РОЗДІЛ VII. ПРИКЛАДНА ЛІНГВІСТИКА: НАПРЯМИ Й АСПЕКТИ ДОСЛІДЖЕННЯ

Oksana Zuban

DOI 10.31558/1815-3070.2019.38.15
УДК 81' 322.3

THE MORPHEMIC SYSTEM STYLOMETRIC ANALYSIS OF THE UKRAINIAN POETS' IDIOSTYLES: CORPUS BASED APPROACH

У статті описано статистичне дослідження ідіостилів чотирьох українських поетів, яке проведено на основі електронних частотних словників Корпусу української мови. У дослідженні введено, обґрунтовано та використано нову метричну модель – морфемну статистичну структуру ідіостилю. Ця модель систематизує кількісні та статистичні характеристики моделей морфемних структуру слів, обчислені у репрезентативних текстових вибірках, і забезпечує отримання вірогідних статистичних висновків для встановлення стилістичних особливостей ідіостилів чотирьох поетів.

Ключові слова: морфемна статистична структура ідіостилю, лексична продуктивність, індекс покриття тексту, відносна частота, модель морфемної структури – ММС, морфемна довжина слова.

1. Introduction

Statistical methods are particularly important in modern stylistics because the lack of quantitative values in functional and author's stylistic research deprives a scientific study of credibility and evidential value of the conclusions reached. The development of statistical stylistics in Ukrainian linguistics was marked by the monograph "Statistical parameters of styles" (Statystyčni parametry styliv) coming out in 1967. However, morphemic structures of words were not the subject of a separate study. The reason of it was a "manual" morphemic and statistical analysis of words in a text. The systematic study of morphemic structure of words in terms of idiostyle parameterization is possible in terms due to both corpus linguistics and computational statistical lexicography. In particular, such research is based on the data from the morpheme frequency dictionaries (FD) compiled from the Ukrainian Language Corpus texts (Zuban' "Častotni morfemni slovnyky", Zuban' "Stylemetryčni oznaky morfemnyx struktur sliv").

The Ukrainian Language Corpus (Korpus ukrajins'koji movy) has a research purpose to solve a wide range of linguistic tasks (Darčuk), especially in the field of morphemics and word formation. Morpheme FDs are compiled from the Ukrainian Language Corpus sample texts (19 FDs).

The purpose of the article is to discuss the stylometric research of the morphemic structure of the Ukrainian language poetic text based upon the data in morpheme FD (Častotni slovnyky Korpusu) of four Ukrainian poets: Lesia Ukrainka; L. Kostenko; V. Stus; T. Shevchenko.

2. Methodological framework of the stylometric research

2.1 Morphemic word structure modelling

The modelling method is employed in the description of a morphemic word structure of electronic dictionaries (Zuban "Automatic Morphemic Analysis" 416): Latin stand for a functional type of a morpheme: P – prefix, R – root, S – suffix, F – ending, I – interfix, X – postfix. The model of a morphemic structure (MMS) of a word is ascribed to every lexeme (lemma) in the text, for example, *несповитий* – PPPRSF. Such model is a unit of measurement in this statistical research. MMS of a word is an invariant symbol model of morphemic word building that presents a number of morphemes in a word and functional types of morphemes. The quantitative, structural and statistical parameters of MMS of a word are regarded as stylometric features in stylistic research.

2.2 Statistical basis of the stylometric analysis

In the stylometric research the following targets were set:

1) to establish style-distinctive quantitative and statistical functional patterns of morphemic word structures in the text of idiostyle;

2) on the basis of the defined statistical characteristics of MMSs and statistical laws, to set up hypotheses-evaluation of the material under study concerning the availability of stylistically marked units in the text: "…because quantitative and qualitative sides of human speech are correlated and interrelated in a certain way; quantitative evaluation may play a role of signals that direct researcher's attention to qualitative peculiarities of an individual or functional style hidden from simple observation" (Piotrovskiy 8).

There are two basic analytical parameters in the stylometric research:

1) quantitative and structural: morphemic word length (MWL – the number of morphemes);

2) statistical: an absolute frequency (f) and a relative frequency (p – %) of MMSs in a sample text and the lexical inventory of an idiostyle.

The relative frequency of MMSs is interpreted as: 1) a specific gravity or lexical productivity in the lexical inventory; 2) a text coverage index in a sample text.

Statistical experiment is based on two comparative approaches:

1) comparison of quantitative, structural and statistical features of MMSs of words in four sample texts;

2) comparison of quantitative, structural and statistical features of MMSs of words in four lexical idiostyle inventories with these characteristics in a lexical inventory of the Ukrainian language system on the basis of Morphemic-Derivational Database of the Ukrainian Language (Klymenko).

The identification of a statistical "behaviour" of a word in the text is related to the notion of a "statistical text structure" which is determined by the ratio of the number of words in the lexical inventory grouped by the rank of an absolute frequency in the text to the total of absolute frequencies of these words in the text. The statistical text structure enables to single out three statistical word groups:

1) high-frequency words, which cover the largest part of a text, but account for the smallest share in the lexical inventory;

2) medium-frequency words, which are characterized by the greatest variation in the correlation with the lexicon productivity;

3) low-frequency words, which cover the smallest part of a text, but account for the largest share in the lexical inventory.

