Stylometric Study of the Fiction Using Sketch Engine

Oksana S. Taran
Lviv Polytechnic National University
Lviv, Ukraine
taran.oksana.serg@gmail.com

Oleksandra S. Palchevska
Lviv State University of Life Safety
Lviv, Ukraine
palch56@ukr.net

Alla A. Luchyk
National University of "Kyiv-Mohyla Academy"
Kyiv, Ukraine
allal@meta.ua

Viktoriia V. Shabunina
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
vshabunina@gmail.com

Oksana V. Labenko
Taras Shevchenko National University of Kyiv
Kyiv, Ukraine
o.v.labenko@gmail.com

ABSTRACT

The paper deals with a stylometric study of I. Asimov's idiostyle considering a corpus-based approach. For the analysis of stylometric features the I. Asimov "Foundation" cycle text corpus was created. The quantitative and statistical processing of the text corpus is done via Sketch Engine tool that enables comparison of phrases and words in the following variants: lemma, token, subcorpus. The last parameter is important for distinguishing individual authorial features, comparing their combinability and identifying the dynamics of idiostyle. The following stylometric features of a text corpus by I. Asimov are described: quantitative morphological and lexical characteristics of the vocabulary, quantitative characteristics of occasionalisms' word formation and statistical estimation of occasionalisms' collocations. It is sated that the frequency of occasionalisms in the cycle of novels undergoes chronological change, as well as their combinability. In this paper, a method of occasionalisms' automated extraction due to keyness score was proposed, however, it requires the subsequent manual verification.

CCS CONCEPTS

• **Applied computing** → **Language translation**.

KEYWORDS

corpus-based approach, corpus, stylometry, Sketch Engine

ACM Reference Format:
Oksana S. Taran, Oleksandra S. Palchevska, Alla A. Luchyk, Viktoriia V. Shabunina, and Oksana V. Labenko. 2021. Stylometric Study of the Fiction Using Sketch Engine. In *Digital Humanities Workshop (DHW 2021), December 23, 2021, Kyiv, Ukraine.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3526242.3526249

DHW 2021, December 23, 2021, Kyiv, Ukraine

ACM ISBN 978-1-4503-8736-1/21/12…$15.00
https://doi.org/10.1145/3526242.3526249

1 INTRODUCTION

With the development of corpus linguistics as well as the corpora and corpus managers emergence, quantitative-statistical, stylometric processing of texts acquired a new format, so as calculations and statistical evaluation became automated.

Stylometry can serve as a practical basis and a toolkit for further stylistic, lexicological, grammatical research allowing: to analyze the peculiarities of the author's language in terms of lexical composition, morphology and syntax (statistical and comparative analysis of lexical structures and syntactic structures in both texts of individual author's corpora as well as of the general language corpora) [12]; solve natural language processing tasks: authorship attribution, authorship verification, style change detection, authorship profiling, and text classification by genre [10]; search for words and collocations, examples of the lexical units use, author's occasionalisms and neologisms, depending on the context of use [14]; conduct research in the stylistics field (consideration of ways to convey stylistic features in texts, determination of techniques and stylistic means that a certain author uses in the text) [6]; develop software for text analysis, automatic search and extraction of equivalents for individual lexemes and collocations; explore phenomena in the language of a corpus [16].

The study of the systematic organization of individual-author lexical innovations (many of which not only became real linguistic findings of individual authors, but also entered the general language fund) is impossible without their statistical explication. However, works on stylometric research of the so-called neological vocabulary are represented only by single samples, despite the fact that recently the general tendency to language stylometric description has significantly increased. This task is still relevant, because modern studies of authorial neologisms are, in fact, a linguistic gap, the completion of which is extremely necessary and urgent.

The researchers agree that a feature of neologism-finding presents a technical challenge and their frequencies, even in very large corpora, will tend to be very low [8, 9]. So, the manuscript presents the new quantitative analysis of the author's style characteristics: for the first time the quantitative morphological and lexical features of I. Asimov's texts corpus are described, which gives the possibility to show the connection between the author's lexical diversity and background. The obtained stylometric parameters are important for further research of I. Asimov's texts and the genre of fiction in

general. Statistical estimation of occasionalisms' collocations based on LogDice score that is carried out gives the possibility to perform the automatic cross-check of the created innovations' list, and to supplement it.

