Елена Юрьевна ЗАХАРОВА[1]
Ольга Юрьевна САВИНА[2]

УДК 81’42

ЛЕКСИЧЕСКОЕ РАЗНООБРАЗИЕ ТЕКСТА И СПОСОБЫ ЕГО ИЗМЕРЕНИЯ

[1] студентка 4 курса кафедры немецкой филологии,
Тюменский государственный университет
helzakh@mail.ru; ORCID: 0000-0002-6511-600X

[2] кандидат филологических наук, доцент
кафедры немецкой филологии,
Тюменский государственный университет
o.y.savina@utmn.ru; ORCID: 0000-0002-4777-3188

Аннотация

В данной статье представлен обзор способов расчета коэффициента лексического разнообразия текста с их последующей классификацией, определены основные преимущества и недостатки способов, рассмотрены основные сферы практического применения коэффициента. Установлено, что самым распространенным способом является соотношение уникальных лексических единиц (тайпов) и всех словоформ (токенов) — TTR (англ. type-token ratio). Однако главной проблемой TTR и нескольких других производных способов является зависимость результата расчета от длины текста, то есть чем больше в тексте лексических единиц, тем ниже значение TTR. Таким образом, сравнение коэффициентов лексического разнообразия текстов разной длины невозможно. В связи с этим были разработаны другие способы расчета. Некоторые представляют собой видоизмененную формулу TTR, модифицированную квадратным корнем, логарифмом или другой математической операцией, однако они не решают проблему TTR. Другая группа способов использует в расчете обычную формулу TTR, дополненную принципом определения выборки, то есть полный текст не исследуется сразу, а разделяется на более удобные для исследования части. Такие

Цитирование: Захарова Е. Ю. Лексическое разнообразие текста и способы его измерения / Е. Ю. Захарова, О. Ю. Савина // Вестник Тюменского государственного университета. Гуманитарные исследования. Humanitates. 2020. Том 6. № 1 (21). С. 20-34.
DOI: 10.21684/2411-197X-2020-6-1-20-34

способы частично или полностью решают проблему зависимости результата от длины текста, но для их применения зачастую требуется дополнительный инструмент. Современные ученые склоняются к отказу от сложных формул и применению независимых способов, поскольку тексты для анализа в большинстве исследований имеют разную длину (особенно это касается текстов, не созданных специально для исследований — художественной литературы или законодательных актов), и здесь зависимые способы не могут дать корректный результат.

Ключевые слова

Лексическое разнообразие, коэффициент лексического разнообразия, лексикометрия, лингвостатистика, квантитативные методы, словарный запас, лексема.

DOI: 10.21684/2411-197X-2020-6-1-20-34

Введение

Как известно, текст, письменный или устный, обладает некоторыми количественными характеристиками. К самым простым характеристикам относят количество слов или предложений, слогов или среднюю длину предложения. Однако существуют также более сложные измерения. Одним из наиболее важных показателей считается лексическое разнообразие текста. В зарубежной науке лексическое разнообразие рассчитывается с 40-х гг. XX в., следовательно, за почти восемьдесят лет появилось большое количество способов его измерения.

В отечественной науке коэффициенту лексического разнообразия (далее — КЛР) уделено недостаточно внимания, и в исследованиях чаще всего используется только один из множества существующих способов. Данная статья является обзором наиболее часто применяемых способов измерения КЛР преимущественно в зарубежной науке с тем, чтобы определить их преимущества и недостатки и обозначить наиболее объективные из них, использование которых следует приветствовать в дальнейших лингвистических исследованиях.

Определение лексического разнообразия

Прежде чем говорить о способах расчета КЛР, необходимо дать ему определение и объяснить некоторые связанные понятия. Итак, лексическое разнообразие обозначает «диапазон и вариативность словарного запаса, который говорящий реализует в тексте» [18, с. 459]. В большинстве способов измерения при расчете учитывается количество всех словоформ (также текстоформ, или токенов (англ. tokens)) и количество уникальных лексических единиц (уникальных словоформ, тайпов (англ. types)).

КЛР может «служить мерой успешности речевого акта, например, при патологии речи или в ситуации говорения на иностранном языке» [7, с. 397]. Он также применим как мера прогресса овладения иностранным языком у учащихся, то есть когда лексическое разнообразие измеряется несколько раз на протяжении определенного срока. Наконец, учеными исследуется лексиче-

ское разнообразие художественных текстов и выявляются особенности авторского стиля.

