Parameterization of the Ukrainian Text Corpus Based on Parsing Results

Nataliia Darchuk[1], Victor Sorokin[1]

[1] *Taras Shevchenko National University, 14, Taras Shevchenko Boulevard, Kyiv, 01601, Ukraine*

Abstract

This paper describes automatic parameterization of the syntactic structure of the sentence represented as a dependency tree. The dependency trees are created by parsing sentences from the Ukrainian Text Corpus. Based on automatically created dependency trees and parameterization of each sentence in these texts, we looked at the features of the author's writing style in the Ukrainian poetic discourse. The developed technique and its software implementation make it possible to systemize graphic structures and discover patterns in the syntactical structure of the sentences, as well as define the author's style and identify the features of the discourse. Lina Kostenko's individual style requires detailed, balanced, in-depth studies. The corpus of Lina Kostenko's texts we created provides a lot of information about the parameters of the author's language; it is convenient to use in various studies, including text creation. This underlines the scientific novelty, theoretical and practical value of our work.

Keywords
Parsing, dependency tree, parameter, coordinate phrases, subordinate phrases, predicative phrases, syntactic structure of the sentence.

1. Introduction

In 1981, I. P. Sevbo published a well-known work on the systematization of linguistic graphics "Graphic representation of syntactic structures and stylistic diagnostics" [1], which describes the theoretical principles, practical significance, and potential of formal syntax represented in form of dependency graphs. In this work, the graph is described as a parameter of an author's style. The monograph contains many interesting ideas and proposals that cannot be implemented without automating the process of text analysis. Therefore, our efforts were aimed at creating a parsing system for the Ukrainian text and parameterizing linguistic information based on parsing results.

Parsing provides an opportunity to catalog syntactic units, creating a foundation for solving many theoretical linguistic problems. Thus, the theoretical necessity to study the co-occurrence of lexical units and syntactic sentence models as linguistic graphs was an epistemological driver to develop Ukrainian language text parsing. The following practical challenges became ontological drivers: linguistic research automation [2], corpus data parameterization to discover features of the individual style of the author [3], automatic identification of phrases and criteria for dividing phrases into syntagms, automatic text summarization, annotation, and keyword extraction [4] based on conjunction criteria, automatic text editing, machine translation, etc. [5] This shows an apparent need in creating parsing systems for the Ukrainian language.

In this paper, we aim to describe the principles of syntactic parsing of Ukrainian language texts based on dependency graphs, as well as showcase text parameterization based on Lina Kostenko's poetry as shown in the Ukrainian Text Corpus on the mova.info portal.

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland.
EMAIL: NataliaDarchuk@gmail.com (N. Darchuk); victor.sorokin@gmail.com (V. Sorokin)
ORCID: 0000-0001-8932-9301 (N.Darchuk) 0000-0002-3637-0535 (V.Sorokin)

2. Related Works

Automation of linguistic research is associated with creating automatic text processing systems. There are four types of such systems: 1) systems without an automatic syntax parser; 2) systems with morphology and syntax parsers; 3) systems in which the syntax parser is a separate unit; 4) systems in which syntax and semantics parser are combined into one unit [2].

The "TREETON" system, which provides morphosyntactical text analysis, represents the second type of mentioned systems. In "TREETON", the dependency and constituency formalisms are combined [6].

The "PSYCHEA" system for automatic indexing of Russian language texts combines the features of the second and the fourth system types [7]. In this system, syntactic parsing was used to disambiguate homonyms. As the developers of "PSYCHEA" aimed to process language in a formal and meaningful way, it led to the combination of syntactic and semantic analysis in the system.

The syntax parser of "ETAP-2" system is an example of the fourth system type [8; 9; 10]. It was used for syntactic and semantic annotation of the Russian language corpus. Each analyzed text is provided in a separate .xml file that contains the morphological information about all words in the text (lemma and a set of its grammatical features) and a representation of the syntactic structure of the sentence as a dependency tree. All tree branches are marked, showing the syntactic relations; there are around 80 types of relations, half of which are described in the traditional "Meaning-Text" theory by I. Melchuk, A. Zholkovsky, and Y. Apresian. "ETAP-2" parsers use a morphological dictionary containing 120 000 lexemes, a combinatorial dictionary with approximately 90 000 lexical items, and a syntagm grammar that includes hundreds of rules. The syntactic and semantic annotation is created automatically, but a manual process is in place to check its results.