The novels of I. Asimov's "Foundation" cycle were chosen as a material for our analysis. The aim is to carry out a corpus-based approach to stylometric study of I. Asimov's idiostyle on the novels of the "Foundation" cycle. The choice of the author working in the genre of science fiction was done due to the fact that the main characteristic feature of the last is the occasional lexical units use, which were mainly studied by manual sampling. In this paper, we propose a method of occasionalisms' automated extraction.

I. Asimov's works have been studied in the following aspects: 1) semantic, 2) cognitive, 3) translational (Allen [2], Bellefontaine [3], Hoppa [7], Nevala-Lee [13], Vainio [18], Westfahl [19]). However, all of these are examples of descriptive research, without taking into consideration quantitative and statistical indicators and the involvement of corpus data, which allow to determine the text stylometric features, to trace the author's style in the dynamics.

2 RESEARCH METHODOLOGY

2.1 Corpus-Based Approach

In modern linguistic studies, corpus-based and corpus-driven approaches are defined. Tognini-Bonelli [17] interprets the corpus-based approach as a method that involves the use of the language corpus to verify, confirm linguistic theory or as a source of illustrations to the formulated theoretical positions. The corpus-driven approach is a method that covers the interpretation of corpus data as a whole, i.e. the obtained results and formulated conclusions are based exclusively on corpus data [17]. Biber [4] proposes to use these two approaches comprehensively, as it will identify not only the phenomena of language but also speech, not yet fixed by grammatical theory [4]. Since our corpus is based on fiction of one cycle by one author, i.e. the corpus is static, it seems appropriate to determine a corpus-based approach with elements of corpus-driven, in particular in the analysis of collocations.

2.2 Creating of a Texts Corpus by I. Asimov in Sketch Engine

For this study, the commercial software Sketch Engine [1] was chosen because it gives the possibility to compare frequencies with the other 25 corpora of the English language and thus identify unique vocabulary, which is one of the indicators of the author's idiostyle.

Sketch Engine software helps to download the corpus, has a wide range of its statistical processing — both texts in general and individual lexical units, has options for creating frequency lists of words, concordances, N-grams, keywords, evaluation of collocations by statistical score, as well as allows to make morphological marking (PoS).

To avoid distorted search results, it is important to have a pre-grapheme analysis, i.e. prepare the corpus so that it does not contain special characters next to words, accidentally separated by a space words, coding errors, etc.

In order to be able to compare the quantitative parameters in different texts of the I. Asimov "Foundation" series we downloaded 7 books in plain text format as separate files. As a result, we have 7 subcorpora, which can be worked out separately and together. The statistics of the created corpus "Foundation_Asimov" is given in figure 1.

The size of the created corpus is 925,029 tokens. This corpus is written, monolingual (English), full-text, by one author (I. Asimov), static, marked (has morphological and syntactic markup), synchronous. It is structured into 7 subcorpora, which are named after the date of publication and the title of the novel. Thus, in each of the Sketch Engine options, queries are possible both for the entire "Foundation_Asimov" corpus as well as for a separate novel, which we select in the Text types section by title.

3 STYLOMETRIC FEATURES OF THE I. ASIMOV'S TEXTS CORPUS

Unlike the free version of NoSketch Engine, the commercial version has a broader functionality, where not only morphological markings are available for English, but also other tools important for stylistic features determination.

3.1 The Quantitative and Morphological Characteristics of Asimov's Texts Corpus

The *Wordlist* tool gives the possibility to sort the corpus lexicon according to the frequency of tokens, lemmas as well as the parts of speech. The last parameter is important for distinguishing individual authorial features. Quantitative ratio of speech parts according to relative frequency (RF[1]) in the I. Asimov's texts corpus are presented in the figure 2.

Despite the expected high frequency of stop words (primarily function words), the highest frequency was given to nouns and verbs: the first indicate the narrative type of texts, and the second describe human activities, some processes. The most frequent verbs in the corpus are: *be* (Fpm[2]=40827.91), *have* (Fpm=13654.71), *do* (Fpm=9173.77). Here, as expected, the most frequent were the verbs, which often act as auxiliary. It correlates with frequency per million of the most frequent verbs in The British National Corpus (v. 2.0) according to Leech et al. [11]: *be* (Fpm=42277), *have* (Fpm= 13655), *do* (Fpm= 5594). But the most frequent nouns in the "Foundation_Asimov" corpus reflect the Asimov's works vocabulary specifics (figure 3) and correlate with the highest score keywords.

