Mariana González · Silvia Susana Reyes ·
Andrea Rodrigo · Max Silberztein (Eds.)

Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities

16th International Conference, NooJ 2022
Rosario, Argentina, June 14–16, 2022
Revised Selected Papers

Springer

Editors
Mariana González
Universidad Nacional de Rosario
Rosario, Argentina

Andrea Rodrigo
Universidad Nacional de Rosario
Rosario, Argentina

Silvia Susana Reyes
Universidad Nacional de Rosario
Rosario, Argentina

Max Silberztein
Université de Franche-Comté
Besancon, France

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-23316-6 ISBN 978-3-031-23317-3 (eBook)
https://doi.org/10.1007/978-3-031-23317-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Prosodic Segmentation of Belarusian Texts in NooJ

Yauheniya Zianouka(✉), Yuras Hetsevich, Mikita Suprunchuk, and David Latyshevich

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus

ssrlab221@gmail.com

Abstract. The article describes the syntactic grammar for automatic text segmentation into syntagms in Belarusian by means of NooJ. It is based on the principle of defining sequences of linguistic elements associated with certain semantic relationships and aimed at searching structural and semantic components of utterances and delimiting them into accentual units. Its implementation is essential for improving the synthetic speech generated by Belarusian text-to-speech systems using prepared syntactic grammars in NooJ.

Keywords: Syntactic grammar · Intonation · Syntagm · Prosodic delimitation · Segmentation · Text-to-speech system

1 Introduction

To date, there are no general rules or mechanisms for automatic prosodic delimitation and an unambiguous definition of syntagms in a written text or speech flow. The study of prosodic speech organization is conducted on the basis of auditory and experimental analyses, with the help of which the parameters of super-segmental means are distinguished. They are the limits of the speech flow segmentation, types of intonation constructions (IC), tonal, dynamic and quantitative signals of the IC center, changes in the speed and intensity of sound. All these components are difficult to transfer as a unified component at the computer level and reproduce identically to natural speech.

This problem is quite common for text-to-speech (which converts arbitrary text into speech) and recognition (which automatically converts a speech signal into written text) systems. Despite the achievements in the field of synthesized speech, the problem of qualitative speech synthesis is only partially developed [1, 2]. Firstly, over time, new technologies are emerging to better solve certain issues of speech synthesis. Secondly, a number of algorithms, as well as many linguistic resources necessary for speech synthesis, are language-dependent and have not yet been developed for all languages, including Belarusian. Therefore, the research in the field of speech synthesis, in particular Belarusian text-to-speech (TTS), is relevant to this day.

This work is a continuation of previous research dedicated to automatic speech delimitation. At previous stages of the research, punctuation was used as the major means of separating analyzed text into syntagms [3, 4]. Now we have applied a technique

M. González et al. (Eds.): NooJ 2022, CCIS 1758, pp. 50–62, 2022.
https://doi.org/10.1007/978-3-031-23317-3_5

for automated phrase segmentation not only at the punctuational level but also at the semantic. The keystone is the number of syntagms in a sentence that can significantly exceed the number of punctuation marks in the text [5]. Morphological and syntactic principle is the main core of the research. The approach is confined to the ability of a particular speech part to match with words of other lexical and grammatical classes and occupy a certain syntactic position. The concept is grounded in a superficial syntactic analysis of a text with an emphasis on grammatical characteristics of speech parts that combine accentual units.

2 Relevance of the Study: Lack of Automatic Prosodic Segmentation in TTS

The perception of oral speech is characterized by the fact of how the listener recognizes the meaning of what he has heard. With any chosen approach to assessing the quality of synthesized speech, the main accepted verification parameters are intelligibility of speech, naturalness of speech, evaluation of individual modules of the system, evaluation of recognition, understanding of meaning, expressiveness, emotionality.

The noted parameters of high-quality speech are indicated not only by the developers of language computer technologies, the problem of localizing intonation boundaries in the voiced text is one of the main tasks of the prosodic processor, which is a mandatory block in any automatic speech synthesis system. Syntagmatic articulation of the speech stream identifies minimal semantic units and displays the structural and semantic components of utterances.

The lack of depth syntactic analysis complicates automatic syntagm allocation. It leads to the search for alternative approaches to the development of machine algorithms, methods and techniques for determining the sequence of language elements associated with certain semantic relationships. The delimitation of speech primarily depends on the structure of the sentence, the word order, homogeneous terms, the nature of the word combinations and other language parameters. These complications should be considered and marked in separate syntagms during the development of such systems.

For us, this direction is quite topical. Text-to-text speech synthesis includes a set of questions to ensure the possibility of high-quality processing of arbitrary text into artificial speech [6]. This technology has a wide potential, as proven by numerous TTS for different languages. For the Belarusian language, the most famous is the system called *"Mulʹtyfon-4"*, on the basis of which the staff of the Speech Synthesis and Recognition Laboratory of the United Institute of Informatics Problems (UIIP) [7] of the National Academy of Sciences of Belarus has created an Internet version of the Belarusian-language synthesizer, which is publicly available and free to use (Fig. 1). It converts electronic text into a speech signal in Belarusian, Russian and English with Belarusian accent [8].