Among the projects devoted to automatic parsing of the Ukrainian text, the project https://mova.institute/ should be mentioned. In it the sentence is presented in the form of direct components. But a significant difference between approaches to automatic parsing by direct components and dependency tree is 1) lack of hierarchical representation of its structure, especially in complex sentences with congruence, subordination, incoherence and inversions, 2) verbocentric approach, when the vertex is the verb. This makes it possible to take into account the distance from the vertex of the graph and the dependent groups of the subject and predicate and in general to parameterize the graph according to the method proposed in this article.

The Ukrainian text parser which we propose belongs to the third system type with a separate syntactic analysis module. This is due to our goal to fully describe the syntax of the sentence, representing the linear morphological sentence structure in a two-dimensional tree-like form. We haven't used the semantic information in the model for the following reasons: 1) the linear order and the observance of the tree-like principle, typical for syntax, are not necessary for semantics; 2) the nature of syntactic rules is such that global semantic problems have to be broken down into even smaller ones, which can be analyzed at a lower, syntactic level; 3) it is important for semantic information to go beyond the sentence level in order to explore the sequence of semantic representation of the sentence as a single representation of the text [11].

In general, the described Ukrainian language parser is a set of operations performed on the input text to establish syntactic connections and syntactic-semantic relations between text units. Sequences of morphological information obtained from automatic morphological analysis are provided as input in this case. The parser outputs relevant information for each word provided as an input.

3. Methods

In corpus studies, there exist several types of syntax models used for automatic text processing: constituency grammar (chain grammar) [5]; dependency grammar [12; 13; 14; 15]; syntactic groups theory by O. V. Gladkyi [16]. One has to combine them all when building an automatic text processing system, as each of them has advantages and disadvantages [2]. For example, an important drawback of constituency grammars is the fact that the linear word order and phrase structure of the sentence must correspond to each other. This does not take into account languages with free word order, in which the phrases may be detached and separated. In constituency grammars, the sentence is represented as a horizontal sequence of phrases and constituents that can be reduced to two main groups: subject group and predicate group. Therefore, when analyzing a complex sentence using this method, different interpretations of syntactic constructions are possible. In a compound sentence, each of the identified groups is analyzed separately. There are different opinions when it comes to the analysis of a complex sentence: either the subject group and the predicate group get distinguished first with the subordinate clause as part of either of them, or the subordinate clause is considered a separate constituent as opposed to the main clause. The same principle also applies to introductory clauses and phrases. This leads to situations where parts of the main clause, though acting as one constituent, are separated by a subordinate clause, and automation in such cases proves to be complicated. In some other cases, the constituency grammars don't show the differences in the structure of separate sentences because formally similar structures can be impossible to distinguish solely based on grammar rules. Besides, there are great difficulties in both analyzing the elliptical structures and trying to distinguish interrogative sentences from affirmative ones. At the initial stages of syntactic analysis, the constituency grammar is used because its rules explain derivation well, and constituents (syntactic groups) are built according to these rules. Dependency grammar, on the other hand, illustrates the hierarchy of the units, which form the foundation to further calculate the information weight of the semantic level units (semantic nodes). That is important for parameterization. As for the A. Gladkyi syntactic group theory, it allows to include the whole dependency groups in the sentence structure. This enables the processing of discontinuous constituents.

Two main approaches are usually used to create syntactic parsers: one is rule-based, the other employs machine learning [17; 18]. The rule-based approach is inherently linguistic, as it represents the linguistic information as formal rules embedded in the code of the program or as a formal language created explicitly for the task. The rules are usually created by linguists. Within the machine learning approach, on the other hand, not the rules are the source of linguistic information, but the selected texts that represent the chosen domain. The training utilizes the general laws inherent in the natural language texts and is based on sample data. Therefore, declarative knowledge (rules) is combined with procedural knowledge (machine learning).

Both methods have advantages and disadvantages. Creating rules is a time-consuming but deeply linguistic process that takes into account even partial complex cases, many of which differ a lot across texts of various styles. The rules are declarative, understandable, and easy to modify depending on desired results. Machine learning does not require manual labor to compile rules, which reduces the time to develop systems. However, the way classifiers function is not easily interpreted linguistically. Also, supervised machine learning requires annotated text corpora, creating which usually involves significant manual labor. The more annotated text the corpus contains, the better the results of the parsing can be [17].

