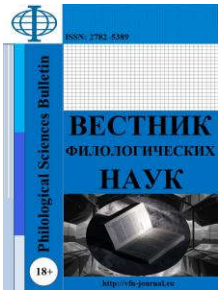

Научно-исследовательский журнал «Вестник филологических наук / Philological Sciences Bulletin»
https://vfn-journal.ru
2024, Том 4, № 6 / 2024, Vol. 4, Iss. 6 https://vfn-journal.ru/archives/category/publications
Научная статья / Original article
Шифр научной специальности: 5.9.8. Теоретическая, прикладная и сравнительно-сопоставительная (филологические науки)
УДК 81'33

[1] Марико М.Л.

[1] Казанский (Приволжский) федеральный университет

Измерение и объяснение сложности текстов организации объединенных наций

Аннотация: *в данном исследовании мы измерили и объяснили сложность текстов ООН на русском и английском языке. Цель текущего исследования – изучить специфику сложности текстов ООН на русском и английском языке, рассматривая такие показатели, как читабельность (FKGL), лексическое разнообразие (TTR), лексическая плотность и среднее значение этих показателей. Мы создали корпус из 40.000 слов, включив в выборку 40 текстов, то есть 20 текстов на русском и 20 текстов на английском языке. После этого тексты были вычислены с помощью таких программ, как Rulingva (https://rulingva.kpfu.ru) и TexInspector (https://textinspector.com). Для интерпретации данных мы использовали методы, которые предложили Ure [11, с. 443-452], Halliday [5, с. 109], Johansson [7, с. 61-79], Templin [12, с. 183], Malvern [13, с. 183] and Danielle McNamara [14, с. 18]. При этом сложность определялась по процентному соотношению, которое показали выбранные метрики. Результаты показали, что русские тексты сложнее для анализа, чем тексты на английском языке. Мы пришли к выводу, что для точной интерпретации значения читабельности или сложности текста необходимо учитывать не только процентное соотношение, но и некоторые другие факторы, такие как выбор слов, синтаксическая сложность и т.д.*

Ключевые слова: *среднее значение, лексическое разнообразие (TTR), лексическая плотность, сложность, читабельность (FKGL), русские и английские тексты ООН*

Для цитирования: Марико М.Л. Измерение и объяснение сложности текстов организации объединенных наций // Вестник филологических наук. 2024. Том 4. № 6. С. 63 – 70.

Поступила в редакцию: 24 мая 2024 г.; Одобрена после рецензирования: 20 июня 2024 г.; Принята к публикации: 30 июня 2024 г.

[1] Mariko M.L.

[1] Kazan (Volga Region) Federal University

Measuring and explaining complexity of the texts of the United Nations

Abstract: *this study measured and explained complexity of Russian and English UN texts. The present study aimed to explore the nature of complexity of Russian and English UN texts by focusing on metrics like readability (FKGL), lexical diversity (TTR), lexical density and the mean of those metrics. We created a corpus of 40.000 words by limiting the sample to 40 texts, that is to say, 20 texts in Russian and 20 texts in English. Afterwards, the texts were computed with such programs like Rulingva (https://rulingva.kpfu.ru) and TexInspector (https://textinspector.com). We applied the methods proposed by Ure [11, p. 443-452], Halliday [5, p. 109], Johansson [7, p. 61-79], Templin [12, p. 183], Malvern [13, p. 183] and Danielle McNamara [14, p. 18] to interpret the data. Being so, complexity was determined by referring to the percentage provided by the selected metrics. In result, the findings revealed that Russian texts are more complex to process than English texts. We concluded that some other factors like word choice, and syntactic complexity, etc. should be considered besides percentage to accurately interpret the meaning of the text's readability or complexity.*

Keywords: *mean, lexical diversity (TTR), lexical density, complexity, readability (FKGL), Russian and English UN texts*

For citation: Mariko M.L. Measuring and explaining complexity of the texts of the United Nations. Philological Sciences Bulletin. 2024. 4 (6). P. 63 – 70.

