SlavVaria 1/2025. 191–202
DOI: 10.15170/SV.1/2025.191

BENDEGÚZ PISCH
(Сегед, Венгрия)

Эксперименты по автоматическому определению уровня сложности русских текстов

Аннотация: Чтение играет очень важную роль в изучении языка, однако выбор подходящего текста зачастую представляет собой непростую задачу. Хотя уже существуют адаптированные книги для изучающих язык, они в основном доступны на английском языке и в ограниченном количестве. Поэтому я попытался разработать алгоритм, способный количественно оценивать уровень сложности большого объёма русскоязычных текстов. Алгоритм и связанный с ним инструментарий были размещены в свободном доступе в интернете. Однако важно отметить, что из-за различий в характеристиках текстов нельзя использовать одну и ту же формулу для оценки текстов разного назначения, поэтому предлагаемая мной формула применима только к художественным текстам. Эта статья описывает лишь процесс разработки формулы и принцип её работы.

Ключевые слова: русский язык, сложность текста, автоматическая обработка текстов, массовая обработка текстов, линейная регрессия

Введение

При изучении иностранного языка чтение играет особенно важную роль в развитии языковых навыков – однако подобрать подходящий текст зачастую бывает непросто. Слишком лёгкий или, наоборот, слишком сложный текст не способствует эффективному обучению, а мотивация учащегося может снизиться.

В наши дни появляется всё больше так называемых адаптированных книг, предназначенных для изучения иностранных языков. Эти материалы содержат ограниченный словарный запас и упрощённые грамматические конструкции. Издательства, как правило, классифицируют такие книги по уровням владения языком. К сожалению, количество адаптированных материалов невелико, и большинство из них написаны на английском языке, поэтому найти подобную литературу на других языках затруднительно. На русском языке адаптированные книги публикуют, например, издательство «Златоуст» и «Русский язык. Курсы» (серия «Класс!ное чтение»).

Чтобы оценить сложность оригинальных текстов на русском языке, я попытался разработать автоматизированный метод, способный анализировать и сравнивать тексты из крупных корпусов. Моей целью было найти

такой алгоритм, который с помощью формулы мог бы количественно определять уровень сложности анализируемых текстов.

Все используемые и созданные инструменты, а также управляющие ими скрипты были выложены в открытый доступ в интернете, что позволяет точно производить как обучение модели, так и оценку с её помощью (PISCH 2025: URL).

Статья сосредотачивается исключительно на разработке и функционировании формулы, не рассматривая детали её практического применения. Целью было создание такой формулы, которая способна быстро давать результат даже при анализе больших текстовых корпусов. Однако при её практическом применении необходимо учитывать и другие факторы, связанные с личностью читателя. Из-за ограниченного числа адаптированных книг для изучающих русский язык потребовалось расширить исходную базу для построения формулы. В итоге формула была разработана на основе обязательной школьной литературы, размеченной по учебным классам, поскольку можно предположить, что тексты, предназначенные для младших классов, имеют более простую структуру.

1. Измерение уровня сложности текстов

Вопрос сложности текстов имеет большое значение в различных сферах повседневной жизни, включая образование. Учебники и учебные материалы, даже если они написаны на родном языке, должны быть понятным для целевой аудитории. Поэтому, например, как словарный запас, так и грамматическая структура значительно различаются между учебниками для начальной школы и для университетов.

Другой важной областью являются медиа: доступность и понятность изложения новостных и информационных материалов могут оказывать влияние на процессы, затрагивающие всю страну. Для государства также важна доступность текста: обеспечение понятного изложения необходимо как для упрощения административных процедур (например, налогообложение), так и для функционирования здравоохранения (например, инструкции по применению лекарств).

Именно с этой целью в Соединённых Штатах Америки в 2010 году был принят закон Plain Writing Act of 2010 (OFFICE OF THE FEDERAL REGISTER 2010), направленный на повышение понятности правительственной коммуникации.

Попытки измерить сложность текстов начались в США ещё до наступления компьютерной эры – в начале XX века (BOGDÁN 2024). Однако тогда можно было использовать лишь такие методы, которые не требовали анализа больших объёмов текста, так как это потребовало бы чрезмерных усилий и сопровождалось бы высокой вероятностью человеческих ошибок.