The percentage ratio of the above-mentioned statistical word groups makes up the statistical model of the text. "The statistical text structure is a kind of a net force that makes it possible to attribute one or another text to some functional style, to define an author, a period of its writing etc." (Perebyjnis "Častotni slovnyky" 130).

In Russian linguistics the term "statistical lexicon structure" of a text was introduced by R. M. Frumkina (Frumkina 78), but it was not used at the morphemic level of text organization in the stylometric research. Therefore, it is worth substantiating its terminological use in stylistics. There is the following sequence of the statistical dependence in the realization of the MMS of a word:

1) one MMS of a word is explicated in a certain number of lexemes (initial forms) constituting a set of units in the lexical inventory;

2) each word of this set may be substituted in the text by a certain number of text tokens. For instance, RSSSF model in the lexical inventory is represented by 31 words, whereas in the text – by 77 text tokens.

Respectively, taking into account the language-speech dichotomy the morphemic statistical structure (MSS) of the idiostyle is defined in two statistical samples: the lexical inventory (language system of idiostyle) and a text (speech system of idiostyle). However, the MSS model of an idiostyle demonstrates another pattern of the text organization that differs from traditional model of lexical statistical text structure.

MSS also expresses the absolute frequency distribution in the lexical inventory and the text. This correlation is based on comparison of the total of absolute frequencies of words in the lexical inventory and in the text grouped in accordance with the same MMS of a word, but not with the rank of absolute frequency in the text. Therefore, MSS reveals not only formal and quantitative features of the idiostyle organization, but also linguistic information about its organization. As a result, statistical patterns of words with different number of morphemes, different functional types of morphemes, different number of root morphemes, etc. are realized in the lexical inventory and the text of the idiostyle.

The MSS model of the idiostyle is considered to be a central notion of the stylometric research. This model is defined as a division of MMS of words into three statistical groups (high-frequency, medium-frequency, low-frequency) through comparison of the relative frequency of one MMS in two rank lists compiled from: 1) the lexical inventory (lexemes) – MSS of the lexicon; 2) a text (tokens) – MSS of a text.

In terms of defined metric notion – the MSS model of the idiostyle – the task of the research is to analyse: 1) the correlation of different statistical groups of words' MMSs; 2) the correlation of MMSs' statistical characteristics and MWL's quantitative characteristics.

3. Stylometric features of four Ukrainian poets' idiostyles at morphemic level of the text organization

The stylometric research of the statistical morphemic system structure of the idiostyles of Lesia Ukrainka, L. Kostenko, V. Stus and T. Shevchenko was based on stylistically homogeneous (poetic speech) and statistically unbalanced samples[1] by the amount of tokens: TSh – 68 295; LU – 36 058; VS – 36 640; LK – 24 238. The size of each sample does not correspond to the recommended use of samples comprising 150,000–200,000 in statistical research (Perebyjnis "Častotni slovnyky" 17), because poetic texts are much smaller than the texts of artistic prose. E.g., the sampling frame of T. Shevchenko has ≈ 68,000 tokens. That is why the stylometric research of idiostyles allows for the decreased samples size. The requirement of statistical balance of samples is also violated, because the volume of a sampling frame of each poet is different, and the stylometric research of morphemic structures implies the need to study the greatest possible text coverage. The statistical data of the relative frequency in percentage are used to obtain reliable conclusions in the research, which is statistically justified in such cases.

This research sets the following goals according to the defined methodological framework:

1) to compile a list of MMSs based on reduced productivity in lexical inventories of the four samples and compare them;

2) to compile a list of MMSs based on reduced text coverage index of the four samples and compare them;

3) to analyse statistical data of MMS lexical productivity and MMS text coverage index in the four poetic samples;

4) to define MWL and compare productivity of words having different morphemic length in lexical inventories of the four samples;

5) to define an average MWL in each poetic sample.

3.1. Morphemic statistical structure of the lexicon

The inventory of MMSs compiles a separate morphemic system for every poet which combines different number of units with different lexical productivity: TSh (82 MMSs – 6 191 lexemes); VS (87 MMSs – 8 029 lexemes); LK (73 MMSs – 5 956 lexemes); LU (61 MMSs – 4 907 lexemes).

[1] In the article the names of samples are indicated by abbreviations: Lesia Ukrainka – LU; L. Kostenko – LK; V. Stus – VS; T. Shevchenko – TSh; language system – LS.

According to the common tradition in linguistic statistics (Perebyjnis "Kil′kisni ta jakisni xarakterystyky"), Table 1 shows a core (75 % of the lexicon) and the base (90 % of the lexicon) of MMS of every poet. The list of MMSs in the lexical inventory of TSh is conditionally considered a standard list for comparison.

