ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА/THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

DOI: https://doi.org/10.60797/RULB.2025.67.5

COMPARATIVE ANALYSIS OF LEXICAL DENSITY, LEXICAL DIVERSITY, AND MULTIWORD EXPRESSIONS IN RUSSIAN, ENGLISH, AND FRENCH LEGAL TEXTS: IMPLICATIONS FOR READABILITY AND UNDERSTANDABILITY

Research article

Mariko M.[1, *]
[1] ORCID : 0009-0006-9036-305X;
[1] Kazan (Volga Region) Federal University, Kazan, Russian Federation

* Corresponding author (mlmariko[at]kpfu.ru)

Abstract

Lexical density is closely related to the notion of information packaging as content words in a text; therefore, texts with a higher proportion of content words are dense as they contain more information as opposed to texts that have a higher proportion of function words [10, P. 61–79]. *Type-token ratio (TTR)*, also known as vocabulary size divided by text length, is a simple measure of lexical diversity. Lexical diversity refers to how varied the vocabulary used in a text is. For texts of similar length, the traditional type-token ratio can be used, which is the number of different words (types) in a text divided by the total number of words (tokens) [1, P. 185–207]. *Multiword expressions* refer to a diverse group of linguistic phenomena, connected by the fact that they do not fit neatly into the word-phrase dichotomy. Like phrases, they appear to be made up of multiple words. In our research, we analyzed text materials with *lexical density, TTR, and multiword expressions* from legal texts (texts of the United Nations). We compared the automatic analysis results to three linguistic measures in Russian, English, and French. A 60,000-word-based corpus was built for the analysis. Our research aimed at examining the *lexical density, TTR, and multiword expressions* of Russian, English, and French UN texts. To reach that goal, we used Rulingva, TextInspector, and LancsBox to compute our data. The results showed that the linguistic features selected for the investigation could impact complexity on account of lexical richness, being a multidimensional concept that encompasses several aspects of lexis use [12, P. 19].

Keywords: French UN texts, English UN texts, Russian UN texts, TTR, lexical density, multi-word expressions, n-grams, readability.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЛЕКСИЧЕСКОЙ НАСЫЩЕННОСТИ, ЛЕКСИЧЕСКОГО РАЗНООБРАЗИЯ И МНОГОСЛОВНЫХ ВЫРАЖЕНИЙ В РУССКИХ, АНГЛИЙСКИХ И ФРАНЦУЗСКИХ ЮРИДИЧЕСКИХ ТЕКСТАХ: ВЛИЯНИЕ НА ЧИТАЕМОСТЬ И ПОНИМАНИЕ

Научная статья

Марико М.[1, *]
[1] ORCID : 0009-0006-9036-305X;
[1] Казанский (Приволжский) федеральный университет, Казань, Российская Федерация

* Корреспондирующий автор (mlmariko[at]kpfu.ru)

Аннотация

Лексическая насыщенность тесно связана с понятием концентрации информации в виде ключевых слов в тексте; поэтому тексты с большей долей ключевых слов являются насыщенными, так как содержат больше информации, в отличие от текстов с большей долей служебных слов [10, С.61–79]. Показатель лексического разнообразия текста (*Type-token ratio, TTR*), также известный как объем словарного запаса, поделенный на длину текста, является простой мерой лексического разнообразия. Лексическое разнообразие определяет, насколько многообразен словарный запас, используемый в тексте. Для текстов одинаковой длины можно использовать традиционное соотношение, которое представляет собой количество различных слов в тексте, разделенное на их общее количество [1, С. 185–207]. *Многословные выражения* относятся к разнообразной группе лингвистических явлений, связанных между собой тем, что они не укладываются в рамки дихотомии «слово-фраза». Как и фразы, они состоят из нескольких слов. В нашем исследовании мы проанализировали юридические текстовые материалы (тексты Организации Объединенных Наций) с *лексической плотностью, TTR* и *многословными выражениями*. Мы сравнили результаты автоматического анализа с тремя лингвистическими показателями на русском, английском и французском языках. Для анализа был создан корпус на основе 60 000 слов. Целью нашей работы было изучение *лексической плотности, TTR* и *многословных выражений* в русских, английских и французских текстах ООН. Для достижения данной цели мы использовали программы Rulingva, TextInspector и LancsBox для расчета данных. Результаты показали, что выбранные для исследования лингвистические признаки могут влиять на сложность за счет лексического богатства, являющегося многомерным понятием, которое охватывает несколько аспектов использования терминов [12, С. 19].