Parsing strategies can be different, namely:

1) sequential analysis, which involves creating a dictionary of reference phrases (syntagms) represented with grammatical word classes;

2) predictive analysis, based on sets of syntactic predictions, hypothetical syntactic functions of individual words in certain types of sentences;

3) reference points method (evolved from predictive analysis), in which typical contexts are determined for words with certain features; this allows to determine the syntactic function of a word in case it can serve different functions;

4) filtering method, which allows to establish word usage restrictions and thus filter out only the information about the word which is relevant to the analyzed text.

Our parser uses all these strategies except the last one.

A grammar of compatibility for all the parts of speech and lexemes showcase the sequential and predictive analyses. The reference points method was directly used for creating the algorithm and the software for syntactic parsing.

The parser for the Ukrainian Text Corpus is deeply linguistic in nature, as it can be used to obtain different information on how syntactic units and their categories function. For example, one can analyze formal syntactic categories such as predicativity, coordination, subordination, as well as take a closer look at subject, predicate, or other constituents.

We have developed a unique novel linguistic product and software which can do the following:

1) detect relations between words and identify word phrases in a simple sentence or a clause;

2) identify constituents (in a complex or compound sentence - clauses);

3) detect relations between clauses.

Thus, as the first step of processing, a full syntactical analysis of the sentence results in creating a dependency tree which can be edited later. Since it is almost impossible to create a precise, mistake-free parsing system for Ukrainian texts, manual processing is necessary to obtain annotated samples of high quality. To proceed with the *second step* of automatic semantic analysis, we need to collect many dependency trees with correct annotation. Then, they can be used as training data for a machine learning system based on vector analysis. The generalized representation of co-occurrence probability for each word can be used to process texts of other discourses present in the Ukrainian language corpus. This data will allow us to create a probabilistic language model to facilitate further research. This showcases both the relevance and the novelty of building a research corpus of Lina Kostenko's texts within the Ukrainian Text Corpus created in the laboratory of computational linguistics of the Research Institute of Philology of the Taras Shevchenko National University of Kyiv.

4. Experiment

The goal of creating a corpus of Lina Kostenko's texts is to develop such a linguistic and software product that would provide extensive information about the author's language and showcase the parameters of her style. Also, it should be convenient to use for other research purposes. In order to reach this goal, we worked on several tasks: linguistic analysis of Lina Kostenko's texts; creation of a database with the linguistic units present in these texts, with their grammatical and quantitative features; development of a convenient user interface to easily search, sort and perform statistical analysis of the database information according to the research purposes [5]. Linguistic processing was carried out in two main ways: 1) the texts were processed automatically with a module responsible for part of speech tagging and grammatical feature recognition; 2) a linguist analyzed the results the system produced, performed quality control, and fixed possible mistakes.

The individual style of Lina Kostenko's works is the object of our research, as it requires a deep and multilevel scientific analysis; the syntactic structure of the sentences of her poetry is the subject of our research. To illustrate the methodology used, Lina Kostenko's ballad poem "Scythian Odyssey" was chosen as an example. We analyzed 987 sentences which contain 8586 words.

Lina Kostenko's individual style requires detailed, balanced, in-depth studies. The corpus of Lina Kostenko's texts we created provides a lot of information about the parameters of the author's language; it is convenient to use in various studies, including text creation. This underlines the scientific novelty, theoretical and practical value of our work.

5. Summary of the research

The parsing system has a few important features.

1) parsing aims to detect all the relation types between words in the phrase (predicative, coordinate, subordinate);

2) grammatical features of the phrase depend on the part of speech of its head. It is well-known that lexical and grammatical features of the word determine its compatibility with other words. Therefore, different types of word phrases exist, as different parts of speech can be its head: noun, adjective, pronoun, numeral, verb, adverb. The syntactic analysis in our system is based on a valency grammar. It includes a subgrammar for verbs (31 206 rules) ([2] Appendix B.1), a subgrammar for nouns (40 023

rules) (Appendix B.4), a subgrammar for adjectives (6 205 rules) (Appendix B.6), and a dictionary of phraseological units (about 2720 units) (Appendix B.8). The valency subgrammars contain information about the lexeme, the governing preposition, and the grammatical case of the governed complement. To encode the part of speech of the complement and its part of speech subclass, a two-character code is used.

