ISSN (en línea): 2707-1642
ISSN (impreso): 2707-1634

Articulo original

Architecture and Research Corpus Indexation of Texts by Yuri Shevelyov

Indexación de textos del Corpus de Arquitectura e Investigación por Yuri Shevelyov

Illia Danyliuk
National Commission for State Language Standards, Ukraine
https://orcid.org/0000-0001-8681-9321

e-mail: i.g.danyluk@gmail.com

Anatolii Zahnitko
Vasyl Stus Donetsk National University, Ukraine
https://orcid.org/0000-0001-7398-6091

e-mail: a.zagnitko@donnu.edu.ua

Hanna Sytar
Vasyl Stus Donetsk National University, Ukraine
https://orcid.org/0000-0001-8806-8322

e-mail: h.v.sytar@donnu.edu.ua

Recibido: 15/2/2023
Aprobado: 25/5/2023

ABSTRACT
The article is devoted to the problem of creating a corpus of texts by Yuri Shevelyov. The main structural components of the text's corpus have been determined, the functions have been singled out, the features of its filling have been characterized, extralinguistic and linguistic marking has been explained.

Keywords: concordance; a corpus of texts; linguo-person; linguo-personology; corpus marking; frequency analysis; Yuri Shevelyov.

RESUMEN
El artículo está dedicado al problema de la creación de un corpus de textos de Yuri Shevelyov. Se han determinado los principales componentes estructurales del corpus de textos, se han señalado las funciones, se han caracterizado las características de su llenado, se ha explicado el marcado extralingüístico y lingüístico.

Palabras clave: concordancia; corpus de textos; linguo-persona; linguo-personología; marcado del corpus; análisis de frecuencia; Yuri Shevelyov.

DOI: https://doi.org/10.47133/NEMITYRA2023d11-2A11

INTRODUCTION

Linguo-personology is based on philosophical understanding of personalism (B. P. Bowne, R. H. Lotze, J. H. Howison, W. James, J. Royce, M. W. Calkins, W. E. Hocking, G. W. Allport, E. Mounierand others), one of the manifestations of which is the thesis of the spiritual renewal of the individual, the longevity of self-improvement. In modern linguistics, the issues of a linguistic portrait of the personality, establishment of structural and functional components of the personality are actualized. The stated and a number of other questions appear to be essential for modern applied linguistics, partly for psycholinguistics, linguo-psychiatrics, neurolinguistic programming, suggestive linguistics, and others. Linguo-personology as a science of linguistic personality studies the individual (monolingo-person) or collective (polylinguo-person) – linguo-person in functional-communicative, verbal-semantic, linguo-psychological, speech-behavioral, discursive-situational, motivational-suggestive, and other dimensions.

The questions of the present interest are the formation of speech needs of the person, determination of their duration, diagnosing changes in language and speech needs in the living space of linguo-persons. The authors pay close attention to internal differentiation of status and functional load of speech and speech transitions, coding, and recoding, formation of sets of linguo-individualizations, ethical, emotional and situational and other factors. Especially important are the studies of prominent linguists, to whom Yuri Shevelyov – a prominent linguist, literary historian, critic belongs. His speech-discursive practices correlate different language elements, unequal discursive practices. He intensifies territorial, social, age, aesthetic, emotional, and other motivators of individual changes. The scientific task of establishing the linguistic-corpus structure of discursive practices of Yuri Shevelyov is highly motivating. The statement correlates with the practical task – the introduction into active use not only of the scientific and creative heritage of the outstanding scientist but also the use of language models and constructions of linguo-individualizations in modern language-codified practice developed by Yuri Shevelyov.

THEORETICAL BASIS

For the first time the concept of lingvo-person was used in the early twentieth century by M. Trubetskoy. It was based on the idea of personalism (E. Mounierand others). In the 2nd half of the XX century in connection with the intensification of the study of linguistic personality, issues of active, associative grammar (J. Karaulov, 1987), the concept of linguistic person has acquired a different dimension and status.

Activation of linguo-personological studies was observed in the late XX – early XXI century. Theoretical and practical developments of communicative strategies and tactics, discourse, text linguistics, attempts to study speech-behavioral stereotypes, expansion of studies in applied linguistics, outlining the theoretical foundations of objective grammar, etc. became significant. Attempts at the linguistic portrayal of the individual and the creation of authorial corpora of texts are no less significant.