Ключевые слова: Французские тексты ООН, английские тексты ООН, русские тексты ООН, TTR, лексическая насыщенность, многословные выражения, n-граммы, читаемость.

Introduction

Legal language does not qualify as a language in the same way as French, Finnish, or Arabic, for example. According to Carles Duarte, the Catalan linguist, it operates as a functional variant of natural language, with its own domain of use and specific linguistic norms such as phraseology, vocabulary, and hierarchy of terms and meanings. Legal language processes several features. These are morphosyntactic, semantic, and pragmatic. This language is used in particular social roles: pleading, claiming, and so on. It is clear to see that legal language is based on ordinary language. For that reason, the grammar and, in general, the vocabulary of legal language are a language for special purposes. This means, first of all, that a large number of legal terms exist whose properties fluctuate according to the branches of law. Additionally, the legal languages of different countries and of different periods possess, to a varying degree, characteristics that distinguish them from ordinary written language. The main goal of this research is comparing Russian, English, and French UN texts, which are based on human rights and therefore considered legal texts. Three linguistic features (lexical density, TTR, and multiword expressions) are used to predict complexity.

Multiword expressions have received tremendous attention from scholars in a variety of disciplines, from theoretical to applied linguistics and psycholinguistics and from lexicography for human users to human language technology. Admittedly, linguists seek to account for their properties and to define typologies thereof; in applied linguistics, multiword expressions of various kinds pose issues for language learning and teaching; issues related to the acquisition and processing of multiword expressions as well as the way they are stored in the mental lexicon are the focal point of psycholinguistic research, whereas lexicographers are well aware of the importance of their presence in dictionaries [4, P. 454] and strive to define optimal representation formats tailored to meet the needs of humans and machines alike. Computational linguists, on the other hand, are concerned with multiword expressions processing, primarily with their identification and discovery in corpora, as well as with cross-lingual equivalence, even though multiword expressions might be of importance in other downstream tasks too.

Moreover, multiword expression identification and discovery are seen as the two facets of multiword expression processing [5, P. 837-892], and lexical resources of all sorts remain at the heart of both: the former could be made easier given a resource lexicon containing them, while the latter could contribute to the enhancement of such a resource [6, P. 250]. Consequently, [7, P. 79–91] suggested the deployment of multiword expression-related lexical resources as a possible solution for improving multiword expression processing; therefore, despite the ever-increasing effort to develop corpora of considerable size as well as language models of all kinds, multiword expression lexicons are still needed.

The term “lexical density” refers to the density of information of text according to how tightly content words have been packed into the grammatical structure [8, P. 86-105]. Content words nouns, verbs, adjectives, and adverbs Halliday [2], while function words are pronouns, determiners, finite verbs, and some classes of adverbs. Lexical density is usually measured by the ratio of the total lexical words to the total ranking clauses [2, P. 135], [1, P. 26–48]. By ranking clauses, Halliday [3, P. 256] means “those that are not embedded and hence have their full status as clauses in the discourse” [3, P. 195].

In legal writing, the lexical density may go much higher, and the language seems complex because it entails a large number of inter-relating technical jargon, each of which has been described and includes information the reader is expected to already comprehend [9, P. 288].