Here it seems interesting to compare the results with the quantitative characteristics of both fiction and scientific style. Among the morpho-marking by Sketch Engine minuses the identification of the negative participle *not* as an adverb, the adverb *yet* – as a conjunction, *that* – as a preposition were noted.

Quantitative ratio of speech parts allows to define *epithetization index* and the *nominalization index*, which are represented by the novels in the table 1. Here and further for calculations of the idiostyle indexes we used the formulas resulted by Buk [5].

The highest is the nominalization index of the first three novels, then, after a 30-year break in the writer's work on the "Foundation" cycle, this index decreases. However, the epithetization index is the largest in the first and last novels of the cycle.

[1]RF – relative frequency, AF – absolute frequency.
[2]Fpm — frequency per million tokens, which is used to compare frequencies between corpora of different sizes.

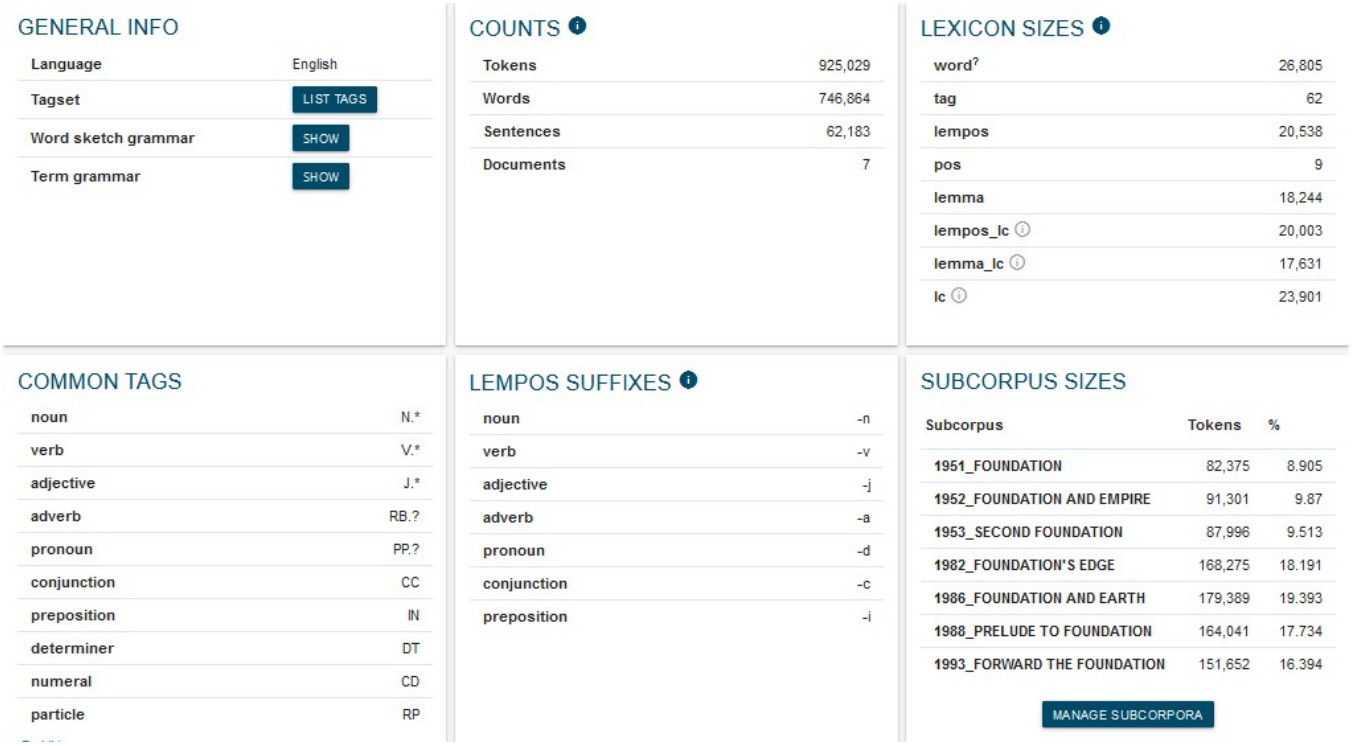

Figure 1: Statistics of the "Foundation" cycle text corpus by I. Asimov.