Table 1. Lexical productivity of MMSs in the inventory of four poets

Specific gravity (%) in lexicon					
MMS		**TSh**	**VS**	**LU**	**LK**
RSF	*плак-а-ти*	22,98	22,42	23,74	24,58
RF	*мій-Ø*	21,80	19,22	19,65	26,46
PRSF	*с-пі-ва-ти*	18,35	19,97	21,38	15,51
RSSF	*дів-ч-ин-а*	5,22	5,57	5,73	5,51
PRF	*ні-хт-о*	5,15	4,70	4,75	4,45
		Core 73,50		**75,25**	**76,51**
PRSFX	*по-див-и-ти-ся*	3,83	3,89	2,55	2,67
			Core 75,77		
R	*де*	2,91	2,32	2,89	2,84
RSFX	*див-и-ти-ся*	2,08	2,04	~~1,39~~	1,48
PRSSF	*на-йм-ич-к-а*	1,91	3,89	2,85	2,13
PRSS	*в-ран-ц-і*	1,83	~~1,05~~	~~0,84~~	1,21
PRS	*з-нов-у*	1,44	1,42	1,63	1,38
RS	*добр-е*	1,32	1,21	1,81	1,12
RSS	*тяж-к-о*	1,28	1,23	2,22	1,21
		Base 90,10			**Base 90,55**
PPRSF	*без-не-вин-н-ий*	1,03	1,23	1,43	0,86
				Base 90,63	
RIRSF	*біл-о-рук-Ø-ий*	0,92	1,13	0,63	0,89
			Base 90,24		
PR	*в-день*	0,84	1,00	0,92	0,82

Table 1 indicates that ≈ 75 % of three poets' lexicon is modelled by 5 common MMSs: RSF, RF, PRSF, RSSF, PRF. The core of V. Stus' morphemic system is formed by 6 MMSs: PRSFX is also included into the core. The next ≈ 15 % of the lexicon (TSh: 8 MMSs; VS: 8 MMSs; LU: 7 MMSs; LK: 8 MMSs) are modelled by different MMSs following: a) models common to the 4 poets: R; PRSSF; PRS; RS; RSS; b) models of separate idiostyles: VS – RIRSF; VS, LK – PPRSF; TSh, LK – PRSS; TSh, VS, LK – RSFX.

In general, according to the 4 lexical inventories, 16 MMSs have ≈ p ≥ 1 %. The rest of MMSs (≈ 10 % of the lexicon) are characterized by productivity ≈ p < 1 %. Automatic determination of low-productive MMSs in FD accomplishes a statistical method of removing individual stylistically marked lexicon from a text. This method may be used with high validity in any kind of stylistic research.

The specific gravity distribution of MMSs in every lexical inventory forms an individual model of MSS of poet's lexicon, which may be visualised in the following way (see Fig. 1).

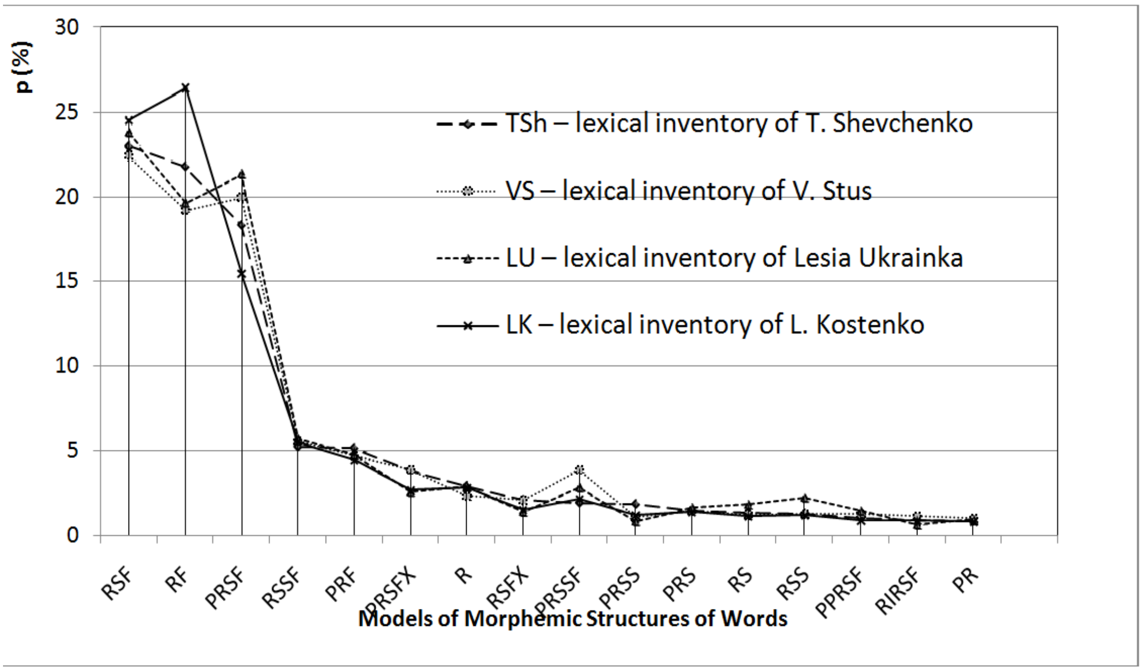

Fig. 1. Graph of the MMSs distribution in four poets' lexicons

The graph demonstrates with what productivity degree 16 MMSs (*x* axis) of the morphemic system of every poet cover the lexical inventory by p % (*y* axis). The curve of every idiostyle displays this dependence. The shape of lines denotes the general and individual features of MMSs' percentage distribution in four lexical inventories. The general features of the percentage distribution define 3 zones of the lexical productivity:

