

Fig.4. The dynamics of average width in 30 years

The pattern replicates the one that was observed in the case of sentence length. As sentence length decreases, the degree of parallel subordination is diminished too.

B. Height (H) of the tree

Table IV presents descriptive statistics for the height of the tree (i. e., the maximum number of sequentially subordinate nodes). This parameter represents the degree of coherent subordination in sentences. The maximum value is relatively small due to the restriction of specifics of annotation done by UDPipe.

TABLE IV. DESCRIPTIVE STATISTICS FOR HEIGHT

Statistics	Value
Minimum	1
Maximum	5.6
Median	3.4
Mean	3.46
SD	0.83
Skewness	0.06
Kurtosis	-0.08
Range	4.6

The minimum average height is also illustrated by Arnold Kolbanovsky’s story (1921) from which the examples of the sentences were previously reviewed.

The maximum average value was found in the story by Yefim Zozulya (1918).

Among the representatives of typical writers in this aspect are Ivan Bunin (1911), Nikolai Garin-Mikhailovsky (1901), Boris Lazarevsky (1912), Valentin Svetsitsky (1906), and Mikhail Basov (1922).

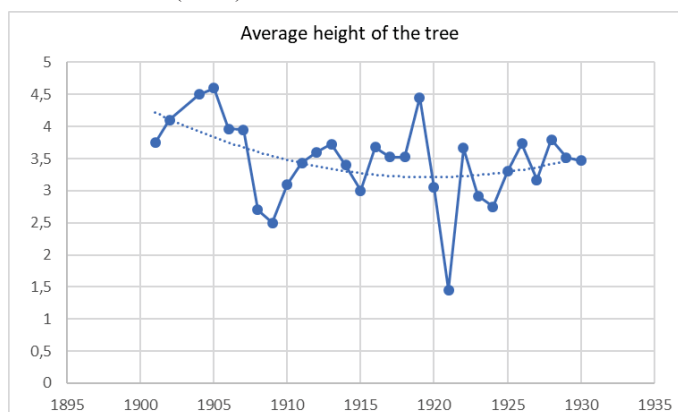


Fig.5. The dynamics of average height in 30 years

The trend of mean height is presented on Fig. 5. The pattern is similar to the sentence length and the width of the tree in a root node. The shorter sentence length means the lesser degree of subordination of both types.

C. Number of left and right subordinates

Table V presents descriptive statistics for the number of left and right subordinates in a root node.

The number of the left and right subordinates is restricted by the width of the tree in a root node. Based on the data obtained, it can be concluded that the mean values of both types of dependents are relatively equal. These data allow us to consider the distribution of left (in the preposition) and right (in the postposition) subordinates in a root node. However, at this stage it would be premature to make any decision in respect of tendencies of prepositive, postpositive or symmetric constructions in the core of a sentence.

TABLE V. DESCRIPTIVE STATISTICS FOR THE NUMBER OF LEFT AND RIGHT SUBORDINATES

Statistics	Left-Value	Right-Value
Minimum	0.5	0.3
Maximum	3.1	2.6
Median	1.5	1.6
Mean	1.54	1.63
SD	0.45	0.43
Skewness	0.68	0.22
Kurtosis	1.78	0.15
Range	2.6	2.3

However, specific stories can be referred to as representatives of one type or another. For example, the minimum average value of left dependents was found in the story by Alexandra Kollontai (1923), on the basis of which we can say that there is such a style pattern when the head of a sentence, often a verb, takes the initial place in a linear order:

Zastegnula koftochku i poshla k vyhodu.

The story by Gaidar Arkady (1927) represents the maximum average value of left subordinates, in the following sentence it equals to 5:

Drugogo by na ego meste davno ordenom nagradili, a Levku net.

The minimal average of right subordinates is illustrated by Arnold Kolbanovsky’s story (1921) that was reviewed previously.

