрациями, близость темы читателю, мотивация чтения и т.д. и эта сторона также очень интересна, однако мы в своей работе будем оперировать более узким пониманием сложности текста, зависящей лишь от лингвистических факторов, самого текстового наполнения.

За более чем 70-летнюю историю изучения возможностей измерения сложности текста данное направление исследований прошло длинный путь от методов ручного подсчета длин слов до анализа больших коллекций документов с применением моделей машинного обучения.

Так, самые ранние попытки оценить сложность текста исходя из средней длины предложения были предприняты ещё в 1880 г. (DuBay, 2006). Опыты на основе чуть более сложных вычислений, таких как средняя длина слова, средняя длина предложения, количество длинных слов и т.д. начали свое широкое распространение в 30-гг. прошлого века. Среди наиболее популярных можно назвать формулы Флэша-Кинсайда (Kincaid et al., 1975), базирующаяся на средней длине слова в слогах и средней длине предложения, индекс Gunning Fog , где признаками являются среднее количество слов в предложении и процент слов длиннее 3 слогов (Gunning, 1968), SMOG (Simple Measure of Gobbledygook).

Плюсом таких формул, безусловно, является простота, прозрачность и быстрота вычисления, однако они основаны лишь на поверхностных характеристиках и не учитывают непосредственного наполнения текста. В то же время такие формулы являются парадоксальными, т.к. рассчитывая степень легкости/трудности текста они никак не задействуют лексику, ключевой аспект передачи информации а, следовательно, и для понимания текста.

Важным подклассом таких традиционных формул являются метрики на основании лексических данных (‘vocabulary-based’). Трудность лексики измеряется здесь её "знакомостью" читателю или на основании частотных списков. Одной из наиболее известных формул такого типа является обновленная формула Дэйла-Чалл, где используется средняя длина предложения и процент "сложных" слов. В качестве сложных слов авторы предлагают считать лексику, не вошедшую в список из 3000 слов английского языка, знакомых 80 процентам английских школьников 4 грейда (Chall and Dale, 1995).

Большинство таких формул были разработаны для английского языка. В основном интерес исследователей этого периода был сконцентрирован на сложности текста в аспекте школьного преподавания и соответствия материала возрасту, поэтому результатом их работы является число, обозначающее класс или возраст, на который рассчитан этот текст. Для определения сложности русского текста необходим пересчет этих параметров: в этой связи следует упомянуть работы И.В.Оборневой (Оборнева, 2006) и проект Readability.io[[1]](#footnote-1) Ивана Бегтина.

Базируясь на исследованиях М.С. Мацковского и Я. Микка, И.В. Оборнева предлагает свой вариант корректировки индексов Флеша, Флеша-Кинсайда и Фога, а также проводит экспериментальные доказательства эффективности методики оценки сложности учебных текстов с целью верификации полученных индексов.

Сначала было необходимо понять и рассчитать различия числовых характеристик английского и русского текста: для этого автор использовала литературные тексты на английском языке и соответствующие переводы на русский язык известных авторов, а также материал словарей. Всего было проанализировано более 100 произведений общим объёмом около 6 млн. слов. Эти тексты сравнивались по каждой из характеристик (среднее кол-во слогов, средняя длина слова, ср кол-во слов в предложении, процент многосложных слов – более 3 слогов, кол-во согласных в слове, сред число знаков препинания).

Сравнение средней длины слова на базе "Толкового словаря русского языка" Ожегова и "Англо-русского словаря" В.К.Мюллера:

|  |  |  |  |
| --- | --- | --- | --- |
| Словарь | Количество слов | Среднее количество слогов | Количество многосложных слов |
| Мюллер | 41977 | 2,97 | 13495 |
| Ожегов | 39163 | 3,29 | 14760 |

На основании полученных данных Оборнева отмечает, что слова в тексте на английском языке содержат слогов меньше чем русские в среднем в 0,71 раза. Количество слов в предложении отражает как синтаксические особенности языка, так и авторский стиль писателя и переводчика, но английские предложения длиннее русских примерно в 1,25 раза. Русскоязычные тексты содержат больше многосложных слов в 1,24 раза.

Далее она вычисляет корректировки формул Флеша и Флэша-Кинсайда для русского языка и для доказательства эффективности предложенной методики оценки сложности были проводит такие эксперименты для проверки коэффициентов как экспертная оценка текстов, анализ по этим формулам сочинений школьников и др.

Работа И.В. Бегтина носит несколько другой характер. Это открытый социальный проект, дающий возможность проверить читабельность любого текста, а также призванный привлечь внимание к самой проблеме непонятности государственных текстов, злоупотребления так называемым канцеляритом и т.д. Как пишет сам автор на заглавной странице сайта, "сервис может быть полезен разработчикам, райтерам, тем кто делает веб-сайты. Сервис нужен всем, кто хочет, чтобы тексты на страницах были понимаемы, журналистам и всем заинтересованным". И. Бегтин разработал оценку текстов исходя из усредненных показателей по 5 известным формулам читабельности:

* Формула Flesch-Kincaid
* Индекс Колман-Лиау
* Automatic Readability Index
* SMOG
* Формула Дэйла-Чейла

Сама работа состояла в том, чтобы подобрать оптимальные коэффициенты по этим формулам для русского языка: "Русский язык и, соответственно, тексты на русском языке отличаются от текстов на английском по слогам, по длине предложений, по числу символов, по тому какие слова можно признавать многосложными. В частности, в русском языке слова, как правило, длиннее, но предложения короче. В общеупотребительной речи больше многосложных слов и коэффициенты в формулах должны быть иными". [[2]](#footnote-2)

По различным данным в мире насчитывается около 200 подобных формул читабельности. С одной стороны, их обвиняют в излишнем упрощении и однобокости видения проблемы сложности, с другой стороны - благодаря своей простоте они уже широко используются для решения различных практических задач, например, две такие формулы (удобочитаемость по Флешу и формула для школьных текстов по Флешу–Кинкейду) встроены в Microsoft Word и помогают быстро оценить пользователю примерную сложность созданного им текста.

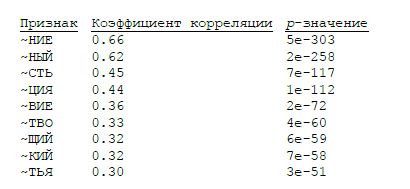
С ростом количества приложений и модулей для обработки естественного языка методами машинного обучения появляются исследования, основанные на работе с большими коллекциями документов, автоматическим извлечением и поиском релевантных признаков и т.д. Проблема автоматического определения сложности текста становится классической задачей построения предсказательной модели на основании обучения на тренировочном корпусе текстов и наборе признаков.

Плюсами такого подхода являются возможность задействования куда большего количества признаков, возможность легко переобучиться на обновленной коллекции данных, а благодаря готовым библиотекам для обработки данных не приходится говорить о сложности вычислений.

Так, например, Константин Дружкин, продолжая работу И.В. Бегтина по адаптации англоязычных метрик удобочитаемости для русского языка, предлагает новый метод обучения модели линейной регрессии на неразмеченном корпусе текстов. В качестве признаков автор выделяет частоты граммем во всем тексте и внутри грамматической категории; одно- двух- и трехбуквенные концовки словоформ и лексем; доли слов, входящих в различные частотные списки. В качестве же "правильных ответов" для тренировки модели он использует среднее от 5 оценок текста по формулам читабельности: "Есть пять “формул сложности”, адаптированных И.В. Бегтиным для русского языка. Их оценки не идеально точны, но они часто согласуются с нашими субъективными ощущениями. Мы возьмём среднее арифметическое между пятью оценками и будем использовать его как ориентир" (Дружкин, 2016). С одной стороны, нам кажется интересной и заманчивой идея автоматической разметки большого корпуса с помощью такого решения, однако анализ корреляции признаков на таком материале может оказаться замкнутым кругом, где мы обучаемся на заведомо непроверенных данных.

В ходе анализа корреляции признаков со сложностью автор высказывает следующие наблюдения:

* В сложных текстах растёт доля родительного падежа (‘род’).
* В сложных текстах увеличивается доля полных форм прилагательных и причастий (*“единый”*, *“объединённый”*, *“объединяющий”*).
* В сложных текстах увеличивается доля существительных.
* В сложных текстах уменьшается доля глаголов в личной форме (‘изъяв’). По-видимому, основная причина уменьшения доли глаголов - это увеличение размера именных групп.
* Очевидно, что признаки на основе концовок словоформ будут сильно коррелировать с грамматическими категориями. Так, среди однобуквенных признаков с положительных корреляцией отмечены:
* ‘Х’: окончание родительного падежа множественного числа (*“президентских”*);
* ‘В’: окончание родительного падежа множественного числа (*“указов”*);
* ‘Й’: окончание родительного падежа множественного числа (*“распоряжений”*);
* Среди трех- и четырехбуквенных концовок лемм уже легко угадываются разные способы образования существительных, а также стандартные окончания прилагательных.



Напомним, что вышеуказанные результаты приведены для русского языка как родного и обучающим корпусом, основанном на усредненном значении 5 формул удобочитаемости. Интересно будет провести параллели с нашими результатами о вкладе некоторых признаков.

Другой подход к решению проблемы определения уровня текста предлагают Sarah E. Schwarm и Mari Ostendorf в своей работе Reading Level Assessment Using Support Vector Machines and Statistical Language Models. (Schwarm, Ostendorf, 2005) Авторы строят статистическую языковую модель на основе корпуса, содержащего около 2500 текстов еженедельного выпуска детского журнала Reader, где уже присутствует деление на грейды (школьные классы) читателей, а также не основе параллельных текстов из энциклопедии Britanica и CNN для взрослых и для детей.

Особенность этой работы состоит в том, что авторы не ставили целью отнесение текста к какому-либо классу сложности, их интересует YES-NO решение, подходит ли этот текст заданному уровню. Предсказательная модель на основе опорных векторов (SVM model) получает на вход как более-менее традиционные признаки (средняя длина предложения, средняя длина слова в слогах, вхождения в частотные списки, частота синтаксических конструкций, оценка текста по Флешу-Кинсайду), так и 12 признаков на основе языковой модели. Статистическая языковая модель (Statistical Language Models) на основе обучающих данных предсказывает вероятность появления конкретной последовательности из n слов.

Метод компонентного анализа признаков для задач определения сложности предлагается в работе S. Sharoff (Sharoff et al, 2008). Эксперимент проводился для английского на материале параллельных статей Википедии и Simple English Wikipedia (этот момент будет рассмотрен подробнее в разделе 2.1.), для русского и китайского языков - на материале сравнения неадаптированных статей на одну и ту же тему в обычных новостных изданиях и на сайте BBC, которые по экспертной оценке являются более простыми. В качестве признаков были предложены:

* доля слов, входящих в частотные списки
* средняя длина предложения
* средняя длина слова в слогах
* индекс Флэша
* доля конструкций, входящих в частотный список частеречных триграмм
* среднее число союзов в предложении
* среднее число полнозначных глаголов на предложение
* среднее число глаголов в пассивной форме на предложение
* среднее число модальных глаголов на предложение
* среднее число предлогов на предложение
* среднее количество знаков пунктуации на предложение

Далее к полученным признакам был применен метод главных компонент (Principal Component Analysis), в результате стало возможным обозначить два главных "измерения" сложности и условно назвать их грамматическим и лексическим. При этом в большинстве случаев простые тексты имеют ожидаемо более низкие показатели по обоим измерениям, однако есть интересные примеры, когда грамматически более просто текст может оказаться сложнее лексически.

Говоря об исследованиях последних лет для русского языка, стоит упомянуть работу Карпова (Karpov et al.,2014), который провел серию экспериментов по определению как сложности текста, так и отдельных предложений. Описаны эксперименты с использованием таких методов, как логистическая регрессия, метод опорных векторов и построение деревьев классификации для бинарной классификации (A1-C2, A2-C2, and B1-C2). Авторы использовали 25 признаков, среди которых количество слов, длина предложения, средняя длина слова и процент присутствия различных частей речи.

Также стоит отметить работу Robert Reynolds (Reynolds, 2014), в которой приводятся результаты эксперимента с классификатором Random Forest на нескольких группах признаков: лексических, морфологических и синтаксических. Автор выдвигает гипотезу, что роль морфологии в формировании сложности текста недооценена из-за того, что огромная часть исследований проводилось для английского языка с относительно бедной морфологией. Действительно, анализируя лучшие признаки, автор сравнивает результаты индивидуального вклада признака и вклада признака в наборе признаков (‘feature set’). В полученных результатах в ‘индивидуальном зачете’ морфологические признаки присутствуют в количестве 7 из 30 лучших, а в наборе из 32 признаков, созданных в помощью алгоритма подбора наилучшего набора признаков (feature subset selection algorithm, CfsSubsetEval) морфологических признаков уже 14. Основываясь на этих результатах, автор делает предположение, что хотя морфологические признаки не такие информативные, как другие, но они предоставляют уникальную информацию о тексте, чем оказываются ценны при создании набора признаков.

Подводя итог, можно обозначить два основных направления работ в области изучения сложности текста: это коррекция и адаптация для других языков уже известных и популярных формул читабельности, а также поиск все новых языковых признаков с помощью современных технологий анализа естественного языка. Наша работа пойдет по второму пути.

Традиционно исследование на основе машинного обучения состоит из трех основных этапов: сбор данных для обучения, сбор и обработка признаков из полученных данных, выбор и настройка параметров математической модели. В случае модели на основе нейронных сетей этапов становится еще меньше: пропадает этап сбора признаков, поскольку модель сама анализирует данные и вычисляет некие признаки, на вход ей требуется лишь большая коллекция размеченных данных, однако такой метод стоит выбирать имея очень большие коллекции данных. Следовательно, перед нами стояли задачи сбора корпуса текстов, поиска релевантных признаков на основе полученного корпуса и обучение модели на этих признаках. Далее мы остановимся подробнее на этих этапах работы.

### 1.2. Особенности определения сложности текста как иностранного

#### 1.2.1. Введение в проблему

Проблема определения сложности текста, безусловно, актуальна и с позиции методики преподавания языков. Подбор текстов для учебных пособий, поиск интересных неадаптированных текстов в интернете для занятий или самостоятельного чтения - всё это делает исследования в области автоматического определения сложности текстов не просто практически применимыми, а даже необходимыми.

Так, Sarah E. Schwarm и Mari Ostendorf говорят о проблеме билингвизма в американских школах: совместно с детьми, для которых родной язык - английский, обучается большое количество детей, для которых английский язык является вторым: в 2001-2002 учебном году в штате Вашингтон 72,215 студентов (что составляет 7.2% от всех студентов) обучались по программе для ограниченного владения английским языком (Limited English Proficient, LEP) (Schworm, Ostendorf, 2005). Легко предположить, что с каждым годом количество таких студентов только увеличивается. Поэтому поиск подходящих материалов для чтения исходя из знаний и уровня владения языком, а также необходимость разработки методов донесения одной информации для разных категорий учащихся, становится все более актуальной проблемой.