Table 1: The quantitative and morphological characteristics of Asimov's vocabulary by the novels

Subcorpus	Noun (AF)	Verb (AF)	Adjective (AF)	The epithetization index	The nominalization index
1951_FOUNDATION	15085	13056	4523	3.335176	1.155407
1952_FOUNDATION AND EMPIRE	16632	14205	5452	3.050624	1.170855
1953_SECOND FOUNDATION	15490	13969	5130	3.019493	1.108884
1982_FOUNDATION'S EDGE	27455	28839	8393	3.271178	0.952009
1986_FOUNDATION AND EARTH	28302	30553	9578	2.954897	0.926325
1988_PRELUDE TO FOUNDATION	25302	28810	8343	3.032722	0.878237
1993_FORWARD THE FOUNDATION	24737	26393	7351	3.36512	0.937256

3.2 The Quantitative and Lexical Characteristics of the Vocabulary

Also, the *Wordlist* tool gives the possibility to determine the author's vocabulary and, accordingly, calculate the *lexical diversity index* (L). It was calculated by the Equation 1:

$$L = \frac{V}{N} \tag{1}$$

where:

- L – the lexical diversity index.
- V – vocabulary volume (number of lemmas in a text).
- N – size of a text.

Since we have 7 subcorpora that reflect the period of Asimov's work during 1951–1993, it is possible to track the lexical diversity index dynamics table 2.

Table 2: The lexical diversity index in "Foundation_Asimov" subcorpora

Subcorpus	The lexical diversity index
1951_FOUNDATION	0.072
1952_FOUNDATION AND EMPIRE	0.071
1953_SECOND FOUNDATION	0.068
1982_FOUNDATION'S EDGE	0.040
1986_FOUNDATION AND EARTH	0.037
1988_PRELUDE TO FOUNDATION	0.040
1993_FORWARD THE FOUNDATION	0.043
In the entire corpus	0.019

As we can see, the lexical diversity index decreased in the novels, which were written after a break of almost 30 years at "Foundation" cycle work.

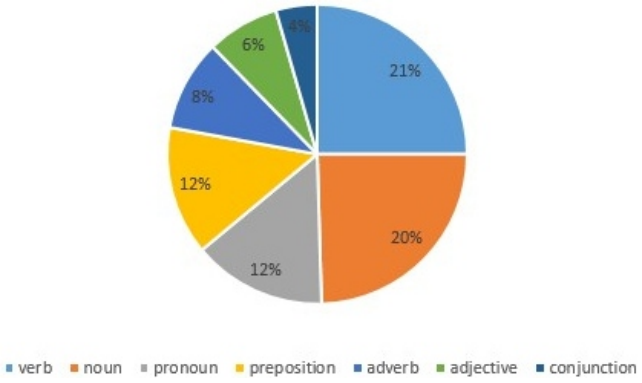

Figure 2: Quantitative ratio of speech parts in the "Foundation_Asimov" corpus.

	Lemma	Frequency ?	Frequency Per Million ?
1	seldon	3,003	3,246.38
2	trevize	2,125	2,297.23
3	time	1,604	1,734.00
4	foundation	1,532	1,656.16
5	man	1,429	1,544.82

Figure 3: The most frequent nouns in the "Foundation_Asimov" corpus.

Since the *Wordlist* tool shows lemmas with AF=1, we can define Hapax Legomena, which is necessary to calculate the exclusiveness index of the vocabulary (Ev) and the exclusiveness index of the text (Et), which are calculated by the equation 2, 3:

$$Ev = \frac{V_1}{V} \tag{2}$$

$$Et = \frac{V_1}{N} \tag{3}$$

where:

- Ev – the exclusiveness index of vocabulary.
- Et – the exclusiveness index of a text.
- V_1 – number of lemmas with AF=1.
- V – vocabulary volume (number of lemmas in a text).
- N – size of a text.

Thus, the exclusiveness index of vocabulary and the exclusiveness index of text also show a decline after a writer's break in the cycle of novels table 3. Here it would be interesting to compare the results with the frequency indicators of fiction style in general and the genre of science fiction in particular, but we will leave it for the future research. The exclusiveness index of vocabulary shows that almost half of the writer's vocabulary is the low-frequency vocabulary.