3) according to theoretical grammar, there are different types of phrases depending on their structure: simple, complex, and combined. Our study is focused only on simple binary phrases; they may be transformed into complex ones, for which semantic analysis is needed to determine the structure. At this stage, the automatic analysis does not take into account the semantics of the words. As the database contains the numbers of each word form, the user can still see complex phrases created from combining simple binary phrases.

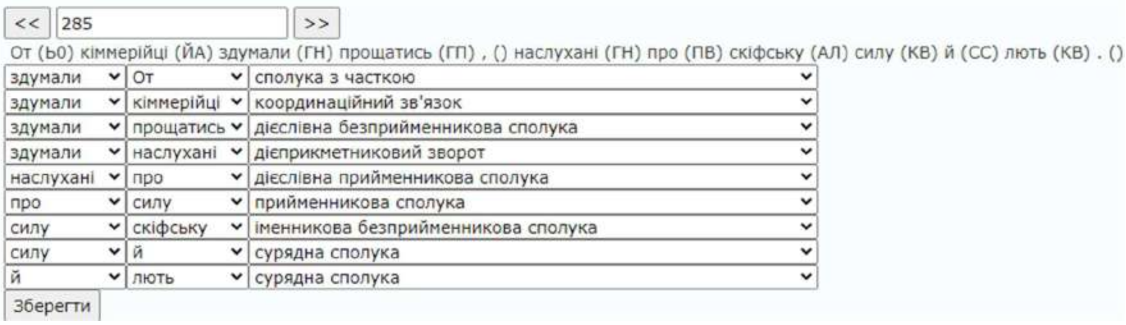
4) We make a clear distinction between "connection" and "relation". By connection we mean a formal connection between the components of a syntactic unit (phrase, simple sentence, complex sentence). And the interaction of lexical meanings and grammatical forms in the composition of phrases is the basis for the formation of semantic syntactic relations. For each word, subordinate, coordinate and predicative relations were established. As part of the general system of connection, they correspond to the components of the situation described in the sentence. We interpret syntactic relations as dependencies between the head word and its dependents and do not use the traditional types of subordinate phrase relations (agreement, government, adjoinment) in this study.

5) the following types of semantic-syntactic relations were automatically established between phrase constituents: subjective relations formed between the subject and the predicate that constitute the nucleus of the sentence; objective relations, in which the direct or indirect object is the dependent; attributive relations, in which the adjectival dependent modifies the head word; adverbial relations, in which the adverbial dependent modifies the head word; completive relations between the components of a complex constituent as opposed to relations between constituents; appositive relations between the appositive and the head word it relates to.

6) as for the semantic relations, it should be possible to use the formal structure of the sentence to determine its semantic structure [19]; syntactic-sematic relations and semantic classification of the words in both head and dependent functions could provide the base for this.

7) subordinate relations in the grammar are divided into two types: core and peripheral. We consider the relation core if the analyzed word is the head of the phrase. In case the analyzed word is a dependent, the relation is peripheral. Predicative relations are established between the subject and the predicate; it is based on their interdependence. Coordinate relations are established between words that are conjuncts. Two words are conjuncts when each of them is a dependent of the same third word, when they are connected by coordinate conjunction or a comma. To detect a sequence of conjuncts, a separate database with word codes is used.

8) thus, automatic analysis of word phrases in the texts of the Ukrainian Text Corpus can produce four types of relations as a result: core, peripheral (adjuncts that showcase subordinate relations), coordinate, predicative.



От (Ь0) кіммерійці (ЙА) здумали (ГН) прощатись (ГП) , () наслухані (ГН) про (ПВ) скіфську (АЛ) силу (КВ) й (СС) лють (КВ) . ()

здумали	От	сполука з часткою
здумали	кіммерійці	координаційний зв'язок
здумали	прощатись	дієслівна безприйменникова сполука
здумали	наслухані	дієприкметниковий зворот
наслухані	про	дієслівна прийменникова сполука
про	силу	прийменникова сполука
силу	скіфську	іменникова безприйменникова сполука
силу	й	сурядна сполука
й	лють	сурядна сполука

Figure 1: Binary phrases and types of syntactic relations in Lina Kostenko's ballad poem "Scythian Odyssey" (sentence 285).

Figure 1 shows a sentence from Lina Kostenko's poem "Scythian Odyssey" and the dependency table automatically created by the program. Morphological annotation is provided for the sentence, including part of speech tags. The table contains head words of the phrases, their dependents, and the syntactic relations between them. This allows creating alphabetically ordered frequency dictionaries of phrases based on relevant works of a specific author. These rules are also used to construct frequency dictionaries for specific lexemes or word classes with their counts and context necessary for illustrative purposes.