Интересный в этом смысле пример представляет собой испано- и англоязычный проект NewSela[[3]](#footnote-3), где преподаватель может найти тексты и задания к ним на одну и ту же тему, но разной степени сложности. "Every article, every assignment, at every level" - так сами авторы проекта описывают его миссию. "Now students can study the same content and learn at their own pace—while teachers save time and headaches" (Теперь учащиеся могут изучать одну и ту же тему в своем темпе - англ.).  
На сайте проекта можно выбрать тексты по тематикам, классам и словарному запасу учащихся. Обычно одна тема содержит 4-5 вариантов уровней. Ниже приведен пример изменения темы статьи от самого простого к сложному:

|  |  |  |
| --- | --- | --- |
| Количество слов | Уровень (grade) | Тема |
| 400 | 3 | "Audience dogs" can help students become better public speakers |
| 593 | 4 | For some students, public speaking is easier when a cute dog is listening |
| 752 | 6 | The key to giving a great public speech? Practice in front of a dog |
| 894 | 7 | These pooches help students perfect their public speaking skills |
| 1100 | 12 | Giving a speech can be terrifying — these dogs are here to help |

К сожалению, нам не удалось найти подробную информацию о самой методике адаптации статей: есть ощущение, что пока это делается вручную. Однако само появление и популярность подобных проектов еще раз говорит об актуальности проблемы.

Надо отметить, что в научной традиции не всегда есть четкое различие между проблемой сложности текстов на родном и иностранном языке. Действительно, в какой-то мере представления о сложном тексте на родном и иностранном языках будут совпадать (длинные слова и предложения, обилие предикатов и определений, причастные/деепричастные обороты кажутся очевидными признаками сложности), однако мы не уверены, что представления о простом тексте также совпадут у носителей и изучающих язык: первым очевидным отличием является *лексика*. Иностранный учащийся сочтет трудным текст, написанный в простой синтаксической форме только потому, что он не знает эти слова. Ниже будут приведены требования к овладению лексикой для каждого уровня, а пока лишь для сравнения скажем: учащийся уровня B1 должен знать около 2, 5 тысяч слов, что примерно соответствует словарному запасу русскоговорящего ученика первого класса школы.

Второе важное отличие - *грамматика и морфология*: понятия падежа, времени и вида глагола совершенно естественны для носителей, но могут представлять определенную сложность для иностранцев, изучающих русский язык. В педагогической практике автор не раз сталкивалась с ситуацией, когда иностранец долго "справляется" с простым словом в непривычной грамматической форме ("так, написал - это совершенный вид, значит он это сделал до конца, письмо написано"). Так, например, Heilman с коллегами в своей работе приводят результаты эксперимента, в ходе которого добавление грамматических признаков принесло бо́льший прирост точности в коллекции текстов как иностранных - 22% против 7% как родных (Heilman et al., 2007). Надо отметить, что данное исследование проводилось для английского языка и требует проверки для текстов на русском языке.

Наконец, существует значительная методологическая разница в изучении проблем сложности текста как родного и иностранного: это *наличие единой понятной шкалы уровней* владения языком и *документов, регламентирующих признаки текста* того или иного уровня (государственные стандарты и лексические минимумы), о чем пойдет речь в следующих двух разделах.

#### 1.2.2. Шкалы (единицы измерения)

Сложность текста складывается из массы параметров, часть из которых связана с самим текстом, другая же - с такими психологическими понятиями, как интерес к теме, мотивированность, знакомство с темой и т.д. Все это влечет за собой трудности при составлении тренировочного корпуса и ставит почти философский вопрос: по какой шкале измерить сложность текста? Большинство традиционных формул читабельности берет в качестве шкалы возраст читателя или класс (англ. grade), для которого текст признается подходящим. Однако вспомнив свой школьный класс мы сразу понимаем условность и неточность этого деления. Наглядно эту ситуацию иллюстрируют авторы проекта по персонификации метрик читабельности и подбору материалов для чтения Lexile: в своем видеоролике они показывают семью, которая пришла в магазин купить ребенку кроссовки, но для поиска подходящей пары родители используют не индивидуальный размер ноги ребенка, а его возраст. Сами авторы проекта предлагают абстрактный числовой индекс, который складывается из метрик текста и словарного запаса конкретного учащегося.

Кроме того, при шкале с классами сразу встает вопрос как измерять сложность текстов вне школьной системы, "взрослых" текстов.

Аниса Катинская в своем эксперименте по автоматическому измерению читабельности текста использует расширенный корпус Ивана Бегтина, однако с сокращением числа групп текстов до 3: " Говоря о шкале сложности, мы решили остановиться на самых «простых» оценках: простой и сложный или простой, средней сложности и сложный" (Селегей et al. 2015).

При анализе сложности иностранного текста обычно берется шкала уровней владения этим языком, уже разработанная методистами и авторами пособий. Мы не будем исключением и предлагаем принять за шкалу сложности текстов их соответствие общепринятым в методике уровням. В настоящее время создана следующая российская система сертификационных уровней общего владения русским языком как иностранным:

*1. А1 (ТЭУ: элементарный уровень)*

Успешное прохождение тестирования означает, что кандидат владеет минимальным уровнем знаний русского языка, достаточным для ограниченного числа ситуаций в повседневном общении. Также сертификат элементарного уровня означает, что кандидат имеет достаточные и необходимые знания для дальнейшего изучения языка и достижения следующего (базового) уровня общего владения русским языком.

*2. А2 (ТБУ: базовый уровень)*

Успешное прохождение тестирования означает, что кандидат владеет начальным уровнем знаний русского языка, достаточным для основных коммуникативных потребностей в ограниченном числе ситуаций бытовой и культурной сфер общения. Базовый уровень является минимальной базой для занятий какой-либо профессиональной деятельностью в ограниченном объеме.

Кроме того, для вступления в гражданство Российской Федерации необходимо подтвердить владение русским языком как иностранным языком на уровне не ниже A2.

*3. B1 (ТРКИ-I: первый сертификационный уровень)*

Успешное прохождение тестирования означает, что кандидат имеет средний уровень владения русским языком, который позволяет кандидату удовлетворить основные коммуникативные потребности в бытовой, учебной и профессиональной сферах общения в соответствии с государственным стандартом русского языка как иностранного языка. Сертификат данного уровня необходим для поступления в российское высшее учебное заведение.

*4. B2 (ТРКИ-II: второй сертификационный уровень)*

Успешное прохождение тестирования означает, что кандидат имеет достаточно высокий уровень владения русским языком, который позволяет кандидату удовлетворить коммуникативные потребности во всех сферах общения. Владение языком на данному уровне позволяет человеку вести профессиональную деятельность на русском языке в качестве специалиста в инженерно-технической, гуманитарной и естественнонаучной областях. Наличие данного сертификата необходимо для получения диплома бакалавра или магистра.

*5.C1 (ТРКИ-III: третий сертификационный уровень)*

Успешное прохождение тестирования означает, что кандидат может вести профессиональную деятельность на русском языке в качестве филолога, переводчика, редактора, журналиста, дипломата, менеджера, работающего в русскоязычном коллективе.

*6. C2 (ТРКИ-IV: четвертый сертификационный уровень)*

Успешное прохождение тестирования означает, что кандидат владеет русским языком на высоком уровне, сравнимым с уровнем носителя языка. Наличие данного сертификата является необходимым для получения диплома магистра-филолога, который дает право на все виды преподавательской и научно-исследовательской деятельности в области русского языка.

Российская многоуровневая система тестирования включена в европейскую структуру языкового тестирования ALTE. Уровни владения русским языком как иностранным соотносятся с уровнями владения другими европейскими языками следующим образом:



К преимуществам выбора шкалы сложности текста, основанной на общепринятой системе уровней владения языка, можно отнести:

1. Наличие регламентирующих документов, подробно описывающих требования к каждому уровню. Это поможет нам на этапе сбора признаков текста.

2. Свободу от таких субъективных категорий,каккласс/возраст/количество лет обучения. Данные уровни имеют под собой конкретный объем языкового материала, который должен знать человек, претендующий на сертификат соответствующего уровня.

3. Наличие маркировки на учебных пособиях, сообщающей нам, для какого уровня они предназначены. Это поможет нам на этапе сбора корпуса.

4. Соотнесенность с жизненными ситуациями (как мы видим, в описании уровней содержится информация о связи уровня владения с возможностями и нуждами реальной жизни - получить работу в России, получить Российское гражданство, получить разрешение на преподавание русского языка). Так уровень сложности становится чуть более объективной категорией.

#### 1.2.3. Регламентирующие документы

Для вышеуказанных уровней существуют специальные нормативные документы (**государственные стандарты владения русским языком как иностранным**), содержащие в себе минимальные обязательные требования, определяющие цели и содержание обучения на каждом конкретном уровне.

Подобные материалы могут быть очень ценны в изучении сложности текста, поскольку в них зафиксированы формальные требования к текстам (количество слов, процент незнакомой лексики), их тематика, уровень знания морфологии, грамматики и синтаксиса на этом уровне. Так, например, выглядит блок "Чтение" Государственного стандарта элементарного уровня владения русским языком (Владимирова, 2001):

Иностранец должен уметь:

— читать текст с установкой на общий охват его содержания; — определить тему текста;  
 — понять достаточно полно и точно основную информацию текста, а также некоторые детали, несущие важную смысловую нагрузку.

Вид чтения: чтение с общим охватом содержания, изучающее чтение.

Тип текста: специально составленные или адаптированные сюжетные тексты (на основе лексико-грамматического материала, соответствующего элементарному уровню).

Тематика текста актуальна для бытовой, социально-культурной и учебной сфер общения.

Объем текста: 250—300 слов.

Количество незнакомых слов: 1—2%.

Интересно проследить изменение требований, предъявляемых к тексту разных уровней:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень | Тип текста | Тематика текста | Объем текста | Количество незнакомых слов |
| А1 | **специально составленные или адаптированные сюжетные тексты** (на основе лексико-грамматического материала, соответствующего элементарному уровню) | актуальна для бытовой, социально-культурной и учебной сфер общения | 250—300 | 1—2% |
| А2 | сообщение, повествование, описание, а также тексты смешанного типа. **Специально составленные или адаптированные сюжетные тексты**, построенные на основе лексико-грамматического материала, соответствующего базовому уровню. | актуальна для сферы повседневного общения, социально-культурной и учебной сфер | 600—700 | 3—4 % |
| B1 | сообщение, повествование, описание, а также **тексты**  **смешанного** **типа с элементами рассуждения**. **Тексты аутентичные (допустима минимальная степень адаптации)** с учетом лексико-грамматического материала данного уровня | актуальна для социально-культурной сферы общения | 900-1000 | 5-7% |
| B2 | тексты описательного и повествовательного характера с элементами рассуждения и эксплицитно выраженной авторской оценкой; **художественный текст** повествовательного характера. | актуальна для социально-культурной, официально-деловой сфер общения | 300-600 | до 10% |
| C1 | полилог, дискуссия с элементами описания и повествования  в качестве аргументирующих элементов, содержащий эксплицитно и имплицитно выраженную оценку; интервью, содержащее элементы устной разговорной речи; текст информационно — описательного и информационно- регламентирующего характера (законы, постановления, информационные сообщения); художественный текст (рассказ, законченный фрагмент повести, романа и т.д.) | актуальна для социально-культурной, официально-деловой сфер общения | 400-750 | до 10% |

Несмотря на некоторую условность представленных данных (например, непонятна ситуация с изменениями объема текста), становится видна общая канва постепенного усложнения текстов и их тематики: переход от специально составленных текстов к аутентичным происходит на третьем уровне (В1). Тогда же добавляется текст с элементами рассуждения как более сложный тип по сравнению с повествованием. Меняется и тематика текстов: от бытовой и учебной, связанной с необходимостью быстрого освоения лексики для повседневных нужд (так называемый survival russian), к появлению на 4 уровне официально-деловой сферы. Также мы видим, что полноценное чтение художественной литературы больших объемов (роман, повесть) появляется лишь к 5 уровню. Тогда же впервые выдвигаются требования к понимаю различных государственных законов и документов.

Еще одним важным документом, регламентирующим уровни владения иностранным языком, является **лексический минимум**, т.е. список слов, знание которых необходимо для успешной сдачи сертификационного тестирования. Кроме того, мы видим, что эти документы взаимосвязаны: в рекомендациях государственного стандарта к текстам фигурирует фраза «с учетом лексико-грамматического материала данного уровня», что и означает вхождение в лексические минимумы.



Отбор лексических единиц производится по следующим критериям:

1) стилистическая немаркированность;

2) способность слова входить в различные словосочетания;

3) семантическая ценность2 (способность слова обозначать часто встречающиеся предметы и явления);

4) высокая словообразовательная способность слова;

5) частотность (при этом учитывались показатели частотности по частотным словарям, по использованию в учебниках РКИ, а также «тематическая» частотность).

Объем лексических минимумов растет вместе с уровнем владения:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Уровень | А1 | А2 | В1 | В2 | С1 | С2 |
| Количество словарных единиц | 591 | 1412 | 2565 | 5547 | 11000 | нет |

Третьим и последним типом документов, входящих в комплекс материалов Российской системы тестирования граждан зарубежных стран по русскому языку, являются **типовые тесты по русскому языку как иностранному**. В них содержатся примеры заданий и текстов, которые необходимо будет выполнить для сдачи экзамена на определенный уровень. Для нашего исследования такие тесты являются ценным источником пополнения обучающего корпуса, т.к. в объективности оценки их уровня не приходится сомневаться.

Таким образом, в этом разделе были представлены основные размышления по поводу возможной разницы подходов к проблеме сложности текста на родном и иностранном языках.   
Так, были озвучены предположения, что лексические и грамматические признаки текста для проблемы определения сложности текста как иностранного могут иметь бОльший вес, чем как родного.   
После обзора некоторых вариантов выстраивания шкалы сложности текста, был предложен свой вариант из 6 групп, связанный с основными уровнями владения русским языком как иностранным, от A1 до С2.  
Также мы рассмотрели основные документы, регламентирующие минимальный набор знаний, соответствующий тому или иному уровню языка, структуру этих документов и информацию, которая поможет нам на этапе формирования признаков: это длина текста в словах, процент незнакомой лексики, грамматические признаки.

# Глава 2. Описание работы по созданию и обучению предсказательной модели

### 2. 1. Сбор и описание корпуса

Важным этапом проведения эксперимента является сбор тренировочных данных, поскольку от них будет зависеть конечный результат. Приведем несколько примеров составления корпусов текстов, ранжированных по сложности. При использовании шкалы школьных классов логично обратиться к учебным материалам, в которых уже содержится информация о классе или возрасте. для которого они предназчены. Например, таким способом воспользовались Sarah Schwarm и Mari Ostendorf: их корпус состоит из более 2500 текстов еженедельного выпуска детского журнала Reader, где уже присутствует деление на грейды (школьные классы) читателей.