Yuri Shevelyov's linguistic personality was considered by R. Trifonov. He clarifies the linguistic and cultural components of the scientist’s individual speech on the basis of essays and letters. M. Moser reveals the main stages of the dynamics of Yu. Shevelyov’s creative personality on the way from Sovietism (Trifonov, 2009; Trifonov, 2015; Moser 2014). The studies of A. Danylenko, M. Kotsyubynska, L. Masenko, L. Tarnovetskaya, and others are no less relevant. Today, most of Yu. Shevelyov’s works are posted on the Internet (http://movahistory.org.ua/wiki/%D0%A8%D0%B5%D0%B2%D0%B5%D0%BB%D1%8C%D0%BE % D0% B2_% D0% AE ._% D0% 9F% D1% 80% D0% B0% D1% 86% D1% 96_% D0% BD% D0% B0% D1% 8F% D0% B2% D0% BD % D1% 96_% D0% B2_% D1% 96% D0% BD% D1% 82% D0% B5% D1% 80% D0% BD% D0% B5% D1% 82% D1% 96

(09.08.2021)) , which in general does not solve the problem of a holistic vision of the creative linguistic persona of Yuri Shevelyov. It is possible on the condition of integral corpus structuring of his heritage, the development of linguo-computer technologies of free navigation within hypertext. Therefore, the purpose of the study is an experimental-research model of text corpus of the linguistic persona of Yuri Shevelyov with the definition of the main components of the corpus, the laws of free linguistic-textual navigation.

METHODOLOGICAL BASIS OF THE STUDY

The integral corpus of the linguistic persona of Yuri Shevelyov, as well as the consideration of the peculiarities of the linguistic portrait of the scientist, requires the use of appropriate methods. Among them are the following methods: the method of discursive analysis using the methodology of corpus content analysis to establish all possible fixations of the studied word contexts, as well as identifying non-random combinations of words-collocations, comparing the latter with regular and quantitatively dominant ones. The use of corpus content analysis can serve as a basis for creating a sketch of a word with intra-corpus differentiation of usual and occasional-authorial (individualized) word combinations. It is also essential for determining their load within the communicative registers, the manifestation of communicative intentions.

INDIVIDUAL-AUTHOR EXPERIMENTAL RESEARCH CORPORA: FUNCTIONAL LOAD IN THE STUDY OF A LINGUO-PERSON

The complete linguistic portrayal is possible provided that the whole set of personal-generated texts is covered with the diagnosis of territorial, social, generational, and other motivations for changes in discursive practices. The set of emphasized questions appears as one of the sides of modern linguo-personology. Its directions are actively studied at the Department of General and Applied Linguistics and Slavic Philology of Vasyl’Stus Donetsk National University (A. Zahnitko, I. Danyliuk, J. Krasnobayeva-Chorna, H. Sytaretc.). Elaboration of levels, categories, and aspects of linguo-personology will make it possible to create an objective and subjective linguistic-societal grammar, at the center of which is a mono- and/or polylinguo-person in an indefinite number of his/her discursive practices. Diagnosing changes in 1) the structure of language personality, 2) in levels of its implementation (verbal-semantic, linguocognitive, motivational (according to Yu. Karaulov), functional (V. Konetska), situational-psychological), 3) the ratio of components (formation of communication skills, the need for communication and proper competence, the formation of language consciousness and language and/or speech behavior) based on: a) the natural need for communication in the native language; b) the communicative need to communicate in a foreign language; c) discursive-motivated practice (pedagogical, medical, etc.) of communication in a non-native language, possible under the condition of creating an integral body of texts of the linguistic personality.

