ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2019. Вип. 70. С. 299–308
Visnyk of the Lviv University. Series Philology. 2019. Is. 70. P. 299–308

УДК 811.161.2'1'324'38(038) І. Франко
doi: http://dx.doi.org/10.30970/vpl.2019.70.9785

DISTINGUISHING QUANTITATIVE PARAMETERS OF AUTHOR'S LANGUAGE AND STYLE (A CASE OF IVAN FRANKO LONG PROSE FICTION)

Solomiya BUK

The Ivan Franko National University of Lviv,
Department for General Linguistics,
1, Universytetska Str., room 343, Lviv, Ukraine, 79000,
tel.: (032) 239 47 56, e-mail: solomija@gmail.com

The article is dedicated to precise analysis of distinguishing quantitative parameters of author's language and style. Such an analysis is made for Ivan Franko long prose fiction for the first time.

The frequency dictionary of all nine Ukrainian novels by Ivan Franko was compiled on the material of an electronic text corpus with an external and internal markup. It can be considered as a statistical combinatory model of Franko's style as well as a lingual statistical portrait of his long prose fiction. The following parameters were obtained: vocabulary sizes, variety, exclusiveness, concentration indexes, the amount of hapax legomena, their occupation of text and vocabulary, amount of words in text with frequency 10 and higher, their occupation of text and vocabulary. They were compared with those of text corpus of Ukrainian general long prose fiction.

Keywords: lingual statistics, text corpus, statistical parameters, quantitative characteristics of text, frequency dictionary, token, lemma, vocabulary richness.

Research problem and the purpose of the article. Ivan Franko (1856–1916) is the famous public figure, poet, writer, translator, ethnographer, philosopher, economist of Western Ukraine when it was a part of the Austro-Hungarian Empire. He had the big influence on the public opinion and formation of national consciousness in Ukrainians in the 19–20 centuries. That is why the modern study of his heritage is a crucial question for the research of language, culture, mentality, self-identification and so on in Ukraine.

We analyzed the works about different writer's language and style and had noticed main tendencies in the world linguistics in the research of the author's language and text [2]: tendency to lexicographical paramertization, special attention to the quantitative description of the texts and the writer's style. One of them is also using electronic text corpus nowadays to compile any kind of dictionary. Writer's text corpora have a huge potential in the study of writer's language both in qualitative and quantitative aspects, moreover some facts of writer's style can be revealed only with the help of a text corpus, such as vocabulary richness [30], indexes of variety and exclusiveness, etc.

The aim of the current research is to specify the distinguishing quantitative features

of author's language and text on the example of Ivan Franko long prose fiction. To reach this purpose, the following statistical parameters of the Ivan Franko long prose fiction will be obtained: the amount of hapax legomena, their occupation of text and vocabulary, exclusiveness index for text and vocabulary, amount of words in text with frequency 10 and higher, their occupation of text and vocabulary, concentration indexes for text and vocabulary.

Relevance of research. Specialists in different branches of linguistics underline the great potential of text corpus for the research of language and text in general and concrete writer's language and style in particular. That is why along with the text corpora of national languages (Croatian, Czech, Polish, Russian, Slovak, Ukrainian and so on) the text corpora of specific writers appear in the world linguistics (Aristotle, Shakespeare, Joyce, etc.), covering Slavic (O. Březina, K. Čapek, F. Dostojevskij, M. Pavić, etc.), and including Ukrainian (H. Skovoroda, T. Shevchenko, V. Shevchuk).

It is important the Corpus of Ivan Franko's long prose fiction (CFP) is currently under development at the Ivan Franko National University of Lviv [4]. Current research is the first attempt in linguistics to make such a database of his works as well as the complex analysis of statistical parameters of Franko's long prose fiction.