Вторым интересным способом является поиск текстов на одну и ту же тему, но с разным уровнем сложности. Так, для английской версии эксперимента в работе S. Sharoff (Sharoff et al, 2008) корпусом послужили параллельные статьи в упрощенной версии Википедии (Simple English Wikipedia) и обычной версии. Simple English Wikipedia - это раздел Википедии, созданный для читателей, которым по разным причинам тяжело понимать обычную английскую Википедию. Для работы с данным языковым разделом достаточно иметь базовые знания английского языка. Прежде всего она ориентирована на изучающих английский язык на начальном уровне и может быть использована во время занятий и для учебного чтения. Также Википедия на упрощённом английском может быть полезна детям для выполнения домашнего задания и для внеклассного чтения[1]. Как правило, статьи в этом разделе Википедии представляют собой значительно упрощённые и укороченные варианты статей английской Википедии. При написании статей в этот раздел, рекомендуется стараться использовать простой синтаксис и лексису, а также постараться использовать лексику из 1000 самых частотных слов английского языка.

Если же мы говорим о сборе корпуса текстов, размеченных по шкале уровней владения иностранным языком, пока почти единственным возможным способом мы видим ориентацию на маркировку учебных пособий. В качестве альтернативного варианта Robert Reynolds предлагает обратиться к текстам портала для изучения языков linqQ[[4]](#footnote-4): на этом сайте желающие могут добавлять понравившиеся им тексты и обмениваться ими с другими студентами. Материалом для корпуса, соответственно, тут может послужить оценка, которую поставил студент, загрузивший материал. Несмотря на то, что идея краудсорсинга нам очень близка, на данном этапе мы отказались от использования этих данных, т.к. а). есть слишком большой риск их субъективности б). там представлена шкала из трех уровней (начинающий, средний и продвинутый), нам же такая шкала кажется слишком размытой.

Для сбора корпуса были использованы тексты учебной текстотеки ЦМО[[5]](#footnote-5) МГУ и учебных пособий, в методической справке которых был указан уровень владения языком, для которого он предназначается. Для удобства пользования корпусом мы присвоили каждому уровню числовой эквивалент от 1 до 6. Таким образом мы собрали 6 уровней от A1-C2 (см. таблицу). В качестве эксперимента был введен еще уровень 0, который включает одиночные элементарные слова ("привет", "спасибо") и 7, т.е. представляющиеся сложными даже для носителей языка (тексты законов, программы политических партий). Далее мы посмотрим, помогут ли эти уровни в улучшении результатов.

|  |  |  |
| --- | --- | --- |
| **Числовой код**  **из корпуса** | **Уровень сложности текста** | **Пример (фрагмент текста)** |
| 0 | Одиночные элементарные слова и фразы русского языка | Меня зовут Антон.  Привет.  Спасибо. |
| 1 | A1 | Сегодня воскресенье. Мы отдыхаем. Папа читает журнал «Спорт». Он очень любит спорт, особенно футбол. Мама в свободное время любит готовить. Сегодня она готовит пельмени. |
| 2 | A2 | Все события фильма происходят 31 декабря. Герои фильма — молодые люди Настя и её жених Коля — готовятся встречать Новый год. У Насти никого нет, кроме Коли. Мама Насти умерла, а своего отца Настя никогда не знала. После смерти мамы Настя прочитала письмо, в котором мама рассказывала, что отца Насти зовут Павел, что познакомились они на море в 1969 году... |
| 3 | B1 | Георгий Гречко, лётчик-космонавт: «Моя мать работала главным инженером завода. Помню, как на следующий день после того, как она ушла на пенсию, она мне сказала: "Первый раз я спала спокойно". До этого она каждую ночь беспокоилась, не случилось ли что-нибудь на заводе, но если бы кто-нибудь предложил моей матери не работать, а только заниматься домашним хозяйством, она бы не согласилась — она любила свой завод, свою работу. |
| 4 | B2 | Всем присутствующим было предложено стать конструкторами придуманной организаторами «мастерской будущего», в которой бы прорабатывались основные проблемы, интересующие молодежь накануне XXI века. Выяснилось, что волнующие россиян и немцев темы во многом схожи. Это бюрократизм и закоснелость мышления, безработица, преувеличение роли денег, равнодушие людей, языковые и культурные барьеры между странами. |
| 5 | C1 | Независимая оценка знаний заключается в том, что студенты сдают экзамены независимым экспертам – преподавателям других вузов, которые не проводили занятия в этих группах. Процедура сдачи экзаменов традиционна – в форме собеседования и письменных ответов на вопросы билетов.  Идею независимой оценки качества образования поддержал Уполномоченный по правам студентов в РФ Артем Хромов, о чем сообщается на его официальном сайте. |
| 6 | C2 | Искусство составляет одну из высших сфер деятельности духа. Эта сторона искусства не есть что-то отвлеченное, мечтательное, безразличное; напротив, она заключает в себе многие области, заселенные живыми, конкретными образами. Чудная страна! В ее светлых садах высятся храмы и колонны, блестят чистые формы статуй, виднеются лица, весельем рдеют щеки, сверкают очи; и горы, и леса, и реки, и утро, и полдень, и вечер блещут отовсюду на живом полотне; не пуст, не безответен там воздух. |
| 7 | Сложные для носителей тексты | Сетевая форма реализации образовательных программ (далее - сетевая форма) обеспечивает возможность освоения обучающимся образовательной программы с использованием ресурсов нескольких организаций, осуществляющих образовательную деятельность, в том числе иностранных, а также при необходимости с использованием ресурсов иных организаций. |

Наш корпус представляет собой текстовый файл, в котором содержится информация об уровне текста, задан его порядковый номер, название пособия и представлен сам текст.

Корпус представляет собой 600 текстов. Он сбалансирован по количеству текстов, представляющих один уровень, следующим образом:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Уровень | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Количество текстов | 20 | 107 | 103 | 108 | 107 | 57 | 67 | 20 |

При составления корпуса мы столкнулись со следующими трудностями:

1. Размытость границ для текстов высоких уровней: многие пособия ориентированы на B1-B2, B2-C1, что затрудняет отнесение текста к одному определенному классу. Дело в том, что эти уровни предполагают уже серьезное владение общей лексикой и грамматикой, и обычно дальнейшие курсы выстраиваются исходя из потребностей студента (русский язык для медиков, для бизнеса, для военных), поэтому понятие сложности текста тут во многом связано с его тематикой.

2. Отсутствие пособий для С2. Поскольку С2 считается наивысшим уровнем владения иностранным языком, предполагается, что человек может читать и понимать практически любые оригинальные тексты общей тематики. Поэтому мы наполнили корпус для уровня текстами из различных научно-популярных и новостных изданий: Русский репортер, Вокруг Света, The Village и др.

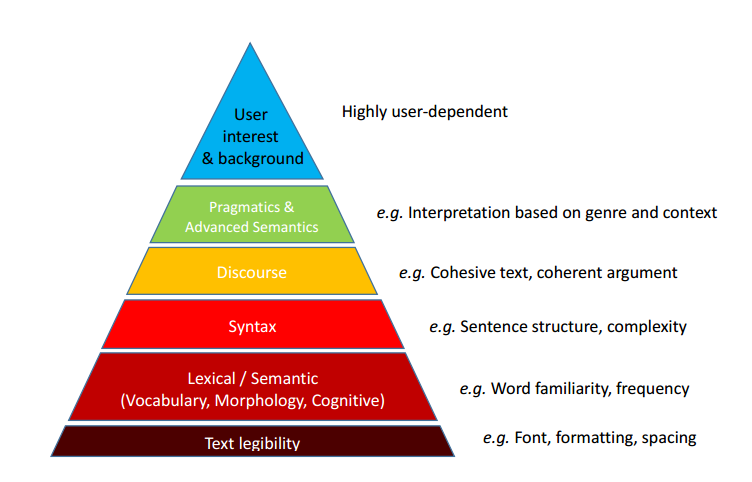
3. Субъективность оценки уровня авторами пособий: к сожалению, пока не существует системы единой независимой оценки уровня при подготовке пособия к печати: авторы вольны сами составлять методическое описание, поэтому представления об уровне рискует быть субъективным.

4. Необязательность и размытость сообщения об уровне: для некоторых пособий уровень не указан вообще или выражен нечетко ("для продвинутых студентов", "для начинающих", "пособие рассчитано на второй семестр первого года обучения")

Оставляя задел для дальнейшей работы скажем также, что идеальный корпус по нашему мнению должен быть построен не на оценках авторов пособий или энтузиастов, а на комбинации нескольких показателей доступности текста, таких как экспертная оценка, тесты на понимание, беседа по тексту, время чтения, результаты экспериментов с айтрекером и др.

### 2.2. Сбор признаков

Дальнейшим и ключевым этапом нашего исследования является анализ полученного корпуса и сбор признаков текстов. Для решения этой задачи был написан программный код на языке Python: текст проходит стадии деления на предложения, токенизации, морфологического анализа с помощью модуля Mystem и непосредственно подсчета признаков.

 Collins-Thompson в своем обзоре исследований, посвященных трудности текстов (Collins-Thompson 2014) предлагает интересную наглядную иллюстрацию классов признаков сложности текста, которые были выделены в ходе экспертной оценки текстов.

Так, первым типом являются все возможные поверхностные признаки текста, связанные с его *разборчивостью, визуальной читабельностью*: текст, шрифт, интервалы. Некоторые исследователи добавляют в этот тип присутствие иллюстраций и графиков. Вспоминая свой читательский опыт, каждый понимает важность этих очевидных признаков. Однако, как мы упоминали выше, наше исследование касается исключительно лингвистических признаков. Второй пласт занимают *лексико-семантические* (знакомость и частотность слов) и *морфологические* признаки (частотность/редкость морфологических форм). На признаках этого класса будет сосредоточен фокус нашего исследования. Далее идут *синтаксические* признаки текста (сложность предложений, особенности грамматической структуры), *дискурсивные* (различные риторические структуры). Следующий класс признаков Collins-Thompson назвал *семантикой высокого уровня*, отнеся сюда различную игру слов, идиомы, некий культурный контекст, необходимый для понимания текста. Сюда же он относит такое явление как сарказм. И, наконец, самый высокий пласт признаков представляет собой некую *персональную информацию* о читателе – его интерес к теме, мотивация к чтению этого текста, уже имеющиеся знания и опыт.

Эта пирамида описывает типы трудностей, т.е. компоненты сложности, которые отмечают респоденты при определении сложности того или иного текста. Очевидно, что они имеют прямую связь с признаками текста, на которых необходимо тренировать предсказательную модель. К непосредственно лингвистическим данным относятся классы признаков от лексики и морфологии до высокой семантики и прагматики.

Фокус нашего исследования сосредоточен во втором и отчасти третьем пластах пирамиды: прежде всего нас интересовала связь лексики, семантики и грамматики со понятием сложности.

Далее выбор признаков для нашего исследования производился исходя из следующих соображений:

1. Признаки должны отражать информацию, представленную в регламентирующих документах
2. На данных признаках должна подтвердиться или опровергнуться наша гипотеза о роли лексики и морфологии в определении сложности иностранного текста
3. Мы не отказываемся от популярных формул и базовых метрик текста и отслеживаем их вклад в сложность текста
4. Нас интересует роль вклад семантических категорий в понятие сложности текста
5. Признаки должны быть достаточны просты и воспроизводимы, поэтому на данном этапе мы отказались от использования синтаксического анализатора ABBYY COMPRENO

Таким образом сформировались следующие классы интересующих нас признаков:

* базовые метрики текста
* признаки на основе формул читабельности
* лексические
* грамматические
* семантические

Некоторая информация о синтаксисе также будет представлена в нашем исследовании, однако в косвенном виде (среднее число знаков пунктуации на предложение, средняя длина предложения, количество подчинительных союзов на предложение, доля существительных и глаголов на предложение), этого недостаточно чтобы говорить об отдельном классе признаков.

#### 2.2.1. Базовые метрики текста

Первая группа признаков представляет собой простые классические вычисления: интуитивно очевидно, что чем длиннее текст и чем больше слов в предложении, тем текст становится сложнее. Кроме того, мы имеем более конкретную информацию из государственных стандартов: например, текст для 1 уровня (А1) должен быть не больше 300 слов.

1. 'words' количество слов в тексте
2. 'sentences' количество предложений в тексте
3. 'mean\_len\_word' средняя длина слова в тексте
4. 'mean\_len\_sentence' средняя длина предложения

В эксперименте по автоматическому определению сложности русских текстов К.Дружкина и А.Катинской было высказано предположение, что медианная длина слова может быть более информативна, т.к. даже в сложных текстах будет находится много коротких слов. Мы проверим это предположение на практике и посчитаем медианные значения для слов и предложений:

1. 'median\_len\_word' медианная длина слова в тексте.
2. 'median\_len\_sentence' медианная длина предложения.
3. 'mean\_len\_word\_in\_syllabels ' средняя длина слова в слогах
4. 'percent\_of\_long\_words' процент слов длиннее 4 слогов

Величина №8 потребуется нам далее для вычисления формул читабельности. В английском варианте сложными считались слова длиннее 3 слогов. Однако мы пользуемся адаптацией формул к русскому языку И.Бегтина, где длинными считаются слова более 4 слогов.

#### 2.2.2. Признаки на основе формул читабельности

Вторую группу признаков составляют формулы читабельности (readability formulas), наиболее широко использующиеся в англоязычном мире для оценки сложности текстов. Эти формулы используют базовые метрики текста, вычисленную выше, в различных вариациях. Поскольку результатом их работы должно стать число, соответствующее возрасту и классу учащихся, для которого он предназначен, в формулах подбираются необходимые коэффициенты альфа и бетта. Мы же перед собой такой задачи не ставим, наш интерес составляет именно новый признак, полученный отношением простейших метрик текста между собой, поэтому мы не используем коэффициенты.

1. 'formula\_flesh\_kinc' Формула Флэша-Кинкайда [Kincaid et al., 1975]. Признаки: количество слогов, количество слов, количество предложений.
2. 'formula\_coleman' Формула Колман-Лиау [Coleman, Liau, 1975]. Признаки: количество букв, количество слов, количество предложений.
3. 'formula\_senter' Automated Readability Index [Senter, Smith, 1967]. Признаки: количество букв, количество слов, количество предложений.
4. 'formula\_dale' Формула Дэйла-Чалл [Dale, Chall, 1948].

Признаки: количество “сложных слов” (не входящих в список из 3000 простых слов), общее количество слов, количество предложений. (В русской адаптации количество “сложных слов” было заменено на количество “длинных” слов, имеющих 4 слога или более).

1. 'formula\_smog' Формула SMOG (Simple Measure of Gobbledygook) [Mc Laughlin, 1969].

Признаки: количество слов длиной более N слогов, количество предложений (для английского N = 3, в русской адаптации N = 4).

#### 2.2.3. Лексические признаки

Третья большая группа признаков - лексические. Тут в качестве признака выступает доля слов в тексте, относящихся к определенным интересующим нас спискам: первая подгруппа - это лексические минимумы, т.е. зафиксированный методистами список слов, которые студент должен знать на определенном уровне. К сожалению, такие минимумы существуют только до уровня B2. Эвристика проста: чем больше в тексте доля знакомых слов - тем он проще и понятнее.

1. 'A1' доля слов, входящих в лексический минимум уровня А1.
2. 'A2' доля слов, входящих в лексический минимум уровня А2.
3. 'B1' доля слов, входящих в лексический минимум уровня B1.
4. 'B2' доля слов, входящих в лексический минимум уровня B2.

Вторую подгруппу лексических признаков составляет доля слов, входящих в списки Частотного словаря современного русского языка (Ляшевская О. Н., Шаров С. А., 2009).