1) high lexical productivity (24,58 – 15,51 %) – the highest peaks of curves – the points RSF, RF, PRSF;

2) medium lexical productivity (5,73 – 1,91 %) – a sharp drop of curves till ≈ 5% (the RSSF model) – the points RSSF, PRF, PRSFX, R, RSFX, PRSSF;

3) low lexical productivity (1,83 – 0,82 %) – low peaks of curves which almost overlap – the points PRSS, PRS, RS, RSS, PPRSF, RIRSF, PR.

The individual features of the percentage distribution of MMSs characterize, first of all, the MSS of L. Kostenko's lexicon: the highest peak is indicated on the second RF model (26,46 %), and the lowest productivity in zone 1 is indicated on the PRSF model (15,51 %).

The curve of the lexical productivity of TSh is distinguished by a drop on the points RF (21,80 %), PRSF (18,35 %); the curves of VS and LU – by a rise on the point PRSF. It means MMS distribution in zone 1 determines the stylistic peculiarities of every sample. In zone 2 the model PRSSF is a strong stylometric feature which has the highest peak in VS' curve (3,89 %). On zone 3 the LU's line on the point RSS is distinguished by the curves' peaks (2,22 %).

I suggest checking the substantiality of the percentage discrepancy on the points RF, PRSF, PRSSF, RSS by Student's t-distribution (see Table 2) for samples percentages as compared to significant percentages discrepancy.

Table 2. Student's *t* value

MMS	Samples under comparison	*t*	Critical *t* value at 99% confidence level
RF	LK:VS	10,20	2,58
	TSh:VS	3,91	
PRSF	LK:LU	7,83	
	LK:TSh	4,71	
PRSSF	LK:VS	5,87	
	LU:VS	3,15	
RSS	LU:TSh	3,48	

The data indicate that Student's *t* value in all four models is higher than the critical t > 2,58. Thus, the data analysis proves that the percentage discrepancy of the RF, PRSF, PRSSF, RSS models is substantial, and their percentage value is the statistical parameter of the idiostyle differentiation.

3.2 Morphemic statistical structure of the text

16 MMSs of every sample cover a different data set of the lexical inventory and the text:

the sample:	% of the lexical inventory:	% of the text:
TSh	92,89	97,42
VS	92,29	96,35
LU	94,44	97,52
LK	93,12	96,15

The comparison of the specific gravity distribution of each model in the lexical inventories with the text coverage index in every sample is designed via two curves on separate graphs (see Fig. 2), where *x* axis – p %, *y* axis – 16 MMSs.