Способы расчета КЛР

Существует несколько основных способов определения значения КЛР. Способы представляют собой формулы, в некоторых случаях дополненные принципом образования выборки.

Самым известным способом измерения КЛР является соотношение уникальных лексических единиц (тайпов) и всех лексических единиц (токенов) **TTR** (англ. The Type-Token Ratio [22]). Вычисляется по формуле:

$$TTR = \frac{V}{N}, \quad (1)$$

где V — количество уникальных лексических единиц, N — общее количество словоформ.

TTR исследовали такие ученые, как Е. Ливен в 1978 г. [14], а также Э. Бейтс, И. Бретертон и Л. Снайдер в 1988 г. [2]. Как пишет В. Йоханссон, проблемой измерения TTR является тот факт, что тексты, содержащие большое количество словоформ, всегда показывают более низкие результаты КЛР, и наоборот [10]. Причина в том, что количество словоформ может расти бесконечно, и даже если ситуация с уникальными лексемами такая же, для пишущего или говорящего часто необходимо использовать функциональные слова несколько раз, чтобы ввести в текст одно новое слово. Это значит, что у длинных текстов TTR априори будет ниже, чем у коротких. Одно из следствий этого — необходимость сравнивать тексты только одинаковой длины.

Несмотря на очевидный недостаток, TTR широко применяется в измерении лексического разнообразия. Х. Торруэлла и Р. Капсада применяют его при исследовании влияния стиля текста на величину КЛР, однако в результате отмечают непригодность TTR именно по причине зависимости результата от количества словоформ. Ученые проанализировали корпус текстов из 418 301 слова, разделив его на семнадцать блоков по 24 606 слов в каждом. Затем они применили формулу TTR для расчета коэффициента к каждому блоку и к корпусу в целом и получили примерно одинаковый результат для блоков (0,15-0,17) и намного меньший результат для корпуса целиком (0,04) [23].

Разные ученые решали проблему TTR двумя способами: либо видоизменяли формулу, либо усложняли принцип определения исследуемой выборки.

Индекс Жиро (англ. Guiraud-index, RTTR) схож с TTR. П. Жиро, автор “Les Caractères Statistiques du Vocabulaire. Essai de méthodologie”, ссылается на закон Ципфа [8]. Согласно этому закону, если все лексемы языка (в конкретном случае — одного текста) расположить в порядке убывания частоты их использования, то частота слова n будет примерно пропорциональна его порядковому номеру n. К примеру, третье слово встречается в тексте в три раза реже, чем

первое, четвертое — в четыре раза реже, чем первое, и т. д. Взяв в качестве материала исследования корпус французской литературы (тексты Ш. Бодлера, А. Рембо, Г. Аполлинера), П. Жиро вывел следующую формулу:

$$c = \frac{V}{\sqrt{N}}, \tag{2}$$

где V — количество уникальных лексических единиц (а именно существительных, глаголов, прилагательных и наречий, исключая служебные слова, напр. артикли и т. д.), а N — общее количество словоформ [6].

В статье «Comparing measures of lexical richness» ее авторы А. Вермеер и Р. ван Хаут исследуют индекс Жиро и другие видоизмененные формулы TTR, чтобы определить наиболее валидный способ, и приходят к выводу, что, хотя данный способ часто является лучшей трансформацией формулы TTR (о других видоизмененных формулах TTR будет сказано ниже), он не предоставляет полной картины качества использованной в исследуемом тексте лексики. Авторы статьи утверждают: «Квадратный корень от TTR — золотая середина между бездействием и применением слишком сложных трансформаций» [25, с. 136].

Скорректированное TTR, или **TTR(c)** (англ. corrected TTR(c)), было описано в работе Дж. Кэролла в 1964 г. [3]. Оно является вариантом TTR и рассчитывается по формуле:

$$TTR(c) = \frac{V}{\sqrt{2N}}, \tag{3}$$

где V — количество отдельных лексических единиц, N — количество всех словоформ.

По мнению А. Вермеера, данное измерение идентично индексу Жиро, а удвоение показателя N не имеет смысла. По этой причине ученый исключает данный способ, исследуя спонтанный дискурс людей, изучающих голландский язык [26]. Среди работ других ученых TTR(c) также непопулярен.

Индекс Хердана (англ. Index of Herdan) предложен и описан Г. Херданом в 1955 г. [9]. Несмотря на наличие логарифма в формуле, индекс Хердана остается идентичным показателю TTR, так как это по-прежнему то же соотношение:

$$C = \frac{\log V}{\log N}, \tag{4}$$

где C — величина данного показателя, V — количество уникальных лексических единиц, N — количество уникальных словоформ.