1. 'fr100' доля слов, входящих в список 100 самых частотных слов
2. 'fr300' доля слов, входящих в список 300 самых частотных слов
3. 'fr500' доля слов, входящих в список 500 самых частотных слов
4. 'fr1000' доля слов, входящих в список 1000 самых частотных слов
5. 'fr3000' доля слов, входящих в список 3000 самых частотных слов
6. 'fr5000' доля слов, входящих в список 5000 самых частотных слов
7. 'fr10000' доля слов, входящих в список 1000 самых частотных слов
8. 'fr20000' доля слов, входящих в список 20000 самых частотных слов
9. 'fr\_33000' доля слов, входящих в список 33000 самых частотных слов

Необходимо уточнить, что для подсчета лексических признаков мы пользовались предобработанными списками слов текста: из них были исключены имена, фамилии, геоназвания и слова, неизвестные морфологическому анализатору (скорее всего они с опечаткой), так мы исключаем возможность низкой доли знакомых слов просто потому, что в тексте много разных имен и названий.

#### 2.2.4. Грамматические признаки

Для подсчета грамматических признаков была использована программа Mystem. Считалась доля того или иного грамматического признака

1) от всех слов текста

2)от всех слов предложения (обозначены постфиксом "\_sentence")

Ожидается, что доля определенных грамматических признаков будет коррелировать со сложностью текста: наиболее очевидный пример - число причастий и деепричастий, количество предлогов и союзов в тексте, число косвенных падежей. Так, в работе К. Дружкина (Дружкин 2016) одним из параметров, сильно коррелирующих со сложностью текста стала доля родительного падежа. Скорее всего это объясняется спецификой его употребления: обычно этот падеж используется в словосочетаниях для выражения партитивных, притяжательных функций и распространяет, а, следовательно, нагружает предложение (ср. 'Я выпил чай' -> 'Я выпил чашку чая', 'Это дом'-> 'Это дом моего отца')

1. ‘A’
2. ‘ADV’
3. ‘ADVPRO’
4. ‘ANUM’
5. ‘APRO’
6. ‘COM’
7. ‘CONJ’
8. ‘INTJ’
9. “NUM’
10. ‘PART’
11. ‘PR’
12. ‘S’
13. ‘SPRO’
14. ‘V’
15. ‘наст’
16. ‘непрош’
17. ‘прош’
18. ‘вин’
19. ‘дат’
20. ‘зват’
21. ‘им’
22. ‘местн’
23. ‘парт’
24. ‘пр’
25. ‘род’
26. ‘твор’
27. ‘ед’
28. ‘мн’
29. ‘деепр’
30. ‘изъяв’
31. ‘инф’
32. ‘пов’
33. ‘прич’
34. ‘кр’
35. ‘полн’
36. ‘притяж’
37. ‘1-л’
38. ‘2-л’
39. ‘3-л’
40. ‘жен’
41. ‘муж’
42. ‘сред’
43. ‘несов’
44. ‘сов’
45. ‘действ’
46. ‘страд’
47. ‘неод’
48. ‘од’
49. ‘нп’
50. ‘пе’
51. ‘A\_sentence’
52. ‘ADV\_sentence’
53. ‘ADVPRO\_sentence’
54. ‘ANUM\_sentence’
55. ‘APRO\_sentence’
56. ‘COM\_sentence’
57. ‘CONJ\_sentence’
58. ‘INTJ\_sentence’
59. “NUM\_sentence’
60. ‘PART\_sentence’
61. ‘PR\_sentence’
62. ‘S\_sentence’
63. ‘SPRO\_sentence’
64. ‘V\_sentence’
65. ‘наст\_sentence’
66. ‘непрош\_sentence’
67. ‘прош\_sentence’
68. ‘вин\_sentence’
69. ‘дат\_sentence’
70. ‘зват\_sentence’
71. ‘им\_sentence’
72. ‘местн\_sentence’
73. ‘парт\_sentence’
74. ‘пр\_sentence’
75. ‘род\_sentence
76. ‘твор\_sentence’
77. ‘ед\_sentence’
78. ‘мн\_sentence’
79. ‘деепр\_sentence’
80. ‘изъяв\_sentence’
81. ‘инф\_sentence’
82. ‘пов\_sentence’
83. ‘прич\_sentence’
84. ‘кр\_sentence’
85. ‘полн\_sentence’
86. ‘притяж\_sentence’
87. ‘1-л\_sentence’
88. ‘2-л\_sentence’
89. ‘3-л\_sentence’
90. ‘жен\_sentence’
91. ‘муж\_sentence’
92. ‘сред\_sentence’
93. ‘несов\_sentence’
94. ‘сов\_sentence’
95. ‘действ\_sentence’
96. ‘страд\_sentence’
97. ‘неод\_sentence’
98. ‘од\_sentence’
99. ‘нп\_sentence’
100. ‘пе\_sentence’

Их полную расшифровку можно найти на сайте Яндекса[[6]](#footnote-6).

#### 2.2.5. Семантические признаки

Признаки, связанные с семантикой слова, представляют собой долю слов, имеющихся в семантических списках, полученных с помощью лингвистического модуля ABBYY COMPRENO. Ещё в работе Ю.А. Томиной (Томина 1985), посвященной объективной оценке языковой трудности текстов, высказывалась мысль, что одним из факторов, влияющим на трудность текста, могут быть показатели абстрактности слова.

К показателям абстрактности слова исследовательница относит:

• абстрактные суффиксы (ость, есть, из, стви, чин, щин, ствие, от(а), ение и мн. др)

• трехбалльная классификация существительных:

* предметный (e.g. карандаш, собака)
* чувственный (e.g. сила, спектр, скорость)
* нечувственный (e.g.закон, протон, условие)

• к абстрактным также относятся слова-термины и научные понятия.

Взяв за основу её предположение, мы взяли списки слов из семантической иерархии ABBYY COMPRENO с семантемами (своеобразными семантическими метками), характеризующими существительные с точки зрения абстрактности/конкретности.

Получилось 4 списка:

'lex\_physical' содержит существительные, обозначающие конкретные материальные объекты, включая людей (e.g. 'котлета', 'стол', 'мама'); 'lex\_virtual' - вполне конкретные, но виртуальные, нематериальные объекты (e.g. 'база', 'интернет'), 'lex\_abstract' - самый большой список, со держащий в себе различные абстрактные понятия, включая термины (e.g. 'авангардизм', 'блажь', 'сглаживание') , и, наконец, в 'lex\_substance' собраны обозначения субстанций(e.g. 'серебро', 'уксус').

Безусловно, это несколько грубое понимание шкалы абстрактности: тут никак не будет работать снятие омонимии ('секрет' есть и в абстрактных понятиях, и как обозначение субстанции), некоторые элементы списка вызывают сомнения ('сидр' относится к субстанции, хотя в нашем контексте исследования напитки, как и еду, логичнее было бы считать физическими объектами), но все же общие закономерности мы надеемся увидеть.

1. 'lex\_physical' доля существительных, входящих в список физических объектов, среди всех существительных текста.
2. 'lex\_virtual' доля существительных, входящих в список виртуальных объектов, среди всех существительных текста.
3. 'lex\_abstract' доля существительных, входящих в список абстрактных понятий, среди всех существительных текста.
4. 'lex\_substance' доля существительных, входящих в список обозначения различных субстанций, среди всех существительных текста.

### 2.3. Отбор релевантных признаков

Далее нам необходимо было измерить для каждого признака степень его корреляции со сложностью текста. В качестве меры был использован коэффициент корреляции Пирсона. Эта величина изменяется от -1 до +1. Чем она ближе к нулю, тем меньше наблюдается связь признака со сложностью. Положительный коэффициент говорит о положительной корреляции (с ростом признака растет и уровень сложности текста), отрицательный – наоборот (с ростом признака сложность текста падает).

Расчеты проводились c помощью модуля для статистических расчетов scipy.stats на Python.

Сначала мы проследим распределение признаков внутри каждого класса:

#### 2.3.1. Подсчет корреляции базовых метрик текста.

|  |  |
| --- | --- |
| Признак | Коэффициент корреляции |
| mean\_len\_sentence | 0.6541 |
| median\_len\_sentence | 0.6157 |
| percent\_of\_long\_words | 0.5932 |
| mean\_len\_word\_in\_syllables | 0.5490 |
| words | 0.5369 |
| mean\_len\_word | 0.5332 |
| median\_punct\_per\_sentence | 0.4731 |
| sentences | 0.4150 |
| median\_len\_word | 0.3707 |
| **Таблица 1. Степень корреляции  традиционных метрик текста** | |

Как мы видим из Таблицы 1, почти все признаки показали достаточно высокую корреляцию со сложностью, при этом лучшими оказались метрики предложения, средние значения показали бо́льшую корреляцию, нежели медианные. Все коэффициенты положительные, т.е. с ростом признака растет и сложность.

#### 2.3.2. Подсчет корреляции признаков на основе формул читабельности.

|  |  |
| --- | --- |
| Признак | Коэффициент корреляции |
| formula\_smog | 0.6863 |
| formula\_senter | 0.6653 |
| formula\_flesh\_kinc | 0.6596 |
| formula\_dale | 0.6565 |
| formula\_coleman | 0.5063 |
| **Таблица 2. Степень корреляции признаков на основе формул читабельности** | |

Формулы читабельности (Таблица 2) показали стабильно высокую положительную зависимость от уровня сложности текста. Лучший результат показала формула SMOG, которая рассчитывается на основе количества слов длиной более N слогов и количества предложений. Стоит заметить, что существенно коэффициент отличается лишь у формулы Колман-Лиау, остальные показатели очень близки между собой, что при необходимости позволяет нам оставить лишь одну из этих формул.

#### 2.3.3. Подсчет корреляции грамматических признаков.

|  |  |
| --- | --- |
| Признак | Коэффициент корреляции |
| им | -0.5163 |
| страд | 0.4816 |
| прич | 0.4712 |
| PR | 0.4659 |
| сред | 0.4534 |
| род | 0.3836 |
| SPRO | -0.3722 |
| твор | 0.3712 |
| действ | 0.3488 |
| ед | -0.3391 |
| **Таблица 3. Корреляция грамматических признаков** | |

Грамматические признаки (Таблица 3) также показывают корреляцию со сложностью, однако по сравнению с первыми двумя классами, меньшую. Видно, что в топ-10 попали 3 признака, связанные с причастиями, что подтверждает наше предположение о связи сложности с причастными конструкциями. Доля признака PR(предлоги) вероятнее всего указывает на синтаксическую сложность предложения. Интересна отрицательная корреляция доли именительного падежа в тексте: скорее всего, это связано с бо́льшим количеством косвенных падежей, а, значит, распространением предложения. Ожидаемо вносит вклад в сложность доля косвенных падежей. Лидирует тут родительный падеж, который используется для выражения партитивных отношений, при согласовании с числами, а также может выражать значения субъекта и объекта действия при конструкциях с номинализацией ("решение правительства", "создание культурных центров" и т.д.), свойственных для публицистических и официальных текстов. Здесь стоит отметить, что в подобном исследовании для текстов русского языка как родного (Дружкин 2016) также наблюдается лидерство родительного среди других падежей. Еще одной интересной особенностью стало то, что все 10 лучших вариантов считают долю признака в тексте, т.е. наша гипотеза о роли доли некоторых грамматических категорий в предложении (им. падеж, количество глаголов в финитной форме, существительных на предложение) не подтвердилась.

#### 2.3.4. Подсчет корреляции лексических признаков.

|  |  |
| --- | --- |
| Признак | Коэффициент корреляции |
| inA2 | -0.8181 |
| inB1 | -0.7972 |
| inA1 | -0.7949 |
| inB2 | -0.7175 |
| infr3000 | -0.6452 |
| infr5000 | -0.6352 |
| infr10000 | -0.6148 |
| infr1000 | -0.5755 |
| infr500 | -0.5459 |
| infr300 | -0.5281 |
| **Таблица 5. Корреляция лексических признаков.** | |

Абсолютными рекордсменами по связи со сложностью стали лексические минимумы для всех уровней и списки слов с "серединной" частотой от 300 до 10000 слов. Все эти признаки имеют отрицательную корреляцию, что означает чем больше слов в тесте покрываются заданным списком слов, тем он проще.

#### 2.3.4. Подсчет корреляции семантических признаков.

|  |  |
| --- | --- |
| Признак | Коэффициент корреляции |
| lex\_abstract | 0.5243 |
| lex\_substance | 0.5221 |
| lex\_physical | -0.1823 |
| lex\_virtual | -0.0711 |
| **Таблица 4. Корреляция семантических признаков** | |

Вполне ожидаемо наблюдается корреляция со сложностью доли абстрактной лексики и субстанций. Неожиданно мал вклад физических объектов (гипотеза состояла в том, что чем больше конкретных материальных предметов - тем проще текст).

#### 2.3.5. Лучшие и худшие признаки: общий зачет.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | inA2 | 23 | infr20000 |
| 2 | inB1 | 24 | formula\_coleman |
| 3 | inA1 | 25 | страд |
| 4 | inB2 | 26 | median\_punct\_per\_sentence |
| 5 | formula\_smog | 27 | прич |
| 6 | formula\_senter | 28 | PR |
| 7 | formula\_flesh\_kinc | 29 | сред |
| 8 | formula\_dale | 30 | sentences |
| 9 | mean\_len\_sentence | 31 | род |
| 10 | infr3000 | 32 | SPRO |
| 11 | infr5000 | 33 | твор |
| 12 | median\_len\_sentence | 34 | median\_len\_word |
| 13 | infr10000 | 35 | infr\_all |
| 14 | percent\_of\_long\_words | 36 | действ |
| 15 | infr1000 | 37 | ед |
| 16 | mean\_len\_word\_in\_syllables | 38 | неод |
| 17 | infr500 | 39 | 1-л |
| 18 | words | 40 | деепр |
| 19 | mean\_len\_word | 41 | изъяв |
| 20 | infr300 | 42 | кр |
| 21 | lex\_abstract | 43 | несов |
| 22 | им | 44 | пр |
| **Таблица 5. Отобранные признаки с коэффициентом  корреляции выше 0.3** | | | |

Таким образом были отобраны лучшие признаки с коэффициентом корреляции выше 0.3). Из 129 признаков таких оказалось 44. Лидирующие позиции в этом списке заняли лексические признаки (лексические минимумы и частотные списки слов), что подтверждает нашу теорию о большом влиянии лексики на сложность иностранного текста. Среди частотных списков наиболее информативными оказались «медианные» списки, от 300 до 10000 слов, слишком маленькие и слишком большие оказались не так эффективны. Большую корреляцию также показали формулы читабельности. Грамматические признаки также доказали свой вклад в понятие сложности текста, они составили 17 из 44, хотя их коэффициенты оказались и не так велики.

### 2.4. Построение предсказательной модели.

Завершающим этапом нашей работы является построение модели машинного обучения, которая получает на вход признаки текста и делает предположение о его уровне сложности. Код выполнен на языке Python, все основные расчеты производятся с помощью библиотеки sklearn.