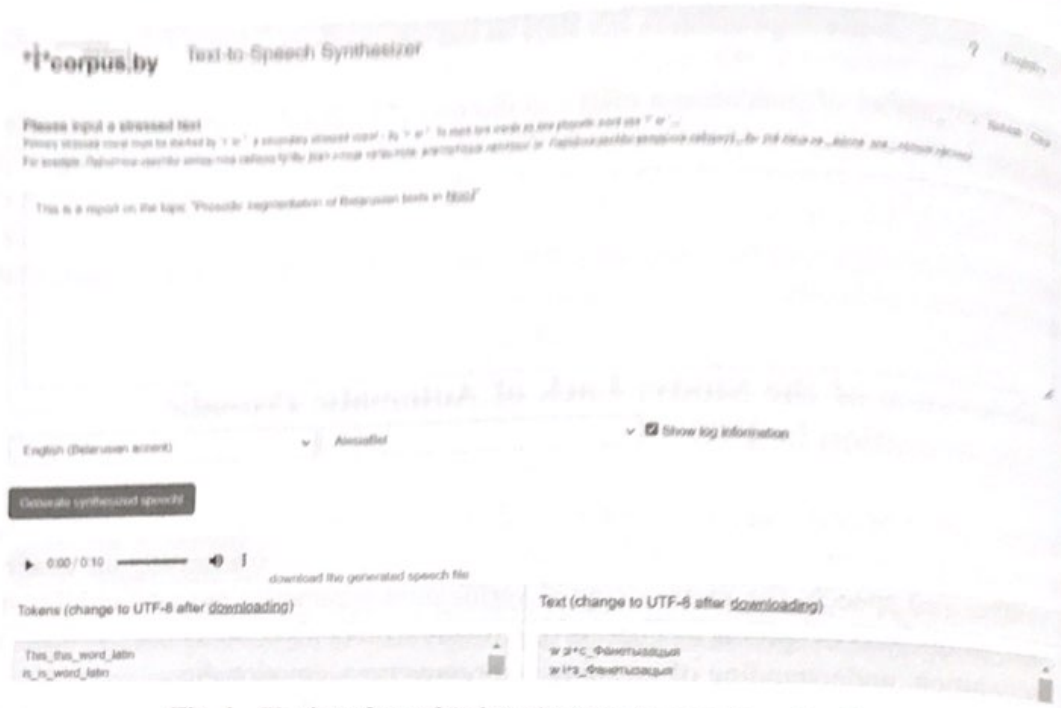


Fig. 1. The interface of Belarusian text-to-speech synthesizer.

Today, new functions are still being incorporated into this synthesizer and already developed ones are being improved. It processes Belarusian texts quite qualitatively, but not perfectly. This is due to a prosodic processor where intonation is almost absent. Belarusian TTS still lacks more or less clear (adequate) intonation – that is, the unity of interrelated components of speech, such as melody, intensity, duration, tempo and timbre, which are inherent in any living utterance. All this reflects a low level of qualified artificial voice and prevents the creation of high-tech national products. In connection with the above, programmers and linguists are faced with the question to consider the phonetic and prosodic characteristics of speech in conjunction with computer technologies where the need for good prosodic processing is the main point to analyze.

3 Text Sources: Corpus of Literary and Medical Texts

In the framework of the research, we composed a text corpus which contains 200 sentences of medical and 200 sentences of literary domains (Fig. 2). The total number is 400 syntactic-accentual units. Stylistic and genre diversity of the selected material is associated with the desire to cover all possible communicative and syntactic types of sentences inherent in the modern Belarusian standard language. Styles differ in the set of linguistic means and their use under the content, tasks and situations of the utterance. So, if a literary text (stylistically enclosed) is focused on evoking an emotional response, influencing the psycho-emotional sphere of the reader/listener, then medical texts are characterized by a strict, almost expressionless nature of scientific and journalistic content using special vocabulary, terminology, abbreviations, a few syntactic constructions. A variety of syntactic constructions of both styles allows considering a variety of word models interaction in the combination of "the main word-dependent components".

Within previous research, the staff of Speech synthesis and recognition laboratory created a text corpus of a medical domain. It was compiled on the basis of medical news published in following medical Internet portals: *Health Committee of Minsk City Executive Committee, Minsk City Gynecological Hospital, 1st Central Regional Clinical Polyclinic of the Central district of Minsk, 4th City Clinical Hospital named after M.J. Saŭčanka.* Our laboratory works on Russian-Belarusian-English translations of these sites. On the rights of authors of bilingual translations, we took the news and formed the corpus. Literary texts are taken from works of Belarusian writers, including the publication *"Belarusian literary heritage: an anthology. In 2 books"*.