The maximum average value of 2.6 was found in texts by Nikolai Shklyar, Lydia Zinovieva-Annibal, and Petr Pil’sky. For example, the following sentence by Lydia Zinovieva-Annibal has the right value of 4:

Doktor posylal ee na jug, na solnyshke katat'sja v svoem kresle, no ona zahotela ostat'sja podol'she v derevne, gde ohotno provela by i vsju zimu v bol'shom, teplom, starom dome.

Representatives of a typical group in respect of the median for left subordinates are texts by Alexey Novikov-Priboy (1917), Konstantin Balmont (1908), Yevlampy Minin (1925), Evgeny Petrov (1927), and Vasily Bashkin (1910), whereas for right subordinates we can mention stories by Alexander

Yakovlev (1922), Victor Goncharov (1927), Nikolai Garin-Mikhailovsky (1901), Yevlampy Minin (1925), and Maxim Gorky (1904).

The dotted line trend shown on Fig. 6 makes it possible to trace the trend of mean average number of left and right subordinates in a root node over time. According to the graph, the average number of left subordinates is relatively stable and correlates with the typical position of the predicate in a sentence. It is interesting that the number of right subordinates iterates the trend similar to the one we observed in distribution of sentence length over time.

D. Symmetry I

Table VI presents tree symmetry index — the ratio of the left subordinate members to the right ones (Symmetry I) [Martynenko, Sherstinova 2019b]. This parameter allows us to characterize core structure of the sentences as postpositive, prepositive or symmetric.

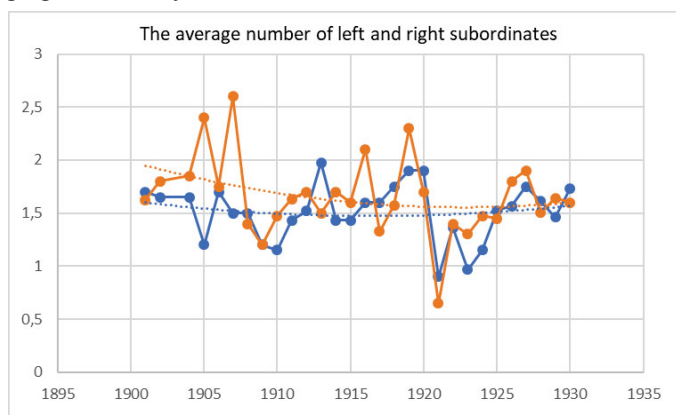


Fig.6. The comparative dynamics of the average number of left and right subordinates in 30 years (the blue line refers to left subordinates, and the orange one – to the right ones).

TABLE VI. DESCRIPTIVE STATISTICS FOR SYMMETRY I

Statistics	Value
Minimum	0.24
Maximum	2.44
Median	0.92
Mean	0.97
SD	0.37
Skewness	0.71
Kurtosis	1.24
Range	2.20

The minimum average of 0.24 was found in the story by Alexandra Kollontai (1923). Symmetry I lesser than 1 is the indication of postpositive constructions, and the lesser its value, the more postpositive is the sentence core structure.

The maximum average symmetry index was observed in Nikolay Tikhonov’s story “Miracle” (1918). The greater value means that the core structure of a sentence is prepositive:

Tochno skazat', chto on podrazumeval pod etim slovom, on ne mog, tak kak sushhnost' etogo chuda, po ego mneniju, nel'zja bylo peredat' slovami.

Among the representatives of the typical texts, we can mention short stories by Leonid Dobychin (1924), Varvara Karacharova (1915), Ivan Bunin (1911), Alexander

Serafimovich (1902), and Peter Pilsky (1903), whose symmetry values are equal (or close) to the median. The constructions in the core of the sentence that these authors prefer to use are close to the symmetrical one.

The dotted line trend of average ratio of the left subordinate members to the right ones over time is shown on Fig. 7.

According to this graph, the average of symmetry index tends to be stable. Based on these data, we can assume that in Russian fiction left subordinates and the right subordinates are balanced. The average values which are close to 1 give us reason to suppose that typically core structures are close to be symmetric.

