

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/346084779>

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ СЛОЖНОСТИ ТЕКСТА ПО РКИ

Conference Paper · November 2018

CITATIONS

0

READS

318

1 author:



[Antonina Laposhina](#)

Pushkin State Russian Language Institute

15 PUBLICATIONS 2 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Readability Metrics for Russian L2 Learners [View project](#)



TIRTEC: Text-Image Russian Textbooks Corpus [View project](#)

Лапошина Антонина Николаевна, Государственный институт
русского языка им. А.С. Пушкина, Москва
antonina.laposhina@gmail.com

АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ СЛОЖНОСТИ ТЕКСТА ПО РКИ

В статье описана концепция системы автоматического определения уровня сложности текстов по РКИ с использованием машинного обучения, её возможности и примеры работы.

Ключевые слова: *сложность текста; удобочитаемость; уровни владения РКИ.*

Введение в проблему

Проблема определения сложности текстов, их категоризации, изучается ещё с начала двадцатого века. Вместе с появлением возможностей оперировать большими объемами данных и задействовать механизмы машинного обучения, интерес к этой теме снова растет, а сфера применения – расширяется: помимо школьного образования, тема определения сложности текста встречается и в исследованиях по доступности и ясности государственных документов [5], и в сфере преподавания иностранных языков [3,4]. В настоящей работе будет описан опыт создания системы автоматической оценки сложности текстов для преподавания РКИ.

Оценка сложности русских текстов как иностранных

Чтение занимает важнейшее место в процессе обучения иностранному языку, правильно подобранные материалы способствуют как усвоению лексики и грамматики, так и повышению интереса к обучению в целом. На наш взгляд

существует несколько особенностей, отличающих понятие сложности текста в контексте преподавания иностранных языков:

1. Наличие понятной шкалы. В нашем исследовании предлагается принять за шкалу сложности текстов их соответствие общепринятым в методике уровням владения иностранным языком, входящих в европейскую структуру языкового тестирования ALTE: 6 уровней от A1 до C2. К плюсам этой шкалы можно отнести свободу от таких субъективных категорий, как класс/возраст/количество лет обучения.

2. Наличие регламентирующих документов. 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 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.

3. Грамматика: понятия падежа, времени и вида глагола совершенно естественны для носителей, представляют сложность для иностранцев, изучающих русский язык. Так, например, Neilman et al. в своей работе [2] приводят результаты эксперимента для английского языка, в ходе которого добавление грамматических признаков принесло бóльший прирост точности в коллекции текстов как иностранных – 22% против 7% как родных).

4. Лексика. Если словарный запас школьника зависит от множества факторов – семьи, интересов и способностей ученика, то лексический запас человека, изучающего иностранный язык, более предсказуем: большинство учебных комплексов ориентируется на лексические минимумы, предназначенные для этого уровня.

Технология создания

Проблема автоматического определения сложности текста с точки зрения компьютерной лингвистики становится классической задачей построения предсказательной модели на основании обучения на тренировочном корпусе текстов и наборе признаков. Для обучения нашей модели был собрана коллекция около 600 текстов, взятых из текстовой ЦМО МГУ и учебных пособий, в методической справке которых был указан уровень владения языком, для которого он предназначен. В качестве лингвистических признаков были использованы:

- Традиционные метрики текстов (такие как средние и медианные длины слов и предложений, процент слов длиннее 4 слогов и др.).

- Признаки на основе формул читабельности. Были выбраны пять наиболее широко используемых в англоязычном мире формул для оценки сложности текстов: формула Флэша-Кинкайда, Колман-Лиау, Дэйла-Чалл, SMOG и Automated Readability Index.

- Лексические признаки. Доля слов в тексте, входящих в лексические минимумы по РКИ и различные списки Частотного словаря современного русского языка[8].

- Грамматические признаки. Для подсчета грамматических признаков была использована программа Mystem. Считалась доля того или иного грамматического признака 1) во всем тексте 2) в предложении. Например, признак «Доля именительного падежа во всем тексте», или «Количество существительных на предложение»

- Семантические признаки. Вслед за Я. Микком [6], берем за основу предположение, что на сложность может влиять доля абстрактной лексики в тексте. Мы использовали списки слов из семантической иерархии ABBYY COMPRENO с семантемами (своеобразными семантическими метками), характеризующими существительные с точки зрения абстрактности/конкретности.