CFP is the first stage of a larger project of the I. Franko text corpus. It consists of nine novel: *Boryslav smijetsja [Boryslav Laughs]* (1880–1881), *Zakhar Berkut* (1883), *Ne spytavšy brodu [Without Asking a Wade]* (1885–1886), *Dlja domašnjoho ohnyšča [For the Hearth]* (1892), *Osnovy suspil'nosty [Pillars of Society]* (1894–1895), *Perekhresni stežky [The Cross-Paths]* (1900), *Boa Constrictor* (1905–1907), *Velykyj šum [The Great Noise]* (1907), *Petriji i Dovbushchuky* (1909–1912). Two of the novels exist in two variants: *Petriji i Dovbushchuky* 1875–1976 and 1909–1912 and *Boa Constrictor* 1884 and 1905–1907. We included the last editions of both works guided by the following principles: 1) these texts are the last will of the author, he changed and reworked the first additions according to his last level of language competency; 2) Ukrainian language of the first edition of "Petriji I Dovbushchuky" I. Franko qualified himself as not perfect, deficient, which contains a lot of old fashioned words not used any more for that time; 3) the inclusion of both editions of these texts would be useful for the research of Franko's style changes however not for the statistical analysis: part of the words are the same in both variants, so it could have influence on such the parameters as: index of richness [1] 4) the text corpus with the last editions of these works is 506 708 tokens, so it perfectly correlate with the text corpus of Ukrainian long prose fiction [16], the difference 6 708 items is only 1,3 % so it means they are correct to compare with each other.

Analysis of recent research and publications. Quantitative parameters of text were studied by many linguists in different countries, such as G. Altmann, R. Köhler [23] (Germany), I.-I. Popescu [25] (Romania), G. Wimmer [31] (Slovakia), Yu. Tuldava [15] (Estonia), A. Pawłowski [24], M. Ruszkowski [26], J. Sambor [27] (Poland), F. Čermák [28], V. Cvrček [20] (Czech Republic), S. Vasić [29] (Serbia), P. Alekseev [18], A. Shaykevich [17] (Russia), V. Perebyjnis [11, 14], N. Darchuk, M. Muravytska [12], Shyrokov [8] etc.

Some statistics were applied to Franko's heritage as well. For example, for his poetry word length was under the analysis in [19]; the list of 35 000 words with their frequencies were compiled in 1990 at the university of Lviv. Some statistical analysis on this base was

made by I. Kovalyk, I. Oshchypko [9], and L. Poljuha [13]. O. Drul compares the number of different words in I. Franko's and B. Hrinchenko's poetry [7].

For his short prose Ye. Nenadkevych compared the number of letters in I. Franko's short story "Hlopska komisija" [Peasant commission] and V. Sfephanyk's one "Zlodij" [Thief] [10]. Yu. Holovach and V. Palchykov had analyzed the frequency-rank distributions of the words of I. Franko's fairy tale "Lys Mykyta", and also applied the Zipf law and the theory of complex networks for this work [21, 22].

As we can see from this short overview the application of statistical and quantitative methods towards I. Franko's works was not systematic neither complex, but rather sporadic.

Presenting of the main material. Many researchers agreed that frequency dictionary is able to present a comprehensive quantitative description of text. With the aim to obtain the complex statistical dimensions of Ivan Franko long prose fiction, nine frequency dictionaries of every mentioned Franko's novel were compiled and every statistical parameter was calculated for every concrete work (see, for example, [3, 5, 6]). Then the combined frequency dictionary of all 9 Ukrainian novels together was compiled (506 708 tokens). It can be considered as a statistical combinatory model of Franko style as well as a lingual statistical portrait of his long prose fiction. The homonyms were resolved in all the texts.

The main basic parameters of Franko long prose fiction are text volume (N = 506 708 tokens) as well as lexeme vocabulary volume (V = 27 192). It means the total number of lemmas in the vocabulary (i. e., the list of different lemmas). It is possible to obtain from such resources the statistical parameters of individual style: variety, exclusiveness, concentration indexes and so on. The common principles and methodology of frequency dictionary compilation make it possible to compare these statistical parameters of different texts.