Individual author’s experimental corpora make it possible to establish the levels and aspects of a linguo-person with reliance on his/her personal texts. Through the lens of the total size of the text corpus, it’s possible to see the identification of certain thematic and key elements of disclosing the duration of linguo-persons, the formation of gaps motivated by territorial, social and other factors. The appropriate linguistic landscape of linguo-persons, reconstruction of creative potential and mechanisms for implementation become visible. It is noteworthy that the results of learning, dissemination, and self-realization of linguo-persons, etc. can be reconstructed. The text corpus will make it possible to trace the patterns of linguistic and social conditioning of the linguo-person and the status of linguo-individualizations as individual self-expression: attraction, desire, inclination, worldview, beliefs. A Linguo-person manifests through relationships with others, where relationships with other individuals are the experience and awareness, and beliefs outline the direction of the individual, his/her language values. Subsequently, knowledge, skills, abilities, and habits to use the language, to differentiate it in different corporate groups, to be able to use different language codes depending on situations, become significant. This confirms the language space of Yuri Shevelyov in different periods of his life – childhood, student years, Kharkiv period, years of emigration in Munich, teaching at Lund University, and later at Harvard and Columbia Universities.

Consistent use of corpus content analysis provides the definition of the emotional background of discursive practices, the expression of feelings and sensations, memory levels, etc. In its entirety, the author's experimental research corpus of texts provides an understanding of the properties of the temperament of the linguo-person. It tracks the change of its speech-behavioral realizations, establishing the typological properties of a linguo-person. In general, the discourse of a linguo-person is his/her living space with different functionally loaded sets of self-realizations in language and in language.

INDIVIDUAL-AUTHOR'S EXPERIMENTAL RESEARCH CORPUS OF A LINGUO-PERSON YURI SHEVELYOV: STRUCTURE, TECHNOLOGIES

Yuri Shevelyov is an outstanding personality who is realized in numerous texts, various discursive practices, unequal linguistic-landscape spaces.

To study the peculiarities of Yuri Shevelyov's speech, it was decided to compose a research corpus of texts. This corpus of texts was designed and implemented by A. Zahnitko, I. Danyliuk, and H. Sytar, teachers of the Department of General and Applied Linguistics and Slavic Philology of Vasyl'Stus Donetsk National University. Within the framework of the internship, 2nd-year students of the specialty "Applied Linguistics" took an active part in the preparation of texts for the corpus.

Yuri Shevelyov's text corpus was created with the help of the free corpus manager NoSketch Engine (https://www.sketchengine.eu/), developed at the University of Masaryk (Brno, Czech Republic) (Rychlý, Smrž, 2004). This corpus, along with others, is available on the server of the Department of General and Applied Linguistics and Slavic Philology of Vasyl'StusDonNU at corpora.donnu.edu.ua.

According to the authors' plan, the corpus contains all the texts of Yuri Shevelyov published by today, in particular:

Shevelyov, Yuri. From the history of the unfinished war. (Compilers Oksana Zabuzhko, LarysaMasenko). Kyiv: Ed. Kyiv-Mohyla Academy, 2009. 471 p. ISBN 978-966-518-519-2; Shevelyov, Yuri. From the history of the unfinished war. (Compilers Oksana Zabuzhko, LarysaMasenko). Kyiv: Ed. Kyiv-Mohyla Academy, 2009. 471 p. ISBN 978-966-518-519-2; Sherekh, Yuri. Outside of books and from books. Kyiv: Chas Publishing House, 1998. 456 p. ISBN 966-95238-3-4; Sherekh, Yuri. Another rug. "Library of Prologue and Modernity Part 130", 1978. 393 p.; Sherekh, Yuri. Not for children. New York: PROLOGUE Publishing House, 1964. 416 p.; Shevelyov, Yuri. Selected works: in 2 books. Book II. Literary Studies. Kyiv: Kyiv-Mohyla Academy Publishing House, 2009. 1151 p. ISBN 978-966-518-496-6; Shevelyov, Yuri. Selected works: in 2 books. Book I. Linguistics. Kyiv: Kyiv-Mohyla Academy Publishing House, 2009. 583 p. ISBN 978-966-518-494-2; Sherekh, Yuri. The third watchman. Baltimore-Toronto: Smoloskyp, 1991. 454 p.; Shevelyov, Yuri. "I, me, me… (and around)". Memoirs. In two volumes. Berezil Magazine Publishing House, MP Kots Publishing House. Kharkiv– New York, 2001; Zabuzhko, Oksana, Shevelyov, Yuri. Selected correspondence against the background of the era: 1992-2002: with appendices, works, comments, reasons for biographies and other documents. Kyiv: VysokaPolitsa, VD Fakt, 2011.

The Ukrainian-language part of the corpus currently covers 104 documents in the Ukrainian language, including 1346424 tokens.