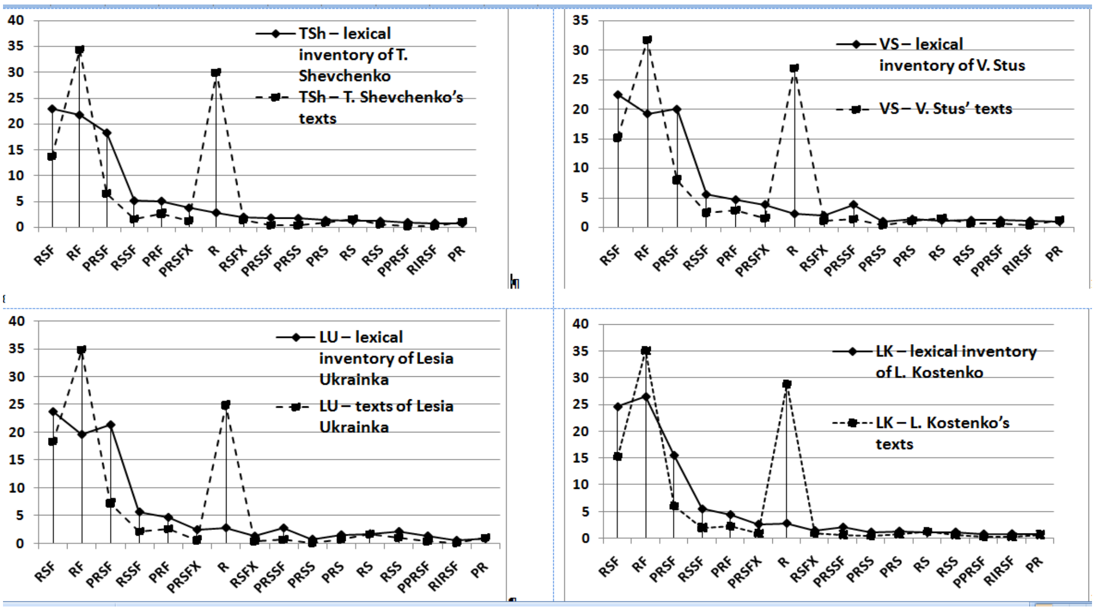

Fig. 2. Graphic models of MSS of four idiostyles

The comparison of two lines on every graph indicates the discrepancy of morphemic structure productivity in language (the lexical inventory) and speech (the text). It also enables the description of the peculiarities of the MSS model of the idiostyle. The curves on four graphs demonstrate a common pattern: each curve of the text coverage index copies the curve shape of the lexical productivity, but the productivity of a great deal of MMSs in the text is lower than in the lexical inventory, except for two MMSs – RF and R (it violates the general downward trend in text coverage index). A higher text coverage index of the R model is due to high frequency of conjunctions, particles and prepositions in the texts – they mostly represent this model in the text, but the RF model has the highest coverage index in the text. The peaks of the text curves RF and R denote them to cover ≈ 58–65 % of every author's text (for comparison, the percentage zone of the RF's lexical inventory: 19,22–26,46 %; R: 2,32–2,99 %). Obviously, the words with such morphemic structure are defined as stylistically neutral in terms of linguistic statistics laws. However, these words have the smallest morphemic length that conveys the stylometric feature of poetic speech on a whole. This phenomenon confirms the importance of introducing the MWL parameter into the MSS model of the idiostyle.

3.3 Morphemic word length

The lexical inventories[2] of poetic samples contain words with different morphemic length (see Table 3): TSh, LU (1–7 morphemes), VS, LK (1–8 morphemes). Having the same number of morphemes, MMSs vary in functional structure; e.g., 4-morpheme structure (4M) may have such functional structures: RSSF – *дів-ч-ин-а*, PRSS – *в-ран-ц-і.* That why each quantitative morphemic model can be represented by a different number of MMSs. Table 3 systematizes the MWL data, the quantity of MMSs having the same number of morphemes, and the relative frequency in the lexical inventory.

Table 3. The ratio of MWL to a number of MMSs and the specific gravity

Number of mor-phemes in a word	T. Shevchenko (TSh)			V. Stus (VS)			L. Kostenko (LK)			Lesia Ukrainka (LU)			% % in Ukrainian language system
	Number of MMSs		% in lexical inventory	Number of MMSs		% in lexical inventory	Number of MMSs		% in lexical inventory	Number of MMSs		% in lexical inventory	
	of simple words	of compound words		of simple words	of compound words		of simple words	of compound words		of simple words	of compound words		
1M	1		2,63	1		2,32	1		2,94	1		2,19	0,22
2M	4	1	24,63	4	1	22,77	4	1	28,93	4	1	22,70	6,87
3M	12	5	31,80	7	7	31,48	7	4	31,48	10	4	33,02	21,24
4M	12	8	30,15	11	8	29,39	12	8	27,44	11	3	31,20	34,51
5M	13	9	9,18	13	11	12,04	13	5	7,79	10	4	8,64	26,13
6M	10	5	1,15	7	7	1,41	7	5	0,96	7	4	1,06	7,67
7M		2	0,065	2	6	0,16	2	2	0,084	1	1	0,041	2,46
8M					2	0,025		2	0,034				
Total	**52**	**30**	**99,61**	**45**	**42**	**99,59**	**46**	**27**	**99,66**	**44**	**17**	**99,45**	**99,1**
	82			87			73			61			

The number of MMSs of simple and compound words in poets' lexicons is represented differently: the largest functional variety is MMSs of simple words in the samples of VS (45 MMSs), LK (46 MMSs), and on MMSs of compound words – in the samples of TSh (30 MMSs) and VS (42 MMSs). 1-morpheme and 2-morpheme words are indicated in the lexicons of four poets with the same number of MMSs that model all possible functional types of 1- and 2-morpheme structures of simple and compound words in the Ukrainian language: R, RR, RF, RS, RX, PR. The quantitative distribution of MMSs (column – a number of MMSs) by the number of morphemes (the first column) shows the variation of the MMS quantity due to increasing number of morphemes in a word. In particular, 4-morpheme MMSs of compound words are the most numerous (8 units) in the samples of TSh, VS, LK. 5-morpheme MMSs of compound words have more functional variability in the samples of TSh (9 MMSs) and VS (11 MMSs). But in the lexicons of LK (5 MMSs) and LU (4 MMSs) the variability of these models is substantially decreasing. 7-morpheme MMSs have a high functional variability in VS' sample (8 MMSs).

Table 3 shows that MMSs are available in words of different morphemic length which is controlled by some patterns of quantitative restriction. According to the data from Morphemic-Derivational Database (Klymenko 150–153), the morphemic length of the Ukrainian word is restricted by an interval of 1–13 morphemes: for simple words – 1–11, for compound words – 2–13. MWL in the samples of TSh and LU is restricted by the interval of 1–7 morphemes; in the samples of VS and LK – 1–8 morphemes. Moreover, MWL impacts differently the lexical productivity of MMSs. The pattern of MWL's impact on MMS's productivity in the lexical systems of the four poetic samples and the Ukrainian language system is demonstrated graphically on Fig. 3.

[2] In order to compare the MWL in poetic samples with the Ukrainian language system the statistical calculations are carried out on the material of the lexical inventories, and not of the texts.