The index of variety V/N = 0.054, is means repetition of a word in the text is thus N/V = 18.6, i. e. on the average every word in the text is used 19 times, which is a bit higher than in long prose fiction in general with the value 15.

The amount of hapax legomena (i.e. the words occurring only once in the text) N_1 = 11 567. The indicator for the vocabulary variability, i. e. exclusiveness index for text (N_1/N = 0.023) and for dictionary (N_1/V = 0.42) are calculated from these data. Hapax legomena occupy 2.3 % of the text and 42.5 % of the vocabulary.

These parameters correspond to the concentration indexes for text and vocabulary. The amount of words in text with frequency 10 and higher (N_{10}) is 453 312 (they cover 89,5 % of text) and in the vocabulary (V_{10}) is 4 681 (they cover only 17,2 % of vocabulary). These characteristics allow the calculation of concentration indexes for text (N_{10}/N = 0.89) and vocabulary (V_{10}/V = 0.17).

From the introduction to the Frequency dictionary of Modern Ukrainian Prose compiled of the works of 25 books of 22 writers (1947–1969) we received the similar information for this literary genre: text volume N = 500 000 tokens, vocabulary volume V = 33 391 lexemes; index of variety V/N = 0.067; repetition of a word in the text N/V = 15.0; The number of hapax legomena N_1 = 14 522; exclusiveness index for text N_1/N = 0.029, which means that the least used words covers 2.9% of text; exclusiveness index for vocabulary N_1/V = 0.43, and most rarely used lemmas covers 43.5% of vocabulary; concentration indexes for text

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2019. Випуск 70

$N_{10}/N = 0.821$, which means the most frequently used words covers 82.1% of text; the most frequently used lemmas with frequency 10 and higher V_{10} = 5 003, they cover 15 % of vocabulary; concentration indexes for vocabulary $V_{10}/V = 0.150$.

The comparison of the achieved data of two text corpora is demonstrated in Table.

Table

The comparison of the quantitative parameters of Ivan Franko long prose fiction with the Ukrainian long prose fiction in general

Quantitative parameter	Text corpus of I. Franko long prose fiction	Text corpus of Ukrainian long prose fiction
N, text volume, tokens	506 708	500 000
V, vocabulary volume, lexeme	27 192	33 391
V/N, index of variety	0.054	0.067
N/V, averige repetition of a word in the text	19	15
N_1, hapax legomena	11 567	14 522
N_1/N, exclusiveness index for text	0.023	0.029
	or 2.3 % of the text	or 2.9 % of text
N_1/V, exclusiveness index for vocabulary	0.42	0.43
	or 42.5 % of the vocabulary	or 43.5 % of vocabulary
N_{10}, frequently used tokens with frequency 10 and higher	453 312	(no exact number is quoted)
N_{10}/N, concentration index for text	0.895	0.821
	89.5 % of text	or 82.1 % of text
V_{10}, frequently used lemmas with frequency 10 and higher	4 681	5 003
V_{10}/V, concentration index for vocabulary	0.172	0.150
	17.2 % of vocabulary	or 15 % of vocabulary

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2019. Випуск 70

As we can see from the table, the same text size of two corpora shows different number for basic quantitative data: vocabulary volume, the number of hapax legomena, the most frequently used tokens in text, and the most frequently used lemmas in vocabulary. As a result for all the other statistical characteristics are different: index of variety and everige repetition of a word in the text, exclusiveness index for text and for vocabulary, concentration indexes for text and vocabulary.

Vocabulary volume for Ukrainian long prose in general (333 91 lexeme) is for 6 199 lexemes (19 %) bigger than vocabulary volume for I. Franko (27 192). Naturally Ukrainian long prose index of variety (0.067) is higher than Franko's (0.054), so everige repetition of a word in the text for Ukrainian long prose (15) is 4 words lower than Franko's (19). The number of **hapax legomena** of Ukrainian general long prose (11 567) predominates this number in Franko's works (14 522) by 2955 words (25.5 %). Concordantly, it has higher exclusiveness index for text 0.029 against 0.022 and exclusiveness index for vocabulary 0.43 against 0.42.