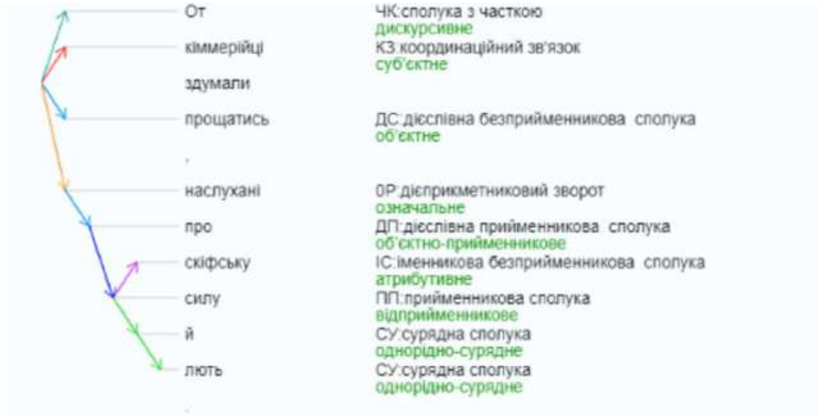
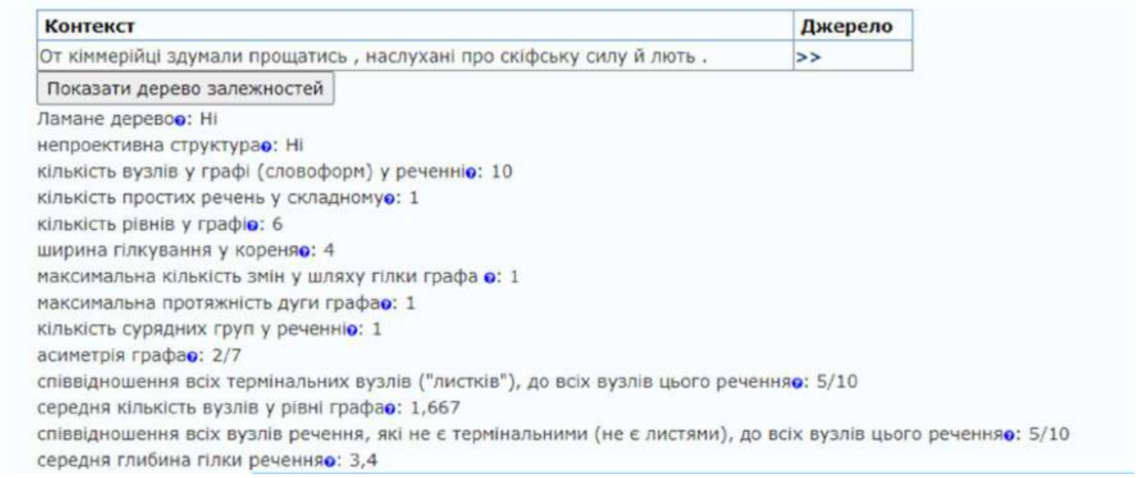


Figure 2: Dependency tree of sentence #285 from Lina Kostenko's ballad poem "Scythian Odyssey"

Figure 2 demonstrates a graphical representation of a dependency tree created by automatically inverting the dependency table. The dependency tree consists of nodes and edges, where nodes represent the words, and edges illustrate the relations between head words and dependents of a phrase. Aside from that, additional information on types of relations between nodes is given. This makes it possible to describe the configuration, form, and outer parameters of the sentence. However, this is not enough to present the structure of the sentence. The information about the type of relations between the constituents of the phrase and semantic-syntactic relations is automatically applied to the set of tree edges. This helps with analyzing complex correlations between semantics and its formal representation, as the text is parsed automatically based on the formal features of its units. Thus, automatic syntactic analysis of the sentence is done on two levels: 1) for each phrase, the program determines its syntactic type based on the morphological features of its head; 2) syntactic relation type is determined for each edge of the graph.