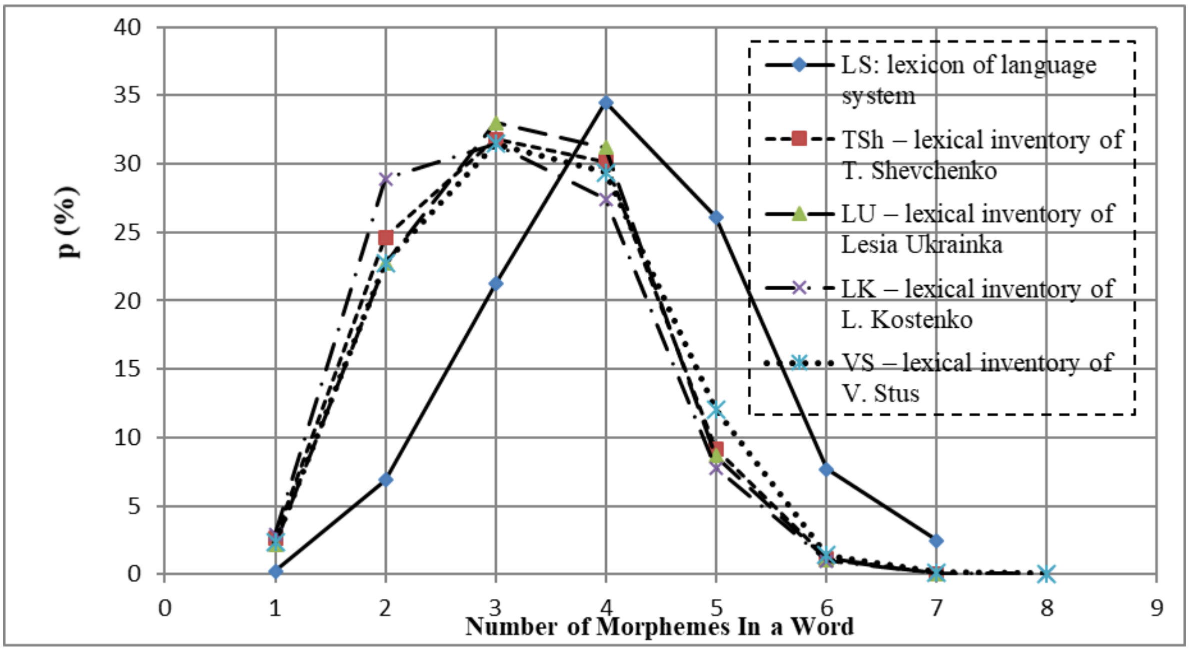

Fig. 3. Graph of MWL distribution in four poets' lexicons and the Ukrainian language system

The top of the language system (LS) curve is formed by 4-morpheme words. The tops of the curves of four poetic samples are formed by 3-morpheme words. 5-morpheme structures, unlike the LS, present a point of sharp productivity drop on the curves of poetic samples. The MWL distribution in poets' lexicons is the most similar in the samples of TSh and VS: the curves almost match. The LU's curve is formed like the previous ones, although the top of this curve conveys the highest specific gravity of 3-morpheme structures (33,02 %). The LK's curve differs most of all and demonstrates the highest specific gravity of 2-morpheme structures (28,93 %), and its sharp drop of the specific gravity (the lowest productivity – 27,44 %) of 4-morpheme structures.

The curves on the graph indicate a quantitative decrease of MWL in poets' lexical inventories, in comparison to LS. According to the data from Morphemic-Derivational Database, an average morphemic length of the Ukrainian word is 3,9 and is restricted by the interval of 4 ± 1 (Klymenko 150). Having determined the average MWL in the lexicon of every poet according to the formula $\overline{\mathrm{X}} = \frac{\sum \mathrm{x_i} \cdot n_i}{\sum n_i}$ ($\mathrm{x_i}$ – a number of MMSs morphemes, $\mathrm{n_i}$ – a number of words in the MMS), such data have been received for every idiostyle: TSh – 3,22; VS – 3,34; LU – 3,23; LK – 3,11. The average MWL of poetic speech is included into the interval of 4 ± 1. However, the poets use morphemically shorter words, which is considered to be the common stylometric feature of poetry.

4 Conclusions

The comparison of the statistical data in the samples of four Ukrainian poets in terms of MWL, the specific gravity in the lexicon, and the text coverage index design the stylometric model which is defined as the MSS model of the idiostyle. The use of such model in the stylistic research systematizes not only specific morphemic phenomena (being some individual speech peculiarities of poet), but also all the structural units at the morphemic level of the text system. This model demonstrates a higher level of generalising the quantitative text model rather than of using the statistical lexicon structure, because the number of MMSs in the text, in comparison with the number of words, is reduced a hundred times. The small inventory of MMSs allows to systematize the statistical characteristics of words in large sample texts.