Frequently used tokens covers less text in the text corpus of Ukrainian general long prose (82.1 %) than in Franko's corpus (89.5 %).

Ukrainian general long prose has 0.074 points lower concentration index for text (0.821) than Franko (0.895) has. Frequent lemmas also cover less vocabulary in Ukrainian general long prose (15 %) than in Franko's vocabulary (17.2 %).

So, for the first glance I. Franko loses from Ukrainian long prose in general for all the parameters: it has 19 % less different lemmas, so his vocabulary richness is lover. It has 25.5 % less hapax legomena and it exclusiveness index is lover, so it has less rich language style than Ukrainian long prose in general. These are objective remote observations of numbers. But to make the correct conclusions we have to take to the consideration the fact that the text corpus of Ukrainian general long prose fiction consist of the works of 22 different writers. Of course this feature has influence on the amount of different topics under the description, writing style and, as a result, on the amount of different lemmas, as well as on the unique vocabulary and the most frequently used words too.

Despite the impression Franko's works have significantly less lexemes in the same text volume (19 %) and essentially less hapaxes (25.5 %), the distance between the text covering by hapaxes however is not so big, its 0, 06 %. Neither the distance between the vocabulary covering by hapaxes: 43.5 % against 42.5 % is 1 %. Concordantly the difference between the exclusiveness indexes doesn't look big: 0.007 points for text and 0.001 points for vocabulary. More over the amount of frequently used lemmas in Franko's works (4 681) is lower than in Ukrainian general long prose fiction (5 003) in 322 lemmas! So Ivan Franko used the same frequent words less often than 22 writers all together!

It seems reasonable to calculate the **percentage ratio of hapax legomena and frequently used lemmas** for both corpora. For text of Ukrainian general long prose fiction 2.9 % (hapaxes) + 82.1 % (frequently used words) = 85 %, so 15 % is covered by the words with the frequency range (2–9). For Franko's texts, 2.3 % (hapaxes) + 89.5 % (frequently used words) = 91.8 %, so 8.2 % (twice less) is covered by the words with the frequency rage (2–9).

For vocabulary of Ukrainian general long prose fiction, 43.5 % (hapaxes) + 15 % (frequently used words) = 58.5 %, so 41.5 % of vocabulary is covered by the words with the

frequency rage (2–9). For Franko's vocabulary 42.5 % (hapaxes) + 17.2 % (frequently used words) = 59.7 %, so 40.3 % is covered by the lemmas with the frequency rage (2–9). So, maybe, percentage ratio of hapax legomena and frequently used lemmas around 40–45 % could be considered as a constant value for text corpora of 500 000 tokens? This hypothesis could be checked in the future research.

Conclusions. The combined frequency dictionary of all 9 Ukrainian novels by I. Franko can be considered as a statistical combinatory model of Franko style. From the writer text corpus one can take a comprehensive quantitative and statistical portrait of writer's language and style.

The following distinguishing quantitative parameters for I. Franko long prose fiction and Ukrainian general long prose fiction were compared: lexeme vocabulary volume, index of variety, the amount of hapax legomena, their occupation of text and vocabulary, exclusiveness index for text and vocabulary, amount of words in text with frequency 10 and higher, their occupation of text and vocabulary, concentration indexes for text and vocabulary. The important similarities and differences were found.

The new statistical parameter for the writer's language and style is offered. It is percentage ratio of hapax legomena and frequently used lemmas. This characteristics helped to find out and to describe more precisely the important features of author's style.

Research prospects. The obtained data would be interesting to compare with the same characteristics for other Ukrainian writers as well as for the foreign Slavic writers. The statistical parameters for direct and author's speech in Franko's texts could be interesting to calculate, as well as to compare them with those in Ukrainian general long prose.