The dependency graph demonstrates important features for stylistic analysis, as it shows the parametric information. In this study, a *parameter* is defined as a quantum of information about the linguistic structure of the sentence. Together with other quanta (parameters), it is represented in the dictionaries, being a specific dictionary representation of structural features of the language. Therefore, syntactic parameterization is an objective representation of the individual style of the author. Based on dependency tree configuration, we suggest analysing the following parameters: node parameter, or mean value of the sentence nodes; tree depth parameter, or mean value of the sentence levels; tree breadth parameter, or mean value of nodes on one sentence level; asymmetry parameter, or the ratio between node counts in subtrees formed by splitting the second tree level; branch parameter, or ratio of terminal node count to the sentence level count; multiplicity parameter, or ratio of nodes with multiple children to the tree count; end-to-end parameter, or mean length of a path from the root node to the terminal node.



Контекст	Джерело
От кіммерійці здумали прощатись , наслухані про скіфську силу й лють .	>>

Ламане дерево: Ні
непроективна структура: Ні
кількість вузлів у графі (словоформ) у реченні: 10
кількість простих речень у складному: 1
кількість рівнів у графі: 6
ширина гілкування у кореня: 4
максимальна кількість змін у шляху гілки графа : 1
максимальна протяжність дуги графа: 1
кількість сурядних груп у реченні: 1
асиметрія графа: 2/7
співвідношення всіх термінальних вузлів ("листків"), до всіх вузлів цього речення: 5/10
середня кількість вузлів у рівні графа: 1,667
співвідношення всіх вузлів речення, які не є термінальними (не є листями), до всіх вузлів цього речення: 5/10
середня глибина гілки речення: 3,4

Figure 3: Parameterization of a dependency tree (sentence 41, "Scythian Odyssey")

Figure 3 shows all the above-mentioned parameters computed for a sentence of the analyzed text (all the sentences of the text are analyzed). This information enables further inspection of word phrases, their syntactic models, the structure of the sentence, and stylistic features of the author's syntax.

6. Results

Tree images model the sentence with a high level of formalization using the graph theory metalanguage. The image allows the user to see the type of the sentence (complex, compound, etc.), the relation between clauses, specific ways in which words co-occur. The co-occurrence of words in the sentence creates its structure and verbalizes its main idea [8, p.66].

Syntactic parameterization results illustrate different features of the text.

Number of Nodes, or the node parameter, can show the conciseness of the sentence if the number is low or prove the sentence is complicated if the number is high. However, it does not represent how structurally complex the phrase is, as the longer the sentence is, the more it can vary stylistically and syntactically. This parameter is computed by counting the words in the tree (we haven't used imaginary words described in the works of I. Sevbo). Figure 3 shows the sentence with ten nodes, while the mean phrase length in the analyzed poem is seven words (standard deviation - 4,74).

Number of Simple Sentences. In general, this parameter demonstrates how discretely/non-discretely the author writes. In this case, discreteness can be associated with segmenting the analyzed situation into atomic facts described by one clause. Both the quantitative parameter and the qualitative characteristic of the arrangement of simple sentences in a complex one are important. This parameter is calculated by counting the clauses in the sentence. Figure 3 shows a sentence with one main clause and one participial clause. On average, the analyzed sentences consist of two clauses (standard deviation - 0,64). A coordinate conjunction usually joins the clauses in the text. Compound sentences with clauses joined by punctuation appear two times less frequently than complex sentences, and compound sentences with coordinate conjunction are three times less frequent.

Number of Root Branches. Previous studies have shown that the number of root branches does not differ a lot across styles and is usually 3. This can be interpreted as a grammatical constant, just as Lucien Tesnière's rule about three actants for the main verb (the sentence root is usually the predicate represented by a finite verb). This parameter is generally computed by counting the number of predicate dependents. To analyze this parameter, we counted the number of edges coming from the root of the tree. On average, there are two root branches in the sentences of the corpus; Figure 3 shows a sentence with four branches.

Root Breadth of the Tree (tree breadth parameter). This parameter illustrates how complex the sentence is. The dependency between the depth (the level of the tree) and breadth (count of nodes on one level) can be established; on average, it is 3-4 edges from the root.

Number of Levels (tree depth parameter). This parameter is computed by counting the number of nodes in the longest path of dependents in the tree. Figure 3 shows the sentence with six levels, while on average, there are four levels in the poem sentences (standard deviation - 1,65).

Maximum Direction Changes. This parameter shows that head words and dependents are disconnected the same number of times as the tree has direction changes. A zigzag pattern can be seen in the image. In Figure 2 and Figure 3, one is the maximum number of direction changes of a graph branch. The structure of the sentence becomes complicated when there are three or more changes in direction. Therefore, it is important to study the reasons for these changes, the average number of changes in different styles, or even research how this parameter changes across the text. This parameter can also be used in automatic text editing. On average, there are two direction changes per sentence in the poem.