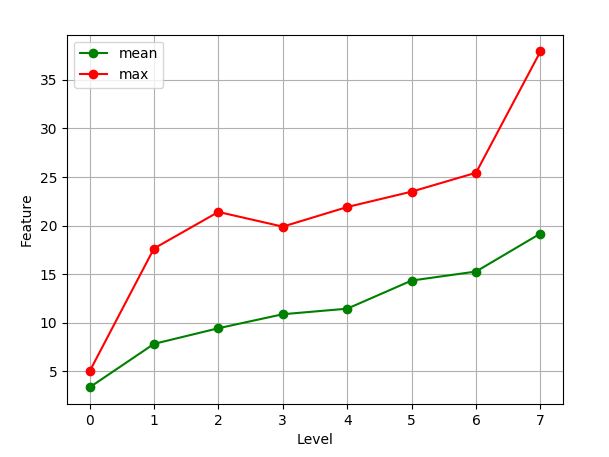
В начале эксперимента предполагалось, что предсказательная модель будет строится на основе линейной регрессии. Во-первых, это совпадает с нашим видением сложности как шкалы, а не закрытого набора классов. Процесс овладения иностранным языком, а, следовательно, и повышение уровня подходящих текстов носит линейный характер: чем больше слов из текста мы знаем, чем длиннее становятся слова и предложения, тем сложнее становится текст, и происходит это постепенно. Иллюстрацией тут может служить Рисунок 1., на котором изображена линейная зависимость сложности текста от средней длины предложения.

Рисунок 1. Распределение средних и максимальных значений для признака средней длины предложения в тексте

Во-вторых, сам результат работы регрессионной модели нам кажется более полезным и информативным для данной задачи и дальнейшего анализа ошибок: несмотря на то, что модель обучается на корпусе с уровнями, обозначенными целыми числами, результат представляет собой дробное число. Таким образом, можно предположить, что текст с результатом 3.7 сложнее текста с результатом 3.1. Если же использовать модель классификации, мы узнаем лишь к какому целому классу система отнесла проверяемый текст.

Однако в ходе анализа признаков были отмечены особенности, которые необходимо учесть в дальнейшей работе. Во-первых, становится очевидно, что полученные признаки во многом коррелируют между собой: частотные списки, формулы читабельности, традиционные метрики текста. Например, очевидно, что рост признака ‘дейст’ будет линейно связан с ростом доли причастий в тексте. Учитывая этот факт, мы добавили в эксперимент модель гребневой регрессии (Ridge Regression), которая хорошо справляется с проблемой мультиколлинеарности признаков. Гребневая регрессия представляет собой одну из техник регуляризации, которые применяются в линейных моделях классификация и регрессии для того, чтобы решить проблемы зависимости признаков друг от друга и переобучения. Для этого вводится дополнительное штрафное слагаемое к основному функционалу регрессии, которое штрафует за избыточное увеличение нормы вектора коэффициентов. Т.е. помимо традиционной оценки по методу наименьших квадратов гребневая регрессия опирается еще и на квадрат нормы весов, который должен быть минимизирован.

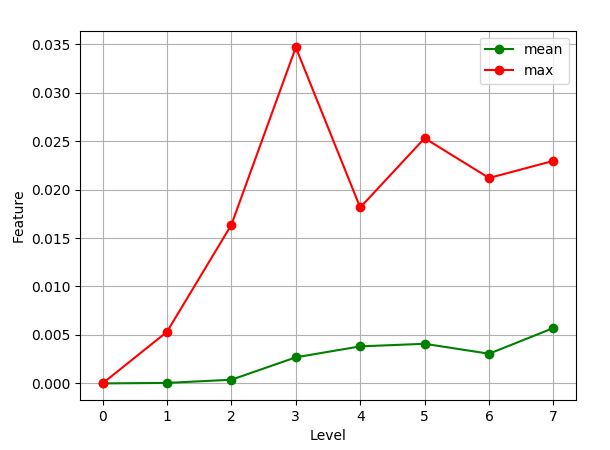
 Второй особенностью является возможная нелинейность грамматических признаков. Так, на Рисунке 2 показан график распределения доли деепричастий в тексте, при этом показаны средние и максимальные значения. По графику становится ясно, что до 3 уровня студенты вообще не знают деепричастия, далее они изучают эту тему и, по всей вероятности, активно тренируют её, отсюда берется скачок максимального значения, далее график выравнивается. В целом мы все же видим рост доли в зависимости от уровня, но подобные факты заставляют нас также осторожнее относится к выбору модели машинного обучения.

Рисунок 2. Распределение средних и максимальных значений доли деепричастий в тексте

Ниже представлены результаты эксперимента по построению модели линейной регрессии и гребневой регрессии со следующими изменяемыми параметрами:

* Обучение проведено на всем корпусе.
* Обучение проведено на корпусе без 0 уровня.
* Обучение проведено на корпусе без 7 уровня.
* Обучение проведено на корпусе без 0 и 7 уровня.

При этом модель строилась:

* Учитывая все признаки (129).
* Учитывая 44 признака с корреляцией более 0.3, отобранных в предыдущей главе.

Деление на обучающие и тестовые данные выполняется автоматически с помощью модуля StratifiedShuffleSplit: случайным образом корпус делится из расчета 80% для обучения модели, 20% - для тестирования. Подобная методика помогает объективно оценить успешность модели вообще, а не на конкретном комплекте тестов. Поскольку при каждом запуске прогона генерируется новое соотношение обучающих и тестовых данных, результат может отличаться, поэтому в таблице мы приводит средний результат от трех прогонов.

В качестве метрик оценки качества мы используем стандартные метрики из модуля sklearn.metrics:

1. Variance score является метрикой успешности регрессионной модели и представляет собой отношение квадрата стандартного отклонения ошибки к квадрату стандартного отклонения правильного ответа. Наивысшее значение - единица.

2. Mean squared error (средний квадрат ошибки) представляет собой среднее значение квадратов всех расстояний предсказания от правильного ответа. Поэтому в идеале должна быть нулем.

3. Mean absolute error (среднее значение ошибки). Это среднее расстояние модуля предсказания от правильного ответа, мера также должна стремиться к нулю.

4. Кроме того, таблица содержит пункт "Завышение/занижение", который отражает анализ тенденции ошибок модели завышать или занижать данные. На начальных этапах исследования нами было замечено, что система склонна занижать уровень текста и даже показывать отрицательные значения. Для борьбы с этим явлением в корпус были введены дополнительные уровни 0, который содержит отдельные фразы русского языка (например, "привет", "меня зовут Маша" и др. простейшие конструкции) и 7 (тексты, заведомо сложные даже для носителей). Подробнее об этом можно прочитать в главе "Сбор и описание корпуса".

|  |  |  |  |
| --- | --- | --- | --- |
| Признаки  Метрики | | Среднее из 3 прогонов  на всех признаках | Среднее из трех прогонов на 44 отобранных признаках |
| Прогон на всем корпусе | | | |
| Linear Regression | Variance score | 0.68 | 0.79 |
| Mean squared error | 0.98 | 0.59 |
| Mean absolute error | 0.78 | 0.59 |
| Завышение/занижение | 102/-125 | 91/-97 |
| Ridge Regression | Variance score | 0.82 | 0.82 |
| Mean squared error | 0.51 | 0.56 |
| Mean absolute error | 0.56 | 0.57 |
| Завышение/занижение | 91/-100 | 87/-94 |
| Прогон без 0 уровня | | | |
| Linear Regression | Variance score | 0. 73 | 0.82 |
| Mean squared error | 0.67 | 0.49 |
| Mean absolute error | 0.84 | 0.56 |
| Завышение/занижение | 105/-102 | 82/-96 |
| Ridge Regression | Variance score | 0.83 | 0.84 |
| Mean squared error | 0.49 | 0.46 |
| Mean absolute error | 0.55 | 0.55 |
| Завышение/занижение | 93/-83 | 80/-93 |
| Прогон без 0 и 7 уровней | | | |
| Linear Regression | Variance score | 0.63 | 0.8 |
| Mean squared error | 0.93 | 0.58 |
| Mean absolute error | 0.79 | 0.54 |
| Завышение/занижение | 128/-97 | 80/-88 |
| Ridge Regression | Variance score | 0.77 | 0.82 |
| Mean squared error | 0.54 | 0.49 |
| Mean absolute error | 0.59 | 0.54 |
| Завышение/занижение | 95/-84 | 77/-87 |
| Прогон без 7 уровня | | | |
| Linear Regression | Variance score | 0.51 | 0.82 |
| Mean squared error | 1.5 | 0.59 |
| Mean absolute error | 0.9 | 0.55 |
| Завышение/занижение | 150/-106 | 93/-86 |
| Ridge Regression | Variance score | 0.78 | 0.82 |
| Mean squared error | 0.61 | 0.5 |
| Mean absolute error | 0.60 | 0.53 |
| Завышение/занижение | 89/-98 | 79/-97 |

Итак, мы видим, что лучший результат показала модель гребневой регрессии (Ridge Regression) с использованием отобранных 44 признаков, обучившаяся на корпусе без 0 уровня.

Мы видим, что при добавлении в корпус 0 и 7 искусственных уровней предсказательная модель чаще ошибается в сторону занижения уровня (отмечено голубым цветом). В остальных случаях можно заметить тенденцию, что при обучении на всех признаках модель склонна завышать свое предположение (отмечено розовым), а при использовании отобранных признаков - занижать.

Хорошо видно, что отбор признаков положительно влияет на результат: по обеим моделям лучшие результаты получились в прогонах с использованием только 44 хороших признаков. При этом стоит отметить, что гребневая регрессия в целом дает гораздо более стабильные и высокие результаты, чем обычная линейная регрессия. Мы предполагаем, что это связано с ее устойчивостью к проблеме мультиколинеарности признаков, о которой речь шла в начале главы.

Кроме того, стоит обратить внимание на прогон линейной регрессии без 7 уровня со всеми признаками: там наблюдается худший за эксперимент результат, высокий квадрат ошибки, сильное завышение результата: это свидетельствует о нестабильности модели линейной регрессии при использовании неотобранных признаков.

### 2.5. Примеры работы предсказательной модели

Ниже приведены примеры результата работы предсказательной модели с лучшими параметрами по результатам эксперимента: модели гребневой регрессии с отобранными признаками, обученной на корпусе без искусственного нулевого уровня.

#### 2.5.1. Пример №1. Элементарный уровень

*Утро в городе.*

*Рамону очень нравится Петербург. Особенно он любит утренний город. Почти каждый день он гуляет по городу и выбирает сюжет новой картины. В Петербурге очень много красивых мест и конечно, это непросто.*

*Сегодня он решил порисовать на Марсовом поле. Он выбрал очень хорошее место: напротив Нева и Троицкий мост, справа Летний сад, слева Мраморный дворец.*

*Город просыпается. На улицах ещё мало людей. Вот идут две девочки-школьницы и едят мороженое. На остановке трамвая стоит старушка, наверное, они едет на рынок. Автобусы, троллейбусы, трамваи ходят ещё редко. А вот молодой человек. У него большой чемодан. Он поднимает руку и останавливает такси. Интересно, куда он едет? В аэропорт? На вокзал? А вот ещё два молодых человека. Да это же Клаус! А кто второй? - Клаус, привет! Что ты здесь делаешь так рано? - Привет, Познакомься, это мой брат Георг. Сегодня он последний день в Петербурге. В 11 часов самолёт, а мы еще не всё посмотрели. - Ну как, Георг, тебе понравился Петербург? Где ты был? - Я в восторге! Два раза я ходил в Эрмитаж, был в Русском музее, видел Исаакиевсккй собор и Петропавловскую крепость. - А на Мойке, 12 ты был? - Конечно, мы не только ходили в музей-квартиру Пушкина, но и в город Пушкин ездили. Ходили там в лицей, видели комнату, где жил Пушкин. - А в Петродворце были? - К сожалению, нет. Я был в Петербурге только 3 дня. Но теперь я точно знаю, что должен увидеть этот город ещё раз. Извини, Рамон, нам надо идти. Через 30 минут наш автобус идет в аэропорт. - До свидания! Счастливого пути!*

Источник: "Жили-были"  
Уровень, заявленный в пособии: 1  
Предсказанный уровень: 1.19  
Информация о тексте:

Слов в тексте: 250  
Предложений в тексте: 38  
Средняя длина предложения: 6.57  
Доля слов, не входящих в лексические минимумы (т.е. потенциально незнакомых):  
A1: 0.232  
A2: 0.155  
B1: 0.081  
B2: 0.077  
Распределение существительных по семантическим классам:  
Физические объекты: 0.566  
Абстрактные объекты: 0.377  
Субстанции: 0.0  
Виртуальные объекты: 0.

Мы видим, что предсказанная сложность в целом адекватно отражает уровень текста, однако с небольшим завышением. Текст соответствует норме по длине предложений и количеству слов. Количество незнакомых слов в лексических минимумах ожидаемо падает, однако для заявленного уровня А1 оно составляет более 20 процентов, тогда как в государственном стандарте рекомендованное количество незнакомых слов 2-3%. При детальном рассмотрении становятся видны особенности входных данных и нашей системы, которые, вероятно, и являются причиной такой неточности.  
Итак, следующая лексика не была найдена в лексическом минимуме уровня А1 и была засчитана как незнакомая:

{'сад', 'непросто', 'останавливать', 'петропавловский', 'до', 'должный', 'дворец', 'напротив', 'летний', 'старушка', 'утренний', 'наверное', 'особенно', 'восторг', 'точно', 'через', 'по', 'последний', 'извинять', 'теперь', 'понравиться', 'надо', 'поле', 'чемодан', 'просыпаться', 'слева', 'рука', 'подымать', 'редко', 'крепость', 'мраморный', 'мост', 'справа', 'мойка', 'счастливый', 'стоять', 'выбирать', 'собор', 'путь', 'лицей', 'самолет', 'сожаление', 'марсовый', 'привет', 'порисовать', 'сюжет', 'же', 'ну', 'почти', 'эрмитаж', 'все', 'рынок'}

1. Возникают некоторые вопросы по содержанию лексических минимумов. Так, в минимуме уровня А1 нет слова "Привет", которое обычно изучается на первом занятии, однако есть слово "школьница", которое интуитивно кажется совсем не так необходимым в пособии для взрослых. Возможно, эти минимумы сложились исторически и требуют пересмотра.

2. Лексические минимумы не всегда последовательны: иногда лексика присутствует в минимуме более раннего этапа обучения, а затем пропадает. Например, та же "школьница" присутствует в минимумах А1, А2, В1 но пропадает из В2. Также из этого минимума пропали слова "редко", "точно", "интересно".

3. К минусам нашей системы на нынешнем этапе развития нужно отнести невозможность учесть словообразовательную информацию: так, слово "утренний" попадает в список незнакомых слов, однако слово "утро" в лексическом минимуме присутствует и, по информации государственного стандарта, учащийся уровня А1 уже владеет начальными знаниями об образовании отыменных прилагательных. Сюда же относятся пары глаголов (рисовать/порисовать, нравиться/понравиться).

4. Также ошибка может произойти и на этапе грамматического анализа текста модулем Mystem: так, начальная форма от "поднимает" получилась "подымать".