Р. ван Хаут и А. Вермеер считают подобную трансформацию формулы TTR нецелесообразной и отмечают, что логарифм не помогает решить проблему TTR [25].

П. Лиссон и Н. Балльер применяют индекс Хердана как средство измерения прогресса студентов, изучающих французский в качестве третьего иностранного языка. Для этого ученые анализируют тексты студентов, написанные ими в течение трех месяцев учебного года. Затем у данных текстов производится расчет КЛР с целью выявить закономерный прогресс. Ученые отмечают, что индекс Хердана наиболее точно помогает отследить изменения в прогрессе во владении иностранным языком с течением времени [15].

Убер-индекс (англ. Uber-Index), разработанный Дугастом в 1978 г., [прив. по 24]. Его также называют «средней частотой употребления слова (англ. the mean word frequency, MWF), так как он показывает среднее количество словоформ, использованных для каждой уникальной лексической единицы в тексте. Очевидно, что средняя частота употребления слова возрастает с количеством словоформ» [17].

Данный показатель имеет следующую формулу:

$$U = \frac{(\log N)^2}{(\log N - \log V)}, \quad (5)$$

где U — значение КЛР, N — количество словоформ, V — количество уникальных лексических единиц.

А. Вермеер пишет: «Ван Хельверт использовала Убер-индекс при изучении спонтанной речи пяти турецких детей, изучающих голландский как второй язык, и обнаружила, что большинство различий между владением языком у детей было сведено к нулю» [26, с. 67]. В собственном исследовании А. Вермеера есть этому подтверждение: он представляет сводную таблицу с данными TTR, количеством всех использованных словоформ и уникальных лексических единиц и т. д. у детей, чей словарный запас голландского языка (как первого и как второго иностранного) равен (в среднем) 1 500, 3 000, 4 500 и 6 000 слов. Uber-index первой группы составляет 15,14, тогда как у второй группы он наиболее высокий — 17,58, а у третьей и четвертой почти не отличается — 17,28 и 17,48 соответственно [26].

Индекс Сомерса (англ. Somers's index) рассчитывается по формуле:

$$S = \frac{\log \log V}{\log \log N}, \quad (6)$$

где V — количество уникальных лексических единиц, а N — общее количество словоформ [21].

П. Лиссон и Н. Балльер, отметившие валидность индекса Хердана при отслеживании прогресса студентов в иностранном языке, пишут, что индекс Сомерса демонстрирует очевидный прогресс КЛР только между первым и вторым текстом [15].

Индекс Мааса (англ. Maas' index):

$$a^2 = \frac{\log N - \log V}{\log N^2}, \quad (7)$$

где V — количество уникальных лексических единиц, N — общее количество словоформ [16].

Индекс Мааса уменьшается с увеличением TTR. Согласно Ф. М. Маккарти и С. Джарвису, индекс Мааса относительно независим от длины текста для разговорных жанров текста и в особенности независим для письменных жанров [18]. Х. Торруэлла и Р. Капсада при исследовании влияния стиля текста на его лексическое разнообразие также отмечают низкую чувствительность индекса Мааса к длине текста [23].

Программа vocd-D (англ. Vocabulary Diversity) [18]. При расчете КЛР vocd-D использует не весь текст целиком, а его отдельные сегменты (без перестановки слов в них). В общем исследуется сто случайных выборок, состоящих из тридцати пяти словоформ каждая. Сначала рассчитывается TTR для каждой отдельной выборки, а затем находится среднее арифметическое всех полученных ранее TTR. Процедура повторяется для выборок длиной от тридцати шести до пятидесяти словоформ. Затем создается эмпирическая кривая для каждого среднего значения TTR.

Параметр D используется как часть формулы для создания теоретической кривой значений TTR, которая наиболее близка к эмпирической кривой TTR, сформированной из средних значений TTR случайных выборок [17]. Наиболее подходящая величина обозначается как D. Так как D формируется из результатов TTR случайных выборок, его величина меняется с каждым новым вычислением. Разница в значениях не так велика, но чтобы достичь наиболее точных результатов, всё вышеперечисленное повторяется трижды, после чего вычисляется среднее значение параметра D. Финальные величины ранжируются от десяти до ста. Наибольшие величины означают наибольшее разнообразие. TTR в данном случае вычисляется по видоизмененной формуле:

$$\text{TTR} = \left(\frac{2}{DN}\right)\left[(1 + DN)^{\frac{1}{2}} - 1\right], \quad (8)$$

где N — количество всех словоформ, D — наиболее подходящая величина теоретической кривой к эмпирической кривой.