Maximum Extent of Link. The previous parameter demonstrates the count of disconnected head-dependent pairs, while this one shows how far away the head and dependent are from one another. It shows the number of unrelated words between the head and the dependent. On the image, the part of the tree under the edge can differ. The most extended link can appear when the sentence is framed by head word and its dependent, while all the other constituents are in between them. This parameter uses only continuous edges. In Figure 3, the maximum link extent is 1, and on average, it is 6 in the poem.

Number of Coordinate Phrases in the Tree. This parameter illustrates a stylistic feature of the author's style. It shows how discrete/non-discrete the author writes, as every coordinate phrase or sequence is an independent part of the tree. The number of coordinate phrases does not provide any information on the structure of the coordinate phrases or their co-occurrence. Different types of coordination are not distinguished at this stage, as it is complicated to do that automatically. Perhaps, the analysis of such phrases could be automated after collecting many sample cases and their manual analysis. The analyzed sentence has only one coordinate phrase, and on average, there are 2-3 coordinate phrases in the sentences of the poem.

The Asymmetry parameter. This parameter shows the ratio between node counts in subtrees formed by splitting the tree in the middle. The resulting image can be symmetrical, which is characteristic of simple sentences. If the sentence is simple, concise, and laconic, the tree should be symmetrical, having the same number of nodes in the left and right parts; in this case, the dependents are evenly distributed throughout the sentence, and the narrative is smooth. There is only 17% percent of such sentences in the poem. More than half of the sentences are complex or compound, so the tree is asymmetrical with more nodes to the right from the root. Obviously, the sentence is more readable if the connections between the dependents are consecutive, and the dependents are situated closer to the head words. As no words split the phrase, the reader does not need to keep them in mind. In general, most trees have more nodes on the right from the root; trees with more nodes on the left are rare. These are mostly sentences with inverted word order or some peculiar stylistic features. Also, the zigzag pattern is more frequently present in such sentences.

7. Discussions

Smooth and rhythmic flow is not characteristic for the ballad poem "Scythian Odyssey", as it is full of complex sentences with asymmetrical clause structures. Simple phrases often frame a complex sentence, while their dependents are situated in the middle of the sentence. The use of ellipsis is also an important feature. Long right-oriented paths with consecutive simple clauses of the same length are also important for the author's individual style. The poetry includes both simple and complex sentences. Even short sentences of Lina Kostenko are diverse: sometimes, symmetrical microstructures appear, and there may be solely right-oriented trees starting with the predicate. In non-projective sentences, the edges cross because of a peculiar word order which is not usual for Ukrainian.

8. Conclusions

The syntax of Lina Kostenko's verse speech is characterized by relative stability, which is motivated by its rhythmic-syntactic organization and covers the entire syntactic organization of the poem - from the smallest unit - phrase to the whole ballad poem "Scythian Odyssey". We follow the approach when the individual style of the poetess is viewed as the choice and arrangement of language elements. The focus is on the qualitative and quantitative characteristics of the grammatical organization of style. The use of statistical analysis data creates a solid basis for distinguishing styles of literary language, to characterize stylistic constants and variables.

Further research prospects involve collecting more statistical information based on the corpus of Lina Kostenko's works, i.e., computing the frequencies of simple, complex, compound sentences, elliptical sentences, interrogative and imperative sentences, all the word phrases types, etc. This will allow us to determine the diagnostic power of the different parameters. A table with information about all the parameters should be compiled as described in [3; 21], paying attention to statistical features, data classification, and various deviations. This table will make it possible to compare texts of different authors or texts of the same author. In addition, specific functions of parameters could be seen, as some parameters show the similarity between the texts, while others highlight the differences in language. Based on the parameterization of the whole corpus of Lina Kostenko's works, an "average" graph of her sentence could be created, which could be interpreted as a constant feature of the author's individual style. The novel software we created adds more features to the Ukrainian Text Corpus and makes conducting linguistic research more convenient.

9. References

[1] Sevbo I.P. Graphic representation of syntactic structures and stylistic diagnostics. Kyiv: Naukova Dumka, 1981. 192 p.

[2] Darchuk N.P. Computer annotation of the Ukrainian text: results and prospects / Darchuk NP - K .: Education of Ukraine, 2013. - 543 p.