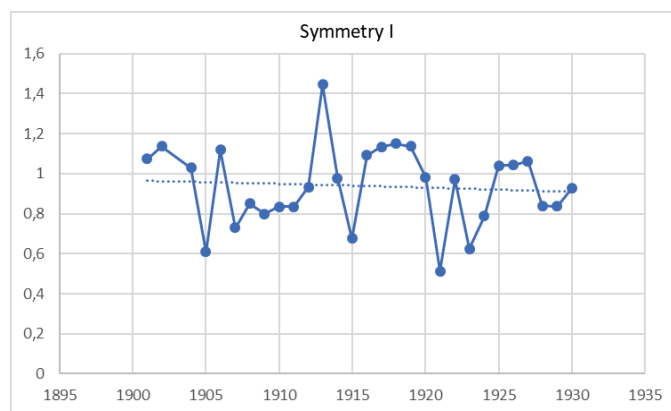


Fig.7. The comparative dynamics of Symmetry I in 30 years.

E. Number of words in left and right positions

Table VII presents descriptive statistics for the number of words in left and right in left and right positions. This parameter represents how many words are used in a sentence relatively to its root.

TABLE VII. DESCRIPTIVE STATISTICS FOR THE NUMBER OF LEFT AND RIGHT WORDS

Statistics	Left-Value	Right-Value
Minimum	0.9	1.3
Maximum	7.3	17.9
Median	3.2	8.4
Mean	3.19	8.81
SD	1.28	3.99
Skewness	0.78	0.45
Kurtosis	0.94	-0.57
Range	6.4	16.6

The minimal average value is also observed in the story by Alexandra Kollontai (1923).

The maximum average is presented in Sergey Auslander’s short story (1912).

Representatives of the typical group by median value are texts by Lydia Avilova (1906), Jerome Yasinsky (1913), Vladimir Korolenko (1901), Lev Urvantsov (1918), and Vladimir Unkovsky (1914).

The minimum value of the number of words in right position was found in Arnold Kolbanovsky’s story (1921), which was already considered earlier.

The maximum average value of the number of words in right position is presented in Vincent Veresaev’s story (1919).

Representatives of a typical group according to the median value of average number of words in the right position are Nikolai Nikitin (1923) Mikhail Chernokov (1916), Maximilian Kravkov (1925), Victor Hoffman (1911), and Boris Nikonov (1906).

The trends of mean average numbers of left and right word numbers over time are presented as the dotted line trends on Fig. 8.

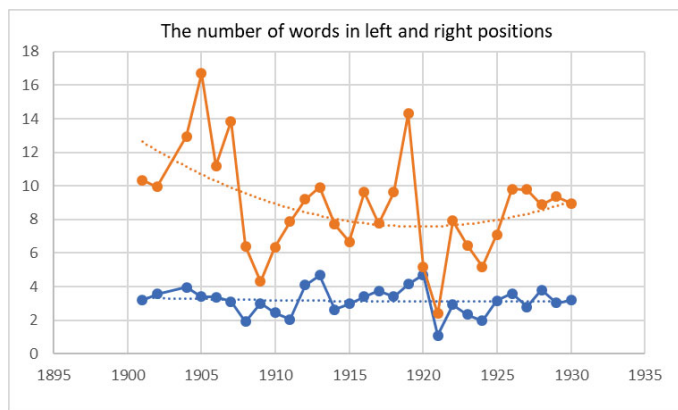


Fig.8. The comparative dynamics of the number of words in left and right positions according to the root in 30 years (the blue line refers to the left position, and the orange one – in the right one).

According to the graph, the average number of words in preposition is rather stable. The average number of words in postposition resembles the trend we observed for the sentence length. Thus, syntactic elements in postposition turned out to be sensitive to the changes of the sentence length.

F. Symmetry II

Table VIII presents descriptive statistics for the ratio of the left subordinate members to the right ones relatively to the root node measured in word numbers (Symmetry II). This parameter characterizes any sentence as prepositive, postpositive or symmetric, depending on whether more words are located to the left of the root, to the right of the root, or equally.