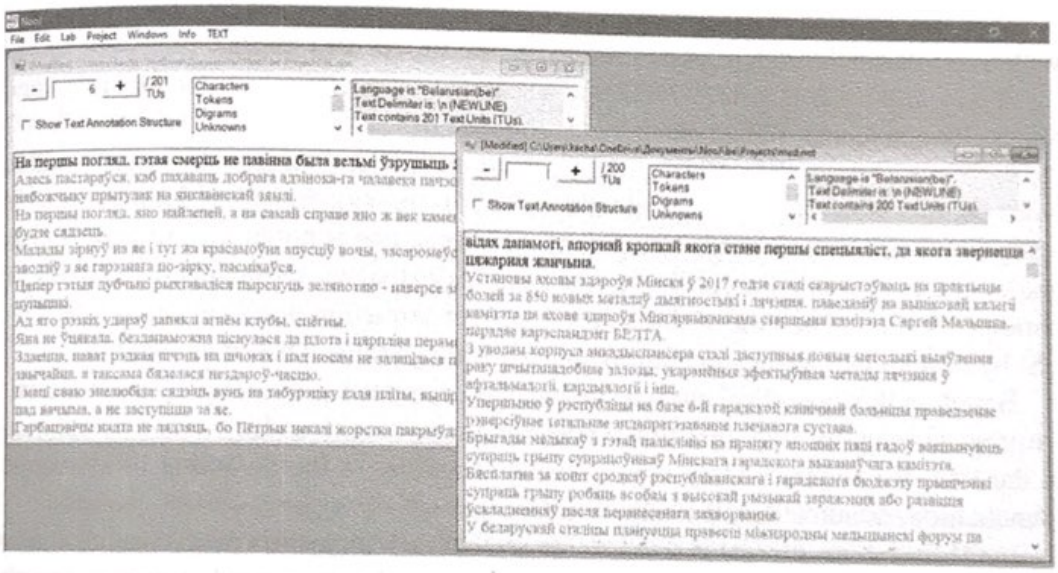


Fig. 2. Text corpus of literary and medical domains.

Initial corpus processing showed the next peculiarities for different styles (journalistic and literary). Literary texts are full of auxiliary parts of speech, have free word order, expressive phrases, and lots of punctuation marks. On the contrary, medical texts are characterized by long sentences, numerous constructions of adjectives and nouns, only nouns, nouns, adjectives and nouns. Auxiliary parts of speech, as a rule, are represented by particles and conjunctions. Punctuation is not so widespread, the most frequent is comma. These regularities were considered during the compilation of the list of rules.

4 Extraction of Syntagms

For the moment, there are no general rules for the syntagm extraction of Belarusian speech. But the results of the statistical analysis based on experimental data give grounds for developing a general algorithm for its delimitation. The system that is planned to find the intonation boundaries of syntagms is based on a superficial syntactic analysis with an emphasis on grammatical characteristics of speech parts. The primary task of this work is to develop rules and an algorithm of formal syntactic grammars that will divide a sentence into syntagms. To develop an algorithm, it is necessary to take into

account all punctuation marks, phraseological units and directly a list of formal rules for dividing a sentence into lexical syntagms.

While delimiting text into syntagms, the next points should be considered: the sentence structure, word order, the presence of homogeneous members, the nature of word combinations and other linguistic parameters. Also, each language has specific rules for syntactic relations and their application. Most of the sentences can be read purely syntactically based on the surface syntactic structure, which in the Belarusian written text is fully displayed by punctuation marks. But sometimes the syntactic information is not enough for the correct delimitation, especially for the ambiguity of the context. This is because of the stylistic and genre diversity.

As it was noted in previous papers [4, 5], three groups of syntagms are distinguished within this research, such as **punctuation, grammatical and lexical.**

A punctuation syntagm (PS) refers to a sentence or part of a sentence that is limited to punctuation marks. Belarusian punctuation includes next marks: ".", ",", ";", ":", "-", "...", "!", "?", "?!", "!!!", "???", "(",")". A *Grammatical Syntagm (GS)* marks stable word combinations (phraseological units and collocations). A *Lexical Syntagm (LS)* is a short sentence of 2–3 words or a part of a sentence that is not limited to punctuation marks and is expressed according to personal lexical signs (through certain words or phrases) or rules. The task of this study is correct extraction of all syntagms (PS, GS, LS) by developing, testing and improving syntactic grammars based on NooJ [9].

Based on the theoretical analysis and applied computer processing of text material, we propose a step-by-step algorithm for determining syntagms and intonation boundaries in the text. It comprises three major blocks according to the definition of syntagms (punctuation, grammatical, lexical).