The article was submitted: May 24, 2024; Approved after reviewing: June 20, 2024; Accepted for publication: June 30, 2024.

Введение

Хотя исследователи политического дискурса не избежали внимания к измерению читабельности [1, с. 63-66], не было особенно большого интереса к адаптации мер к специфическим политическим контекстам. Это порождает две широкие группы вопросов, которые позволяют нам немного поразмышлять: во-первых, теоретические проблемы, связанные с тем, что использование таких мер подразумевает в отношении элементов, определяющих сложность текста, и их соответствующих значений; и, во-вторых, общее отсутствие интереса со статистической точки зрения.

Читабельность, не путать с разборчивостью, можно определить, как характеристику письменного текста, которая подразумевает адекватное понимание сообщения, задуманного писателем. С другой стороны, разборчивость относится к расположению и шрифту письменного текста [2, с. 74]. Согласно [6, с. 109], наиболее полным является следующее определение читабельности, данное в [3, с. 11-28]: "Это совокупность (включая все взаимодействия) всех тех элементов в данном материале, которые влияют на успешность работы с ним группы читателей. Успех – это понимание материала, оптимальная скорость его чтения и интерес к нему" [3, с. 23]. Формулы читабельности призваны предсказать требуемую способность к чтению, которой должен обладать человек, чтобы понять смысл текста [10, с. 132-137].

Лексическая плотность, лексическое разнообразие или TTR [4, с. 197-222] – это термины, которые относятся к статистическим показателям, измеряющим лексическое богатство текстов, а также могут использоваться для оценки общего прогресса читателей. Лексическое разнообразие текста учитывает, сколько различных слов используется в тексте, а лексическая плотность позволяет определить соотношение лексических единиц, например, существительных, глаголов, прилагательных и некоторых наречий в тексте [7, с. 61-79].

Термин "средний" обычно относится к средней длине текста или среднему количеству слов в предложении. Это мера читабельности или сложности текста [7, с. 61-79].

В данной статье мы рассматриваем читабельность (FKGL), лексическую плотность, лексическое разнообразие (TTR) и среднее значение этих показателей для измерения и объяснения сложности текстов ООН на русском и английском языках. Мы выбрали эти метрики, потому что они были доказаны несколькими исследователями, такими как [11, с. 443-452, 5, с. 109] и др. для определения сложности текста.

Материалы и методы исследований

В этой статье мы измерили и объяснили сложность русских и английских текстов с помощью двух вычислительных программ. Это Rulingva (https://rulingva.kpfu.ru) и Text Inspector (https://textinspector.com). Мы создали корпус из 40.000 слов, выбрав сорок (40) параллельных текстов из документов Организации Объединённых Наций (https://www.ohchr.org/en/ohchr_homepage), каждый текст состоял из 1000 слов. Стоит отметить, что документы Организации Объединенных Наций считаются политическими текстами. Поэтому мы сосредоточились на четырех следующих метриках: Flesch-Kincaid Grade Level (FKGL), лексическая плотность, лексическое разнообразие (TTR) и среднее значение этих метрик.

Результаты и обсуждения

Лексическая плотность русских текстов варьировалась от 68% до 72%, а среднее значение показало 69% (см. рисунок 4 ниже). Соотношение типов и слов в русских текстах варьировалось от 0,48 до 0,60, а среднее значение показало 0,5 (см. рис. 6 ниже). Метрика Flesch-Kincaid Grade Level основана на длине слов и длине предложений. Показатель читабельности (FKGL) русских текстов варьировался от 13 до 22 (рис. 5), а среднее значение показало четырнадцать (14). С другой стороны, лексическая плотность английских текстов варьировалась от 36% до 45%, а среднее значение - от 36% до 42% (рис. 1). Показатель читабельности (FKGL) английских текстов варьировался от 15 до 20, а среднее значение составило 17 (рис. 2). Соотношение типов и слов (TTR) варьировалось от 0,33 до 0,42, а среднее значение составило 0,38 (рис. 3).