The MMSs distribution in the lexical inventories and in the texts is carried out by Zipf's law (Zipf): 16 MMSs of the four poetic samples cover ≈ 92–94 % of the lexical inventory and ≈ 96–97 % of the text; 45 MMSs (LU), 71 MMSs (VS), 57 MMSs (LK), 66 MMSs (TSh) cover ≈ 6–8 % of the lexical inventory and ≈ 3–4 % of the text. Nevertheless, the ratio of the number of words of one MMS in the lexical inventory to the total of its absolute frequencies in the text does not follow Zipf's law, because the word distribution in the lexicon and in the text is implemented not in accordance with the mathematical model – by a variant of the absolute frequency, but according to the linguistic feature – by morphemic word building. Although Zipf's law does not work in this distribution, it explains the rising peaks of the relative frequency of 1- and 2-morpheme MMSs (R, RF) in the text: high-frequency words have a minimal morphemic length.

The reduction of the average MWL in poetic samples under study confirms the statement by V. A. Moskovych about the influence of style functions on this phenomenon: the basis of poetic style is defined not as a substantive, but as an aesthetic function.Therefore, in poetry special attention is paid to the word structure: "In poetry this explains the reduction to a minimum of words having intensity and length close to maximum" (Moskovich 29).

The MSS model of idiostyle may be considered as a typological statistical model in the study of different text styles. The statistical data introduced in the article are to be interpreted by means of qualitative methods of linguistic analysis that is going to be conducted in the future.

Література

1. Дарчук Н. Комп’ютерне анотування тексту: результати і перспективи: монографія. Київ: Освіта України, 2013. 543 с.
2. Зубань О. Стилеметричні ознаки морфемних структур слів у поетичному мовленні Т. Шевченка (на матеріалі Корпусу української мови). *Мовні і концептуальні картини світу*. 2014. Вип. 48. С. 165–179.
3. Зубань О. Частотні морфемні словники в Корпусі української мови – джерело стилеметричних досліджень. *Acta Universitatis Palackianae Olomucensis Philologica 104 – 2016: UCRAINICA VII: Současná ukrajinistika Problémy jazyka, literatury a kultury*. 2016. № 104. S. 22–33
4. Клименко Н. Ф. Основи морфеміки сучасної української мови: навч. посіб. Київ: ІЗМН, 1998. 182 с.
5. Москович В. А. Глубина и длина слова в естественных языках. *Вопросы языкознания.* 1967. № 6. С. 17–33.
6. Перебийніс В. С. Кількісні та якісні характеристики системи фонем сучасної української літературної мови: монографія. Київ: наукова думка, 1970. 272 с.
7. Перебийніс В. С., Муравицька М. П., Дарчук Н. П. Частотні словники та їх використання: монографія. Київ: Наукова думка, 1985. 203 с.
8. Пиотровский Р. Г. Информационные измерения языка: монография. Ленинград: Наука, 1968. 116 с.
9. Статистичні параметри стилів / ред. В.С. Перебийніс. Київ: Наукова Думка, 1967. 260 с.
10. Фрумкина Р. М. Статистическая структура лексики Пушкина. *Вопросы языкознания*. 1960. № 3. С. 78–81.
11. Zipf George K. Human Behavior and the Principle of Least Effort. Cambridge (Maas): Addison-Wesley, 1949. 573 p.
12. Zuban O. Automatic Morphemic Analysis in the Corpus of the Ukrainian Language: Results and Prospects. *Jazykovedný časopis (Journal of Linguistics)*. 2017. № 68 / 2. P. 415–426.

Список джерел

1. Корпус української мови. URL: http://www.mova.info/corpus.aspx (дата звернення 20.09.2019).
2. Частотні словники корпусу. URL: http://www.mova.info/article.aspx?l1=210&DID=5215 (дата звернення 20.09.2019).

References

Darčuk, Nataliya. *Komp"juterne anotuvannja tekstu: rezul'taty i perspektyvy (Computer Text Annotation: Results and Perspectives).* Kyiv: Osvita Ukrayiny, 2013. Print.

Frumkina, Revekka. *“Statisticheskaya struktura leksiki Pushkina (The Statistical Structure of Pushkin’s Vocabulary)”. Voprosy yazykoznaniya (Questions of Linguistics)* 3 (1960): 76–81. Print.

Klymenko, Nina. (1998). *Osnovy morfemiky sučasnoji ukrajins'koji movy (Morphemics Frameworks of Modern Ukrainian).* Kyjiv, 1998. Print.

Moskovich, Vol'f. *“Glubina i dlina slova v estestvennykh yazykakh (The Intensity and Length of a Word in Natural Languages)”. Voprosy yazykoznaniya (Questions of Linguistics)* 6 (1967): 17–33. Print.

Perebyjnis, Valentyna, and Darčuk N., Muravyts'ka M. *Častotni slovnyky ta jix vykorystannja (Frequency Dictionaries and Their Use*). Kyjiv: Naukova dumka, 1985. Print.

Perebyjnis, Valentyna. *Kil'kisni ta jakisni xarakterystyky systemy fonem sučasnoji ukrajins'koji literaturnoji movy (Quantitative and Qualitative Characteristics of the Phoneme System of Modern Ukrainian Literary Language)*. Kyjiv: Naukova dumka, 1970. Print.