TABLE VIII. DESCRIPTIVE STATISTICS FOR SYMMETRY II

Statistics	Value
Minimum	0.06
Maximum	3.24
Median	0.57
Mean	0.70
SD	0.51
Skewness	2.37
Kurtosis	7.77
Range	3.18

To illustrate the minimum value, we can consider an example from the story by Vladimir Zazubrin (1923). In this example it equals to 0:

Stoit na uglu Oktjabr'skoj i Kommunisticheskoy ulic.

The sentence is postpositive, when this value is less than 1.

The maximum average value belongs to the story by Andrei Platonov (1926). In following case, we can say that the author's style is prone to prepositive sentences:

Pojetomu, kogda v odno utro, daleko ne prekrasnoe (tumanom i sljakot'ju ljubit pugnut' surovij Peterburg naivnogo provinciala), v perednej Kirilla Platonovicha okazalas' zheltaja, perevjazannaja tolstennymi bechevkami, korzina, a v stolovoj, priderzhivajas' bol'she temnyh uglov, skonfuzhenno prohazhivalsja nikomu nevedomyj molodoj chelovek, vo vsem dome srazu pochuvstvovalos' kakoe-to razdrashenie.

Representatives of the typical class in respect with Symmetry II include texts by Nikolai Nikitin (1923), Vladimir Unkovsky (1914), Boris Chetverikov (1929), Nikolai Shklyar (1916), Vasily Ryakhovsky (1924), and Sergey Garin (1917).

The dotted line trend of average Symmetry II is shown on Fig. 9.

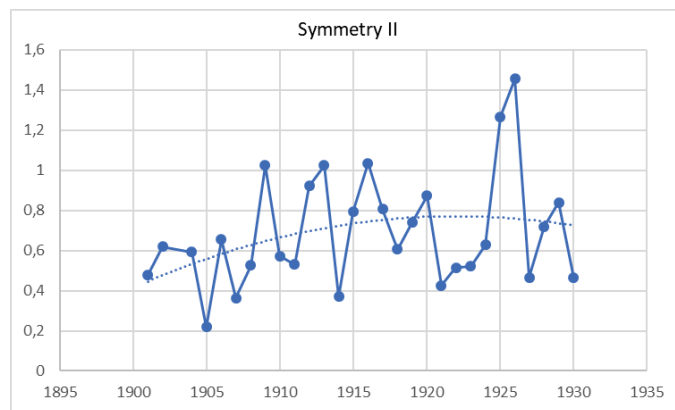


Fig.9. The comparative dynamics of Symmetry II in 30 years.

According to the graph, the average value tends to increase from the beginning of the century, since the average sentence length and the number of words in right position decrease. We can say that sentences become more symmetric: left- and right-branches of a tree are more balanced. Further, with increase of sentence length, the ratio begins to decrease by the end of the period.

III. RESULTS: EXTENSIVE SYNTACTIC PARAMETERS

In previous sections we considered average values of each parameters and how they differ among authors. Further, it seems worth to compare the mean values between the time periods provided by the corpus. Thus, the corpus contains the following subcorpora, referring to the main historical periods of the era in question [Sherstinova, Martynenko 2020].

- **Period I.** Short stories of the beginning of the 20th century (1900–1913).
- **Period II.** Short stories of the era of war and the acute social upheaval (1914–1922) – World War I, the February and October Revolutions, and the subsequent Civil War.
- **Period III.** Short stories of the post-revolutionary era (1923–1930).

Table IX presents the mean values for each parameter in concern for each of the three periods.

It can be pointed out that the values of the most parameters, with the sole exception of Symmetry II, tend to decrease with different significance. Thus, we can assume, that during the period under review syntactic structures are simplified over time.