Extralinguistic and linguistic markup is used in the created case. **Extralinguistic corpus markup** combines:

a) metatext data. They include:

• *area* with possible meanings of literary studies, linguistics, general works;

• *author* - in this building only Yuri Shevelyov, created for the possibility of combining with other buildings;

• *genre*: essay, article, monograph, interview, preface, speech, report, memoirs, article, speech, article, report, letter, introductory word;

• *name* (title of the work);

• *source*;

• *style*: journalistic, scientific and epistolary;

• *type*: in the original language, translation from English, translation from German, translation from French;

• *year*.

b) *structural markup*. The corpus contains data:

• about text borders in <doc> - </doc>tags;

• about paragraph boundaries in <p> - </p>tags;

• sentence boundaries in <s> - </s>tags;

• special tag <g /> indicates punctuation marks that are not separated by a space from the previous token.

Linguistic marking of the analyzed corpus of Yuri Shevelyov's texts today is the result of automatic morphological analysis and lemmatization performed with the help of the author's tools. The structure of the standard tag for each token is as follows: in the first position is the mark of the grammatical class of the word, then the mark of subclasses, all marks – single characters in Latin or numbers, each subclass is assigned a position that does not change for different classes.

For example, for a *conference* word, the tag looks like Izzooin1m (Fig. 1).

Scheme 1. Deciphering the tag for the word form *"conference"*

Noun		*feminine*		*Ablative case*		*inanimate*		*Soft group*
I	Z	Z	o	O	i	n	1	m
	General name		*Singular*		*Noun type of declension*		*First declension*	

The system of tags is described in detail in the study (Zahnitko, Danyliuk, 2013). In particular, the classification of grammar classes has the following form (Table 1):

Table 1. Classes of words in the corpus of texts by Yuri Shevelyov

Noun	I.*
Verb	D.*
Participle	Dk.*
Adverb	Ds.*
Adjective	K.*
Pronoun	Z.*
Adverb	S.*
Preposition	J.*
Conjunction	P.*
Numeral	C.*
Particle	T.*
Interjection	W.*
Abbreviation	A.*
The rest	R.*

The gender category is described by the following tags (Table 2):

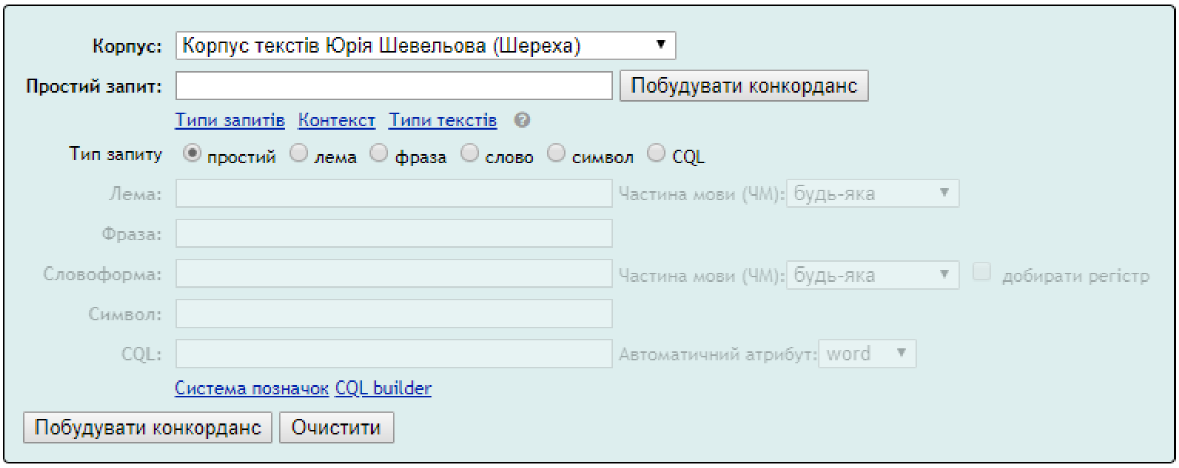
Table 2. Grammar gender in the corpus of texts of Yuri Shevelyov

masculine	..c.*
feminine	..z.*
neutral	..s.*

The quantitative parameters of the created corpus are as follows: 1,346,424 tokens were collected in 104 documents, of which 1,037,949 were words in 66,039 sentences. The general lexicon includes 138,187 different word forms and punctuation marks described by 157 original tags, and 32,172 lemmas.