In the perspective we plan to calculate some other important distinguishing features for Franko's works as following: Zipf's Law, Menzerat-Altman's Law, h-point and k-point, sequences distribution of parts of speech, "temperature" of the text as so on. Such the results together with qualitative interpretation will show the wider picture of writer's works.

Список використаної літератури

1. *Бук С.* Кількісне зіставлення текстів (на матеріалі редакцій 1884 та 1907 років повісті Івана Франка "Boa constrictor") / С. Бук // Українське літературознавствою. – 2012. – Вип. 76. – С. 179–192.
2. *Бук С.* Сучасні методи дослідженні мови письменника у слов'янознавстві / С. Бук // Проблеми слов'янознавства. – 2012. – Вип. 61. – С. 86–95.
3. *Бук С.* Частотний словник роману Івана Франка "Основи суспільності" : Інтерпретація твору крізь призму статистичної лексикографії / С. Бук ; Ф. С. Бацевич (наук. ред). – Львів, 2012. – 264 с.
4. *Бук С.* Корпус текстів Івана Франка: спроба визначення основних параметрів / С. Бук // Прикладна лінгвістика та лінгвістичні технології: MegaLing–2006: зб. наук. пр. / НАН України. Укр. мовн.-інформ. фонд, Таврійськ. нац. ун-т ім. В. І. Вернадського ; за ред. В. А. Широкова. – Київ : Довіра, 2007. – С. 72–82.
5. *Бук С.* Частотний словник роману Івана Франка "Boa constrictor" / С. Бук // Стежками Франкового тексту (комунікативні, лінгвосеміотичні, когнітивні та лінгвостатистичні виміри прози) / С. Бук ; Ф. С. Бацевич (наук. ред.). – Львів : Видавничий центр ЛНУ імені Івана Франка, 2013. – С. 202–501.