Как уже было сказано выше, лексическая плотность и лексическое разнообразие (TTR) русских текстов оказались высокими, среднее значение 69% и 0,56 для TTR означает, что русские тексты сложнее для

анализа, чем английские тексты, лексическая плотность которых оказалась низкой (в диапазоне от 36 % до 45 %), а среднее значение TTR составило 42. Наиболее известным показателем лексического разнообразия является соотношение типов и слов (TTR) Templin [12, с. 183], это количество уникальных слов в тексте, деленное на общее количество слов в тексте. Лексическое разнообразие повышает сложность, поскольку каждое уникальное слово представляет новую информацию, которую необходимо перевести и интегрировать в контекст дискурса. С другой стороны, низкое лексическое разнообразие подразумевает большее количество повторов слов и избыточность. Лексическое разнообразие также связано с лексической сложности текста, поскольку оно указывает на то, что автор текста может использовать более широкий спектр слов. Текст является плотным, если он содержит много лексических слов по отношению к общему количеству слов, то есть лексических и функциональных. Более длинный текст обычно дает меньшее значение TTR, чем более короткий [7, с. 61-79]. В [8, с. 307-322] также отмечается, что лексическая плотность не обязательно измеряет лексику, поскольку она зависит от синтаксических и связных свойств текста. Опора на лексическое разнообразие была признана неадекватной для измерения развития словарного запаса несколькими авторами, которые утверждают, что TTR неизбежно падает с увеличением размера выборки лексем и, следовательно, не является показателем лексического разнообразия. Таким образом, любое отдельное значение TTR не является надежным, так как оно будет зависеть от длины в словах используемой единицы языков [9, с. 323-337]. Диапазон находится между теоретическим 0 (бесконечное повторение одного типа) и 1 (полное неповторение, обнаруженное в согласовании). В редких случаях исследователи выражают этот TTR в процентах, умножая коэффициент на 100 [13, с. 288].

Как правило, тексты с меньшей плотностью легче понять, и устные тексты имеют более низкий уровень лексической плотности, чем письменные [11, с. 443-452], [5, с. 109]. Однако, как утверждается в [7, с. 61-79], текст может иметь высокое лексическое разнообразие (содержать много различных типов слов), но низкую лексическую плотность (содержать много местоимений и вспомогательных слов, а не существительных и лексических глаголов) или наоборот. Таким образом, четвертая метрика, которую мы использовали в этой статье, называется читабельностью (FKGL). Читабельность по шкале Flesch-Kincaid Grade Level рассчитывалась автоматически с помощью программ *Rulingva* (https://rulingva.kpfu.ru) и *TextInspector* (https://textinspector.com). Обычно в некоторых показателях читабельности более низкий процент может означать более простой или более читабельный текст, а более высокий - на более сложный или трудный. Тем не менее, в зависимости от конкретной метрики или используемого анализа, порог, который считается "средним", может меняться. Важно учитывать и другие факторы и показатели, такие как длина предложения, выбор слов и синтаксическая сложность, чтобы получить более полное представление о читабельности или сложности текста.