A significant number of single words prompted a check of their occasionality.

4 OCCASIONALISMS IN THE STYLOMETRIC STUDY

4.1 Automatic Extraction of Occasionalisms From the Corpus

Since the use of neologisms in general and author's neologisms (occasionalisms) in particular is a genre feature for science fiction, the most practical Sketch Engine tool for the occasionalisms search is Keywords, which compares the created I. Asimov's texts corpus (focus corpus) with that of the reference – any from the Sketch Engine database, after what keyword lists according to the specified parameters are displayed. For comparison, both synchronous corpora of Internet texts (*English Web Corpus 2018* or *enTenTen2020*) and diachronic corpora (*Transhistorical Corpus of Written English*) are available on the site, but in order to work properly with some corpora it is necessary to make tokenization or markup. Accordingly, keyword lists may differ after comparison with different reference corpora. Prior to the corpora comparing, there is the need to determine various parameters, which will, obviously, also affect the final results. One can select by lemmas, word forms, phrases, parts of speech, specify frequency ranges or focus on rarer or more general words, as well as the set rules for selecting words with or including words that begin with a capital letter or contain special characters, and so on.

In this study, the English corpus *enTenTen2020* was chosen as the largest one (38 149 437 411 tokens) to be the reference corpus. The obtained results partially coincided with the hypothesis of correlation between frequencies in the two corpora, where on the first positions of the list focused on rare words, there were occasionalisms observed during the pilot manual selection. These were mostly the names of science fiction inventions and phenomena or scientific terms, for example: *psychohistory*, *jet-down*, *electro-clarifier*, *visisonor*, etc., as well as the author's proper names — the names of fictional characters. Additionally, a list of verbose concepts was obtained, most of which are relevant to this study: *neuronic whip*, *psychic probe*, *meteorological vessel*. The list include not only nouns, but also other parts of speech that are also occasional: *trimensional*, *beblistered*, *offensify*.

The display of lists can be enriched by individual columns with numerical values of important characteristics, such as frequency in the focus or reference corpus, relative frequency, typicality score, which shows the difference between relative frequencies in the two corpora and, finally, the uniqueness of the word. Analysis of these characteristics allows one to select the occasionalisms themselves, and not rare words, as evidenced by the zero frequency in the reference corpus or a high keyness score.

Thus, the corpus-based approach using Sketch Engine tool made it possible to 1) automate extraction of occasionalisms, which, however, requires subsequent manual verification; 2) determine occasionalisms with the help of the keyness score. The obtained results were uploaded as a separate xml-file for further processing and sorting in MS Excel. The final occasionalisms' list consist of 1,115 units.

Table 3: Hapax Legomena and the exclusiveness index in subcorpora "Foundation_Asimov"

Subcorpus	Hapax Legomena	The exclusiveness index of vocabulary	The exclusiveness index of text
1951_FOUNDATION	2870	0.481	0.043
1952_FOUNDATION AND EMPIRE	3071	0.473	0.042
1953_SECOND FOUNDATION	2826	0.472	0.039
1982_FOUNDATION'S EDGE	3005	0.437	0.022
1986_FOUNDATION AND EARTH	2875	0.423	0.020
1988_PRELUDE TO FOUNDATION	2759	0.416	0.021
1993_FORWARD THE FOUNDATION	3062	0.463	0.025

4.2 The Quantitative Characteristics of Occasionalisms' Word Formation

The occasionalisms list was manually marked according to the word formation type. The seven creative types with uneven distribution were obtained:

- `One-component` (21.1%): *achaotic, achaotism, actionist, compor, comporellian, deducer, demoiry, demurity, desperance*
- `Compounds` (15.7%): *balancecard, baleyworld, cobwebbery, dethreading, earthlump, earthpeople.*
- `Jukstapositions` (43.8%): *after-blood, a-glimmer, air-car, air-jet, air-machine, air-taxi, air-vessel, all-but-complete, area-increasing, argyropol, arm-end, arm-stopping, auto-propel, auto-sweep, baby-smasher, baby-smashing, back-crawl, back-fringe, bad-expecting, blood-blind, blood-debt, blubber-head.*
- `Two-component` (17.8%): *arrogant metal, arrowed highway, cycloidic pathway, cylindered communication.*
- `Three-component` (1.3%): *emergency alarm box, expanding society set, galactic social pressure.*
- `Four-component` (0.3%): *mathematics of human behavior, course of future history.*
- `Five-component` (0.1%): *numerical probability of total destruction.*