In the first stage, the text is divided into sentences. Punctuation marks, which define the end of a sentence (a period, a question mark, an exclamation mark, a question mark with an exclamation mark, three exclamation marks) are used for this. The next step is partitioning the sentence into syntagms, namely the sequential extraction of syntagms (punctuational, grammatical and lexical). After determining each type of syntagm, the syntagm boundary is arranged with the input of the corresponding marker. PS are distinguished according to punctuation marks that characterize syntactic relations within a sentence (comma, semicolon, dash, colon, brackets, quotation marks). Numbers, abbreviations and proper names are allocated in a separate syntagm precisely because the problem of their decoding has not been solved in the Belarusian TTS. This is done separately by the system user. The search for stable word combinations is carried out on the basis of Belarusian digitized dictionary of phraseological units by I. Liepiešaŭ. They form a separate syntagm (GS). Next, the search for conjunctions and placing the marker of the syntagm boundary before conjunction according to their category by functional meaning: connective (combinative, enumerative-distributive, comparative, gradational) and subordinate (explanatory, temporary, conditional, causal, target, introductory, final, comparative, of place, mode of action, measures, of degrees). The last point is applying the list of formal rules for dividing the text into lexical syntagms in accordance with the computer legend. The result is an output of all sentences delimited by intonation boundaries with their formal markers.

Figure 3 shows the algorithm for determining syntagms and intonation boundaries in sentences.

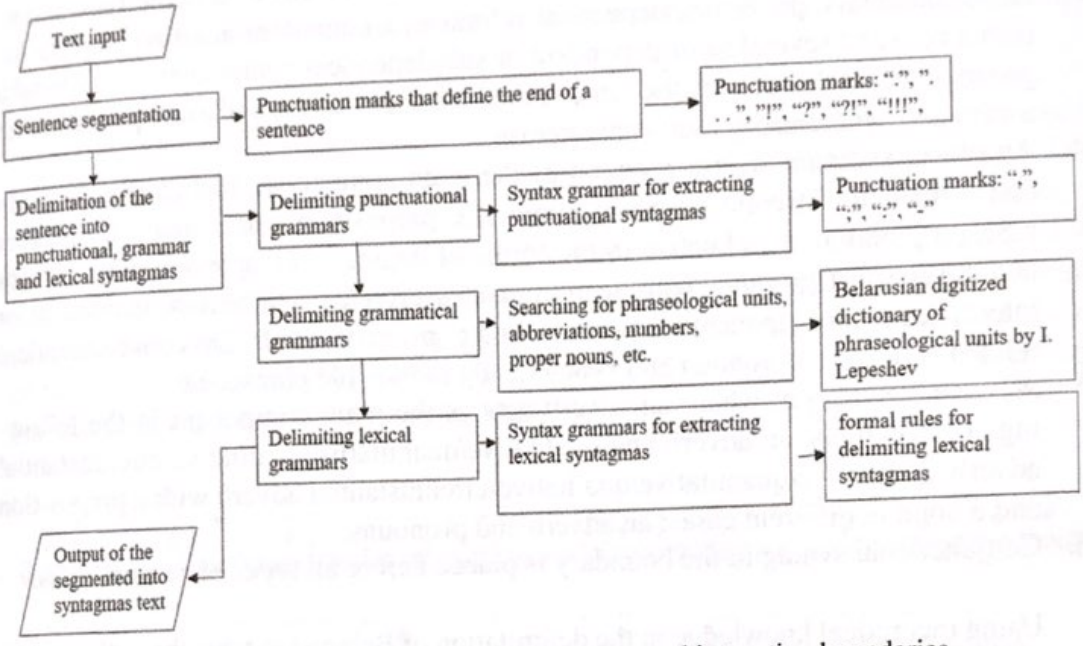


Fig. 3. The algorithm for extracting syntagms and intonation boundaries.

Separating lexical syntagms is the most difficult problem, which are interconnected at the semantic and syntactic levels. This group of lexemes can be determined on the basis of creating general syntactic grammars for a computer expert system, which will search for similar syntactic constructions in a database or a corpus. Each grammar must be presented with a personal syntactic rule to isolate their intonation boundaries in a specific sentence.

In the research, following syntactic rules were used for the arrangement of syntagm boundaries. They are based on the semantic and formal union of two (or more) full-meaning words connected by subordinate relations of the Belarusian standard language. According to the number of principal parts of speech that can serve as the main component, there are 6 types of phrases. They are extracted according to the main word/component and some subordinate members of the sentence:

1. An attributive syntagm that includes a noun and several dependent or interdependent components; a noun and a group of words that convey the same related concepts; a noun and a compound name; a noun and a syntactically indivisible phrase. In these combinations, the main component is the noun.
2. Attributive syntagm with prepositional arrangement separates combinations of prepositions, adjectives with nouns; combinations of prepositions, nouns or a verb group.

3. Predicative syntagm, where the verb or adverbial part is the grammatical and semantic core of the sentence which enters subordinate relations with numerous subordinate members of the sentence: a verb with adverbs; a verb with nouns in different cases; a verb and preposition with nouns in different cases; a verb with participles, subordinate numerals, pronouns, dependent infinitive, a consistent auxiliary verb *to be*; also a verb and several semi-dependent or sub-dependent components; a verb and a group of words that convey the same related concepts; a verb and a compound name; a verb and syntactically indivisible phrase.
4. An object syntagm with a pronoun as the main component, including a pronoun and a noun in different cases; a pronoun, a preposition and a noun in different cases; a pronoun, an adjective in the forms of degree of comparison with adverbs and particles; a pronoun with an infinitive; a pronoun and several dependent or interdependent components; a pronoun and a group of words that convey identical related concepts; a pronoun and syntactically indivisible phrase, etc.
5. Adverbial Syntagm, where an adverb acts as the main component in the following combinations: an adverb and qualitative/quantitative/qualitative-circumstantial adverb; qualitative/quantitative/qualitative-circumstantial adverb with a preposition and a noun in different cases; an adverb and pronouns.
6. Conjunctional syntagm: the boundary is placed before all types of conjunctions.