[3] Buk S., Rovenchak A. Simple definition of distances between texts from rank–frequency distributions. A case of Ukrainian long prose works by Ivan Franko // Glottometrics. 2019. No. 46. P. 1–11.

[4] Bisikalo O..V Application of the method of syntactic analysis of sentences to determine the keywords of Ukrainian-language content / O.V. Bisikalo, V.A. Vysotska // Radio Electronics, Informatics, Management. 2016. - № 3. - P. 54–65.

[5] Bisikalo, Oleg. The Method of Modelling the Mechanism of Random Access Memory of System for Natural Language Processing / Oleg Bisikalo, Ilona Bogach, Vladyslava Sholota // Proceedings of 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), Lviv-Slavske, Ukraine, February 25 – 29, 2020. – Pp. 472-477. – DOI: 10.1109/TCSET49122.2020.235477.

[6] Malkovsky M.G. Syntax model in the system of morphosyntactic analysis "TREETON" / Malkovsky M.G., Starostin A.S. // Computer Linguistics and Intellectual Technologies: Tr. International Conf. "Dialogue 2006" (Bekasovo, May 31 - June 4, 2006) / ed. A. Narignani. - M., 2006. - S. 481—492.

[7] Rybakov F. I. Automatic indexing in natural language / F. I. Rybakov, E. A. Rudnev, V. A. Petukhov. - M .: Energy, 1980. - 160 p.

[8] Grigoriev N. V. Emergency mechanisms for the syntactic component of the ETAP-3 system / N. V. Grigoriev // Word in the text and in the dictionary / Ros. acad. Sciences, Institute of Rus. lang. them. VV Vinogradova, Institute of Information Transmission Problems. - M., 2000. - S. 485-490.

[9] Linguistic support of the ETAP-2 system / Yu. D. Apresyan [and others]. — M.: Nauka, 1989. — 294, [1] p.

[10] Tsinman L. L. Linguistic processor "ETAP": procedures for weakening syntactic rules and their use / L. L. Tsinman, V. G. Sizov // Word in the text and in the dictionary / Ros. acad. Sciences,

Institute of Rus. lang. them. VV Vinogradova, Institute of Information Transmission Problems. - M., 2000. - S. 485-490.

[11] Kudryashova I. M. Interaction of syntactic and semantic structures in the process of linguistic analysis / Kudryashova I. M., Sokolova E. G. // Scientific and technical information. Series 2. - 1984. - No. 6. - P. 58-62.

[12] Langenbach, M. Automatic parsing of sentences on the principle of grammar of dependencies - Scientific Bulletin of the Lesia Ukrainka East European National University, 2015. - P. 249-254. - (Philological sciences. Linguistics). 6.

[13] Lozynska, O..V, M.V. Davydov, V.V. Pasichnyk. Transformation of grammar trees of components into dependence trees for grammatical analysis of Ukrainian sentences - Lviv: Lviv Polytechnic National University, 2016. - P. 22-31.

[14]. Masytska, Tatiana. Dependence theory in modern syntax. - Volyn: Actual problems of modern linguistics, 2012. - P. 133-144. - (Volyn Philological: text and context).

[15] Testelec Y.G. Introduction to General Syntax. - Moscow, 2001.

[16] Gladkiy A. V. On the procedure for constructing systems of syntactic groups // Moscow Linguistic Journal. 1998. V. 4. S. 32-45.

[17] Leontyeva N. N. Automatic understanding of texts: systems, models, resources. Moscow, 2006.

[18]. Manning K., Raghavan P., Schütze C., An Introduction to Information Retrieval. Moscow, 2011.

[19] Lanhenbakh, Marharyta. (2017). Corpus-Based Semantic Models of the Noun Phrases Containing Words with ‘Person’ Marker. Journal of Linguistics/Jazykovedný casopis. - vol. 68. - p. 249-157. Doi 10.1515/jazcas-2017-0034.

[20] Darchuk N. Compiling of the Electronic Dictionary of Models of the Ukrainian Language Multicomponent Complex Sentences / Ukrainian Linguistics, № 49, 2019 c. 117 – 129.

[21] Buk S., Krynytskyi Y., Rovenchak A. Properties of autosemantic word networks in Ukrainian texts // Advances in Complex Systems. 2019. Vol. 22, No. 6. Article 1950016 (22 pages). DOI: https://doi.org/10.1142/S0219525919500164