6. *Бук С.* Частотний словник повісті І. Франка "Перехресні стежки" / С. Бук, А. Ровенчак // Стежками Франкового тексту (комунікативні, стилістичні та лексичні виміру роману "Перехресні стежки") / С. Бук ; Ф. С. Бацевич (наук. ред.). – Львів, 2007. – С. 138–369.
7. *Друль О.* Кількісна оцінка пуризму Бориса Грінченка / О. Друль // Україна Модерна 2019 (у друці).
8. Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна та ін. – Київ, 2005. – 471 с.
9. Лексика поетичних творів Івана Франка. Методичні вказівки з розвитку лексики / І. І. Ковалик, І. Й. Ощипко, Л. М. Полюга. – Львів, 1990. – 264 с.
10. *Ненадкевич Є. О.* Із студій над стилем Франкової й Стефаникової новели / Є. О. Ненадкевич // Записки Волинського інституту народної освіти ім. Івана Франка. – Житомир, 1927. Кн. II.
11. *Перебийніс В. С.* Статистичні методи для лінгвістів : навчальний посібник / В. С. Перебийніс. – Вінниця : Нова книга, 2002. – 168 с.
12. *Перебийніс В.С.* Частотні словники та їх використання / В. С. Перебийніс, М. П. Муравицька, Н. П. Дарчук. – Київ : Наукова думка, 1985. – 204 с.
13. *Полюга Л.* Статистичний аналіз лексики поетичних творів І. Франка / Л. Полюга // Іван Франко і національне відродження. – Львів, 1991. – С. 164–166.
14. Статистичні параметри стилів / за ред. В. С. Перебийніс. – Київ, 1967. – 260 с.
15. *Тулдава Ю. П.* Проблемы и методы квантитативно-системного исследования лексики / Ю. П. Тулдава. – Таллинн : Валгус, 1987. – 203 с.
16. Частотний словник сучасної української художньої прози : у 2 т. / за ред. Перебийніс В. С. – Київ, 1981.
17. *Шайкевич А. Я.* Статистический словарь языка Достоевского / А. Я. Шайкевич, В. М. Андрющенко, Н. А. Ребецкая. – Москва, 2003. – 880 с.
18. *Alekseev P. M.* Frequency dictionaries / P. M. Alekseev // Quantitative Linguistik : ein internationales Handbuch = Quantitative linguistics : an international handbook / edited by Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski. – Berlin ; New York: Walter de Gruyter, 2005. – P. 312–324.
19. *Best K.-H.* Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung / K.-H. Best, S. Zinenko // Zeitschrift für Slavische Philologie. – 1998. – Bd. 58. – S. 107–123.
20. *Cvrček V.* Kvantitativní analýza kontextu. Praha: Nakladatelství Lidové noviny / V. Cvrček // Ústav Českého národního korpusu. – 2013. – 283 s.
21. *Holovatch Y., Palchykov V.* Complex Networks of Words in Fables / Y. Holovatch, V. Palchykov // Maths Meets Myths: Complexity-science approaches to folktales, myths, sagas, and histories / R. Kenna, M. Mac Carron, P. Mac Carron (Editors). – Springer, 2016.
22. *Holovatch Y., Palchykov V.* Lys Mykyta and Zipf Law / Y. Holovatch, V. Palchykov // Statistical Physics 2005: Modern Problems and New Applications, August 28–30, 2005, Lviv, Ukraine: Book of abstracts. – P. 136.
23. *Köhler R., Altmann G.* Aims and Methods of Quantitative Linguistics / R. Köhler, G. Altmann // Problems of Quantitative Linguistics. – Chernivci: Ruta, 2005. – P. 12–42.
24. *Pawłowski A.* Metody kwantytatywne w sekwencyjnej analizie tekstu / A. Pawłowski. – Warszawa : UW, 2001. – 168 s.
25. *Popescu I.-I.* Some aspects of word frequencies / I.-I. Popescu // Glottometrics. – 2006. – No. 13. – P. 23–46.
26. *Ruszkowski M.* Statystyka w badaniach stylistyczno-składniowych / M. Ruszkowski. – Kielce : Wydawnictwo Świętorszyskiej, 2004. – 144 s.
27. *Sambor J.* Językoznawstwo statystyczne dla pracowników informacji naukowej / J. Sambor. – Warszawa, 1987. – 98 s.

28. Slovník Karla Čapka / Hl. ed. František Čermák. – Praha : Nakladatelství Lidové noviny; Ústav Českého národního korpusu, 2007. – 715 s.
29. *Vasić S.* Polazne osnove novije srpske proze – frekvencijski rečnik romana Oslobodioci i izdajnici Milorada Danojlića / S. Vasić. – Beograd : Kultura, 2002. – 85 s.
30. *Wimmer W.* Review Article: On vocabulary richness / W. Wimmer, G. Altmann // Journal of Quantitative Linguistics 1999. – Vol. 6. – No. 1. – P. 1–9.
31. *Wimmer G.* Úvod do analýzy textov / G. Wimmer, G. Altmann, L. Hřebíček, S. Ondrejovič, S. Wimmerová. – Bratislava : Veda, 2003. – 343 s.