5. Автор пособия использует слово "старушка", которого нет ни в одном лексическом минимуме, при этом синоним "бабушка" известно уже на уровне А1. В дальнейшем можно решать подобные проблемы, подключив словарь синонимов и предлагать заменить слово на более частотный вариант.  
  
5. Также этот текст изобилует названиями различных достопримечательностей Петербурга, которых нет ни в лексических минимумах, ни в частотных списках. Это тоже делает общую оценку текста выше.

#### 2.5.2. Пример №2. Работа с текстами уровня В1.

Два текста из тестов для подготовки к государственному сертификационному тестированию по русскому языку на уровень В1.

Текст№1, раздел "Чтение". Задание: прочитайте текст — фрагмент из энциклопедии «Великие памятники природы». Выполните задания после него. Выберите вариант, который наиболее полно и точно отражает содержание текста.

*Байкал — древнейшее озеро на Земле: ему 20—25 миллионов лет. Глубина Байкала — 1620 метров. Таких глубоких озёр в мире больше нет. Когда Байкал спокоен, на глубине 40 метров видны разноцветные камни... Вода в нём пресная (несолёная) и очень холодная. Только в августе её температура поднимается до 15 градус в.*

*В народных песнях Байкал называют «славным морем». И это неудивительно. Его длину можно сравнить с расстоянием от Москвы до Петербурга (636 километров), хотя в мире есть озёра, гораздо большие по площади.*

*336 рек несут свои воды в Байкал, и только одна Ангара берёт своё начало в озере и несёт свои воды в Енисей, крупнейшую реку Сибири. Байкал — уникальное создание природы. Известно, что в озере имеется 600 видов растений и 1200 видов животных, из них 75 % встречается только здесь, в Байкале.*

*Байкал — озеро загадка. До сих пор учёные не могут понять, как по явилась в Байкале рыба из северных морей. Непонятно, как и почему в Байкале сохранились рыбы и растения, которые исчезли в других озёрах и морях.*

*Но Байкал не только загадочное озеро. Это одно из самых красивых озёр нашей планеты. И неудивительно, что об этом прекрасном и загадочном озере рассказывают легенды. Вот одна из них.*

*Много дочерей было у старого Байкала. Но особенно он любил красавицу Ангару. И решил Байкал никому не отдавать в жёны свою любимую дочь. Но услышала Ангара о прекрасном и сильном Енисее и захотела уйти к нему. Рассердился отец и по ставил на её пути высокие горы. Тогда все 336 сестёр Ангары помогли ей убежать к Енисею. Увидел это Байкал и бросил громадный камень, чтобы остановить её. Но Ангара убежала и нашла с Енисеем своё счастье. С того времени несёт она свои воды в Енисей. А камень, который бросил Байкал, и сейчас стоит на том же месте.*

После текста представлены 4 тестовых задания на понимание, например:

20. В легенде рассказывается ....

(A) об один оком Байкале

(Б) о помощи Ангары Енисею

(B) о побеге Ангары к Енисею

Текст №2, представленный в разделе "Говорение". Подразумевает ознакомительное чтение со словарем 15 мин и подготовка краткого пересказа.

*Я хочу рассказать, как одна встреча с великим человеком сыграла огромную роль в моей жизни.*

*Когда мне был о во семнадцать лет, я приехал в Петербург и поступил учиться в университет. Я всегда очень любил музыку. Моим любимым композитором был Пётр Ильич Чайковский. Поэтому в свободное время я часто ходил в оперный театр и с удовольствием слушал все оперы Чайковского, смотрел его балеты.*

*Однажды мои друзья пригласили меня в гости в одну семью. Хозяйка этого дома была прекрасной певицей и часто выступала на концертах. Я с удовольствием пошёл к ним. Это был для меня счастливый вечер, потому что в тот вечер к ним в гости пришёл Пётр Ильич Чайковский. Хозяйка дома пела арии из его опер. Петру Ильичу понравилось её пение. Чайковский говорил очень мягко и просто, все внимательно слушали его. Начался интересный разговор о музыке, о литературе, об искусстве.*

*Поздно вечером мы вместе с Чайковским вышли из дома, и он спросил, где живу. Узнав, что я живу недалеко от его дома, он предложил мне пойти пешком. Я был счастлив, ведь я не только познакомился с великим композитором, и мог поговорить с ним во время нашей прогулки.*

*Мы пошли по набережной реки Невы. Была прекрасная лунная ночь. Сначала мы шли молча. Потом Пётр Ильич спросил меня:*

*—Я слышал, что Вы хотите стать художником. Это правда?*

*—Да, — ответил я.*

*Мы помолчали, а потом я спросил его:*

*— Пётр Ильич, говорят, что гении создают свои произведения, пишут музыку, картины только в те минуты, когда они работают легко и свободно, как будто кто-то помогает им. В общем, когда к ним приходит вдохновение. Что Вы лаете об этом?*

*— Ах, молодой человек, не говорите глупости! Нельзя ждать вдохновения, должен прежде всего труд, труд и труд! Нужно не ждать вдохновения, а серьёзно работать, трудиться каждый день. Помните, молодой человек, одного вдохновения мало, даже гений или очень талантливый человек ничего не добьётся в жизни, не сделает ничего значительного, если не будет трудиться. Я, например, считаю, что я самый обыкновенный человек.*

*Я не согласился с ним и хотел поспорить, но он остановил меня и продолжил;*

*— Нет, нет, не спорьте, я знаю, что говорю. Советую Вам, молодой человек, запомнить на всю жизнь, что вдохновение приходит только к тому человеку, который серьёзно и много работает, вдохновение рождается только из труда и во время труда. Я каждое утро в 8 или 9 часов начинаю работать и пишу музыку. Если мне не нравится, что я написал сегодня, завтра я буду делать эту же работу, буду писать всё сначала. Так я пишу день, два, десять дней. Вы сможете сделать больше и лучше, чем талантливые, но ленивые люди.*

*— Значит, Вы думаете, что нет абсолютно неталантливых, неспособных людей?*

*— Я думаю, что таких людей не так много. Но есть очень много людей, которые не хотят или не умеют работать, и тогда они говорят, что у них сегодня нет вдохновения.*

*Когда мы остановились около дома, где жил Чайковский, я решил задать ему ещё один вопрос, который очень волновал меня.*

*— Я согласен с Вами, Пётр Ильич. Очень хорошо работать для себя и по своему желанию. Но что делать, если приходится писать то, что ты должен, работать по заказу? Если ты пишешь то, что тебе заказали и за что ты получаешь деньги?*

*Я задал этот вопрос, потому что думал о себе, о своих картинах, которые тогда писал только по заказу.*

*— Ну что ж, я сам часто работаю по заказу, — ответил Чайковский. — И это очень неплохо, иногда результат бывает даже лучше, чем когда работаешь по своему желанию. А вспомните великого Моцарта. Он часто писал музыку по заказу. Или таких художников, как Микеланджело, Рафаэль... Они тоже писали по заказу.*

*Мы попрощались. Чайковский ушёл. Я пошёл домой. Я шёл и думал о том, что сказал мне Чайковский. Слова Петра Ильича помогли мне найти свой путь в жизни.*

|  |  |  |
| --- | --- | --- |
|  | Текст №1 (Байкал) | Текст №2 (Чайковский) |
| Источник | Типовой тест по русскому языку как иностранному. Первый сертификационный уровень. Общее владение. Второй вариант. | Типовой тест по русскому языку как иностранному. Первый сертификационный уровень. Общее владение. Второй вариант. |
| Раздел теста | Чтение | Говорение |
| Уровень, заявленный в пособии | 3 | 3 |
| Предсказанный уровень | 2.8 | 2.6 |
| Слов в тексте | 279 | 615 |
| Предложений в тексте | 27 | 55 |
| Средняя длина предложения | 10.3 | 11.8 |
| Процент слов, не входящих в лексические минимумы (т.е. потенциально незнакомых): | | |
| A1 | 0.492 | 0.321 |
| A2 | 0.291 | 0.193 |
| B1 | 0.177 | 0.107 |
| B2 | 0.102 | 0.088 |

Даже такие простые расчеты показывают при сравнении интересные результаты: текст №2 почти в два раза длиннее первого, однако оказывается ниже по шкале сложности. Дело в том, что первый текст призван проверить непосредственно навыки чтения, поэтому там задействована более трудная лексика и жанр статьи из энциклопедии. Второй же текст должен лишь дать материал для пересказа, поэтому должен быть понятен, интересен и содержать идею для дискуссии.

#### 2.5.3. Пример №3. Работа модели с неадаптированными текстами.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Источник текста | Уровень | Средняя длина предложения | Средняя длина слова | Процент слов, в лексиче ском минимуме В2 | Процент слов в 33000 частот ных слов | Процент абстрактной лексики |
| Народная сказка  "Маша и медведи" | 3.2 | 8.5 | 4.8 | 80% | 96% | 29% |
| Статья из блога про путешествия  (ок. 1 тыс. слов) | 3.9 | 12.17 | 5.1 | 82% | 96% | 60% |
| А.П.Чехов.  "Общее образование" | 4.1 | 11 | 4.8 | 78% | 94% | 44% |
| А.С.Пушкин. "Капитанская дочка" (отрывок ок. 3 тыс. слов) | 5.1 | 11.4 | 4.9 | 75% | 92% | 46% |
| Типовой договор на аренду квартиры | 5.5 | 9.4 | 6.3 | 63% | 89% | 77% |
| Л.Н. Толстой.  "Анна Каренина"  (отрывок ок. 3 тыс. слов) | 5.8 | 22.9 | 5 | 79% | 93% | 48% |
| Правила пользования московским метрополитеном | 6.5 | 10.2 | 6.8 | 67% | 94% | 66% |
| Алексей Навальный.  Расследование "Он вам не Димон"  (отрывок ок. 3 тыс. слов) | 6.6 | 15.6 | 6.2 | 71% | 90% | 48% |
| В.Набоков.  "Лолита"  (отрывок ок. 3 тыс. слов) | 6.9 | 23.4 | 5.5 | 71% | 91% | 54% |

Как мы видим из этой таблицы, тексты в неадаптированном и несокращенном виде должны появляться в обучении начиная с середины третьего уровня (B1), чаще же всего такие тексты оцениваются 5 уровнем (С1) и выше. Это соответствует информации, представленной в требованиях государственного стандарта: впервые упоминается о введении в обучение аутентичных текстов в стандарте 3 уровня, а к пятому уровню предъявляется требование понимать и адекватно интерпретировать тексты официально-деловой сфере общения, представленной текстами постановлений, законов, официальными сообщениями. Также на основании этих примеров мы можем предположить два основных направления усложнения текста: лексическое (как, например, в договоре на аренду квартиры или правилах пользования метрополитена), которое характеризуется высоким процентом абстрактной и редкой лексики, и грамматическое (как например, известное тяготение Толстого к длинным языковым конструкциям), выраженное, в основном длинной предложения.

### 2.6. Анализ ошибок

Мы проанализировали тексты, на которых наша программа ошибалась более чем на единицу (т.е. на уровень). В среднем таких текстов за один прогон оказывается 15%. Анализ показал, что больше всего трудностей вызывают тексты 4, 5 и 6 уровней. Предполагаем, это связано в первую очередь с недостатками нашего корпуса и особенностями обучения на таких высокий уровнях.

Для высоких уровней достаточно часто авторы пособий указывают два возможных уровня (В1-В2, В2-С1), либо указывают абстрактное "для студентов продвинутого уровня". Во многом они правы, т.к. становится сложно провести границу между уровнями: студенты достаточно свободно владеют грамматическим и лексическим инструментарием и дальнейшая работа заключается в совершенствовании стиля письма и говорения, практике чтения и обсуждения оригинальных текстов, поэтому сложность тут измеряется более сложными параметрами нежели на первых уровнях. Она во многом будет связана с жанровыми особенностями текста, его темой, соотнесенностью с интересами и специальностью студента.

Присутствует явная тенденция завышения сложности для текстов 5 уровня. Проанализировав повторно корпус, мы предполагаем, что это связано в первую очередь с типами пособий: во-первых, они сложны тематически и жанрово, содержат в себе много научных статей по специальности. Кроме того, на таких высоких уровнях уделяется внимание различным функциональным стилям русского языка, оборотам речи, фразеологии, поэтому иногда там наблюдается такая концентрация сложных оборотов русского языка, какую не всегда встретишь в обычной жизни.

Одновременно с этим явлением особенно интересна тенденция занижения уровня для текстов С2. Дело в том, что из-за дефицита корпусного материала, для уровня С2 мы использовали материалы русских СМИ разных тематических направленностей: политика, путешествия, науки, социальные исследования. Наша модель достаточно часто занижает уровень этих текстов до 4-5 уровня, но пока не будет проведена комплексная экспертная оценка этих текстов, трудно судить кто прав - субъективное мнение автора или математическая модель.

Основными путями исправления ошибок 5 и 6 уровней мы видим расширение корпуса и его ручную разметку несколькими респондентами.

# Заключение и направления дальнейшей работы

Данная работа посвящена поиску и отбору релевантных признаков для автоматического определения сложности текста как иностранного. В первой части работы мы остановились на теории и основных методах изучения проблемы сложности текста, а также сделали предположения о разнице составляющих понятия сложности для родного и иностранного текста.

Во второй главе мы проследили основные этапы работы над созданием предсказательной модели с использованием машинного обучения. Первым важным этапом был сбор корпуса обучающих данных, далее были выделены несколько типов признаков и рассмотрена их корреляция со сложность.

Полученные результаты доказывают существенную корреляцию признаков на основе лексических минимумов и частотных списков и формул читабельности со сложностью русских текстов как иностранных. Среди семантических признаков стоит выделить списки абстрактных слов и субстанций. Некоторые морфологические признаки также показали связь со сложностью - лидерами этой категории признаков стали именительный и родительный падежи, средний род, количество предлогов, а также признаки, связанные с долей причастий и деепричастий в тексте.

Заключительным этапом работы стал эксперимент по обучению предсказательной модели и сравнение результата её работы на всех признаках и только отобранных 44 с наилучшей корреляцией. Эксперимент показал, что лучшие результаты получила модель гребневой регрессии, обученная на отобранных признаках и на корпусе без 0 уровня.

В качестве основных направлений дальнейшей работы мы рассматриваем расширение корпуса, подключение синтаксической и словообразовательной информации, добавление новых семантических групп глаголов (глаголы движения, эмоциональная лексика), дальнейшая отладка методов машинного обучения. Среди первоочередных задач хотелось бы отметить добавление признаков, связанных с долей приставок и суффиксов - эти морфемы помимо прочего отвечают за образование видовых пар (рисовать - нарисовать, закончить - заканчивать), групп глаголов движения с приставками (перейти, пройти, зайти), пар одно- и разнонаправленных глаголов (везти - возить, лететь - летать), что традиционно представляет трудность для иностранцев, изучающих русский язык и, скорее всего, сильно отличается от представлений о сложности русского текста как родного. Также интересно было бы сравнить вклад вышеописанных признаков в сложность текстов на других языках.