The role of affixation is quite important in the production of new one-component authors units. At the level of word formation, affixation is realized as adding semantically significant prefixes and suffixes to the base of the word. Word formation (formation of compound and juxtaposition) is understood as the formation of a new word on the basis of two words. As a rule, compounds consist of two bases where the first basis specifies the meaning of the second (carrier of a generic sign). Two-component and multi-component occasionalisms are formed according to certain models by means of adjoining as well as the grammatical connection.

4.3 Statistical Estimation of Occasionalisms' Collocations

Word Sketch tool analyzes the search word's grammatical relationship with other parts of speech. This is interesting for studying of collocations and especially for occasional adjectives collocations determination. Firstly, this tool allows to perform the automatic cross-check of the created list of occasionalisms, and secondly, to supplement it. For example, one of the most productive concretizing elements of occasional compounds and jukstapositions is *hyper-* (30 units): *hyperatomic, hyper-engine, hypernuclear, hypernuclear motor, hyper-plan, hyper-raced, hyper-radiational, hyper-region, hyper-relay, hypershift, hypership, hyperspatial, hyperspatiality, hyperspatially, hyperthrust, hypertracer,hypertracking, hyper-video, hyperwarp, hyperwave. Word Sketch* tool gave the possibility to identify the following 21 collocations with the adjective *hyperspatial* with LogDice>5: *hyperspatialtravel* (12.4), *hyperspatial contact* (10.8), *hyperspatial distance* (10.3), *hyperspatial separation* (10.2), *hyperspatial pursuit* (10.2), *hyperspatial communication* (9.9), *hyperspatial tracking* (9.4), *hyperspatial missile* (9.4), *hyperspatial expedition* (9.2), *hyperspatial velocity* (9.2), *hyperspatial engine* (9.1), *hyperspatial flight* (8.9), *hyperspatial technology* (8.9), *hyperspatial image* (8.8), *Hyperspatial Theory* (8.8), *hyperspatial line* (8.7), *hyperspatial Galaxy* (8.1), *hyperspatial life* (7.1), *hyperspatial ship* (6.8), and also differently spelled: *hyperspatial jump* (10.3), *hyperspatial Jump* (9.2). Word sketch visualization can be seen in figure 4.

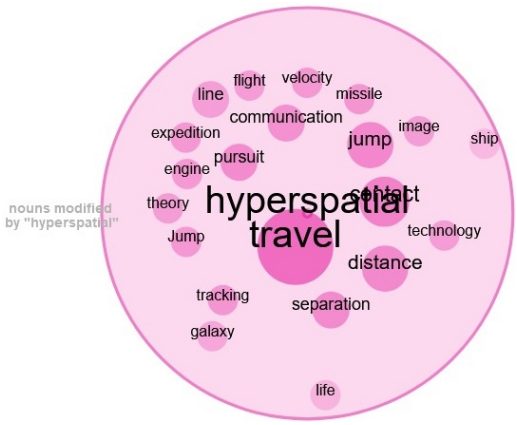

Figure 4: Word sketch for *hyperspatial.*

In all these cases, the AF is low: in the range 1–15, i.e. RF≤0.0000019, but also they have the high LogDice score, which indicates how strong the collocation is. These collocations are important for the conceptual analysis of the text, as well as for lexicographic practice. By the way, Asimov didn't use this adjective in his trilogy "Foundation" — only starting with his "Foundation's Edge" (1982). "Historical Dictionary of Science Fiction" dates back the use of occasionalism *hyperspatial* in the Asimov's work 1954 [15].

For nouns among the different grammatical relations *Word Sketch* tool shows the following construction "Noun is Noun", that has

the practical value for compiling a dictionary of the writer's language. For example, the word *psychohistory* which is one of the most frequent occasionalisms in the corpus "Foundation_Asimov" (AF=483), introduced by I. Asimov, has high LogDice score in the constraction "*psychohistory* is/is not …":

- `science`(12.3): *I quite understand that **psychohistory is a statistical science** and cannot predict the future of a single man with any accuracy* (1951). *Psychohistory **was not yet an experimental science*** (1993).
- `game`(12.0): *And, Seldon, do not tell me that **psychohistory is just a game**, that it does not exist* (1993).
- `study`(12.0): ***Psychohistory is just an abstract study*** (1988).
- `tool`(11.8): *...**psychohistory is a most valuable tool** to be used for the preservation of our culture* (1993).