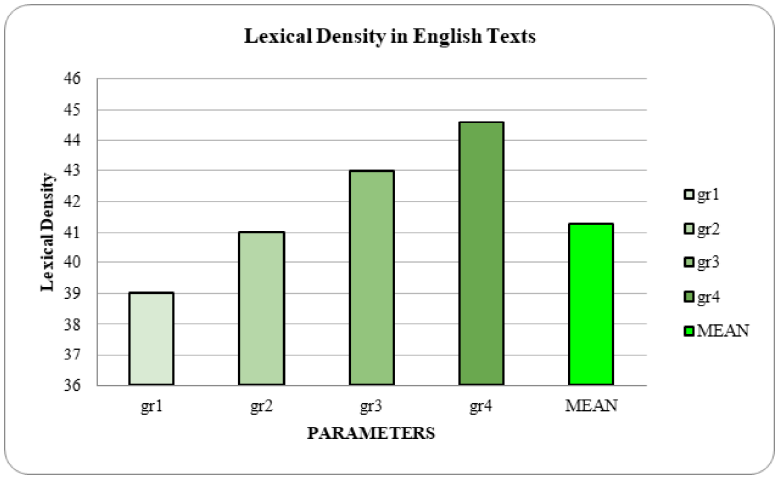

Рис. 1. Лексическая плотность и среднее значение 20 текстов на английском языке.
Fig. 1. Lexical density and average value of 20 texts in English.

Лексическая плотность первой группы варьируется между 36% и 39%; второй группы - между 36% и 41%; третьей группы - между 36% и 43%; четвертой группы - между 36% и 45%. Среднее значение по всем группам варьируется между 36% и 42%.

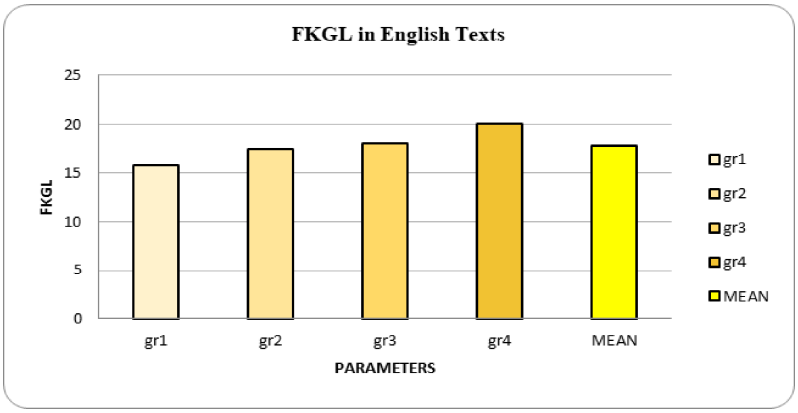

Рис. 2. Показатели читабельности (FKGL) и среднего значения 20 текстов на английском языке.
Fig. 2. Readability indices (FKGL) and average value of 20 texts in English.

Читабельность первой группы составляет 15, второй - 17, третьей - 18, четвертой - 20. Среднее значение всех групп составляет 17.

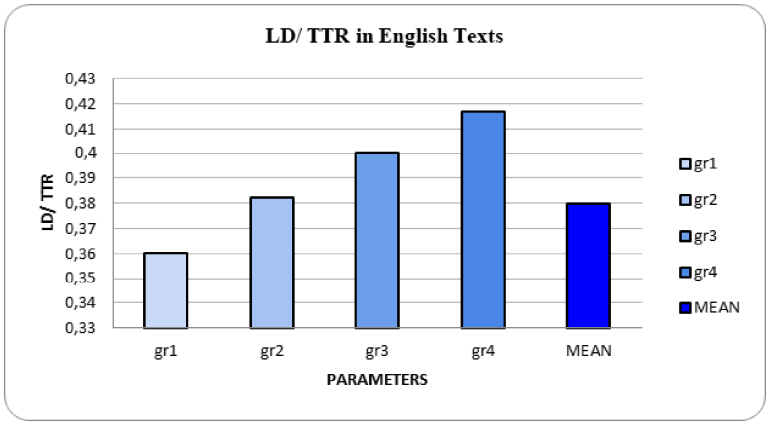

Рис. 3. Лексическое разнообразие (TTR) и среднее значение 20 текстов на английском языке.
Fig. 3. Lexical diversity (TTR) and average value of 20 English texts.