Typical functions are available in the corpus manager, such as *building a concordance* based on a simple search, searching in lemmas, searching for a phrase, word form, symbol, or a specific template built using a regular expression.

Figure 1. Interface for selecting the type of query in the case



Корпус: Корпус текстів Юрія Шевельова (Шереха)

Простий запит: Побудувати конкорданс

Типи запитів Контекст Типи текстів

Тип запиту ● простий ○ лема ○ фраза ○ слово ○ символ ○ CQL

Лема: Частина мови (ЧМ): будь-яка

Фраза:

Словоформа: Частина мови (ЧМ): будь-яка □ добирати регістр

Символ:

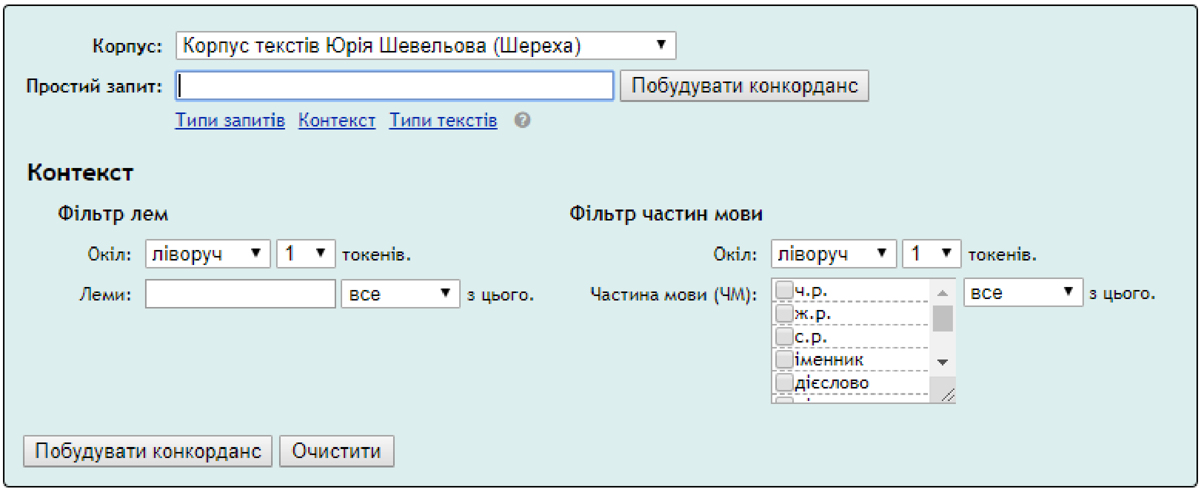
CQL: Автоматичний атрибут: word

Система позначок CQL builder

Побудувати конкорданс Очистити

The query can be based on an additional search in the context of filtering the desired lemmas or word forms at a distance of up to 15 tokens to the right or left of the main word.

Figure 2. Interface for selecting the context of the query in the case



Корпус: Корпус текстів Юрія Шевельова (Шереха)

Простий запит: Побудувати конкорданс

Типи запитів Контекст Типи текстів

Контекст

Фільтр лем

Окіл: ліворуч 1 токенів.

Леми: все з цього.

Фільтр частин мови

Окіл: ліворуч 1 токенів.

Частина мови (ЧМ): □ ч.р. □ ж.р. □ с.р. □ іменник □ дієслово все з цього.