Research methods and principles

In this article, we built up a 60,000-word-based corpus extracted from Russian, English, and French legal texts (the documents of the United Nations). The texts are available on the site of the Office of the High Commissioner for Human Rights (UN Human Rights) — the leading UN entity on human rights. They represent the world’s commitment to the promotion and protection of the full range of human rights and freedoms set out in the Universal Declaration of Human Rights.

We selected 20 texts in Russian, 20 texts in English, and 20 texts in French. The texts were about 1,000 words in length. After collecting the data, *Rulingva, TextInspector, and LancsBox* were used for the analysis. Furthermore, we focus on *lexical density, TTR,* and *5-grams* (multi-word expressions) of the texts of the United Nations in three languages (Russian, English, and French). In our previous research, we shed more light on the lexical density and TTR of both Russian and English UN texts. The main goal of the present paper is to use LancsBox to compute multi-word expressions, lexical density, and TTR of French UN texts for the sake of comparison and better understanding of their complexity. LancsBox (LancsBox X) is a free desktop tool that can quickly process very large corpora (millions and billions of words) and can consist of simple texts or richly annotated XML documents. It produces concordances, summary tables, collocation graphs and tables, wordlists, and keyword lists.

Main results

Table 1 - Data analysis using Lancsbox, Rulingva and TextInspector

DOI: https://doi.org/10.60797/RULB.2025.67.5.1

PARAMETERS	TTR	LD	-	TTR	LD	-	TTR	LD
FT1	0,75	55,41	RT1	0,57	71,37	ET1	0,37	38,97
FT2	0,80	82,28	RT2	0,58	71, 2	ET2	0,40	44,13
FT3	0,79	73,57	RT3	0,59	70,2	ET3	0,40	42,52
FT4	0,80	84,68	RT4	0,6	69,29	ET4	0,36	38,9
FT5	0,78	65,64	RT5	0,53	72,4	ET5	0,40	43,57
FT6	0,79	71,79	RT6	0,54	71,11	ET6	0,36	41,29

PARAMETERS	TTR	LD	-	TTR	LD	-	TTR	LD
FT7	0,78	71,20	RT7	71,37	69	ET7	0,37	38,39
FT8	0,76	65,07	RT8	0,52	69,7	ET8	0,4	41,26
FT9	0,78	76,57	RT9	0,56	69,24	ET9	0,41	43,38
FT10	0,83	106,06	RT10	0,57	68,49	ET10	0,35	39,72
FT11	0,81	96,06	RT11	0,53	69,42	ET11	0,37	40,57
FT12	0,84	119,94	RT12	0,51	68,68	ET12	0,41	42,77
FT13	0,77	62,96	RT13	0,55	71,05	ET13	0,39	41,26
FT14	0,80	84,68	RT14	0,56	67,12	ET14	0,38	40,17
FT15	0,79	74,51	RT15	0,56	74,64	ET15	0,38	44,26
FT6	0,79	71,79	RT16	0,6	69,61	ET16	0,33	39,3
FT17	0,78	75,29	RT17	0,62	71,14	ET17	0,43	45,05
FT18	0,82	102,65	RT18	0,61	70,53	ET18	0,38	41,53
FT19	0,80	81,99	RT19	0,55	71,14	ET19	0,37	38,98
FT20	0,77	71,66	RT20	0,52	68,48	ET20	0,40	45,09

Note: this table represents the type-token ratio (TTR) and lexical density (LD) of French, Russian, and English UN texts. The categorization of the corpus ranged from text 1 to text 60, in particular, French texts were classified as FT1 to FT20; the same was done with Russian and English texts ranging from RT1 to RT20 and ET1 to ET20. The texts were computed with such tools as Lancsbox, Rulingva, and TextInspector. In addition to TTR and lexical density, we also computed multi-word expressions, more particularly 5-grams of the three languages, using LancsBox. We provided ample information about 5-grams of French, Russian, and English UN texts below

Discussion

The lexical density of Russian texts varied from 68% to 72%; the type-token ratio of Russian texts varied from 0.48 to 0.60. On the other hand, the lexical density of English texts varied from 36% to 45%, and the type-token ratio varied from 0.33 to 0.42. As far as French is concerned, its lexical density varied between 55% and 96%, and the TTR fluctuated between 0.75 and 0.84.