Word Sketch Difference is a tool that is intended for comparison of phrases in the following variants: lemma (comparison of two lemmas due to the collocates), word form (comparison of two word forms of the same lemma due to the collocates), subcorpus (comparison one lemma usage in different subcorpora through the collocates). For example, let's compare the already mentioned I. Asimov's occasionalism *psychohistory* in novels by years of publication: the first "Foundation" (1951) and the last "Forward the Foundation" (1993). In the last novel of the cycle the frequency of lexical unit's usage by the author has increased 16 times: from AF=16 in "Foundation" to AF=260 in "Forward the Foundation".

Secondly, the last novel shows a compatibility that is missing in the first novel of the cycle:

- prepositional constructions ``psychohistory + in + NOUN'' (*way*, *place*, *time*, *detail*): *Foundation of psychohistorians only-mentalists, mind-touching psychohistorians-who will be able to work on **psychohistory in** a multiminded **way**, advancing it far more quickly than individual thinkers ever could.*
- prepositional constructions ``VERB + with + psychohistory'' (*do*, *interfere*, *satisfy*, *breakdown*, etc.): *Isn't that what you hope **to do with psychohistory**?*
- prepositional constructions ``at + psychohistory'': *… you must keep working **at psychohistory**.*
- constructions where the lexeme is used in the function of an object ``VERB + out + psychohistory (object)'': *...we're going **to work out psychohistory** in time to prevent the Fall of the Empire*e.
- constructions``VERB + without + psychohistory'': *In other words, psychohistory simply tells you what you **would know without psychohistory**.*
- constructions ``psychohistory + as + NOUN'': *It is possible to use **psychohistory as a tool** to manipulate the emotions of the people and achieve short-term effects.*

Totally 20 constructions which are not presented in the first novel were found. All this testifies to the development of the writer's idiostyle.

5 CONCLUSIONS

A corpus-based approach to stylometric study of I. Asimov's idiostyle on the novels of the "Foundation" cycle gave the possibility to make certain conclusions. For the analysis of stylometric features the I. Asimov "Foundation" cycle text corpus was created with the help of Sketch Engine. The "Foundation_Asimov" corpus is structured into 7 subcorpora according to 7 novels of the cycle. This enabled the quantitative parameters of each description, the idiostyle dynamics comparison and identification: the frequency of occasionalisms in the cycle of novels undergoes chronological change, as well as their collocational combinability, which was obviously influenced by extralinguistic factors. A comparison of the lexical diversity index, Hapax Legomena and the exclusiveness index of text in I. Asimov's 7 novels, i.e. for 40 years of his work, showed the decrease of the last after almost 30 years of work on the "Foundation" cycle. The exclusiveness index of vocabulary shows that almost half of the writer's vocabulary is the low-frequency vocabulary. The lower lexical diversity in the last I. Asimov's books may be connected with the fact that during this period he worked more in the genre of popular science literature. However, to test this hypothesis, it is necessary to conduct a stylometric study of the entire work of the author and rank the results over the years.

Analysis of the speech part frequency distribution revealed the highest frequency of nouns and verbs: the former indicate the narrative text type, and the latter the human activity description due to the theme and genre of novels.

This study considers the occasionalisms' extraction method using Sketch Engine tool based on the keyness score. In the created I. Asimov "Foundation" cycle text corpus 1115 occasional units were traced. The obtained base of occasionalisms can be analyzed by word-formation structure, semantics, etc., using the actual linguistic methods. Statistical estimation of occasionalisms' collocations based on LogDice score allowed to perform the automatic cross-check of the created list of occasionalisms, and to supplement it. Also, the found collocations are important for the conceptual analysis of the text, as well as for lexicographic practice.

Thus, the stylometry of I. Asimov "Foundation" cycle gives the possibility to identify certain patterns and features of the writer's idiostyle, which are especially valuable in the system of American literature, in comparison with other science fiction authors, to determine the influence of I. Asimov's work on his successors in the literature, as well as to identify the genre patterns.