Всего было использовано около 150 признаков для каждого текста. В Таблице 1 приведены 10 наиболее успешных признаков из разных категорий. Лидирующие позиции в этом списке заняли лексические признаки (лексические минимумы и частотные списки слов), что подтверждает нашу гипотезу о большом влиянии лексики на сложность русского текста как иностранного. Среди частотных списков наиболее информативными оказались «медианные» списки, от 300 до 10000 слов, слишком маленькие и слишком большие оказались не так эффективны.

Признак	Коэффициент
Доля слов, входящих в лексический минимум A2	0.81
Формула SMOG	0.66
Средняя длина предложения	0.62
Доля слов, входящих в топ-10000 по частотности	0.59
Доля слов длиннее 4 слогов	0.58
Доля слов с частотой более 5 IPM	0.56
Доля слов в именительном падеже	0.55
Доля абстрактных слов	0.55
Средняя длина слова	0.52
Количество знаков пунктуации в предложении	0.48
Таблица 1. Подсчет корреляции лучших в своих категориях признаков (коэффициент рассчитан по Пирсону)	

Большую корреляцию также показали формулы читабельности и традиционные метрики текста. Грамматические признаки показали свой вклад в понятие сложности текста, хотя их коэффициенты оказались и не так велики.

Пример работы модели

Для демонстрации работы предсказательной модели был выбран текст из сборника «Тесты по русскому языку как иностранному. Первый сертификационный уровень», текст №2 о Ю.А. Гагарине [7].

Уровень текста: [2.71334516] [A2-B1]
 Слов: 352, норма 900
 Комментарий: чтение должно занять до 10 минут
 Средняя длина слова: 5.57
 Средняя длина предложения: 8.38
 Лексическая сложность: 4.5
 Структурная сложность: 2
 Слов нет в словаре, может быть, опечатка? : set(цуп)
 Низкочастотные слова: set('военно-воздушный', 'байконур', 'объехать', 'космодром', 'выносливость')
 Процент незнакомых слов: 18%
 Слова, не вошедшие в лексический минимум для B1: {'риск', 'виток', 'существо', 'всемирный', 'космодром', 'белка', 'полет', 'рост', 'подавать', 'военно-воздушный', 'вес', 'любимец', 'авиационный', 'проживать', 'красота', 'цуп', 'понадобиться', 'слава', 'конструктор', 'реакция', 'гибель', 'удачный', 'трагически', 'стрелка', 'благополучно', 'объехать', 'байконур', 'прерываться', 'ракета', 'выдерживать', 'учитываться', 'выносливость', 'управление', 'отряд', 'выращивать', 'парень', 'старт', 'доброволец', 'солнечный', 'потрясать', 'действовать', 'настоящий'}

В целом модель адекватно оценивает текст как чуть ниже B1 из-за маленького объема и простоты синтаксиса, сложность же лексики, напротив, выше нормы, для B1 нормой является 5-7 % незнакомых слов. Там, где результат сравнивается с нормой, используется информация из государственных стандартов для данного уровня. К недостаткам нашей системы можно отнести отсутствие учета словообразовательной информации (например, слово «солнечный» ученики такого уровня должны понять, зная «солнце», имеющееся в лексическом минимуме). Кроме того, автоматический анализатор морфологии может ошибаться (например, не распознал Белку и Стрелку как имена собственные). Эти недостатки планируется исправить в будущем. Ниже приведены несколько примеров оценки неадаптированных текстов из сети Интернет:

Источник текста	Уровень	Средняя длина предложения	Средняя длина слова	Процент слов из лексич. минимума B2	Процент абстрактной лексики
Народная сказка "Маша и медведи"	3.2 (B1)	8.5	4.8	80%	29%
Статья из блога про путешествия (ок. 1 тыс. слов)	3.9 (B1-B2)	12.17	5.1	82%	60%
А.П.Чехов. "Общее образование"	4.1 (B2)	11	4.8	78%	44%