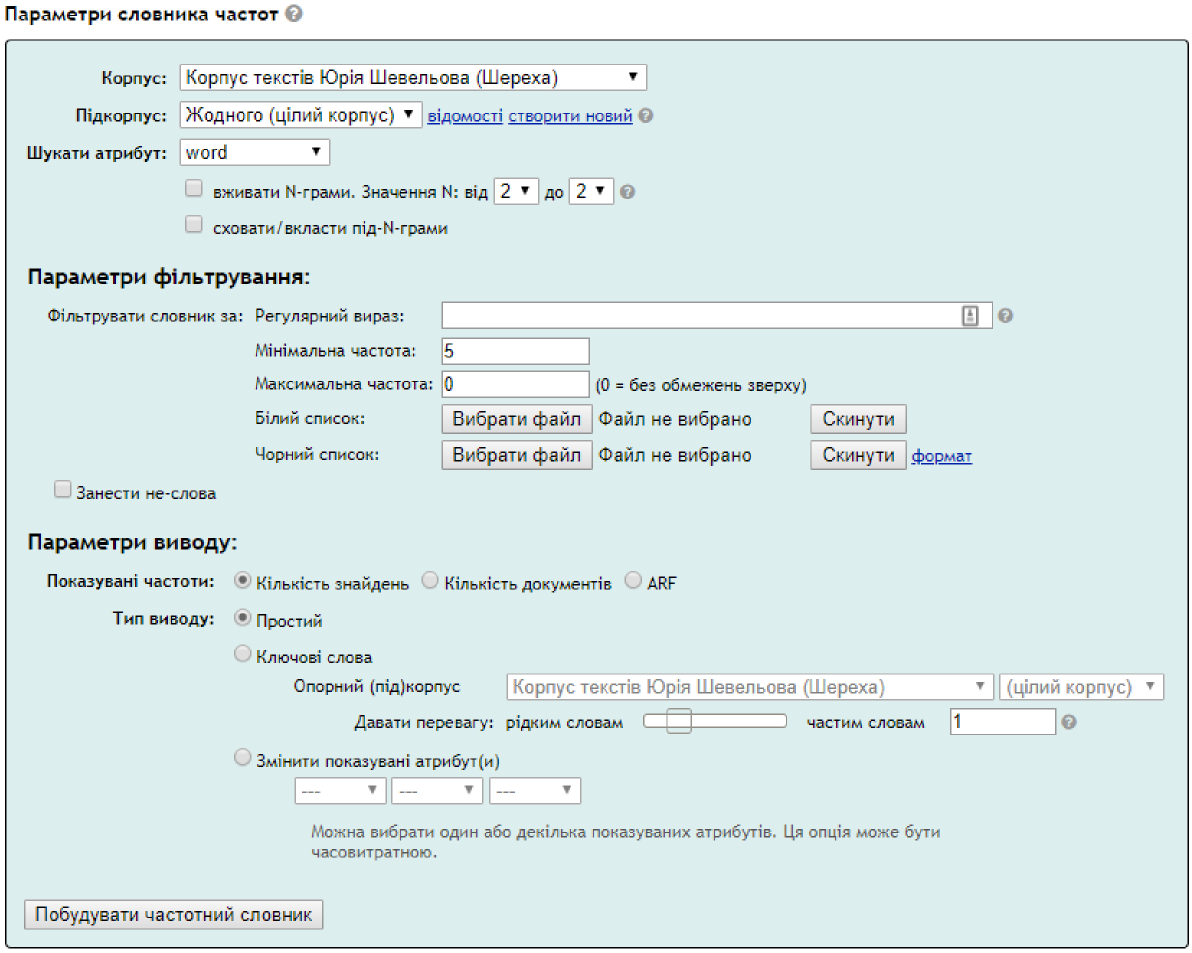
Побудувати конкорданс Очистити

Finally, the search can be limited to the different types of texts provided by extralinguistic markup.

In the constructed concordance different types of data sorting and filtering, frequency analysis of morphological symbols or word forms for the lemma, etc. are possible.

Another typical function of the corpus manager is frequency analysis with the ability to select the minimum or maximum frequency, part-of-speech filters, N-grams.

Figure 3. Interface for building a frequency dictionary in the case



Параметри словника частот

Корпус: Корпус текстів Юрія Шевельова (Шереха) ▼

Підкорпус: Жодного (цілий корпус) ▼ відомості створити новий

Шукати атрибут: word ▼

□ вживати N-грами. Значення N: від 2 ▼ до 2 ▼

□ сховати/вкласти під-N-грами

Параметри фільтрування:

Фільтрувати словник за: Регулярний вираз:

Мінімальна частота: 5

Максимальна частота: 0 (0 = без обмежень зверху)

Білий список: Вибрати файл Файл не вибрано Скинути

Чорний список: Вибрати файл Файл не вибрано Скинути формат

□ Занести не-слова

Параметри виводу:

Показувані частоти: ● Кількість знайдень ○ Кількість документів ○ ARF

Тип виводу: ● Простий

○ Ключові слова

Опорний (під)корпус Корпус текстів Юрія Шевельова (Шереха) ▼ (цілий корпус) ▼

Давати перевагу: рідким словам частим словам 1

○ Змінити показувані атрибут(и)

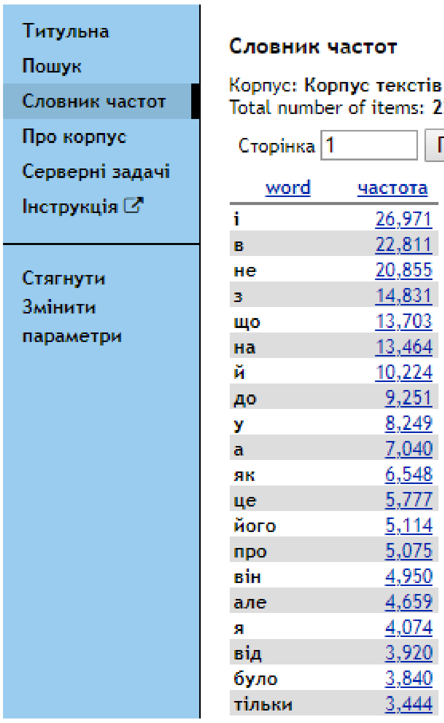
--- ▼ --- ▼ --- ▼

Можна вибрати один або декілька показуваних атрибутів. Ця опція може бути часовитратною.

Побудувати частотний словник

The frequency analysis tool also allows you to select all lemmas and all word forms.

Figure 4. Fragment of the frequency of word forms from the corpus by Yuri Shevelyov



Словник частот

Корпус: Корпус текстів Юрія Шевельова (Шереха)
Total number of items: 22,218

Сторінка 1 Перейти Наступна >

word	частота
і	26,971
в	22,811
не	20,855
з	14,831
що	13,703
на	13,464
й	10,224
до	9,251
у	8,249
а	7,040
як	6,548
це	5,777
його	5,114
про	5,075
він	4,950
але	4,659
я	4,074
від	3,920
було	3,840
тільки	3,444

CONCLUSIONS

Thus, the created by Yuri Shevelyov corpus is research, full-text and dynamic, and it has extralinguistic and linguistic markup. Among the important functions are a) building a concordance based on a simple search, search in lemmas, search for a phrase, word form, symbol, or pattern created using a regular expression, b) frequency analysis for word forms, lemmas, and tags.

Now the team of authors is working to eliminate errors of automatic morphological analysis and filling the English-language part of the corpus. The prospect of the study is to create a comprehensive body of texts by Yuri Shevelyov, which will cover documents in Ukrainian and English and will be a reliable basis for studying the features of the speech of this prominent scholar.

REFERENCES

Zahnitko, A., Danyliuk, I. (2013). Corpus of texts of grammatical service. Applied Linguistics and Linguistic Technologies: Mega Ling 2012 (pp. 102-112). Kyiv: UMIF.

Karaulov, Y. (1987). Russian language and linguistic personality. Moscow: URSS.

Moser, M. (2014). The Language Behavior of Galician Russophiles during the Interwar Period. Russian Linguistics 38, 2014, 315–339

Moser, M. (2014a). Yuri Shevelyov on the road of reviva

http://www.historians.in.ua/index.php/en/ukrayinska-mova/1066-mikhael-mozer-yurii-shevelov-na-dorozi-vidradianshchennia

Трифонов, Р.А. (2009). Metalanguage fragments of Yuri Shevelyov's memoirs are representatives of the individual picture of the linguist's world. Bulletin of Kharkiv National University named after V.N. Karazina, 843, 55,19-26.

Trifonov, R. (2015). Linguo-cultural components of Yuri Shevelyov's individual speech (based on essays and letters)

http://www.historians.in.ua/index.php/en/ukrayinska-mova/1697-roman-tryfonov-linhvokulturni-skladnyky-indyvidualnoho-movlennya-yuriya-shevelova-na-materiali-ese-ta-lystiv

Rychlý, P., Smrž, P. (2004). Manatee, Bonito and Word Sketches for Czech. Proceedings of the Second International Conference on Corpus Linguisitcs (pp. 124-132). Saint-Petersburg: Saint-Petersburg State University Press.