Первичные измерения часто основывались на предположении, что результаты анализа одного фрагмента текста (например, одного абзаца) бу-

дут характерны и для остальной части текста. Однако – особенно в случае художественной литературы – это предположение далеко не всегда оказывается справедливым.

Одним из таких методов является разработанный в 1952 году Робертом Ганнингом индекс сложности текста – Gunning Fog Index (GUNNING 1968), который показывает степень трудности текста по шкале, соответствующей ожидаемым навыкам чтения на определённом школьном уровне. Согласно этому индексу, текст с уровнем 12 соответствует уровню понимания учащегося, окончившего 12-й класс. Метод был изначально разработан для англоязычных текстов. Основной целью алгоритма было повышение читаемости газетных и журнальных публикаций.

При анализе текстового блока объёмом 100 слов формула оперирует двумя параметрами: первый – это среднее количество слов в предложении, а второй – доля так называемых длинных слов по отношению к общему количеству слов в тексте. Под длинными словами обычно понимаются слова, состоящие из трёх и более слогов, однако это определение довольно субъективно. Технические и общеупотребительные термины часто не учитываются, поэтому такую категорию сложно чётко определить с помощью компьютера.

Другим важным тестом является уровень читаемости по Флешу–Кинкаиду (Flesch–Kincaid readability level). Эта формула также учитывает два параметра: первый – среднее количество слов в предложениях, а второй – среднее количество слогов в словах. Метод был также разработан для английского языка, но Ирина Оборнева попыталась адаптировать его к русскому языку (ОБОРНЕВА 2005). Её целью было измерение текстов, используемых в образовательной сфере, поэтому среди 100 проанализированных текстов были материалы, предназначенные для детской аудитории.

Оборнева откалибровала формулу, сравнивая английские оригиналы с их русскими переводами. Однако Соловьёв и соавторы (SOLOVYEV et al. 2018) указали, что такая модифицированная формула применима только к литературным текстам и непригодна для анализа научных материалов. В связи с этим они предложили другую формулу для оценки сложности научных текстов.

Хотя в обоих упомянутых исследованиях приводятся числовые значения, полученные результаты невозможно воспроизвести из-за неизвестных этапов предварительной обработки текста и калибровки формул. Эти методы могут быть полезны, но они недостаточно прозрачны. В своей работе я стремлюсь провести чётко задокументированный, воспроизводимый эксперимент по оценке сложности литературных текстов.

В качестве исходной точки я выбрал индекс Ганнинга из-за его простоты, однако при адаптации метода к русскому языку возникло множество трудностей. Например, в русском языке часто используются аффиксы, состоящие из двух (а иногда и трёх) слогов. Поэтому надёжное определение так называемых «длинных слов» возможно только при наличии

программы, способной точно выявлять корень или словарную форму слова. Однако поскольку число аффиксов в русском языке значительно выше, чем в английском, их ручное выделение (лемматизация) представляет собой более сложную задачу.

Дополнительную сложность представляет тот факт, что словарная форма слов в русском языке может содержать аффиксы – например, у глаголов, или у существительных с уменьшительно-ласкательными суффиксами. Эти особенности искажают результаты, получаемые по формуле. В настоящий момент не существует общедоступного морфологического анализатора, который мог бы с достаточной точностью и гибкостью обрабатывать подобные случаи.

Для русского языка доступен морфологический анализатор MyStem, разработанный компанией «Яндекс» (MYSTEM: URL), однако он тоже не всегда даёт точные результаты. Например, MyStem автоматически интерпретирует дефисы как пробелы, вследствие чего не обрабатывает сложные слова с дефисом как единое целое. Также он недостаточно надёжно учитывает контекст, в котором слово употребляется, и в случае омонимии может предоставлять неверный разбор.

Все эти факторы указывают на необходимость разработки нового алгоритма, основанного на других принципах и учитывающего особенности русского языка. Хотя на сложность текста влияет множество факторов, например синтаксис, проведённые эксперименты показали, что наибольшую точность, наилучшую проверяемость и наименьшую вероятность ошибки при анализе дают именно самые простые параметры. Кроме того, существуют аспекты, которые невозможно измерить с помощью компьютера – например, смысловые уровни художественного произведения. Поскольку эти аспекты не поддаются количественной оценке, формула не может их учитывать.