Отношение тип-токен в первой группе варьируется между 0,33 и 0,36; во второй группе - между 0,33 и 0,38; в третьей - между 0,33 и 0,40; в четвертой - между 0,33 и 0,42. Среднее значение по всем группам составляет от 0,33 до 0,38.

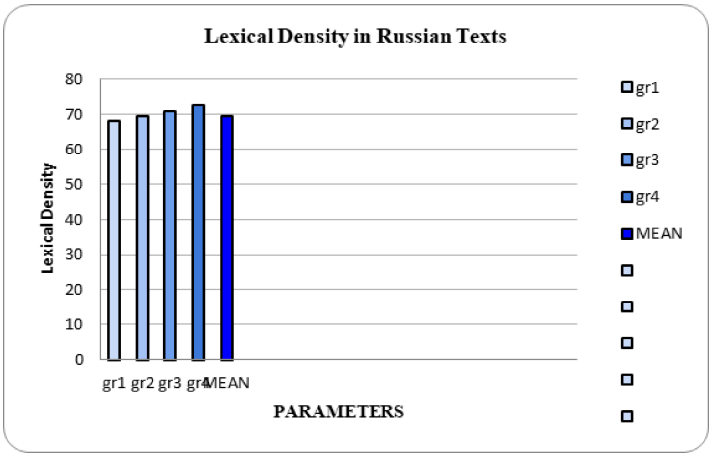

Рис. 4. Лексическая плотность и среднее значение 20 текстов на русском языке.
Fig. 4. Lexical density and average value of 20 texts in Russian.

Лексическая плотность первой группы составляет 68%, второй - 69%, третьей - 70%, четвертой - 72%. Среднее значение по всем группам составляет 69%.

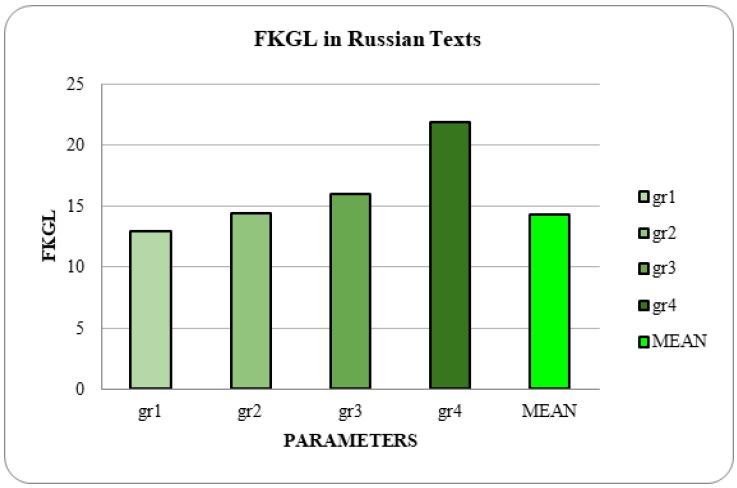

Рис. 5. Читабельность (FKGL) и среднее значение 20 текстов на русском языке.
Fig. 5. Readability (FKGL) and average value of 20 texts in Russian.

Читабельность первой группы составляет 13, второй - 14, третьей - 16, четвертой - 22. Среднее значение по всем группам составляет 14.