Using theoretical knowledge in the delimitation of Belarusian texts the authors have developed a list of formal rules for determining lexical syntagms based on the corpus of literary and medical domains (see Fig. 4).

```
I+N+I+N→I/V//PUNKT/
I+N+I+J+N+V→I/C//PUNKT/
I+N+D+J+J+N→I/C/V//PUNKT//L/PART2
I+N+N→I/C/V//PUNKT/L/J
I+N+N+I+N→C/V//PUNKT/L/MV
I+N+N+J+N→I/C/V//PUNKT//L/PART2/R
I+N+N+N/NPG→NPN/I/V//PUNKT/C
I+N+N→I/C/V//PUNKT/L/PART2
```

Fig. 4. The fragment of a list of formal rules for determining lexical syntagms.

Each line describes a combination of speech parts that are included in one syntactic rule. The computer system must consistently analyze each rule until it finds the item that matches the combinations of certain words in the sentence and automatically sets the boundaries of syntagms. The main principle is to consider the right and the left contexts that separate syntagms. Uppercase of Latin letters marks a part of speech and its case, the "+" signs a combination, the right arrow "→" indicates the parts of speech that separates previous and subsequent syntagms (starts a new syntagm), forward slash "/" suggests possible variants of those parts of speech that begin the next syntagm. The "/PUNKT/" symbol describes any of the punctuation marks that possibly separates punctuational syntagms. It is important to note that syntactic grammars are designed for

the computer processing of syntactic-accent units at the machine level. For the moment, the list consists of 300 formal rules. However, their number may increase during the analysis of a larger volume of material and testing the system for defining new types of syntagms.

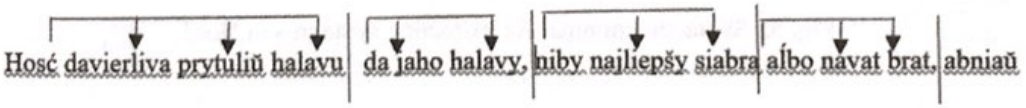
For a clearer understanding, consider two sentences of the literary and medical domains.

An example of medical sentence:

Haradskaja kliničnaja baĺnica chutkaj dapamohi horada Minska zjaŭliajecca viadučaj ustanovaj u krainie pa akazanni kruhlasutačnaj spiecyjalizavanaj ekstranaj dapamohi roznym katehoryjam nasieĺnictva: [Haradskaja kliničnaja baĺnica (JN+JN+N→JR+NR+NG+NG)] [chutkaj dapamohi horada Minska (J+N+N+N→I/C/V//PUNKT/L/PART2)] [zjaŭliajecca viadučaj ustanovaj u krainie (V+J+N+I+N→I)] [krainie pa akazanni kruhlasutačnaj spiecyjalizavanaj ekstranaj dapamohi (I+N+JG+JG+NG+NG→I/J/C/V//PUNKT/L/PART2/J)] [roznym katehoryjam nasieĺnictva (J+N+N→V/R/I//PUNKT)].

An example of a literary sentence:

Hosć davierliva prytuliŭ halavu da jaho halavy, niby najliepšy siabra aĺbo navat brat, abniaŭ.: [Hosć davierliva prytuliŭ halavu (N+R+V+N→I/R/P/INF)] [da jaho halavy (I+P+N→I/C/V//PUNKT/)] [niby najliepšy siabra (R+J+N→C//I//PUNKT/)] [aĺbo navat brat (C+R+N→C/I//PUNKT)] [abniaŭ (/PUNKT/→V→/PUNKT/)].



5 Syntactic Grammar for Extracting Syntagms in NooJ

The work carried out on prosodic segmentation makes it possible to automate the created and systematized resources based on NooJ [10]. The developed algorithm and formal rules for determining syntagms form the basis for creating a syntactic grammar for computer processing of grammatical structure of syntagm. Based on the segmentation methods proposed above, we improved syntactic grammar (Fig. 5), which represents the initial stage of prosodic text processing for Belarusian-language speech synthesis systems. It consists of 8 graphs which search syntagms according to parts of speech. They are adjective, prepositional, pronoun, noun, verb, adverbial, conjunctional, particle groups. They are considered as the first component in the sequence of parts of speech that form a new syntagm.