As mentioned above, the lexical density and lexical diversity (*TTR*) of Russian texts proved to be high (72 %), and 0.56 for *TTR* is an indication that Russian texts are more difficult to process than English texts, which proved to be low (varying between 36% and 45%) in lexical density. Traditionally, lexical diversity has been measured using the *TTR*. A text is dense if it contains many lexical words relative to the total number of words, that is, lexical and functional. A longer text usually gives a lower *TTR* value than a shorter text [7, P. 61–79], [8, P. 86–105] also point out that lexical density does not necessarily measure lexis, since it depends on the syntactic and cohesive properties of the composition. Relying on lexical diversity has been reported as inadequate to measure vocabulary development by several authors, who claim that *TTR* inevitably falls with the increasing size of the token sample and consequently is not indicative of lexical richness. Thus, any single value of *TTR* lacks reliability, as it will depend on the length in words of the language sample used [9, P. 288].

As a rule, texts with a lower density are more easily understood, and spoken texts have lower lexical density levels than written texts [11, P. 185–207], [5, P. 837–892]. However, as argued by [7, P. 79–91], a text may have high lexical diversity (contain many different word types) but low lexical density (contain many pronouns and auxiliaries rather than nouns and lexical verbs), or vice versa.

We also measured multi-word expressions of Russian, English, and French UN texts using LancsBox by focusing on 5-grams. We investigated the frequency of 5-grams in our corpus and construed that English UN texts have the most frequent multi-word expressions; English was followed by French and then Russian. Multi-word expressions work like words that are commonly used together, and can cause comprehension difficulties when readers are not familiar with them. A higher frequency of multi-word expressions in our corpus is an indication of the complexity of UN texts. Both English and French UN texts were found to contain more 5-grams than Russian UN texts, though this could be explained by the fact that English has a higher analytism than French and that Russian, being a synthetic language, has the lowest analytism of the three.

Conclusion

French texts revealed to be denser than both Russian (0.51 to 0.61) and English texts (0.33 to 0.42). The TTR of French texts varied from 0.75 to 0.84. The idea behind measuring lexical diversity (TTR) is that a varied vocabulary is a sign of high proficiency, whereas frequent repetitions of a limited range of words are typical for low-skilled readers. According to P. Nation [13, P. 1-16], repetition has, unsurprisingly, been shown to play a crucial role in vocabulary learning. The more often learners encounter a word, the higher the chance they will recall that word and integrate it into their repertoire of linguistic resources.

Russian and English texts having shown a low percentage of TTR means that the texts are more difficult to process. In relation to lexical density, French texts (55%–96%) were more highly dense than both Russian (68%–72%) and English (36%–

45%) texts. The idea behind the measure of lexical density is that a high percentage of lexical words indicates a high degree of lexical richness. According to J. Ure [14, P. 443–452], a higher percentage of lexical density reveals the complexity of a text.

Based on the previous works (Cobb & Horst, 2015 [11]; Constant et al., 2017 [5]; O'Loughlin, 1994 [1]; Savary, 2019 [7]; etc.) done in the field of complexity and the current results, we can say that French UN texts (55%–96%) are more complex than both Russian and English UN texts in terms of lexical density. The last parameter measured in our research is multi-word expressions, in particular 5-grams. English UN texts revealed to have more 5-n-grams than Russian and French. This is probably due to the fact that the analytism of English is higher than that of French, and the Russian analytism is the lowest of the three, as it is a synthetic language. Based on these results, we can draw a conclusion that the linguistic features like lexical density, TTR, and n-grams can impact complexity.