2. Отбор текстов для анализа

Поскольку адаптированные книги для изучающих русский язык не были легко доступны, текстовые корпуса пришлось формировать на основе других исходных материалов. Для проведения экспериментов я использовал две разные, вручную собранные текстовые коллекции. В ходе работы стало ясно, что формулы дают разные результаты для текстов, принадлежащих к разным стилистическим слоям (например, художественным и научным), поэтому на данный аспект также необходимо было обратить внимание как при составлении формул, так и при последующем обучении математической модели с целью повышения эффективности её работы.

Первая коллекция текстов, использованная в экспериментах с алгоритмами на основе индекса Ганнинга, включала 50 вручную отобранных текстов. Это были произведения XIX, XX и XXI веков различных жанров и для разных целевых аудиторий, изначально написанные на русском

языке или переведённые на русский. Ещё одним критерием отбора было то, что авторы должны быть хорошо известны широкой аудитории. Таким образом, в коллекции оказались, например, «*Преступление и наказание*» Фёдора Достоевского, первый том серии «*Метро 2033*» Дмитрия Глуховского и «*Анна Каренина*» Льва Толстого. Коллекция также включала переводы, например: «*Кристина*» Стивена Кинга, «*Мальчишки с улицы Пала*» Ференца Молнара, две книги Джорджа Оруэлла, две русские версии первого тома серии о Гарри Поттере, а также русский перевод романа «*Золотой автомобиль*» Йенё Рейтё.

В коллекцию был также включён один текст, сгенерированный компьютером. Произведение Исаака Бабеля «*Конармия*» на основе предварительно собранной статистики считается трудным текстом. В связи с быстрым развитием искусственного интеллекта всё легче становится автоматически генерировать тексты, поэтому я предпринял попытку использовать ИИ для автоматического упрощения текстов. В эксперименте я использовал модель GPT-4O, известную как ChatGPT. Так как данная модель может обрабатывать только небольшие тексты, для анализа я выбрал рассказ «*Эскадронный Трунов*» из сборника «*Конармия*». Полученный упрощённый текст вместе с оригиналом был включён в коллекцию для последующего анализа.

Поскольку данная коллекция текстов не была размечена по школьным уровням, в качестве итогового результата работы я представил классификацию этих текстов по школьным классам, с тем отличием, что включил также два учебных пособия (под обозначениями *Bog* и *Nik*), использованных Соловьёвым и его соавторами. Обозначения *Bog* и *Nik* относятся к учебникам по обществоведению: под обозначением «Bog» имеются в виду учебники, подготовленные под руководством Л. Н. Боголюбова, а под обозначением «Nik» – учебники, автором которых является А. Ф. Никитин.

Таким образом можно увидеть, как формула измеряет сложность научных текстов.

Концепция оценки сложности текстов базируется на том, что тексты, рекомендованные учащимся различных классов, обладают определёнными свойствами, которые, разумеется, не являются идентичными, но могут быть исследованы автоматически. В ходе анализа с применением различных алгоритмов осуществляется попытка определить, к какой из выделенных характеристик классов наиболее близок анализируемый текст.

Когда стало очевидно, что простые формулы не дают результатов и оптимизацию необходимо производить математическими методами, я ввёл вторую коллекцию текстов, которую использовал в качестве обучающей выборки для модели. Для этого потребовалась более крупная текстовая коллекция, содержащая рекомендованные произведения, размеченные по школьным классам. Хотя в России, по-видимому, нет единых

предписаний для всех школ, Федеральный институт педагогических измерений опубликовал список рекомендованных произведений, составленный в рамках основной образовательной программы основного общего образования[1]. Важно отметить, что из-за региональных особенностей и этнического разнообразия трудно создать единый централизованный список литературы – вероятно, именно поэтому официального перечня нет. Из этого также следует, что если бы кто-то другой составлял подобный список, то в него, вероятно, вошли бы иные произведения, а классификация по классам отличалась бы. По этой причине и моя коллекция текстов претерпела определённые изменения.