TABLE IX. DISTRIBUTION OF SYNTACTIC PARAMETERS BY PERIODS

Syntactic parameter	Period I	Period II	Period III
Mean sentence length (MS)	10.95	8.99	8.93
Mean paragraph length (MP)	2.77	2.3	2.42
The width of the tree in a root node (W)	3.28	3.17	3.03
The ratio of the left subordinate members to the right ones relatively to the root node measured in word numbers (SII)	0.67	0.65	0.75
The ratio of the left subordinate members to the right ones (SI)	1.01	1.01	0.89
The height of the tree (H)	3.69	3.46	3.29
The number of left subordinate members in a root (NLRR1)	1.59	1.59	1.45
The number of right subordinate members in a root node (NLRR2)	1,71	1,6	1,58
Number of words in left position (WM1)	3,36	3,29	2,91
The number of words in right position (WM2)	9,83	8,45	8,15

VII. CONCLUSION

Modern computing and information technologies open fundamentally new opportunities for solving theoretical and practical problems, which in the recent past seemed absolutely utopian, but today they have every chance to be realized. In particular, it became possible to form large corpora of texts of different genres and process these texts with the most modern technologies. This allows us to solve the problems of stylistic diagnostics at a new level, using strict statistical analysis and to solve the problems of attribution, taxonomy, typology, periodization and other types of ordering and systematization of textual data.

In this paper, ten syntactic parameters have been considered whose numerical indicators can be used to carry out statistical analysis of a large texts volume and which can reflect text syntactic structure and its complexity. The data obtained have the potential to classify authors' style according to stylistic features of their texts or to identify the most frequent syntactic structures in texts. It is worth noting that this study should be considered as an exploratory one, and its results should be treated as preliminary. However, the obtained results show the effectiveness of different measures for estimating syntactic complexity of texts and revealing their correlation, which may be used for NLP software development.

ACKNOWLEDGMENT

The research is supported by the Russian Foundation for Basic Research, project # 17-29-09173 "The Russian language on the edge of radical historical changes: the study of language and style in prerevolutionary, revolutionary and post-revolutionary artistic prose by the methods of mathematical and computer linguistics (a corpus-based research on Russian short stories)".