Типовой договор на аренду квартиры	5.5 (C1-C2)	9.4	6.3	63%	77%
Л.Н. Толстой. "Анна Каренина" (отрывок ок. 3 тыс. слов)	5.8 (C1-C2)	22.9	5	79%	48%
Правила пользования московским метрополитеном	6.5 (C2)	10.2	6.8	67%	66%
В. Набоков. "Лолита" (отрывок ок. 3 тыс. слов)	6.9 (>C2)	23.4	5.5	71%	54%

В настоящее время создан рабочий прототип модели для автоматического определения сложности текста по РКИ. Такая модель может использоваться для повышения удобства подготовки материалов к уроку, подбору текстов для пособий и сертификационных тестов (здесь наиболее важно соответствовать государственному стандарту), а также для помощи в проверке существующих пособий на соответствие заявленному уровню.

В качестве направлений дальнейшей работы, мы рассматриваем подключение словообразовательной информации, идентификации идиом и коллокаций, поиск синтаксических признаков, оказывающих влияние на сложность русского текста как иностранного. Также планируется создание браузерной версии нашего анализатора текста для открытого использования.

Литература

1. Collins-Thompson K. Computational assessment of text readability: a survey of current and future research. Special issue of International Journal of Applied Linguistics, 2014, pp. 97-135
2. Heilman, M., Collins-Thompson, K. Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. In Proceedings of HLT-NAACL.- 2007. p. 460–467.

3. Nasser Zalmout at all. 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.

4. Sharoff S., 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.

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

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

7. Тесты по русскому языку как иностранному. Первый сертификационный уровень. Общее владение. – Екатеринбург, 2007.

Laposhina Antonina Nikolaevna, Pushkin State Russian Language Institute, Moscow, Russia

AUTOMATIC APPROACH TO TEXT DIFFICULTY MEASUREMENT FOR RFL

This paper presents a concept of an automatic reading difficulty measurement for RFL, based on machine learning algorithms. Examples of use are discussed.

Key words: *readability; text complexity; reading difficulty; RFL.*

*Международная научно-практическая интернет-конференция
«Актуальные вопросы описания и преподавания русского языка
как иностранного/неродного»*

Фонд содействия продвижению русского языка
и образования на русском

Государственный институт русского языка им. А.С. Пушкина

**АКТУАЛЬНЫЕ ВОПРОСЫ
ОПИСАНИЯ И ПРЕПОДАВАНИЯ
РУССКОГО ЯЗЫКА КАК
ИНОСТРАННОГО/НЕРОДНОГО**

Сборник материалов

Международной научно-практической
интернет-конференции

(Москва, 27 ноября – 1 декабря 2017 г.)

Москва

2018

M43

Под общей редакцией доктора пед. наук, профессора Н.В. Кулибиной

M43 Международная научно-практическая интернет-конференция «Актуальные вопросы описания и преподавания русского языка как иностранного/неродного» (Москва, 27 ноября – 1 декабря 2017 г.): Сборник материалов / Под общ. ред. Н.В. Кулибиной. – М., 2018. – 1074 с.: ил. [Электронное издание].

ISBN 978-5-98269-173-6

Данный сборник включает материалы Международной научно-практической интернет-конференции «Актуальные вопросы описания и преподавания русского языка как иностранного/неродного», проходившей на платформе «Образование на русском» с 27 ноября по 1 декабря 2017 г. Организаторами интернет-конференции выступили: Фонд содействия продвижению русского языка и образования на русском, Государственный институт русского языка им. А.С. Пушкина.

Цель мероприятия состояла в анализе, обобщении и распространении в профессиональном сообществе научных достижений в области описания русского языка как иностранного/неродного и методики его преподавания.

Конференция организована с использованием гранта Президента Российской Федерации на развитие гражданского общества № 08.P27.11.0039 от 18.09.2017, предоставленного Фондом президентских грантов.

ISBN 978-5-98269-173-6

© Государственный институт русского языка
им. А.С. Пушкина, 2018