Дж. Т. Макки и Дж. Р. Брайан, авторы «Measuring Vocabulary Diversity Using Dedicated Software», называют следующие преимущества параметра D:

1) он не зависит от количества слов в выборке;
2) он использует все возможные данные;
3) он более информативен: представляет то, каким образом TTR меняется в зависимости от диапазона словоформ для каждого говорящего или пишущего [20].

Ф. Маккарти и С. Джарвис исследуют валидность voc-d и утверждают, что «случайная выборка приводит к большим различиям при оценке результатов текстов с очень большим разнообразием» [19, с. 390].

TTR сегмента (англ. mean segment type-token ratio, MSTTR) рассчитывается путем разделения текста на сегменты, состоящие из фиксированного количества слов, и вычисления TTR для каждого сегмента с последующим нахождением среднего значения TTR [11]. Этот способ также независим от длины исследуемого текста. М. А. Ковингтон и Дж. Д. Макфолл отмечают, что MSTTR не позволяет в полной мере проследить отличия в TTR в пределах текста.

MSTTR применяется в работе Х. Торруэллы и Р. Капсады при определении различий лексического разнообразия текстов разных жанров. Авторы исследования отмечают валидность данного способа при подобном исследовании [23].

Скользящее среднее значение TTR (англ. moving average type-token ratio, MATTR) является откорректированным вариантом MATTR [4]. Если сегменты в MSTTR строго определены, то при вычислении MATTR для одного сегмента также устанавливается определенная длина, но сегмент движется, например: если в тексте тысяча слов, сначала рассчитывается TTR для слов с первого по сто первое, затем со второго по сто второе и т. д. (длина сегмента при этом равна ста словам). После этого производится расчет среднего значения всех полученных TTR (среднее арифметическое). Как утверждают авторы способа, MATTR не зависит от длины текста. В свободном доступе существуют программы для вычисления MATTR, написанные на языке C#.

В статье «Measuring Lexical Diversity in Narrative Discourse of People With Aphasia» ее автор Г. Фергадиотис испытывает валидность MATTR в области афазиологии. Он называет MATTR убедительным показателем лексического разнообразия дискурса людей с афазией. «Значительным преимуществом MATTR является его достоверность, потому что он эквивалентен TTR и достаточно прост для понимания и объяснения. Достоверность является весьма желательным свойством, особенно для профессионалов, работающих с людьми с языковыми и речевыми нарушениями в лабораторных условиях» [7, с. 406].

Профиль лексической частотности (англ. Lexical Frequency Profile, LFP). В основе данного способа лежит анализ частотных списков слов. Слова из текста разделяются на четыре списка: первая и вторая тысяча наиболее часто употребленных слов, слова из академического словаря Э. Коксхед (англ. the Academic Word List) [5] и слова, не подошедшие ни одной из предыдущих групп [13]. LFP работает только с английским языком.

LFP показывает, сколько процентов слов из каждой группы есть в данном тексте. Для расчета LFP используется программа RANGE, которая сопоставляет частотные списки со словами из анализируемого текста. Например, если эссе учащегося содержит двести слов, и из них сто шестьдесят из первой группы, двадцать из второй, десять из AWL и десять из последней, LFP будет равен 80%–10%–5%–5%.

Опыт применения LFP демонстрирует группа ученых из Ирана, исследуя лексическое разнообразие продуктивного словарного запаса студентов, изучающих английский язык как второй иностранный. Согласно их исследованию, «знание 2 000 самых часто употребляемых слов является необходимым минимумом для базовой устной коммуникации и составляют 87 процентов письменных текстов и около 80 процентов стандартных академических текстов. Знание 3 000 слов необходимо для минимального восприятия и 5 000 слов для чтения на досуге». Ученые также доказали, что студенты шестого и восьмого семестров используют больше сложной (не самой частотной) лексики, чем студенты второго и четвертого семестров. Они также повышают свои знания лексики при прохождении разных курсов и могут применять изученную лексику в речи [1, с. 1846].

Мера текстового лексического разнообразия (англ. The Measure of Textual Lexical Diversity, MTLD) означает «среднюю длину последовательных строк текста, сохраняющую заданный TTR», а именно, по определению автора способа, 0,72 [19, с. 384]. TTR рассчитывается для каждого слова последовательно. Как только TTR становится меньше, чем 0,72, строка заканчивается и начинается новая. Строка не должна быть короче десяти слов. Конечное значение MTLD рассчитывается по формуле:

$$MTLD = \frac{L}{n}, \tag{9}$$

где L — количество словоформ, n — общее количество строк.