Piotrovskiy, Raymúnd. *Informatsionnyye izmereniya yazyka (Informational Dimensions of Language).* Leningrad: Nauka, 1968. Print.

Statystyčni parametry styliv (Statistical Parameters of Styles) / za red. Perebyjnis, V. Kyjiv: Naukova dumka, 1967. Print.

Zipf, George. *Human Behavior and the Principle of Least Effort*. Cambridge (Maas): Addison-Wesley, 1949. Print.

Zuban, Oksana. *“Automatic Morphemic Analysis in the Corpus of the Ukrainian Language: Results and Prospects”. Jazykovedný časopis (Journal of Linguistics),* 68 / 2: 415–426. Print.

Zuban', Oksana. “*Častotni morfemni slovnyky v Korpusi ukrajins'koï movy – džerelo stylemetryčnych doslidžen' (Morpheme Frequency Dictionaries in the Corpus of the Ukrainian Language Are a Source of Stylometric Research)”. Acta Universitatis Palackianae Olomucensis, Philologica 104 – 2016: UCRAINICA VII: Současná ukrajinistika, Problémy jazyka, literatury a kultury,* (2016): 22–33. Print.

Zuban', Oksana. *“Stylemetryčni oznaky morfemnyx struktur sliv u poetyčnomu movlenni T. Ševčenka (na materiali Korpusu ukrajins'koji movy) (The Stylometric Features of Morphemic Word Structures in T. Shevchenko’s Poetic Speech (Based on the Corpus of the Ukrainian Language))”. Movni i konceptual'ni kartyny svitu (Linguistic and Conceptual Worldviews)* 48 (2014): 165–179. Print.

List of Sources

ČSK: *Častotni slovnyky Korpusu (Frequency Dictionaries of the Corpus).* Accessible at: http://www.mova.info/article.aspx?l1=210&DID=5215. Web. 20 Sept. 2019.

KUM: *Korpus ukrajins'koji movy (Corpus of the Ukrainian Language).* Accessible at: http://www.mova.info/corpus.aspx. Web. 20 Sept. 2019.

List of Abbreviations

f – absolute frequency;
F – ending;
FD – frequency dictionaries;
I – interfix;
LK – L. Kostenko;
LS – language system;
LU – Lesia Ukrainka;
MMS – model of morphemic structure;
MSS – morphemic statistical structure;
MWL – morphemic word length;
p – relative frequency;
P – prefix;
R – root;
S – suffix;
t – Student's value;
TSh – T. Shevchenko;
VS – V. Stus;
X – postfix.

Надійшла до редакції 4 жовтня 2019 року.

THE STATISTICAL STRUCTURE OF THE MORPHEMIC SYSTEM OF IDIOSTYLE (BASED ON THE DATA OF THE UKRAINIAN LANGUAGE CORPUS)

Oksana Zuban

Department of Ukrainian Language and Applied Linguistics of Institute of Philology, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

Abstract

Background: The development of Ukrainian corpus linguistics opens up new prospects for researchers in the application of statistical techniques for studies of Ukrainian text; it requires a development and implementation of new statistical models in organizing a linguostatistical experiment.

Purpose: to provide the theoretical justification of the model of morphemic statistical structure of idiostyle and apply this model to stylometric research of poetic texts of the four Ukrainian poets (Lesia Ukrainka; L. Kostenko; V. Stus; T. Shevchenko).

Results: Stylometric research was carried out based on the electronic morpheme frequency dictionaries that had been automatically compiled in the Ukrainian Language Corpus. A unit of the dictionaries' registries is a model of morphemic structure (MMS) of a word. This model is a unit of measurement in this statistical research and an object under study of quantitative and structural organization of morphemic word building.

On the basis of the defined methodological framework, and using tabular and graphical forms for presenting statistical data this research provides: the analysis of lexical productivity of MMSs and the text coverage index of MMSs in the four poetic samples; the analysis of productivity of words having different morphemic length in the lexical inventories of the 4 poetic samples and in the Ukrainian language system; the calculations of average morphemic word length in every poetic sample.

Discussion: The use of the model of morphemic statistical structure of idiostyle in the stylistic research demonstrates a higher level in generalising the quantitative text model than in using the statistical lexicon structure, because the number of MMSs in the text, in comparison with the number of words, is reduced a hundred times. The small inventory of MMSs allows to systematize the statistical characteristics of words in large sample texts and define the statistical parameters of idiostyle at morphemic level of the text organization. The statistical data introduced in the article are to be interpreted by means of qualitative methods of linguistic analysis that is going to be fulfilled in the future.

Keywords: morphemic statistical structure of idiostyle, lexical productivity, text coverage index, sample, relative frequency, model of morphemic structure of a word – MMS of a word, morphemic word length.

Vitae

Oksana Zuban, PhD in Philology, Associate Professor, Associate Professor of the Department of Ukrainian Language and Applied Linguistics of Institute of Philology, Taras Shevchenko National University of Kyiv. Her areas of research interests include semantics, grammar, linguistic statistics, forensic linguistics, SEO-copywriting, natural language processing, lexicography.

Correspondence: oxana.mell.zuban@gmail.com