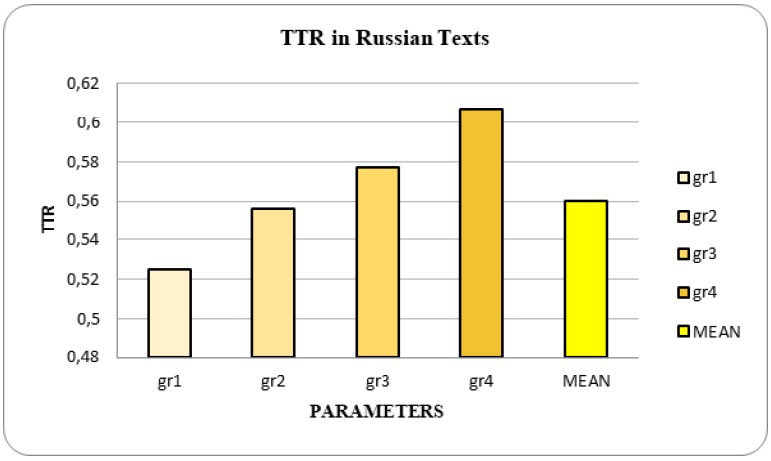

Рис. 6. Лексическое разнообразие (TTR) и среднее значение для 20 текстов на английском языке.
Fig. 6. Lexical diversity (TTR) and average value for 20 English texts.

Отношение тип-токен в первой группе варьируется между 0,48 и 0,52; во второй группе - между 0,48 и 0,56; в третьей - между 0,48 и 0,57; в четвертой - между 0,48 и 0,60. Среднее значение по всем группам варьируется между 0,48 и 0,56.

Следует отметить, что тексты Организации Объединенных Наций изобилуют техническими терминами, и использование технической лексики может способствовать усложнению текста по нескольким причинам:

- Конкретное и точное значение: техническая лексика часто включает термины, которые имеют точное значение в определенной области или сфере. Эти термины могут быть не общеизвестны или не понятны людям за рамками данной области. Когда используются такие термины, это может создать барьер для читателей, не знакомых с терминологией, и затруднить понимание текста.

- Требуются специальные знания: техническая лексика обычно ассоциируется со специальными знаниями или опытом в определенной предметной области. Она предполагает, что читатель обладает определенным уровнем фоновых знаний или знаком с темой. Если читатель не обладает такими знаниями, ему будет трудно понять текст, а это приведет к сложности.

- Отсутствие повседневного использования: техническая лексика не часто используется в разговорной речи. Она состоит из терминов, характерных для определенной профессии, отрасли или учебной дисциплины. В результате читателям, которые не сталкиваются с этими терминами регулярно, они могут показаться незнакомыми и сложными для восприятия, что повышает сложность текста.

- Интерпретация и контекстуальное понимание: техническая лексика часто требует более глубокого понимания предмета и связанных с ним понятий. Чтобы интерпретировать и правильно применять технические термины, читатель должен понимать контекст этой темы. Без такого понимания текст становится более сложным для восприятия, и его сложность повышается.

При использовании технической лексики важно учитывать аудиторию и цель текста. Хотя она может быть необходима в некоторых специализированных областях, ее следует использовать разумно и сопровождать четкими пояснениями или контекстом, чтобы текст оставался доступным для более широкого круга читателей.

Выводы

Данное исследование стало попыткой сделать некоторые выводы из наших предыдущих исследований. В наших предыдущих статьях мы упоминали, что лексическая плотность, лексическое разнообразие (TTR), уровень сложности по Flesch-Kincaid Grade Level (FKGL) могут служить предикторами сложности, однако в данной работе мы сосредоточились на средних показателях, чтобы понять, что лежит в основе сложности. Средний показатель 69% может означать, что текст имеет среднюю длину предложения или количество слов, которые считаются средними. Однако важно отметить, что конкретный расчет или метрика, используемая для определения среднего значения, может меняться в зависимости от контекста или проводимого анализа.

В контексте читабельности или сложности текста средний показатель в 69% не дает достаточно информации для вынесения однозначного заключения. Не зная конкретного расчета или метрики, использованной для определения "среднего значения", трудно точно интерпретировать его смысл. Различные системы или алгоритмы могут по-разному интерпретировать "среднее значение".