Как показывают результаты исследования, отраженные в статье «Relationships between text length and lexical diversity measures: can we use short texts of less than 100 tokens?» [12], MTLD, в сравнении с TTR, индексом Жиро и vocd-D, действительно наименее зависим от длины текста, однако эта зависимость всё-таки прослеживалась среди текстов длиной от пятидесяти до ста словоформ, от ста до двухсот словоформ и от пятидесяти до двухсот словоформ. Х. Торруэлла и Р. Капсада отмечают, что MTLD может быть чувствителен к длине текста [23]. Считается, что наилучшим объемом текста для анализа с помощью MTLD является текст в по крайней мере сто словоформ. Следовательно, для анализа коротких текстов данный способ не подойдет. Для расчета MTLD была разработана программа Gramulator, также существует онлайн-инструмент Textinspector.

Как следует из представленного обзора, особой важностью при характеристике способов обладают принципы расчета КЛР (формула с дополнением или только формула), зависимость результата от длины текста, простота расчета и необходимость специального программного обеспечения. На основании указанных критериев были составлены две сводные таблицы, которые позволяют увидеть все характеристики в сжатой форме. В таблице 1 представлены способы, зависимые от длины текста, в таблице 2 — независимые способы.

Данные таблицы позволяют нам сделать несколько выводов. Все способы, построенные по принципу видоизменения формулы TTR, как и сам TTR, зависят

от длины исследуемого текста. Такие способы не требуют специальной программы для расчета, так как могут быть рассчитаны при помощи обычного калькулятора. Однако результаты, полученные при использовании этих способов, априори необъективны, например, при сравнении КЛР текстов разной длины. Таблицы сформированы по принципу зависимости результатов вычисления от длины текста, так как именно он был поводом для разработки новых способов вычисления КЛР.

Другой важный вывод заключается в том, что исследования в области способов вычисления КЛР на данный момент строятся вокруг изначального соотношения количества уникальных лексических единиц и общего количества словоформ, но исследуется не полный текст в первоначальном виде, а его отдельные фрагменты, определяемые по разным принципам (как это делается, например, при вычислении MATTR). Впрочем, попытки исследовать текст иначе, чем от начала и до конца без каких-либо изменений (например, MSTTR), предпринимались учеными еще в 40-х гг. XX в., однако не получили тогда широкого распространения.

Современные ученые склоняются к применению независимых способов расчета КЛР. Ученые, которые с помощью КЛР отслеживают прогресс в овладении учащимися иностранными языками, аргументируют свой выбор зависимостью размера текста от уровня языковых знаний. При лонгитюдных исследованиях в области английского языка достаточно распространен LFP. Ученые-афазиологи называют MTLD и MATTR наиболее подходящими способами

Таблица 1

Основные характеристики способов вычисления КЛР, зависимых от длины текста

Table 1

The main characteristics of LD measures dependent on the text length

Способ	Вид способа	Простота расчета	Необходимость специального ПО	Примечание
TTR	Формула	Да	Нет	Наиболее известен и широко применяется
c/RTTR	Измененная формула TTR	Да	Нет	—
TTR(c)	Измененная формула TTR	Да	Нет	—
C	Измененная формула TTR	Да	Нет	—
U	Измененная формула TTR	Да	Нет	—
S	Формула	Да	Нет	—

Таблица 2

Основные характеристики способов вычисления КЛР, независимых от длины текста

Table 2

The main characteristics of LD measures independent on the text length

Способ	Вид способа	Простота расчета	Необходимость специального ПО	Примечание
a^2	Формула	Да	Нет	Чувствителен к длине текста менее, чем MTLD, MSTTR и vocd-D
D/vocd-D	Измененная формула TTR, способ образования выборки	Нет	Да	—
MSTTR	Формула TTR, способ образования выборки	Нет	Да	—
MATTR	Формула TTR, способ образования выборки	Нет	Да	Вариант MSTTR
LFP	Анализ частотных списков слов	Нет	Да	ПО только на английском; позволяет проверять только длинные тексты
MTLD	Формула TTR, способ формирования выборки	Нет	Нет	Зависимость прослеживалась в текстах из 50-100, 100-200, 50-200 словоформ

вычисления КЛР. Наконец, при изучении влияния стиля текста на его лексическое разнообразие главную роль играют индекс Мааса и MSTTR, поскольку они обладают наименьшей чувствительностью к размеру текста.