REFERENCES

[1] Lexical Computing CZ s.r.o 2021. *SketchEngine.* Lexical Computing CZ s.r.o. https://www.sketchengine.eu

[2] L. David Allen. 1977. *CliffsNotes on Asimov's Foundation Trilogy & Other Works.* Houghton Mifflin Harcourt.

[3] Eric Norris Bellefontaine. 1973. *Science Fiction and the Work of Isaac Asimov.* Ph. D. Dissertation. University of Maine.

[4] Douglas Biber. 2009. Corpus-Based and Corpus-driven Analyses of Language Variation and Use. In *The Oxford Handbook of Linguistic Analysis* (1 ed.), Bernd Heine and Heiko Narrog (Eds.). Oxford University Press, 368–394. https://doi.org/10.1093/oxfordhb/9780199544004.013.0008

[5] S. N. Buk. 2021. *Corpus lexicographic and linguistic statistical dimensions of ivan Franko's long prose fiction: vocabulary and text.* Thesis for the Degree of Doctor of Sciences in Philology, specialty 10.02.01 – Ukrainian Language. Taras Shevchenko National University of Kyiv. http://scc.univ.kiev.ua/upload/iblock/8e7/wlg53oro9zm7mobzltnv8n6uzntsdlah/dis_Buk%20S..pdf

[6] Maciej Eder, Maciej Tomasz Piasecki, and Tomasz Walkowiak. 2017. *An open stylometric system based on multilevel text analysis.* https://doi.org/10.11649/cs.1430

[7] Jocelyn Hoppa. 2007. Isaac Asimov: Science Fiction Trailblazer. In *Authors Teens Love.* Authors Teens Love Series, Vol. 6. Enslow Publishers.

[8] Alicja Kacprzak and Weronika Woźniak. 2020. Les néologismes récents en -ing en polonaise [Recent Neologisms in -ing in the Polish Language]. *Acta Universitatis Carolinae Philologica* 2020, 4 (2020), 133–151 pages. https://doi.org/10.14712/24646830.2021.7

[9] Adam Kilgarriff, Jan Busta, and Pavel Rychlý. 2015. *Diacran: a framework for diachronic analysis.* https://www.sketchengine.eu/wp-content/uploads/Diacran_CL2015.pdf

[10] Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P. G. Demidov. 2019. A Survey on Stylometric Text Features. In *2019 25th Conference of Open Innovations Association (FRUCT)*. 184–195. https://doi.org/10.23919/FRUCT48121.2019.8981504

[11] Geoffrey Leech, Paul Rayson, and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* Routledge, New York, NY.

[12] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.* 50, 6, Article 86 (nov 2017), 36 pages. https://doi.org/10.1145/3132039

[13] Alec Nevala-Lee. 2019. *Astounding: John W. Campbell, Isaac Asimov, Robert A. Heinlein, L. Ron Hubbard and the Golden Age of Science Fiction.* Dey Street Books.

[14] Cécile Poix. 2018. Neology in children's literature: A typology of occasionalisms. *Lexis* 12 (2018). https://doi.org/10.4000/lexis.2111

[15] Jesse Sheidlower. 2021. *Historical Dictionary of Science Fiction.* Retrieved February 22, 2021 from https://sfdictionary.com

[16] Jan Švec and Jan Rygl. 2016. Building Corpora for Stylometric Research. In *Text, Speech, and Dialogue*, Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (Eds.). Springer International Publishing, Cham, 20–27. https://doi.org/10.1007/978-3-319-45510-5_3

[17] Elena Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. Studies in Corpus Linguistics, Vol. 6. John Benjamins Publishing Company, Amsterdam / Philadelphia. https://doi.org/10.1075/scl.6

[18] Johanna Vainio. 2008. *Ghost in the machine. Androids in search of humanity in Isaacs Asimovs "The Bicentennial Man" and Philip K. Dicks Do Androids Dream of Electric Sheep?* Pro Gradu Thesis. University of Tampere. https://moam.info/ghost-in-the-machine-androids-in-search-of-humanity-in-isaac-_5a09406c1723dd0d9e091f25.html

[19] Gary Westfahl. 1993. The Words That Could Happen: Science Fiction Neologisms and the Creation of Future Worlds. *Extrapolation* 34, 4 (1993), 290–304.