The principle of grammar is as follows: the system consistently analyzes the formal grammatical indicators of words/phrases in a sentence. If it finds a coincidence of morphological and syntactic characteristics of speech parts according to the formal rules of each subgraph, it encloses them in a syntagm. It also notes the graph to which the syntagm corresponds. The characters *SYNT, (,),* in the subgraph reflect the beginning and the end of the syntagm, which the system automatically highlights in the sentence. The right and left contexts are indicated before and after these symbols. Then combinations of speech parts should coincide with a separate formal rule given in the list of formal rules. After finding the correct subgraph corresponding to a certain rule from the list, the system analyzes the right context: formal markers indicating the boundary between syntagms (main/auxiliary parts of speech or punctuation marks). This marker is an indicator of the next syntagm. Thus, a boundary is drawn between the combination of words of one subgraph and certain markers that begin a new syntagm. In accordance with this principle, syntactic grammar works.

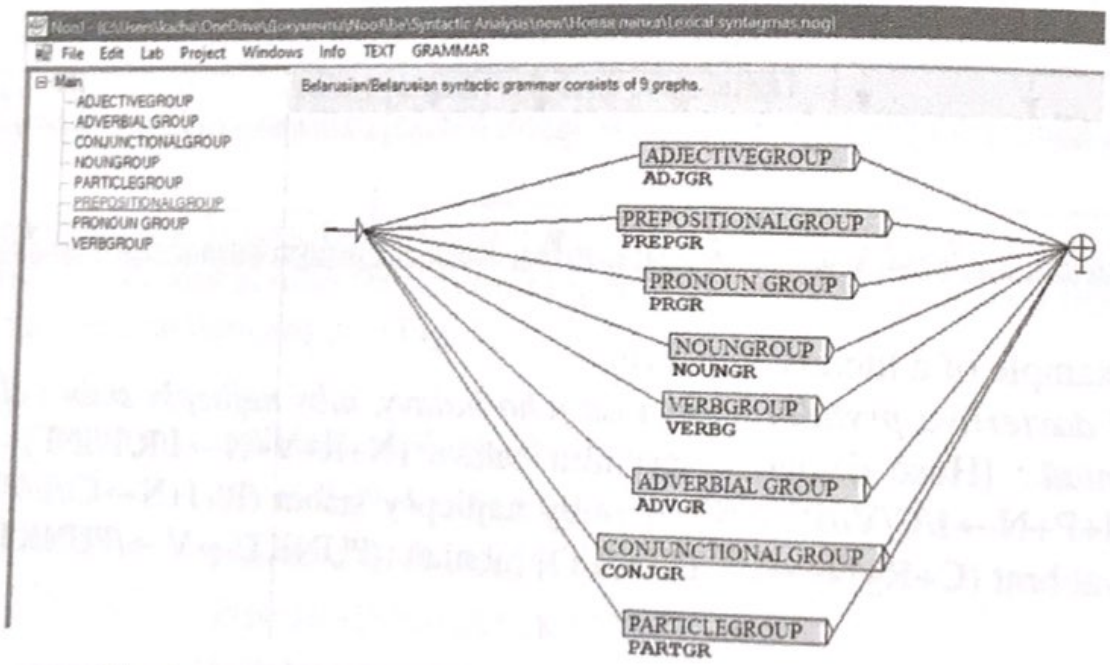


Fig. 5. Syntactic grammar for extracting syntagms in NooJ.

For instance, the adverbial graph consists of nine subgraphs (Fig. 6). The main principle of this grammar is the combination of adverbs as the first component of a syntagm and their subordinate components.

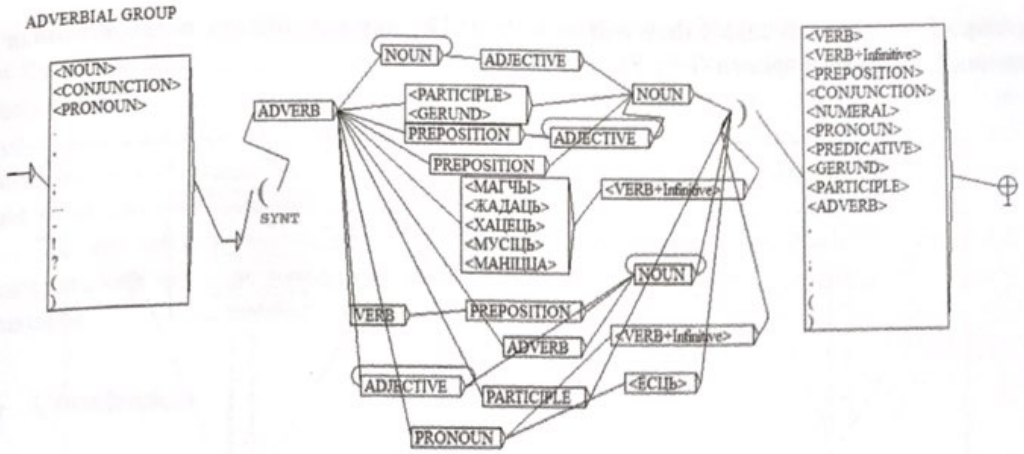


Fig. 6. The graph for extracting adverbial syntagms.