REFERENCES

- [1] V.G. Admoni, "Razmer predlozheniya i slovosochetaniya kak yavlenie sintaksicheskogo stroya" ["The Length of Sentences and Phrases as a Phenomenon of Syntactic Structure"], *Voprosy yazykoznanija*, 1966 (4), pp. 111–118.
- [2] G.N. Akimova, "Razmer predlozheniya kak faktor stilistiki i grammatiki" ["Sentence Length as a Factor of Stylistics and Grammar"], *Voprosy yazykoznanija*, 1973 (2), pp. 67–79.
- [3] A.N. Baranov, *Lingvisticheskaya ekspertiza tekstov [Linguistic examination of texts]*, Moscow: Flinta, Nauka, 2013.
- [4] S. Bhat, S.-Y. Yoon, "Automatic assessment of syntactic complexity for spontaneous speech scoring", *Speech Communication*, 67, 2015, pp. 42–57.
- [5] N. Bogdanova-Beglarian, G. Martynenko, T. Sherstinova, "The 'One Day of Speech' corpus: Phonetic and syntactic studies of everyday spoken Russian", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9319, 2015, pp. 429–437.
- [6] C. Buongiovanni, F. Gracci, D. Brunato, F. Dell'Orletta, "Lost in Text. A Cross-Genre Analysis of Linguistic Phenomena within Text", *CLiC-it*, 2019, Web: <http://ceur-ws.org/Vol-2481/paper10.pdf>.
- [7] H. Choi, "Ability of sentence comprehension according to syntactic complexity and speech rate in patients with Dementia of Alzheimer's Type", *Communication Sciences and Disorders*, 24(3), 2019, pp. 673–682.
- [8] [CoNLL 2018 Shared Task] Web: <https://universaldependencies.org/conll18/results.html> (last accessed on 02.05.2020).
- [9] H. Delage, U.H. Frauenfelder, "Syntax and working memory in typically developing children: Focus on syntactic complexity", *LIA Language, Interaction and Acquisition*, 10(2), 2019, pp. 141–176.
- [10] [DotEditor] Web: <http://vincenthee.github.io/DotEditor/> (last accessed on 02.05.2020).
- [11] [ETAP-4] Web: <http://cl.iitp.ru/ru/etap4download> (last accessed on 31.08.2019).
- [12] R. Evans, C. Orasan, "Identifying signs of syntactic complexity for rule-based sentence simplification", *Natural Language Engineering*, 25(1), 2019, pp. 69–119.
- [13] F.M.G. França, A.F. de Souza (eds.), *Intelligent Text Categorization and Clustering*, Springer-Verlag Berlin Heidelberg, 2009.
- [14] M. Hosseinia, A. Mukherjee, "A Parallel Hierarchical Attention Network for Style Change Detection: Notebook for PAN at CLEF", 2018, Web: http://ceur-ws.org/Vol-2125/paper_91.pdf.
- [15] R. Huxtable, "Sentence Length", *Science*, 197 (4300), 1977, pp. 208–208.
- [16] F. Kuiken, I. Vedder, A. Housen, B. De Clercq, "Variation in syntactic complexity: Introduction", *International Journal of Applied Linguistics*, 29(2), 2019, pp. 161–170.
- [17] T. Larsson, H. Kaatari, "Syntactic complexity across registers: Investigating (in)formality in second-language writing", *Journal of English for Academic Purposes*, 45, 2020, 100850.
- [18] G.A. Lesskis, "Nekotorye statisticheskie zakonomernosti kharakteristiki prostogo i slozhnogo predlozheniya v russkoj nauchnoj i khudozhestvennoj proze XVIII–XX vv." ["Some Statistical Laws of the Characteristics of Simple and Compound Sentences in Russian Scientific and Fiction Texts of 18–20th centuries"], *Russkij yazyk v nacional'noj shkole [Russian language in the national school]*, 1968 (2), pp. 67–80.
- [19] O.N. Lyashevskaya, T.O. Shavrina, I.V. Trofimov, N.A. Vlasova, "GramEval 2020 shared task: Russian full morphology and universal dependencies parsing", *Komp'yuternaja Lingvistika i Intellektual'nye Tehnologii*, Issue 19(26), 2020, pp. 553–569.
- [20] M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre and C. D. Manning. "Universal Stanford dependencies: A cross-linguistic typology." *LREC* (2014). Web: https://nlp.stanford.edu/pubs/USD_LREC14_paper_camera_ready.pdf.
- [21] G. Martynenko, T. Sherstinova (2020), "Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century", *R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019)*, Saint Petersburg, Russia, November 27, 2019, CEUR Workshop Proceedings. Vol. 2552, 2020, pp. 105–120.