References

1. Buk, S. (2012). Kilkisne zistavlennia tekstiv (na materiali redaktsii 1884 ta 1907 rokiv povisti Ivana Franka "Boa constrictor"). *Ukrainske literaturoznavstvoiu, 76*. 179–192.
2. Buk, S. (2012). Suchasni metody doslidzhenni movy pysmennyka u slovianoznavstvi. *Problemy slovianoznavstva, 61,* 86–95.
3. Buk, S. (2012). *Chastotnyi slovnyk romanu Ivana Franka "Osnovy suspilnosti": Interpretatsiia tvoru kriz pryzmu statystychnoi leksykohrafii*. F. S. Batsevych (nauk. red). Lviv.
4. Buk, S. (2007). Korpus tekstiv Ivana Franka: sproba vyznachennia osnovnykh parametriv. *Prykladna linhvistyka ta linhvistychni tekhnolohii: MegaLing, 2006: Zb. nauk. pr*. / NAN Ukrainy. Ukr. movn.-inform. fond, Tavriisk. nats. Un-t im. V. I. Vernadskoho; za red. V. A. Shyrokova. Kyiv: Dovira, 72–82.
5. Buk, S. (2013). Chastotnyi slovnyk romanu Ivana Franka "Boa constrictor". *Stezhkamy Frankovoho tekstu (komunikatyvni, linhvosemiotychni, kohnityvni ta linhvostatystychni vymiry prozy)*. F. S. Batsevych (nauk. red.). Lviv : Vydavnychyi tsentr LNU imeni Ivana Franka, 202–501.
6. Buk, S., Rovenchak, A. (2007). Chastotnyi slovnyk povisti I. Franka "Perekhresni stezhky". *Stezhkamy Frankovoho tekstu (komunikatyvni, stylistychni ta leksychni vymiru romanu "Perekhresni stezhky")*. F. S. Batsevych (nauk. red.). Lviv, 138–369.
7. *Drul, O.* (2019). Kilkisna otsinka puryzmu Borysa Hrinchenka. *Ukraina Moderna* (u drutsi).
8. Shyrokov, V. A., Buhakov, O. V., Hriaznukhina, T. O. ta in. (2005). *Korpusna linhvistyka*. Kyiv.
9. Kovalyk, I. I., Oshchypko, I. Y., Poliuha, L. M. (1990). *Leksyka poetychnykh tvoriv Ivana Franka. Metodychni vkazivky z rozvytku leksyky.* Lviv.
10. Nenadkevych, Ye. O. (1927). Iz studii nad stylem Frankovoi y Stefanykovoi novely. *Zapysky Volynskoho instytutu narodnoi osvity im. Ivana Franka*. Zhytomyr, Kn. II.
11. Perebyinis, V. S. (2002). *Statystychni metody dlia linhvistiv: Navchalnyi posibnyk*. Vinnytsia: Nova knyha.
12. Perebyinis, V. S., Muravytska, M. P., Darchuk, N. P. (1985). *Chastotni slovnyky ta yikh vykorystannia*. Kyiv: Nauk. dumka.
13. Poliuha, L. (1991). Statystychnyi analiz leksyky poetychnykh tvoriv I. Franka. *Ivan Franko i natsionalne vidrodzhennia*. Lviv, 164–166.
14. *Statystychni parametry styliv*. (1967). Za red. V. S. Perebyinis. Kyiv.
15. Tuldava, Yu. P. (1987). *Problemy y metody kvantytatyvno-systemnoho yssledovanyia leksyky*. Tallynn : Valhus.
16. *Chastotnyi slovnyk suchasnoi ukrainskoi khudozhnoi prozy*: u 2 t. (1981). Za red. Perebyinis V. C. Kyiv.