Коллекция не включает произведения, рекомендованные для первых четырёх классов, так как они часто представляют собой слишком короткие тексты, на основе которых нельзя провести надёжные измерения. Из текстов были удалены стихи и драматические произведения, поскольку параметр, связанный с предложениями, в этих случаях трудно поддаётся измерению. В тех случаях, когда произведение относилось сразу к нескольким классам, всегда принимался в расчёт младший класс. Когда в списке по какому-либо автору не указывались конкретные произведения, а приводились лишь примеры, в коллекцию включались именно эти примеры. Если по рекомендации предусматривалось три произведения, но были указаны только два примера, то в коллекцию также включались только эти два текста. В случае, если в списке значилось лишь имя автора, то в коллекцию случайным образом включалось одно его произведение. Если автор писал не только на русском, но и на других языках (например, Владимир Набоков), то в коллекцию попадало произведение, изначально написанное на русском языке. Что касается произведений иностранных авторов, у которых имелось несколько переводов, то в коллекцию включался только один вариант перевода. Там, где в списке были указаны только определённые главы из книги, в коллекцию всё равно включалась вся книга. Наконец, балансировка количества текстов по классам производилась вручную путём случайного удаления отдельных произведений.

Тексты прошли предварительную обработку. На первом этапе были удалены те тексты, в которых не менее 10% слов, встречающихся хотя бы один раз, или 20% всех словоформ не распознавались морфологическим анализатором MyStem – это позволило отсеять тексты с большим количеством иноязычных фрагментов или содержащие многочисленные ошибки оцифровки. На втором этапе, для облегчения анализа, тексты

[1] См., например: Список рекомендованной в соответствии с примерными основными общеобразовательными программами литературы, разработанный Русской школьной библиотечной ассоциацией по заказу Минобрнауки России в 2015 г. (Источник: информационное письмо Департамента государственной политики в сфере общего образования Минобрнауки России от 14 апреля 2016 года № 08-709 «О списках рекомендуемых произведений»)

были разделены на предложения с помощью Perl-скрипта. Используемый мною скрипт сегментации (WOLOSZ 2025: URL) считал границей предложения любой фрагмент текста, в котором за словом, заканчивающимся на знаке препинания, следовало слово, начинающееся с заглавной буквы и отделённое пробелом – при условии, что это слово не входило в список сокращений, включённый в скрипт. Последнее условие важно, например, в тех случаях, когда за сокращением следует имя, поскольку в таких случаях скрипт мог бы ошибочно определить границу предложения. В настоящее время известен лишь один случай, когда скрипт может допустить ошибку, однако эта ошибка пока не устранена.Информация об издании, названия и подзаголовки, содержащиеся в начале книг, как правило, представляют собой единицы без знаков препинания в конце, для распознавания которых необходимо учитывать контекст. Поскольку формулировки таких данных не стандартизированы, на них сложно (или невозможно) построить программу, которая бы работала безошибочно во всех случаях. В настоящий момент программа обрабатывает такие данные как одно предложение, длящееся до первого встреченного в тексте знака препинания, обозначающего конец предложения. На выходе скрипт генерирует файл, в котором каждая строка соответствует одному предложению. Это облегчает последующий анализ, так как скрипт, осуществляющий статистические расчёты, не нуждается в дополнительной обработке текста.

Третьим этапом стала дополнительная фильтрация, в результате которой была получена финальная словарная выборка для анализа. В текстах могут встречаться последовательности символов, формально являющиеся словами, но не имеющие значения для анализа: к таким относятся, например, числа или слова, написанные не кириллическими буквами. Для составления списка слов скрипт сначала разбивал текст по пробелам, после чего из полученного списка удалялись квадратные и фигурные скобки, использующиеся для сносок и ссылок, а также содержащиеся в них числа. Затем удалялись все знаки препинания и прочие символы, прилегающие к словам (например, запятые, двоеточия, скобки, знаки препинания в конце предложения). Однако дефис («-») сохранялся, так как он часто является частью слова (например, в слове что-то). На четвёртом этапе все символы *Ё* и *ё* заменялись на строчную е, поскольку их употребление в текстах не является единообразным. Далее все прописные буквы кириллицы заменялись строчными. Это облегчало сравнение словоформ, поскольку для компьютера заглавные и строчные буквы – это разные символы. На последнем этапе скрипт сохранял в итоговом списке только те элементы, которые содержали исключительно 32 буквы русского алфавита (без учёта буквы *ё*) и дефис («-»).