Заключение

Итак, обзор основных способов измерения КЛР показал, что все их разнообразие можно классифицировать на две группы: способы, результаты которых зависят от длины текста, и независимые способы. К первой группе относится самый известный способ, называемый TTR, и его производные, такие как индекс Жиро, индекс Хердана и другие. Все они представляют собой уравнение и могут быть рассчитаны без применения специального программного обеспечения. Независимые способы, такие как MTLD, индекс Мааса и другие, состоят из формулы TTR и принципа определения выборки. Следовательно, для расчета КЛР с их помощью необходима специальная программа.

В настоящее время ученые склоняются к выводу, что использовать способы расчета КЛР, независимые от длины текста, правильнее и логичнее. Однако, поскольку для расчета с их помощью необходима программа, в экспериментах создаются условия, когда все участники намеренно создают тексты с примерно одинаковым количеством слов. Это применяется при отслеживании прогресса в иностранном языке с течением времени. Однако если исследование касается текстов, написанных не в экспериментальных целях, например, исследования о литературных трудах или научных работах, зависимые способы уступают независимым. Данный обзор призван помочь определить, какой из представленных способов вычисления коэффициента лексического разнообразия необходимо применять в том или ином исследовании.

СПИСОК ЛИТЕРАТУРЫ

1. Azodi N. Measuring the lexical richness of productive vocabulary in Iranian EFL University students' writing performance / N. Azodi, F. Karimi, R. Vaezi // Theory and Practice in Language Studies. 2014. Vol. 4. No. 9. Pp. 1837-1849.
2. Bates E. From first words to grammar: individual differences and dissociable mechanisms / E. Bates, I. Bretherton, L. Snyder. Cambridge: Cambridge University Press, 1988. 326 p.
3. Carroll J. B. Language and Thought / J. B. Carroll. Englewood Cliffs N. J.: Prentice Hall, 1964. 118 p.
4. Covington M. A. Cutting the Gordian Knot: the Moving-Average Type-Token Ratio (MATTR) / M. A. Covington, J. D. McFall // Journal of Quantitative Linguistics. 2010. Vol. 17. No. 2. Pp. 94-100.
5. Coxhead A. A new academic word list / A. Coxhead // TESOL Quarterly. 2000. Vol. 34. No. 2. Pp. 213-238.
6. Daller M. Guiraud's index of lexical richness / M. Daller // UWE Bristol Research Repository. 2011. URL: http://eprints.uwe.ac.uk/11902/ (дата обращения: 12.12.2019).
7. Fergadiotis G. Measuring lexical diversity in narrative discourse of people with aphasia / G. Fergadiotis, H. W. Heather, M. W. Thomas // American Journal of Speech-Language Pathology. 2013. Vol. 22. No. 2. Pp. 397-408.
8. Guiraud P. Les Caractères Statistiques du Vocabulaire. Essai de méthodologie / P. Guiraud. Paris: Presses Universitaires de France, 1954. 116 p.
9. Herdan G. A. New derivation and interpretation of Yule's "Characteristic" K / G. A. Herdan // Zeitschrift für angewandte Mathematik und Physik. 1955. Vol. 6. Pp. 332-334.
10. Johansson V. Lexical diversity and lexical density in speech and writing: a developmental perspective / V. Johansson // Working Papers. 2008. Vol. 53. Pp. 61-79.
11. Johnson W. I. A program of research / W. I. Johnson // Psychological Monographs. 1944. Vol. 56. No. 2. Pp. 1-15.
12. Koizumi R. Relationships between text length and lexical diversity measures: can we use short texts of less than 100 tokens? / R. Koizumi // Vocabulary Learning and Instruction. 2012. Vol. 1. No. 1. Pp. 60-69.