17. Shaikevych, A. Ya., Andriushchenko, V. M., Rebetskaia, N. A. (2003). *Statystycheskyi slovar yazыka Dostoevskoho.* Moskva.
18. Alekseev, P. M. (2005). Frequency dictionaries. In *Quantitative Linguistik : ein internationales Handbuch = Quantitative linguistics : an international handbook*. Edited by Reinhard Köhler, Gabriel Altmann, Rajmund G. Piotrowski. – Berlin ; New York: Walter de Gruyter, 312–324.
19. Best, K.-H., Zinenko, S. (1998). Wortkomplexität im Ukrainischen und ihre linguistische Bedeutung. *Zeitschrift für Slavische Philologie, 58*, 107–123.
20. Cvrček, V. (2013). *Kvantitativní analýza kontextu*. Praha: Nakladatelství Lidové noviny / Ústav Českého národního korpusu.
21. Holovatch, Y., Palchykov, V. (2016). Complex Networks of Words in Fables. In: *Maths Meets Myths: Complexity-science approaches to folktales, myths, sagas, and histories*. R. Kenna, M. Mac Carron, P. Mac Carron (Editors). – Springer.
22. Holovatch, Y., Palchykov, V. (2005). Lys Mykyta and Zipf Law. *Statistical Physics 2005: Modern Problems and New Applications*. August 28–30. Lviv, Ukraine: Book of abstracts, 136.
23. Köhler, R., Altmann, G. (2005). Aims and Methods of Quantitative Linguistics. *Problems of Quantitative Linguistics*. Chernivci: Ruta, 12–42.
24. Pawłowski, A. (2001). *Metody kwantytatywne w sekwencyjnej analizie tekstu*. Warszawa: UW.
25. Popescu, I.-I. (2006). Some aspects of word frequencies. *Glottometrics, 13*, 23–46.
26. Ruszkowski, M. (2004). *Statystyka w badaniach stylistyczno-składniowych*. Kielce: Wydawnictwo Świętorszyskiej.
27. Sambor, J. (1987). *Językoznawstwo statystyczne dla pracowników informacji naukowej*. Warszawa.
28. *Slovník Karla Čapka*. (2001). Hl. ed. František Čermák. Praha: Nakladatelství Lidové noviny; Ústav Českého národního korpusu.
29. Vasić, S. (2002). *Polazne osnove novije srpske proze – frekvencijski rečnik romana Oslobodioci i izdajnici Milorada Danojlića*. Beograd: Kultura.
30. Wimmer, W., Altmann, G. (1999). Review Article: On vocabulary richness. *Journal of Quantitative Linguistics*, *6,No. 1,* 1–9.
31. Wimmer, G., Altmann, G., Hřebíček, L., Ondrejovič, S., Wimmerová, S. (2003). *Úvod do analýzy textov*. Bratislava: Veda.

Стаття: надійшла до редакції 15.10.2018
прийнята до друку 13.11.2018

ISSN 2078-5534. Вісник Львівського університету. Серія філологічна. 2019. Випуск 70

РОЗРІЗНЮВАЛЬНІ КВАНТИТАТИВНІ ПАРАМЕТРИ АВТОРСЬКОЇ МОВИ ТА СТИЛЮ (НА МАТЕРІАЛІ ВЕЛИКОЇ ПРОЗИ ІВАНА ФРАНКА)

Соломія БУК

Львівський національний університет імені Івана Франка,
кафедра загального мовознавства,
вул. Університетська, 1, к. 343, Львів, Україна, 79000,
тел.: (032) 239 47 56, e-mail: solomija@gmail.com

Статтю присвячено детальному аналізу розрізнювальних кількісних параметрів мови та стилю автора. Таке дослідження проведено вперше для великої прози Івана Франка.

На матеріалі електронного корпусу текстів зі зовнішньою та внутрішньою розміткою укладено частотний словник усіх дев'яти українських романів Івана Франка. Його можна розглядати як статистичну комплексну модель стилю І. Франка, а також як мовно-статистичний портрет його великої прози. Отримано такі параметри: обсяг словника лексем, індекси різноманітності, винятковості, концентрації, кількість hapax legomena та їх відсоткову частку у тексті й словнику, кількість слів з частотою 10 і вище та їх відсоткову частку у тексті й словнику. Ці характеристики зіставлено з відповідними даними "Частотного словника сучасної української художньої прози".

Запропоновано новий розрізнювальний статистичний параметр для мови та стилю письменника. Це процентне співвідношення hapax legomena і найчастотнішої лексики. Ця характеристика допомогла з'ясувати і детальніше описати важливі особливості авторського стилю. У перспективі заплановано також обчислити Закон Ціпфа, закон Менцерата-Альтмана, h-point і k-point, розподіл послідовностей частин мови, "температуру" тексту.

Ключові слова: лінгвостатистика, корпус текстів, статистичні параметри, квантитативні характеристики тексту, частотний словник, слововживання, лема (лексема), лексичне багатство словника.