Как мы уже упоминали выше, лексическая плотность, лексическое разнообразие (TTR) и FKGL могут служить предикторами сложности. Результаты нашего исследования по лексической плотности подтверждают теорию Ure [11, с. 443-452.], уточненную Halliday [5, с. 109]. Согласно их исследованиям, плотность текста выше, если она превышает 40%. Как видно, плотность и русского, и английского текстов превысила 40%. Когда речь заходит о лексическом разнообразии (TTR), наши результаты также подтверждают исследования Johansson [7, с. 661-79, Malvern [13, с. 288] и Templin [12, с. 183]; они подтвердили, что длинные тексты имеют более низкий показатель TTR, а текст с более низким показателем TTR является показателем сложности. В отношении уровня сложности по Flesch-Kincaid Grade Level (FKGL), наши результаты показали, что и тексты на русском и английском языке длиннее. В связи с этим Danielle McNamara и др. [14, с. 18] утверждают, что более длинные предложения создают большую нагрузку на рабочую память и тем самым повышают сложность понимания. Ure, Halliday, Johansson, Templin, Malvern, Danielle McNamara и др. являются теми авторами, которые изучали различные категории текстов в области сложности и доказали своими исследованиями, что лексическая плотность, лексическое разнообразие (TTR) и уровень сложности по Flesch-Kincaid Grade Level являются предикторами сложности, как это было показано в нашем исследовании. Выбранные нами метрики могут служить предикторами сложности в юридических документах, в более конкретной области - в текстах ООН на русском и английском языке.

Тем не менее, важно отметить, что для точной интерпретации значения читабельности или сложности текста необходимо учитывать и некоторые другие факторы, помимо процента:

- Выбор слов: использование технической или специализированной лексики, незнакомых терминов или жаргона может способствовать повышению сложности текста.
- Синтаксическая сложность: расположение и структура предложений, включая использование подчиненных клаузул, могут повлиять на общую сложность текста.
- Словарный запас: уровень сложности слов, используемых в тексте, а также частота использования необычных или специализированных терминов могут повлиять на читабельность.
- Структура текста: организация и связность текста, включая использование заголовков, подзаголовков и переходов, может повлиять на удобочитаемость.
- Контекст и предмет: сложность обсуждаемой темы может потребовать более высокого уровня знаний или опыта для ее понимания.
- Подготовка и знания читателя: знакомство читателя с темой, его предыдущие знания и опыт могут повлиять на восприятие сложности текста.

Главный вывод, который мы можем сделать, заключается в том, что следует уделять больше внимания всем вышеупомянутым факторам и не полагаться только на один процент или метрику для точной оценки читабельности или сложности текста. Различные метрики и инструменты могут отдавать предпочтение

разным факторам, поэтому для получения полного понимания полезно использовать несколько подходов и учитывать общий контекст.

Список источников

1. Канн, Дэймон, Грег Гельцхаузер, Кейли Джонсон. Анализ сложности текста в исследованиях политической науки // PS: Political Science & Politics. 2014. № 47 (3). С. 63 – 66.
2. Дубай У.Х. Принципы читаемости. Коста-Меса: Impact Information, 2004. 74 с.
3. Дейл Э., Чалл Дж.С. Формула прогнозирования читаемости // Бюллетень образовательных исследований. 1948. № 27 (1). С. 11 – 28.
4. Даллер Х., Ван Хаут Р., Трефферс-Даллер Дж. Лексическое богатство спонтанной речи билингвов // Прикладная лингвистика. 2003. № 24/2. С. 197 – 222.
5. Холлидей М.А.К. Разговорный и письменный язык. Geelong Victoria: Deakin Univ. Press, 1985. 109 с.
6. Heydari P. The Validity of Some Readability Formulas // ACM Journal of Computer Documentation. Mediterranean Journal of Social Sciences. № 3 (2). P. 423 – 435.
7. Johansson V. Lexical variation and lexical density in speech and writing: a developmental perspective // Lund University, Dept. of Linguistics and Phonetics Working Papers. 2008. № 53. P. 61 – 79.
8. Laufer B., Nation P. Vocabulary Size and Use: Lexical Richness in L2 Written Production // Applied Linguistics. 1995. № 16.3. P. 307 – 322.
9. Макки Г., Малверн Д., Ричардс Б. Измерение словарного разнообразия с помощью специального программного обеспечения // Литературные и лингвистические вычисления. 2000. № 15. С. 323 – 337.
10. Редиш Дж. У формул читаемости есть еще больше ограничений, чем у Клэра. // Журнал компьютерной документации ACM. 2000. № 24 (3). С. 132 – 137.
11. Юре Дж. Лексическая плотность и дифференциация регистров. Приложения лингвистики: избранные доклады Второго международного конгресса прикладной лингвистики (Кембридж, 1969). ред. Г.Э. Перрин, Дж. Л.М. Трим. Кембридж: Cambridge Univ. Press, 1971. С. 443 – 452.
12. Templin M.C. Certain Language Skills in Children: Their Development and Interrelationships. Vol. 10. Minneapolis, MN: University of Minnesota Press, 1957. 183 p.
13. Malvern D., Richards B., Chipere N. et al. Lexical variation and language development: quantification and assessment. Basingstoke, UK: Palgrave Macmillan, 2004. 288 p.
14. Danielle S. McNamara, Arthur C. Graesser Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing. Cambridge University Press, 2011. 18 p.