- [22] G. Martynenko, T. Sherstinova, "Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture", Alexandrov, D., Boukhanovsky, A., Chugunov, A., Kabanov, Y., Koltsova, O. (eds.), *Digital Transformation and Global Society. DTGS 2018. Communications in Computer and Information Science*, Vol. 859, 2018, pp. 299–309.
- [23] G.Ya. Martynenko, T.Yu. Sherstinova, T.I. Popova, A.G. Melnik, E.V. Zamirajlova, "O printsipakh sozdaniya korpusa russkogo rasskaza pervoy trety XX veka" ["On the principles of creation of the Russian short stories corpus of the first third of the 20th century"]. In: Proceedings of the XV International Conference on Computer and Cognitive Linguistics 'TEL 2018', Kazan, 2018, pp. 180–197.
- [24] G. Martynenko, T. Sherstinova (2019a) "Analytical Distribution Model for Syntactic Variables Average Values in Russian literary Texts", *Proc. of the Int. Conf. DTGS-2019. Digital Transformation and Global Society. 4th International Conference, DTGS 2019, St. Petersburg, Russia, June 19–21, 2019, Revised Selected Papers. Communications in Computer and Information Science*, Vol. 1038, Springer International Publishing, pp. 719–731. <https://www.springer.com/gp/book/9783030378578>
- [25] G. Martynenko, T. Sherstinova (2019b), "Simmetrika sintaksicheskikh figur v khudozhestvennoy proze: na materiale russkogo rasskaza XX veka", *IMS-2019*, St. Petersburg, 2019, Web: <http://ojs.itmo.ru/index.php/CLCO/article/view/1040/883>.
- [26] G. Martynenko, *Osnovy stilemetrii [The Foundation of Stylometrics]*. St. Petersburg State University, 1988, St. Petersburg.
- [27] G. Martynenko, *Metody matematicheskoy lingvistiki v stilisticheskikh issledovaniyakh [Methods of mathematical linguistics in stylistic studies]*, 2019, St. Petersburg: Nestor-Istoriya.
- [28] M.A. Marusenko et al. *In search of the lost author*, St. Petersburg State University Publishing House, 2001.
- [29] M.A. Nippold, M.W. Frantz-Kaspar, L.M. Vigeland, "Spoken language production in young adults: Examining syntactic complexity", *Journal of Speech, Language, and Hearing Research*, 60(5), 2017, pp. 1339-1347.
- [30] D. Olmsted, "On Some Axioms about Sentence Length", *Language* 43 (1), 303–305 (1967).
- [31] L. Ortega, "Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing", *Applied Linguistics*, 24(4), Dec. 2003, pp. 492-518.
- [32] E. Pimonova, O. Durandin, A. Malafeev, "Doc2vec Or Better Interpretability? A Method Study for Authorship Attribution", *Komp'yuternaya Lingvistika i Intellektual'nye Tehnologii*, Issue 19, 2020 International Conference on Computational Linguistics and Intellectual Technologies, Dialogue 2020; Moscow; Russian Federation, pp. 606–614.
- [33] K. Rudnicka, "Variation of sentence length across time and genre. Influence on syntactic usage in English", *Studies in Corpus Linguistics*, 2018, 85, pp. 219-240.
- [34] G. Scontras, W. Badecker, L. Shank, E. Lim, E. Fedorenko, "Syntactic complexity effects in sentence production", *Cognitive Science*, 39(3), 2015, pp. 559-583.
- [35] J.E. Sung, S. Choi, B. Eom, J.K. Yoo, J.H. Jeong, "Syntactic complexity as a linguistic marker to differentiate mild cognitive impairment from normal aging", *Journal of Speech, Language, and Hearing Research*, 63(5), 2020, pp. 1416-1429.
- [36] R. Saha Roy, S. Agarwal, N. Ganguly, M. Choudhury, "Syntactic complexity of Web search queries through the lenses of language models, networks and users", *Information Processing and Management*, 52(5), 2016, pp. 923-948.
- [37] [Stanford NLP] Web: <https://github.com/stanfordnlp> (last accessed on 03.04.2020).
- [38] [Universal Dependencies] Web: <https://universaldependencies.org/> (last accessed on 10.04.2020).
- [39] [UD Guidelines] Web: <https://universaldependencies.org/guidelines.html> (last accessed on 02.05.2020).
- [40] [Universal Dependency Relations] Web: <https://universaldependencies.org/u/dep/index.html> (last accessed on 10.04.2020).
- [41] [UDPipe] Web: <http://ufal.mff.cuni.cz/udpipe> (last accessed on 02.05.2020).
- [42] Y.-H. Yang, W.D. Marslen-Wilson, M. Bozic, "Syntactic complexity and frequency in the neurocognitive language system", *Journal of Cognitive Neuroscience*, 29(9), 2017, pp. 1605-1620.
- [43] G. Yule, "On Sentence-Length as a Statistical Characteristic of Style in Prose: With Application to Two Cases of Disputed Authorship", *Biometrika* 30 (3/4), 1939, pp. 363–390.
- [44] P. Zhou, "A study on the subjectival position and the syntactic complexity in Spoken English", *Glottometrics*, 47, 2019, pp. 83-103.
- [45] H. Zhang and H. Liu, "Interrelations among Dependency Tree Widths, Heights and Sentence Lengths", *Quantitative Analysis of Dependency Structures*, Series: Quantitative Linguistics [QL], 72, De Gruyter Mouton, 2018, pp. 31-52.