As a reminder, we have already investigated both lexical density and TTR of Russian and English UN texts; in this research, we made an attempt to measure the complexity of French UN texts to comprehend the similarities and dissimilarities lying between these three languages. The findings enabled us to utilize and test a tool called X LancsBox to analyze French UN texts and to be able to confirm the results provided by Rulingva and TextInspector, which we used to examine Russian and English UN texts. The results of this research could be exploited by all the specialists in legal fields or by the students whose field of interest is human rights.

Дополнительные материалы

Дополнительные материалы доступны на онлайн-странице статьи.

Конфликт интересов

Не указан.

Рецензия

Гурская О.А., Университет Сан Диего для Интегративных Исследований, Сан Диего Соединенные Штаты Америки
DOI: https://doi.org/10.60797/RULB.2025.67.5.2

Supplementary materials

Supplementary materials are available online on the article's webpage.

Conflict of Interest

None declared.

Review

Hurskaya V., San Diego University for Integrative Studies, San Diego USA
DOI: https://doi.org/10.60797/RULB.2025.67.5.2

Список литературы на английском языке / References in English

1. O'Loughlin K. Lexical density in candidate output on two versions of an oral proficiency test / K. O'Loughlin // Melbourne Papers in Language Teaching. — 1995. — Vol. 1, № 3. — P. 26–48.

2. Halliday M.A.K. Spoken and written language / M.A.K. Halliday. — Waurn Ponds, Vic : Deakin University, 1985. — 135 p.

3. Halliday M.A.K. The Language of Science / M.A.K. Halliday. — New York ; London : Continuum, 2004. — 256 p.

4. Evert S. The statistics of word cooccurrences: Word pairs and collocations: doctoral dissertation / S. Evert. — Saarbrücken : Universität des Saarlandes, 2004. — 454 p.

5. Constant M. Multiword expression processing: A survey / M. Constant, G. Eryiğit, J. Monti et al. // Computational Linguistics. — 2017. — Vol. 43, № 4. — P. 837–892.

6. Ramisch C. Multiword expressions in computational linguistics: Down the rabbit hole and through the looking glass / C. Ramisch. — Aix Marseille Université (AMU), 2023. — 250 p.

7. Savary A. Without lexicons, multiword expression identification will never fly: A position statement / A. Savary, S. Cordeiro, C. Ramisch // Proceedings of the joint workshop on Multiword Expressions and WordNet (MWE-WN 2019). — Florence, 2019. — P. 79–91.

8. Halliday M.A.K. The construction of knowledge and value in the grammar of scientific discourse: Charles Darwin's The Origin of the Species / M.A.K. Halliday // Writing science: Literacy and discourse power / Ed. by M.A.K. Halliday & J.K. Martin. — Washington ; London : Falmer, 1993. — P. 86–105.

9. Halliday M.A.K. Writing science. Literacy and discourse power / M.A.K. Halliday, J.R. Martin (Eds.). — London : Flamer Press, 1993. — 288 p.

10. Johansson V. Lexical diversity and lexical density in speech and writing: A developmental perspective / V. Johansson // Lund Working Papers in Linguistics. — 2009. — Vol. 53. — P. 61–79.

11. Cobb T. Learner corpora and lexis / T. Cobb, M. Horst // The Cambridge handbook of learner corpus research / Ed. by S. Granger, G. Gilquin, F. Meunier. — Cambridge : Cambridge University Press, 2015. — P. 185–207.

12. Lisson P. Investigating lexical progression through lexical diversity metrics in a corpus of French L3 / P. Lisson, N. Ballier // Discours. — 2018. — № 23. — P. 19.

13. Nation P. How much input do you need to learn the most frequent 9,000 words? / P. Nation // Reading in a Foreign Language. — 2014. — Vol. 26, № 2. — P. 1–16.

14. Ure J. Lexical density and register differentiation / J. Ure // 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.