It's very important to consider the right and the left context (words/expressions) which surround this syntagm for delimiting its boundaries. According to this subgraph, the left context can be a noun, a conjunction, a pronoun or some punctuation marks represented in this figure. The right context can be represented by any main part of speech, some auxiliary parts of speech or punctuation marks.

Figure 7 demonstrates the search of verbal syntagms where a verb is the main component.

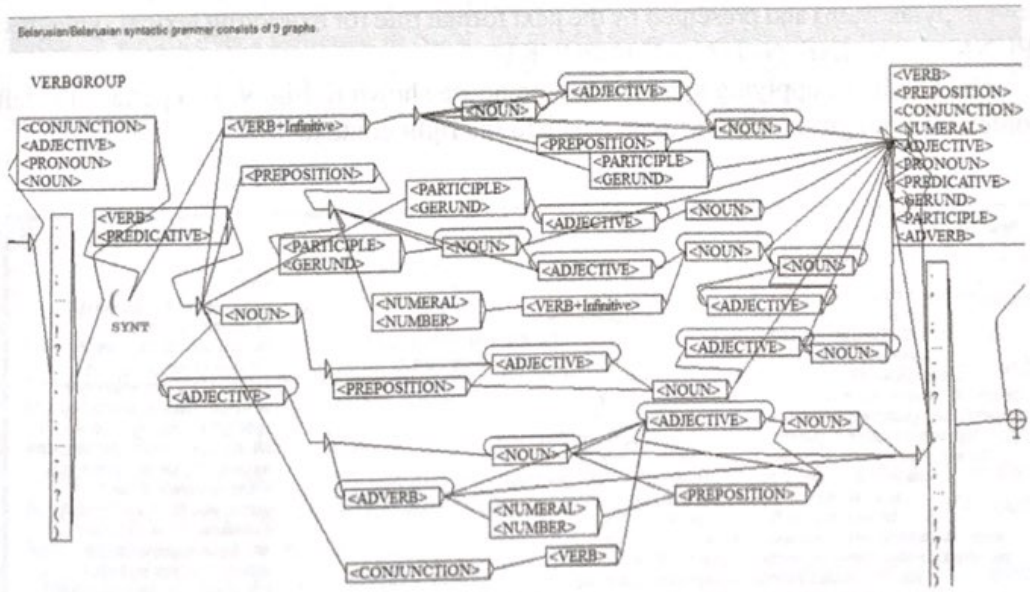


Fig. 7. The graph for extracting verbal syntagms.

For a clearer understanding of grammar, let's analyze the first line of the graph: a verb in combination with an infinitive, noun, adjective and noun forms the first verb

group of syntagms in case if they will be followed by any punctuation, notional word or an auxiliary part of speech (Fig. 8).

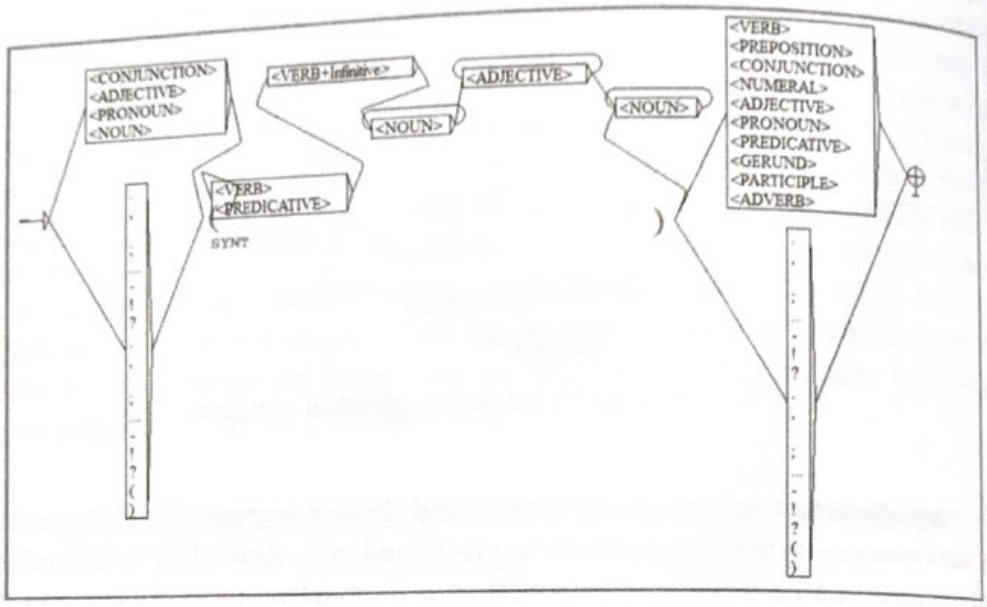


Fig. 8. The graph for extracting verbal syntagms according to the rule /PUNKT/→V+INF+N+J+N→C/Л///PUNKT/

It is realized in the sentence *"Павел, вярнуўшыся з вуліцы, вырашыў прыбраць кватэру новым пыласосам"*. (Paviel, viarnuŭšysia z vulicy, vyrašyŭ prybrać kvateru novym pylasosam) and presented by the next formal rule for extracting lexical syntagms: /PUNKT/→V+INF+N+J+N→C/Л///PUNKT/

The results of applying syntactic grammar are shown in Fig. 9. It separates the left context, a syntagm marked with its type and the right context.