13. Laufer B. Vocabulary size and use: lexical richness in L2 written production / B. Laufer, P. Nation // Applied Linguistics. 1995. Vol. 16. No. 3. Pp. 307-322.
14. Lieven E. V. M. Conversations between mothers and young children: individual differences and their possible implication for the study of child language learning / E. V. M. Lieven // The Development of Communication / N. Waterson, C. E. Snow (eds.). Chichester: Wiley, 1978.
15. Lissón P. Investigating lexical progression through lexical diversity metrics in a Corpus of French L3 / P. Lissón, N. Ballier // Discours. 2018. Vol. 23. URL: https://www.researchgate.net/publication/333723678_Investigating_Lexical_Progression_through_Lexical_Diversity_Metrics_in_a_Corpus_of_French_L3 (дата обращения: 22.12.2019).
16. Maas H. D. Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes / H. D. Maas // Zeitschrift für Literaturwissenschaft und Linguistik. 1972. Vol. 2. No. 8. Pp. 73-96.
17. Malvern D. Lexical Diversity and Language Development: Quantification and Assessment / D. Malvern, B. Richards, N. Chipere, P. Durán. Hampshire: Palgrave Macmillan, 2004. 272 p.
18. McCarthy P. M. Voc-D: A theoretical and empirical evaluation / P. M. McCarthy, S. Jarvis // Language Testing. 2007. Vol. 24. No. 4. Pp. 459-488.
19. McCarthy P. M. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment / P. M. McCarthy, S. Jarvis // Behavior Research Methods. 2010. Vol. 42. No. 2. Pp. 381-392.
20. McKee G. T. Measuring vocabulary diversity using dedicated software / G. T. McKee, J. R. Brian // Literary and Linguistic Computing. 2000. Vol. 15. No. 3. Pp. 323-337.
21. Somers H. H. Statistical methods in literary analysis / H. H. Somers // The Computer and Literary Style // J. Leeds (ed.). Kent, OH: Kent State University, 1966. Pp. 128-140.
22. Templin M. Certain Language Skills in Children: Their Development and Inter-Relationships / M. Templin. Minneapolis, MN: University of Minnesota Press, 1957. 208 p.
23. Torruella J. Lexical statistics and tipological structures: a measure of lexical richness / J. Torruella, R. Capsada // Social and Behavioral Sciences. 2013. Vol. 95. Pp. 447-454.
24. Tweedie F. J. How variable may a constant be? Measures of lexical richness in perspective / F. J. Tweedie, R. H. Baayen // Computers and the Humanities. 1998. Vol. 32. No. 5. Pp. 323-352.
25. Van Hout R. Comparing Measures of Lexical Richness. Modelling and Assessing Vocabulary Knowledge / R. Van Hout, A. Vermeer. Amsterdam: Benjamins, 2007.
26. Vermeer A. Coming to grips with lexical richness in spontaneous speech data / A. Vermeer // Language Testing. 2000. Vol. 17. Pp. 65-83.

Elena Yu. ZAKHAROVA[1]
Olga Yu. SAVINA[2]

UDC 81'42

LEXICAL DIVERSITY MEASURES' REVIEW AND CLASSIFICATION

[1] Undergraduate Student,
Department of German Philology,
University of Tyumen
helzakh@mail.ru; ORCID: 0000-0002-6511-600X

[2] Cand. Sci. (Philol.), Associate Professor,
Department of German Philology,
University of Tyumen
o.y.savina@utmn.ru; ORCID: 0000-0002-4777-3188

Abstract

This paper reviews various lexical diversity (LD) measures and their classification. The authors define the most significant advantages and disadvantages of the measures and investigate the main scopes of LD application. They include measuring LD in the speech of children and people with aphasia, checking progress in learning a foreign language, and investigating different writing styles of certain authors. Results show that the most frequently used measure is the type-token ratio (TTR), which means the ratio of different words (types) to the total number of words (tokens).

The most important problem of TTR and other measures based on TTR is that the more tokens a text has, the less is the TTR value. This has led to the development of other measures; some of them are based on a TTR formula, thus, they do not solve the problem and the calculation result is also affected by the text length. In that case, the texts with different length cannot be compared.

Another group of measures rests upon the TTR formula supplemented by a principle of sample forming. These measures solve the problem of the TTR partially or completely, though they often require some extra instruments. Fortunately, these instruments are

Citation: Zakharova E. Yu., Savina O. Yu. 2020. "Lexical Diversity Measures' Review and Classification". Tyumen State University Herald. Humanities Research. Humanitates, vol. 6, no. 1 (21), pp. 20-34.
DOI: 10.21684/2411-197X-2020-6-1-20-34

available on the Internet and demand no particular knowledge on their working principle or in programming.

Contemporary researchers tend to use independent measures, because texts mostly have different length and the dependent measures cannot give proper results.

Keywords

Lexical diversity, lexical diversity measures, lexicometry, statistical linguistics, quantitative methods, vocabulary, lexeme.