References

1. Cann, Damon, Greg Goelzhauser, Kaylee Johnson. “Analyzing Text Complexity in Political Science Research. PS: Political Science & Politics. 2014. No. 47 (3). P. 63 – 66.
2. Dubay W.H. The principles of readability. Costa Mesa: Impact Information, 2004. 74 p.
3. Dale E., Chall J.S. A formula for Predicting Readability. Educational Research Bulletin. 1948. No. 27 (1). P. 11 – 28.
4. Daller H., Van Hout R., Treffers-Daller J. Lexical richness in the spontaneous speech of bilinguals. Applied Linguistics. 2003. No. 24/2. P. 197 – 222.
5. Halliday M.A.K. Spoken and written language. Geelong Victoria: Deakin Univ. Press, 1985. 109 p.
6. Heydari P. The Validity of Some Readability Formulas. ACM Journal of Computer Documentation. Mediterranean Journal of Social Sciences. No. 3 (2). P. 423 – 435.
7. Johansson V. Lexical diversity and lexical density in speech and writing: a developmental perspective. Lund University, Dept. of Linguistics and Phonetics Working Papers. 2008. No. 53. P. 61 – 79.
8. Laufer B., Nation P. Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics. 1995. No. 16.3. P. 307 – 322.
9. McKee G., Malvern D., Richards B. Measuring vocabulary diversity using dedicated software. Literary and Linguistic Computing. 2000. № 15. P. 323 – 337.
10. Redish J. Readability Formulas Have Even More Limitations Than Klare Discusses. ACM Journal of Computer Documentation. 2000. No. 24(3). P. 132 – 137.
11. Ure J. Lexical density and register differentiation. Applications of linguistics: selected papers of the Second International Congress of Applied Linguistics (Cambridge 1969). ed. G.E. Perren, J.L. M. Trim. Cambridge: Cambridge Univ. Press, 1971. P. 443 – 452.

12. Templin M.C. Certain Language Skills in Children: Their Development and Interrelationships, Vol. 10. Minneapolis, MN: University of Minnesota Press, 1957. 183 p.
13. Malvern D., Richards B., Chipere N. et al. Lexical diversity and language development: quantification and assessment. Basingstoke, UK: Palgrave Macmillan, 2004. 288 p.
14. Danielle S., McNamara Arthur C., Graesser Coh-Metrix: An Automated Tool for Theoretical and Applied Natural Language Processing. Cambridge University Press, 2011. 18 p.

Информация об авторах

Марико М.Л., *старший преподаватель кафедры теории и практики преподвавния иностранных языков Казанский (Приволжский) федеральный университет "КФУ"*