3. Рассмотренные подходы и полученные результаты

3.1. Измерение текстов по числу слогов и средней длине предложений

Первый алгоритм, основанный на количестве слогов, работал по принципу, схожему с индексом Ганнинга, однако вместо понятия «длинное слово», присутствующего в оригинальной формуле, скрипт анализировал среднее количество слогов в словах русскоязычных текстов. Это решение было принято из-за вышеупомянутого субъективного характера определения длинных слов. Таким образом, первым компонентом формулы стало среднее количество слогов в словах очищенного списка, представленного ранее. Количество слогов в слове скрипт определял по количеству содержащихся в нём гласных букв. Слова без гласных – обычно это предлоги (например, *в*, *с*, *к*) – обрабатывались как слова с нулевым количеством слогов. Вторым компонентом формулы было среднее количество слов в предложениях текста. В некоторых текстах, по авторскому замыслу, одно предложение могло длиться на протяжении сотен страниц (см., например, Ежи Анджеевского «*Врата рая*»), поэтому программа игнорировала 5 самых длинных предложений, так как они могли привести к искажённым результатам.

Уже на раннем этапе анализа стало ясно, что среднее количество слогов в словах мало отличается между текстами и не даёт ожидаемой информации, поэтому данный параметр оказался малополезным. В то же время среднее количество слов в предложениях давало более точное представление об уровне сложности текстов. Например, результаты анализа первого корпуса показывают, что «*Анну Каренину*» характеризуют значительно более длинные предложения, чем, скажем, «*Метро 2033*» или переводы «*Гарри Поттера*».

Следующим шагом стало преобразование метода, основанного на слогах, с введением взвешенной системы оценки: слова с большим количеством слогов получали больше очков, что влияло на итоговый расчёт. Слова из 1 и 2 слогов не получали очков, 3-сложные – 1 очко, 4-сложные – 2, 5-сложные – 3, 6-сложные – 4, а слова с 7 и более слогами – 5 очков. Результаты показали, что даже при таком подходе существенных различий между текстами не наблюдается, однако он выявил важную проблему, связанную с массовым анализом текстов. Качество оцифрованных текстов может значительно различаться, и в них могут встречаться ошибки, возникшие при оцифровке. Одной из таких ошибок является отсутствие пробела между двумя словами или ситуация, когда программа распознавания текста считает одним длинным словом два коротких. Наибольшие значения в измерениях давали именно такие тексты. Поскольку в рамках массового анализа текстов нет возможности отфильтровать подобные ошибки, использование этого алгоритма оказалось нецелесообразным.

3.2. Измерение текстов по средней длине слов и средней длине предложений

Следующим подходом стала формула, использующая среднее количество букв в словах, при этом анализ вёлся на основе очищенного списка слов. Однако и здесь не наблюдалось значительных различий между текстами, поэтому следующим шагом было внесение модификации в формулу: вместо подсчёта символов скрипт стал рассчитывать среднее количество звуков в словах. Для этого необходимо было ввести несколько правил. Во-первых, определённые гласные в русском языке (*е, ё, ю, я*) могут обозначать один или два звука в зависимости от своей позиции в слове; во-вторых, следовало исключить символы, не обозначающие звука (ъ, ь). Полученные результаты вновь показали лишь незначительные различия между текстами, а автоматизированное определение звуков оказалось неочевидным и неоднозначным – таким образом, и этот метод не оказался надёжным.