DOI: 10.21684/2411-197X-2020-6-1-20-34

REFERENCES

1. Azodi N., Karimi F., Vaezi R. 2014. "Measuring the lexical richness of productive vocabulary in Iranian EFL university students' writing performance". Theory and Practice in Language Studies, vol. 4, no. 9, pp. 1837-1849.
2. Bates E., Bretherton I., Snyder L. 1988. From first Words to Grammar: Individual Differences and Dissociable Mechanisms. Cambridge: Cambridge University Press.
3. Carroll J. B. 1964. Language and Thought. Englewood Cliffs N.J.: Prentice Hall.
4. Covington M. A., McFall J. D. 2010. "Cutting the Gordian knot: the moving-average type-token ratio (MATTR)". Journal of Quantitative Linguistics, vol. 17, no. 2, pp. 94-100.
5. Coxhead A. 2000. "A new academic word list". TESOL Quarterly, vol. 34, no. 2, pp. 213-238.
6. Daller M. 2011. "Guiraud's index of lexical richness". UWE Bristol Research Repository. Accessed 12 December 2019. http://eprints.uwe.ac.uk/11902/
7. Fergadiotis G., Heather H. W., Thomas M. W. 2013. "Measuring lexical diversity in narrative discourse of people with aphasia". American Journal of Speech-Language Pathology, vol. 22, no. 2, pp. 397-408.
8. Guiraud P. 1954. Les Charactères Statistiques du Vocabulaire. Essai de méthodologie. Paris: Presses Universitaires de France.
9. Herdan G. A. 1955. "New derivation and interpretation of Yule's 'Characteristic' K". Zeitschrift für angewandte Mathematik und Physik, vol. 6, pp. 332-334.
10. Johansson V. 2008. "Lexical diversity and lexical density in speech and writing: a developmental perspective". Working Papers, vol. 53, pp. 61-79.
11. Johnson W. I. 1944. "A program of research". Psychological Monographs, vol. 56, no. 2, pp. 1-15.
12. Koizumi R. 2012. "Relationships between text length and lexical diversity measures: can we use short texts of less than 100 tokens?". Vocabulary Learning and Instruction, vol. 1, no. 1, pp. 60-69.
13. Laufer B., Nation P. 1995. "Vocabulary size and use: lexical richness in l2 written production". Applied Linguistics, vol. 16, no. 3, pp. 307-322.
14. Lieven E. V. M. 1978. "Conversations between mothers and young children: individual differences and their possible implication for the study of child language learning". In: Waterson N., Snow C. E. (eds.). The Development of Communication. Chichester: Wiley.

15. Lissón P., Ballier N. 2018. “Investigating lexical progression through lexical diversity metrics in a corpus of french L3”. Discours, vol. 23. Accessed 22 December 2019. https://www.researchgate.net/publication/333723678_Investigating_Lexical_Progression_through_Lexical_Diversity_Metrics_in_a_Corpus_of_French_L3
16. Maas H. D. 1972. “Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes”. Zeitschrift für Literaturwissenschaft und Linguistik, vol. 2, no. 8, pp. 73-96.
17. Malvern D., Richards B., Chipere N., Durán P. 2004. Lexical Diversity and Language Development: Quantification and Assessment. Hampshire, Palgrave Macmillan.
18. McCarthy P. M., Jarvis S. 2007. “Voc-D: a theoretical and empirical evaluation”. Language Testing, vol. 24, no. 4, pp. 459-488.
19. McCarthy P. M., Jarvis S. 2010. “MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment”. Behavior Research Methods, vol. 42, no. 2, pp. 381-392.
20. McKee G.T, Brian J.R. 2000. “Measuring vocabulary diversity using dedicated software”. Literary and Linguistic Computing, vol. 15, no. 3, pp. 323-337.
21. Somers H. H. 1996. “Statistical methods in literary analysis”. In: Leeds J. (ed.). The Computer and Literary Style. Kent, OH: Kent State University.
22. Templin M. 1957. Certain Language Skills in Children: Their Development and Inter-relationships. Minneapolis, MN: University of Minnesota Press.
23. Torruella J., Capsada R. 2013. “Lexical statistics and tipological structures: a measure of lexical richness”. Social and Behavioral Sciences, vol. 95, pp. 447-454.
24. Tweedie F. J., Baayen R. H. 1998. “How variable may a constant be? Measures of lexical richness in perspective”. Computers and the Humanities, vol. 32, no. 5, pp. 323-352.
25. Van Hout R., Vermeer A. 2007. Comparing Measures of Lexical Richness. Modelling and Assessing Vocabulary Knowledge. Amsterdam: Benjamins.
26. Vermeer A. 2000. “Coming to grips with lexical richness in spontaneous speech data”. Language Testing, vol. 17, pp. 65-83.