# Приложение 1. Программный код для анализа текста и сбора признаков.

**import** re

**import** pymystem3

**from** collections **import** defaultdict

**import** json

**import** statistics

m = pymystem3.Mystem(entire\_input=False)

whole\_texts = open('text\_all.txt','r', encoding = 'utf\_8').read()

result\_list\_of\_features = open('result\_features\_text\_all.txt', 'w'

'', encoding='utf-8')

all\_features\_for\_all\_texts = dict()

## словники ##

slovnik\_A1 = open('slovnik\_A1.txt','r', encoding = 'utf\_8').readlines()

slovnik\_A1\_list = [f.replace('\n','') **for** f **in** slovnik\_A1]

slovnik\_A2 = open('slovnik\_A2.txt','r', encoding = 'utf\_8').readlines()

slovnik\_A2\_list = [f.replace('\n','') **for** f **in** slovnik\_A2]

slovnik\_B1 = open('slovnik\_B1.txt','r', encoding = 'utf\_8').readlines()

slovnik\_B1\_list = [f.replace('\n','') **for** f **in** slovnik\_B1]

slovnik\_B2 = open('slovnik\_B2.txt','r', encoding = 'utf\_8').readlines()

slovnik\_B2\_list = [f.replace('\n','') **for** f **in** slovnik\_B2]

## Списки частотных слов ##

fr\_100 = open('fr\_100.txt','r', encoding = 'utf\_8').readlines()

fr\_100\_list = [f.replace('\n','') **for** f **in** fr\_100]

fr\_300 = open('fr\_300.txt','r', encoding = 'utf\_8').readlines()

fr\_300\_list = [f.replace('\n','') **for** f **in** fr\_300]

fr\_500 = open('fr\_500.txt','r', encoding = 'utf\_8').readlines()

fr\_500\_list = [f.replace('\n','') **for** f **in** fr\_500]

fr\_1000 = open('fr\_1000.txt','r', encoding = 'utf\_8').readlines()

fr\_1000\_list = [f.replace('\n','') **for** f **in** fr\_1000]

fr\_3000 = open('fr\_3000.txt','r', encoding = 'utf\_8').readlines()

fr\_3000\_list = [f.replace('\n','') **for** f **in** fr\_3000]

fr\_5000 = open('fr\_5000.txt','r', encoding = 'utf\_8').readlines()

fr\_5000\_list = [f.replace('\n','') **for** f **in** fr\_5000]

fr\_10000 = open('fr\_10000.txt','r', encoding = 'utf\_8').readlines()

fr\_10000\_list = [f.replace('\n','') **for** f **in** fr\_10000]

fr\_20000 = open('fr\_20000.txt','r', encoding = 'utf\_8').readlines()

fr\_20000\_list = [f.replace('\n','') **for** f **in** fr\_20000]

fr\_33000 = open('fr\_33000.txt','r', encoding = 'utf\_8').readlines()

fr\_33000\_list = [f.replace('\n','') **for** f **in** fr\_33000]

##семантические списки##

lex\_physical = open('lex\_physical.txt','r', encoding = 'utf\_8').readlines()

lex\_physical\_list = [f.replace('\n','') **for** f **in** lex\_physical]

lex\_abstract = open('lex\_abstract.txt','r', encoding = 'utf\_8').readlines()

lex\_abstract\_list = [f.replace('\n','') **for** f **in** lex\_abstract]

lex\_substance = open('lex\_substance.txt','r', encoding = 'utf\_8').readlines()

lex\_substance\_list = [f.replace('\n','') **for** f **in** lex\_substance]

lex\_virtual = open('lex\_virtual.txt','r', encoding = 'utf\_8').readlines()

lex\_virtual\_list = [f.replace('\n','') **for** f **in** lex\_virtual]

modal\_words = ['хочется','нужно','надо','кажется','казаться','пожалуй','хотеть','должный',

'хотеться']

##делим документ на тексты и начинаем работу##

whole\_texts\_list = whole\_texts.split("<border>")

whole\_texts\_list\_len = len(whole\_texts\_list)

gram\_features = ['A', 'ADV', 'ADVPRO', 'ANUM', 'APRO', 'COM', 'CONJ', 'INTJ', 'NUM', 'PART',

'PR', 'S', 'SPRO', 'V', 'наст', 'непрош', 'прош', 'вин', 'дат', 'зват',

'им', 'местн', 'парт', 'пр', 'род', 'твор', 'ед', 'мн', 'деепр', 'изъяв',

'инф', 'пов', 'прич', 'кр', 'полн', 'притяж', '1-л', '2-л', '3-л', 'жен',

'муж', 'сред', 'несов', 'сов', 'действ', 'страд', 'неод', 'од', 'нп',

'пе']

##бьем на предложения##

**def** split\_into\_sentences(text):

caps = "([А-ЯA-Z])"

prefixes = "(Mr|St|Mrs|Ms|Dr|прим|т|д)[.]"

acronyms = "([A-Z][.][A-Z][.](?:[A-Z][.])?)"

websites = "[.](com|net|org|ru|gov|ру|орг)"

text = " " + text + " "

text = text.replace("\n"," ")

text = re.sub(prefixes,"\\1<prd>",text)

text = re.sub(websites,"<prd>\\1",text)

text = re.sub("\s" + caps + "[.] "," \\1<prd> ",text)

text = re.sub(caps + "[.]" + caps + "[.]" + caps + "[.]","\\1<prd>\\2<prd>\\3<prd>",text)

text = re.sub(caps + "[.]" + caps + "[.]","\\1<prd>\\2<prd>",text)

text = re.sub(" " + caps + "[.]"," \\1<prd>",text)

**if** "”" **in** text: text = text.replace(".”","”.")

**if** "\"" **in** text: text = text.replace(".\"","\".")

**if** "!" **in** text: text = text.replace("!\"","\"!")

**if** "?" **in** text: text = text.replace("?\"","\"?")

text = text.replace("...",".")

text = text.replace("!!!","!<stop>")

text = text.replace("?!","!<stop>")

text = text.replace(". ",".<stop>")

text = text.replace("?","?<stop>")

text = text.replace("!","!<stop>")

text = text.replace("<prd>",".")

sentences = text.split("<stop>")

**if** len(sentences) > 1:

sentences = sentences[:-1]

sentences = [s.strip() **for** s **in** sentences]

**return** sentences

**for** t **in** range(len(whole\_texts\_list)):

start\_one\_text = time.clock()

text\_number = str(t)

level = whole\_texts\_list[t][0]

**if** level == '\n':

**print**('ошибка в уровне(пустая строка)')

sentences = split\_into\_sentences(whole\_texts\_list[t]) # список строк из предложений

whole\_analized\_text = m.analyze(whole\_texts\_list[t]) #весь текст одним списком

whole\_lemmas\_list = []

noun\_list = []

bastard\_list = []

geo\_imen\_list = []

conj\_adversative\_list = [] #противительные союзы

modal\_words\_list = []

words\_length\_list = []

number\_of\_syllables\_list = []

long\_words\_list = [] #слова более чем 4 слога

dict\_of\_features = defaultdict(int)

**for** i **in** gram\_features:

dict\_of\_features[i] = 0

#считаем слоги и буквы

**def** count\_syllables(element):

i\_text = element.get('text')

i\_text\_syl\_counter = 0

syllables = ['а','е','ё','и','о','у','ы','ь']

**for** ii **in** i\_text:

**if** ii **in** syllables:

i\_text\_syl\_counter += 1

**if** i\_text\_syl\_counter == 0:

i\_text\_syl\_counter += 1

number\_of\_syllables\_list.append(i\_text\_syl\_counter)

words\_length\_list.append(len(i\_text))

**if** i\_text\_syl\_counter >= 4:

long\_words\_list.append(i\_text\_syl\_counter)

**return** True

#чистка от имен, гео объектов и бастардов

**def** clean\_from\_name\_geo\_bastard(element):

whole\_lemmas\_list.append(element.get('analysis')[0]['lex'])

gr\_info = element.get('analysis')[0]['gr']

**if** 'qual' **in** element.get('analysis')[0]:

**if** element.get('analysis')[0]['qual'] == 'bastard':

bastard\_list.append(element.get('text'))

**if** (gr\_info.find('имя') > 0 **or** gr\_info.find('гео') > 0):

geo\_imen\_list.append(element.get('analysis')[0]['lex'])

**if** (gr\_info.find('фам') > 0 **or** gr\_info.find('отч') > 0):

geo\_imen\_list.append(element.get('analysis')[0]['lex'])

**if** (element.get('analysis')[0]['lex'] == 'но' **or**

element.get('analysis')[0]['lex'] == 'а' **or**

element.get('analysis')[0]['lex'] == 'однако' **or**

element.get('analysis')[0]['lex'] == 'зато'):

conj\_adversative\_list.append(element.get('analysis')[0]['lex'])

**if** (element.get('analysis')[0]['lex'] **in** modal\_words):

modal\_words\_list.append(element.get('analysis')[0]['lex'])

**return** True

##подсчет грам. информации##

**def** count\_gram(element):

gr\_info = element.get('analysis')[0]['gr']

gr\_info = gr\_info.replace(',','<b>')

gr\_info = gr\_info.replace('=','<b>')

gr\_info = gr\_info.split('<b>')

**for** i **in** gram\_features:

**if** i **in** gr\_info:

dict\_of\_features[i] += 1

**if** 'S' **in** gr\_info:

noun\_list.append(element.get('analysis')[0]['lex'])

**return** True

##общий цикл просмотра анализа слов##

**def** gram\_analyze(element):

**for** i **in** element:

count\_syllables(i)

**if** len(i.get('analysis')) > 0:

clean\_from\_name\_geo\_bastard(i)

count\_gram(i)

**return** True

##подсчет медианного грам. признака на предложение##

**def** gram\_value\_per\_sentence(element, gr\_feature):

list\_of\_gram\_value\_per\_sentence = []

**for** i **in** element:

counter = 0

**for** ii **in** i:

**if** len(ii.get('analysis')) > 0:

gr\_info = ii.get('analysis')[0]['gr']

gr\_info = gr\_info.replace(',','<b>')

gr\_info = gr\_info.replace('=','<b>')

gr\_info = gr\_info.split('<b>')

**if** gr\_feature **in** gr\_info:

counter +=1

list\_of\_gram\_value\_per\_sentence.append(counter)

**return** len(list\_of\_gram\_value\_per\_sentence)

##считаем пунктуацию по предложениям##

**def** punctuation\_per\_sentence(element):

list\_punctuation\_score = []

punctuation = [',','-',':',';','—']

**for** i **in** sentences:

counter = 0

**for** ii **in** i:

**if** ii **in** punctuation:

counter += 1

list\_punctuation\_score.append(counter)

**return** list\_punctuation\_score

## Вычисляем процент слов из разных словников и частотных списков ##

**def** percent\_of\_known\_words(element, list\_of\_words):

known\_words = [w **for** w **in** element **if** w **in** list\_of\_words]

unknown\_words = [f **for** f **in** element **if** f **not** **in** list\_of\_words]

percent = len(known\_words)/len(element)

#print('�звестные слова: ', known\_words)

#print('Незнакомые слова: ', unknown\_words)

**return** percent

##запускаем функцию со всеми грам. анализами##

gram\_analyze(whole\_analized\_text)

##меняем значения в словаре с простых счетчиков на процент встречаемости в тексте##

**for** i **in** gram\_features:

dict\_of\_features[i] = dict\_of\_features[i] / len(whole\_lemmas\_list)

**for** i **in** gram\_features:

dict\_of\_features[i+'\_sentence'] = dict\_of\_features[i] / len(sentences)

clean\_lemmas\_list = [f **for** f **in** whole\_lemmas\_list **if** f **not** **in** geo\_imen\_list **and** f **not** **in** bastard\_list]

#clean\_lemmas\_list = list(set(clean\_lemmas\_list)) #список лемм без имен и гео

noun\_unic\_list = list(set(noun\_list)) #список уникальный сущ.

all\_words = len(whole\_analized\_text)

all\_sentences = len(sentences)

all\_syllables = sum(number\_of\_syllables\_list)

all\_letters = sum(words\_length\_list)

long\_words = len(long\_words\_list)

all\_len\_words = [len(f) **for** f **in** whole\_lemmas\_list]

all\_len\_sentences = [len(f.split(' ')) **for** f **in** sentences]

#цифры про текст

dict\_of\_features['words'] = (len(whole\_analized\_text)) #всего слов в тексте

dict\_of\_features['sentences'] = (len(sentences)) #всего предложений в тексте

dict\_of\_features['mean\_len\_word'] = (sum(words\_length\_list))/all\_words #средняя длина слова в тексте

dict\_of\_features['median\_len\_word'] = statistics.median(all\_len\_words)

dict\_of\_features['median\_len\_sentence'] = statistics.median(all\_len\_sentences)

dict\_of\_features['mean\_len\_sentence'] = all\_words/all\_sentences #средняя длина предложения в тексте

dict\_of\_features['mean\_len\_word\_in\_syllables'] = all\_syllables/all\_words

dict\_of\_features['percent\_of\_long\_words'] = long\_words/all\_words

##формулы читабельности (адаптированные, из Бегтина)##

dict\_of\_features['formula\_flesh\_kinc'] = (all\_words/all\_sentences) + (all\_syllables/all\_words)

dict\_of\_features['formula\_coleman'] = (all\_letters/all\_words) + (all\_sentences/all\_words)

dict\_of\_features['formula\_senter'] = (all\_letters/all\_words) + (all\_words/all\_sentences)

dict\_of\_features['formula\_dale'] = (long\_words/all\_words) + (all\_words/all\_sentences)

dict\_of\_features['formula\_smog'] = (30\*(long\_words/all\_sentences))\*\*0.5

##Доля слов, входящих в лексические минимумы##

dict\_of\_features['inA1'] = percent\_of\_known\_words(clean\_lemmas\_list,slovnik\_A1\_list)

dict\_of\_features['inA2'] = percent\_of\_known\_words(clean\_lemmas\_list,slovnik\_A2\_list)

dict\_of\_features['inB1'] = percent\_of\_known\_words(clean\_lemmas\_list,slovnik\_B1\_list)

dict\_of\_features['inB2'] = percent\_of\_known\_words(clean\_lemmas\_list,slovnik\_B2\_list)

##Доля слов, входящих в частотные списки##

dict\_of\_features['infr100'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_100\_list)

dict\_of\_features['infr300'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_300\_list)

dict\_of\_features['infr500'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_500\_list)

dict\_of\_features['infr1000'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_1000\_list)

dict\_of\_features['infr3000'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_3000\_list)

dict\_of\_features['infr5000'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_5000\_list)

dict\_of\_features['infr10000'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_10000\_list)

dict\_of\_features['infr20000'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_20000\_list)

dict\_of\_features['infr\_all'] = percent\_of\_known\_words(clean\_lemmas\_list,fr\_33000\_list)

##Доля абстрактных/конкретных сущ от всех сущ##

dict\_of\_features['lex\_physical'] = percent\_of\_known\_words(noun\_unic\_list,lex\_physical\_list)

dict\_of\_features['lex\_abstract'] = percent\_of\_known\_words(noun\_unic\_list,lex\_abstract\_list)

dict\_of\_features['lex\_substance'] = percent\_of\_known\_words(noun\_unic\_list,lex\_substance\_list)

dict\_of\_features['lex\_virtual'] = percent\_of\_known\_words(noun\_unic\_list,lex\_virtual\_list)