3.3. Формулы, использующие частотность словоформ

Поскольку измерение вышеописанных параметров не дало надёжных результатов, пришлось искать другой метод. В подходе, основанном на лексике, частота появления слов также может служить возможным параметром (SOLOVYEV et al. 2018), который используется и в другом индексе – индексе читаемости Dale-Chall. Первая составляющая формулы, применяемой в индексе, снова представляет собой среднее количество слов в предложениях. В качестве второго параметра был введён новый подход – понятие так называемых «*знакомых слов*» (*familiar words*). В данном методе использовался заранее составленный список из 3000 слов. Слова, входящие в этот список, определялись как знакомые. При измерении сложности текста вторая составляющая формулы рассчитывалась как процент незнакомых слов, встречающихся в тексте.

Поскольку морфологическая система русского языка гораздо более сложная, чем у английского, возникла необходимость расширить список из 3000 слов. Предполагается, что словоформы, встречающиеся чаще, являются более важными элементами данного языка.

Поэтому можно провести такой расчёт, который покажет, какой процент слов в данном тексте принадлежит к группе, определяемой как важная (то есть знакомая) с точки зрения словарного запаса. Однако из-за отсутствия надёжной процедуры лемматизации в случае русского языка необходимо работать со словоформами, так как у одной и той же лексемы может быть несколько форм.

Для измерения потребовались списки наиболее частотных словоформ русского языка. Эти списки (WOLOSZ 2024: URL) также прошли описанную выше процедуру очистки текста. Каждый список содержит определённое количество наиболее частотных словоформ, полученных на основе анализа большого текстового корпуса. В своём исследовании я ис-

пользовал следующие списки, содержащие заданное количество наиболее частотных словоформ: 1000, 1500, 2000, 3000, 5000, 10000, 20000, 30000, 40000, 50000, 100000. По результатам измерений видно, что списки, содержащие 1000 и 5000 наиболее частотных словоформ, уже дают достаточно разнообразные результаты – и также выяснилось, что увеличение длины списка негативно влияет на полученные значения. Я пытался выразить это значение в виде простых формул, однако вручную настроить формулу оказалось довольно сложно. Поэтому для оптимизации формулы я использовал математический метод, называемый линейной регрессией, при котором анализируется корреляция между параметрами.

В качестве обучающей выборки для модели я использовал второй корпус текстов, описанный выше, разбив каждый текст на предложения и случайным образом их перемешав. Затем файлы были разделены на 11 равных частей (на основе количества предложений), в которых после перемешивания предполагалось наличие фрагментов текста с примерно одинаковыми характеристиками. Из каждого файла первые 10 частей попали в обучающую выборку, а оставшаяся 11-я часть – в валидационную. Таким образом, при обучении модели использовалась проверочная выборка, схожая с обучающими данными.

В качестве выходного значения модели было выбрано значение, основанное на школьных классах, как в индексе читаемости Ганнинга. После проведения нескольких экспериментов наилучший результат был получен с использованием следующих трёх параметров: среднее количество слов в предложениях, а также доля слов, не входящих в списки из 5000 и 1000 наиболее частотных словоформ. На основе этого была получена итоговая формула. Новая формула может быть проверена и воспроизведена, так как скрипты находятся в свободном доступе в интернете.

Точность формулы можно выразить с помощью так называемой средней квадратичной ошибки, причём чем меньше это значение, тем точнее формула. Изначально это значение превышало 5, однако в окончательной формуле его удалось снизить до 2,3370.

Формула:

DC_NEW_FICTION = 0.0093 * DWP1 + 0.0569 * DWP2 + 0.2629 * ASL + 2.0093

где:

DWP1 – процент словоформ, не входящих в список из 5000 наиболее частотных словоформ,

DWP2 – процент словоформ, не входящих в список из 1000 наиболее частотных словоформ,

ASL – среднее количество слов в предложениях.

Несколько примеров полученных результатов на основе первой, предназначенной для тестирования коллекции текстов:

Если упорядочить результаты по возрастанию, то на первом месте оказывается «*Маленький принц*» с оценкой 5,9879. В начале списка также находятся следующие произведения: «*Ночной дозор*» – 6,2478, *«Пригла-*

сите доктора на свидание» – 6,2982, «*Винни-Пух*» – 6,5679, а учебник серии *Nik*, предназначенный для учеников 5-го класса, получил оценку 6,9606.

В конце списка – «*Жизнь Арсеньева*» с самым высоким показателем: 10,2904. Эту книгу опережают учебники серии *Bog* и *Nik* для 10-го и 11-го классов (например, «*Nik 11-й класс*» – 9,3965), а также «*Мёртвые души*» – 9,0972.

Если сравнить оригинальный рассказ об «*Эскадронном Трунове*» и его упрощённую, переработанную с помощью искусственного интеллекта версию, то упрощённый вариант находится на 11-м месте с оценкой 6,8523, в то время как оригинал – на 56-м месте с оценкой 8,9814.

Хотя формула, судя по значению ошибки, точнее предыдущих, в некоторых случаях она всё же не даёт достаточно точной оценки. Тем не менее, на основе тестов можно утверждать, что именно эта формула работает наилучшим образом, поэтому её использование целесообразно. Важно, однако, отметить, что из-за большого разнообразия художественных текстов дальнейшая оптимизация формулы за пределами определённой границы погрешности становится невозможной.

4. Заключение

Как видно, измерение сложности текстов представляет собой довольно сложную задачу. Существует множество различных подходов, которые в разной степени поддаются автоматизации и адаптации к языкам, отличным от английского. Проведение экспериментов может быть затруднено рядом факторов: качеством и объёмом используемых текстов, надёжностью разметки данных, применимостью различных параметров к конкретному языку.

Предложенная мной формула ориентирована прежде всего на измерение сложности художественных текстов. Наиболее эффективными параметрами оказались среднее количество слов в предложениях, а также процент слов, не входящих в списки из 5000 и 1000 самых частотных словоформ.

Важно, однако, отметить, что даже эти параметры работают с довольно широкой погрешностью. Поскольку художественные тексты обладают значительным разнообразием, точность формулы с определённого момента уже не может быть улучшена.

Литература

ОБОРНЕВА 2005 = ОБОРНЕВА И.В. Автоматизация оценки качества восприятия текста // Вестник Московского городского педагогического университета, Серия: Информатика и информатизация образования, 2005. 86–91.

BOGDÁN 2024 = BOGDÁN P. Angol nyelvű olvashatósági formulák magyar nyelvi adaptálásának lehetséges irányai // Iskolakultúra, 2024. 34(1). 69–77. DOI: 10.14232/iskkult.2024.1.69

GUNNING 1968 = GUNNING R. The technique of clear writing. McGraw-Hill, 1968.

MYSTEM = Программа MyStem производит морфологический анализ текста на русском языке. https://yandex.ru/dev/mystem/

OFFICE OF THE FEDERAL REGISTER 2010 = Office of the Federal Register, National Archives and Records Administration. Public Law 111 – 274 – Plain Writing Act of 2010. U.S. Government Printing Office, 2010. https://www.govinfo.gov/app/details/PLAW-111publ274

PISCH 2025 = russian-text-complexity. https://github.com/bpisch/russian-text-complexity

SOLOVYEV et al. 2018 = SOLOVYEV V., IVANOV V., SOLNYSHKINA M. Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics // Journal of Intelligent & Fuzzy Systems, 2018. 34. 3049–3058. DOI: 10.3233/JIFS-169489

WOLOSZ 2024 = WOLOSZ R. Szavakból álló n-gramok gyakorisága orosz szövegekben. https://www.wolosz.hu/russian_n-grams.html

WOLOSZ 2025 = WOLOSZ R. Orosz nyelvű szövegek mondatokra szegmentálása (Perl script). https://www.wolosz.hu/russian_segmentation.html

Experiments on the Automated Determination of Difficulty Levels in Russian Texts. Reading plays a very important role in language learning; however, selecting the appropriate text is often not an easy task. Although there are already books adapted for learners, they are mostly available only in English and in limited numbers. Therefore, I attempted to develop an algorithm capable of quantitatively expressing the difficulty level of large volumes of Russian-language texts. I have made the algorithm and its associated toolkit freely available on the internet. It is important to note, however, that due to differing characteristics of texts, a single formula cannot be used to evaluate all types of texts. Thus, the formula I propose is applicable only to literary texts.

Keywords: Russian language, text complexity, automated text processing, large-scale text processing, linear regression