##Доля названий и бастардов и др.##

dict\_of\_features['lex\_names\_and\_geo'] = percent\_of\_known\_words(whole\_lemmas\_list,geo\_imen\_list)

dict\_of\_features['lex\_bastards'] = percent\_of\_known\_words(whole\_lemmas\_list,bastard\_list)

dict\_of\_features['conj\_adversative'] = percent\_of\_known\_words(clean\_lemmas\_list,conj\_adversative\_list)

dict\_of\_features['modal\_verbs'] = percent\_of\_known\_words(clean\_lemmas\_list,modal\_words\_list)

##грам. значения на предложение##

dict\_of\_features['median\_punct\_per\_sentence'] = statistics.median((punctuation\_per\_sentence(sentences)))#медианное пунктуации на предложение

all\_features\_for\_all\_texts[level+' '+text\_number] = dict\_of\_features

json.dump(all\_features\_for\_all\_texts, result\_list\_of\_features)

# Приложение 2. Программный код для настройки работы предсказательной модели.

**import** numpy **as** np

**import** math

**import** json

**import** matplotlib.pyplot **as** plt

**from** sklearn **import** linear\_model

**from** sklearn.cross\_validation **import** StratifiedShuffleSplit

**from** sklearn.metrics **import** mean\_absolute\_error

**from** sklearn.metrics **import** r2\_score

**from** sklearn.metrics **import** mean\_squared\_error

regr = linear\_model.LinearRegression()

ridge = linear\_model.Ridge(alpha=0.1)

##все признаки##

feature\_names\_list = ['inA1', 'inA2','inB1','inB2',

'infr100','infr300', 'infr500','infr1000','infr3000','infr5000','infr10000','infr20000','infr\_all',

'lex\_physical', 'lex\_abstract','lex\_substance','lex\_virtual','lex\_names\_and\_geo',

'lex\_bastards','conj\_adversative','modal\_verbs',

'A', 'ADV', 'ADVPRO', 'ANUM', 'APRO', 'COM', 'CONJ', 'INTJ', 'NUM', 'PART',

'PR', 'S', 'SPRO', 'V', 'непрош', 'прош', 'вин', 'дат', 'зват',

'им', 'пр', 'род', 'твор', 'ед', 'мн', 'деепр', 'изъяв',

'инф', 'пов', 'прич', 'кр', 'полн', 'притяж', '1-л', '2-л', '3-л', 'жен',

'муж', 'сред', 'несов', 'сов', 'действ', 'страд', 'неод', 'од', 'нп', 'пе',

'A\_sentence', 'ADV\_sentence', 'ADVPRO\_sentence', 'ANUM\_sentence', 'APRO\_sentence',

'COM\_sentence', 'CONJ\_sentence', 'INTJ\_sentence', 'NUM\_sentence', 'PART\_sentence',

'PR\_sentence', 'S\_sentence', 'SPRO\_sentence', 'V\_sentence', 'непрош\_sentence', 'прош\_sentence', 'вин\_sentence',

'дат\_sentence', 'зват\_sentence','им\_sentence', 'пр\_sentence', 'род\_sentence', 'твор\_sentence', 'ед\_sentence',

'мн\_sentence', 'деепр\_sentence', 'изъяв\_sentence', 'инф\_sentence', 'пов\_sentence', 'прич\_sentence',

'кр\_sentence', 'полн\_sentence', 'притяж\_sentence', '1-л\_sentence', '2-л\_sentence', '3-л\_sentence',

'жен\_sentence','муж\_sentence', 'сред\_sentence', 'несов\_sentence', 'сов\_sentence', 'действ\_sentence',

'страд\_sentence', 'неод\_sentence', 'од\_sentence', 'нп\_sentence', 'пе\_sentence',

'words','sentences','mean\_len\_word', 'mean\_len\_sentence', 'median\_len\_word',

'percent\_of\_long\_words','mean\_len\_word\_in\_syllables','median\_punct\_per\_sentence',

'median\_len\_sentence',

'formula\_flesh\_kinc', 'formula\_senter', 'formula\_coleman','formula\_dale', 'formula\_smog']

##считываем данные

open\_features = open('my\_features.txt', 'r', encoding='utf-8')

data = json.load(open\_features)

result\_list\_of\_features\_x\_train = []

result\_list\_of\_levels\_y\_train = []

result\_list\_of\_text\_numbers = []

**for** i **in** data:

i\_list = i.split(' ')

i\_level = int(i\_list[0])

i\_number = int(i\_list[1])

**if** i\_level == 0:

**continue**

result\_list\_of\_levels\_y\_train.append(i\_level)

result\_list\_of\_text\_numbers.append(i\_number)

data\_for\_one\_text = data[i]

features\_for\_one\_text = []

**for** ii **in** feature\_names\_list:

features\_for\_one\_text.append(data\_for\_one\_text[ii])#словарь с признаками.

result\_list\_of\_features\_x\_train.append(features\_for\_one\_text)

**print**('Всего признаков', len(result\_list\_of\_features\_x\_train[0]))

##уже готовенькие features и targets

x = np.array(result\_list\_of\_features\_x\_train)

y = np.array(result\_list\_of\_levels\_y\_train)

#список с номерами текстов#

z = np.array(result\_list\_of\_text\_numbers)

#список названий призаков#

feature\_names = np.array(feature\_names\_list)

##делим данные на обуч. и тест##

stratSplit = StratifiedShuffleSplit(y, 1, test\_size=0.2, random\_state=3)

StratifiedShuffleSplit(y, n\_iter=1, test\_size=0.2)

**for** train\_idx, test\_idx **in** stratSplit:

#print(list(train\_idx), 'test', list(test\_idx))

x\_train = x[train\_idx]

y\_train = y[train\_idx]

z\_train = z[train\_idx]

x\_test = x[test\_idx]

y\_test = y[test\_idx]

z\_test = z[test\_idx]

##основная функция##

**def** fit\_and\_predict(model):

model.fit(x\_train, y\_train)

**print**(str(model))

**if** coefi == True:

# The coefficients

coef = model.coef\_

coef\_list = [f **for** f **in** coef]

bag\_of\_coef = dict()

**for** i **in** range(len(coef\_list)):

bag\_of\_coef[feature\_names\_list[i]] = "{0:.3f}".format(coef\_list[i])

##печатаем в порядке самых больших значений по модулю##

**print**(sorted(bag\_of\_coef.items(), key=**lambda** x: math.fabs(float(x[1])), reverse=True))

prediction = model.predict(x\_test)

#print('Prediction', prediction)

#Explained variance score: 1 is perfect prediction

**print**('Variance score: %.2f' % model.score(x\_test, y\_test))

##R^2 (coefficient of determination) regression score function: 1 is perfect##

#print('R2 score: %.2f' % r2\_score(y\_test, model.predict(x\_test)))

# The mean squared error

**print**("Mean squared error: %.2f"

% mean\_squared\_error(y\_test, model.predict(x\_test)))

##mean\_absolute\_error##

**print**("Mean absolute error: %.2f"

% mean\_absolute\_error(y\_test, model.predict(x\_test)))

##test error смотрим на распределение ошибок, нет ли странных выбросов##

test\_error = list(model.predict(x\_test) - y\_test)

count\_error = []

counter\_plus = []

counter\_minus = []

#print("Test error:", test\_error)

**for** i **in** test\_error:

**if** i > 0:

counter\_plus.append(i)

**if** i < 0:

counter\_minus.append(i)

**if** math.fabs(i) >= 1:

i\_index = test\_error.index(i)

count\_error.append(list(y\_test)[i\_index])

**print**('Подозрительный голубчик: ', list(z\_test)[i\_index], 'Правильно:', list(y\_test)[i\_index], 'Предсказано:', list(model.predict(x\_test))[i\_index])

**print**(count\_error,len(count\_error),len(test\_error))

fit\_and\_predict(regr)

fit\_and\_predict(ridge)

# Список литературы

Coleman M., Liau T. L. A computer readability formula designed for machine scoring. //Journal of Applied Psychology. — 1975. — Т. 60, № 2.

Collins-Thompson, K. and Callan, J. Predicting reading difficulty with statistical language models. //Journal of the American Society for Information Science and Technology, 56(13), 2005:1448–1462.

Collins-Thompson, K. and Callan, J. A language modeling approach to predicting reading difficulty. In Proceedings of HLT//NAACL, volume 4, 2004.

Collins-Thompson K. Computational assessment of text readability: a survey of current and future research. In: François, Thomas and Delphine Bernhard (eds.), Recent Advances in Automatic Readability Assessment and Text Simplification. //Special issue of International Journal of Applied Linguistics, 2014: 97-135

Dale E., Chall J. S. A formula for predicting readability: Instructions // Educational research bulletin, 1948: 37—54.

DuBay W. H. Smart Language: Readers, Readability, and the Grading of Text. — ERIC, 2007.

DuBay, W. The principles of readability. Impact Information, 2004, pages 1–76.

Flesch, R. A new readability yardstick. //Journal of applied psychology, 32(3), 1948: 221.

Gunning, R. The technique of clear writing. McGraw-Hill NewYork, NY, 1968.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. An analysis of statistical models and features for reading difficulty prediction. //In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, 2008: 71–79.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. //In Proceedings of HLT-NAACL’07, 2007: 460–467.

Kincaid J. [et al.] Derivation of new readability formulas for navy enlisted personnel: тех. отч. / /DTIC Document, 1975.

Joachims T.  Text categorization with support vector machines:   
Learning with many relevant features. //Technical Report 23, Universitat Dortmund, LS VIII, 1997.

Mc Laughlin G. H. SMOG grading a new readability formula // Journal of reading, Т. 12, № 8, 1969: 639—646.

Nasser Zalmout, Hind Saddiki and Nizar Habash. Analysis of Foreign Language Teaching Methods: An Automatic Readability Approach. //Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016), 2016.

Schwarm, S.E. & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. //In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL ’05), USA, 2005: 523–530.

Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. // MLMTA. — Citeseer. 2003: 273—280.

Senter R., Smith E. Automated readability index: тех. отч. // DTIC Document, 1967.

Sharoff Serge, Svitlana Kurella, and Anthony Hartley. Seeking needles in the web’s haystack: Finding texts suitable for language learners. //In Proceedings of the 8th Teaching and Language Corpora Conference (TaLC-8), 2008.

Stenner, A. J., Burdick, H., Sanford, E. E. & Burdick, D. S. The Lexile Framework for Reading Technical Report. //MetaMetrics, Inc, 2007.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. How Accurate are Lexile Text Measures // Journal of Applied Measurement. 7(3), 2006: 307-322.

Wright, B. D. & Stenner, A. J. Readability and Reading Ability. //Paper presented to the Australian Council on Education Research, 1998.

[Андрюшина](http://www.ozon.ru/person/3319677/) Н., [Галина Битехтина](http://www.ozon.ru/person/3206169/), [Татьяна Владимирова](http://www.ozon.ru/person/3163327/), [Анна Иванова](http://www.ozon.ru/person/981491/), [Любовь Клобукова](http://www.ozon.ru/person/3325273/), [Любовь Красильникова](http://www.ozon.ru/person/3609209/), [М. Нахабина](http://www.ozon.ru/person/1925226/), [Наталия Соболева](http://www.ozon.ru/person/981489/), [В. Стародуб](http://www.ozon.ru/person/5858792/),[В. Степаненко](http://www.ozon.ru/person/3609229/). Требования по русскому языку как иностранному. Первый сертификационный уровень. Общее владение. Профессиональный модуль. // — М.: Златоуст, 2011.

[Андрюшина](http://www.ozon.ru/person/3319677/) Н., [Галина Битехтина](http://www.ozon.ru/person/3206169/), [Татьяна Владимирова](http://www.ozon.ru/person/3163327/), [Любовь Клобукова](http://www.ozon.ru/person/3325273/), [Любовь Красильникова](http://www.ozon.ru/person/3609209/), [Майя Нахабина](http://www.ozon.ru/person/7555472/), [В. Степаненко](http://www.ozon.ru/person/3609229/) Программа по русскому языку как иностранному. Первый сертификационный уровень. Общее владение. // — М.: Златоуст, 2013

[Андрюшина](http://www.ozon.ru/person/3319677/) Н., [Татьяна Козлова](http://www.ozon.ru/person/352104/). Лексический минимум по русскому языку как иностранному. Элементарный уровень. Общее владение. // — М.: Златоуст, 2014

[Андрюшина](http://www.ozon.ru/person/3319677/) Н., [Татьяна Козлова](http://www.ozon.ru/person/352104/). Лексический минимум по русскому языку как иностранному. Базовый уровень. Общее владение. // — М.: Златоуст, 2011.

[Афанасьева](http://www.ozon.ru/person/5237583/) И., [Галина Битехтина](http://www.ozon.ru/person/3206169/), [Любовь Клобукова](http://www.ozon.ru/person/3325273/), [Ирина Яценко](http://www.ozon.ru/person/3497393/) Лексический минимум по русскому языку как иностранному. Второй сертификационный уровень. Общее владение. // — М.: Златоуст, 2014

Владимирова Т.Е., Нахабина М.М., Соболева Н.И. Андрюшина Н.П., Государственный стандарт по русскому языку как иностранному. Элементарный уровень. //— М.: Златоуст, 2001.

Дружкин К.Ю. Метрики удобочитаемости для русского языка. Выпускная квалификационная работа, Национальный исследовательский университет «Высшая школа экономики», Москва, 2016.

Криони Н.К., НикинА.Д. , Филиппова А.В. Автоматизированная система анализа сложности учебных текстов. //Вестник УГАТУ, 2008. №1.

Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). //— М.: Азбуковник, 2009.

Микк, Я.А.Оптимизация сложности учебного текста : в помощь авторам и редакторам. //М. : Просвещение, 1981.

[Нахабина](http://www.ozon.ru/person/1925226/) М., [В. Антонова](http://www.ozon.ru/person/1925228/), [А. Толстых](http://www.ozon.ru/person/1925229/). Типовые тесты по русскому языку как иностранному. Элементарный уровень. Общее владение. //— М.: Златоуст, 2014

Норейко Л. Лексический минимум по русскому языку как иностранному. Первый сертификационный уровень. Общее владение. //— М.: Златоуст, 2014

Оборнева, И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров : диссертация ... к.п.н.: Москва, 2006.

Селегей В.П. et al. Отчет о научно-исследовательской работе по договору № 081-R приложению А1 от 01.10.2013. Тема исследования «Центр по исследованию разума и машин». Москва, 2015.

Томина Ю.А . Объективная оценка языковой трудности текстов (описание, повествование, рассуждение, доказательство) : диссертация ...к.п.н.: Москва, 1985.

1. http://ru.readability.io [↑](#footnote-ref-1)
2. https://habrahabr.ru/company/infoculture/blog/238875 [↑](#footnote-ref-2)
3. https://newsela.com/about/ [↑](#footnote-ref-3)
4. https://www.lingq.com [↑](#footnote-ref-4)
5. http://texts.cie.ru [↑](#footnote-ref-5)
6. https://tech.yandex.ru/mystem/doc/grammemes-values-docpage [↑](#footnote-ref-6)