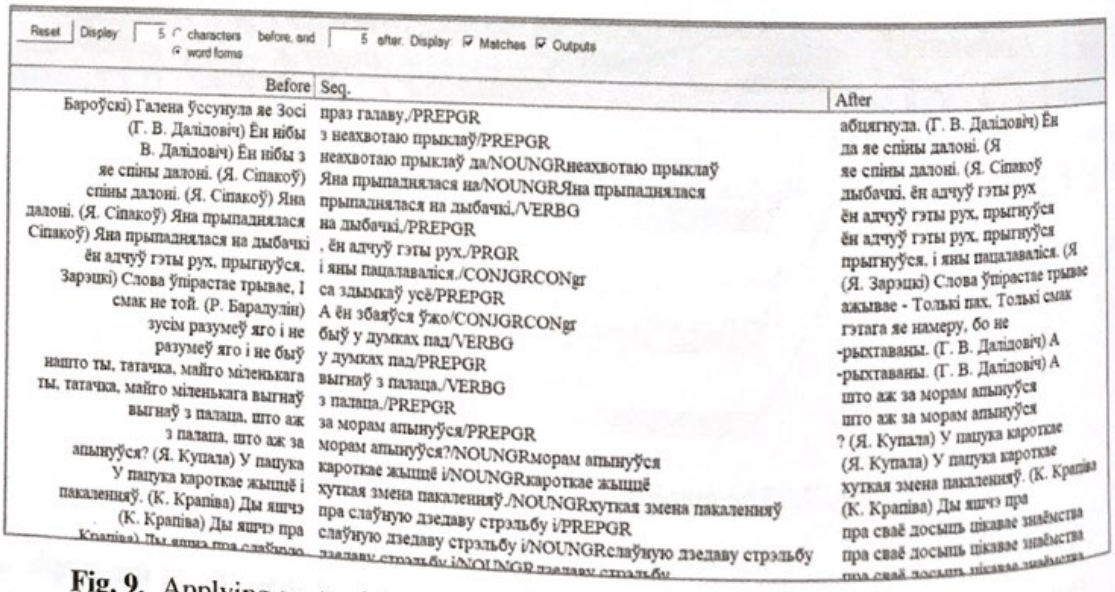


Fig. 9. Applying syntactic grammar in the corpus of medical and literary domains.

The grammar has some flaws and demands follow-up revision. The main hypothesis of the grammar is sequential processing of each subgraph from the most complex to the simplest. The same is necessary for graphs. Unfortunately, for today this problem is not resolved. The system analyzes the corpus randomly, the same sentences are checked according to different formal rules. Also, there is a problem with homonyms, numbers and abbreviations which are not taken into account in the grammar.

The next step of the research is detailed grammar testing on the whole corpus for searching new word combinations into syntagms, adding them into graphs and correcting mistakes.

6 Conclusion

The article presents the syntactic grammar for automatic extraction of syntagms and highlighting the intonation boundaries between syntagms at the syntax level in NooJ. The main core is the morphological and syntactic principle that lies in the ability of a particular speech part to match with other words or word forms and occupy certain positions in the sentence. The concept is grounded in a superficial syntactic analysis of different texts (based on the corpus of literary and medical domains) with the emphasis on grammatical characteristics of speech parts that combine accentual units. The purpose of preparing the corpus of different domains is to search and define syntactic constructions in the Belarusian language, not separated by punctuation marks and conjunctions. Their detailed analysis (mostly manual processing of every sentence) provides the source for compiling a list of formal syntactic rules that will later be used by an expert system as the means of searching for identical structures in the input text and determining intonation boundaries within every sentence in NooJ. Identified prosodic aspects estimate the value of intonation peculiarities of Belarusian. Obtained results will be used for further research in automatic processing of prosodic structure of Belarusian, in particular for computer systems with voice accompaniment.

References

1. Lobanov, B., Levkovskaya, T.: Multi-stream word recognition based on a large set of decision rules and acoustic features. In: Proceedings of the 5th International Title Suppressed Due to Excessive Length 9th Workshop Speech and Computer SPECOM 2000. Revised Selected Papers, St.-Petersburg, pp. 75–78 (2000)
2. Lobanov, B., Tsirulnik, L., Zhadinets, D., Karnevskaya, E.: Language- and speakerspecific implementation of intonation contours in multilingual TTS synthesis. In: Speech Prosody: Proceedings of the 3rd International Conference, Dresden, Germany, 2–5 May, Revised Selected Papers, vol. 2, pp. 553–556 (2006)
3. Okrut, T., Hetsevich, Y., Lobanov, B., Yakubovich, Y.: Resources for identification of cues with author's text insertions in Belarusian and Russian electronic texts. In: Monti, J., Silberztein, M., Monteleone, M., di Buono, M.P. (eds.) Formalising Natural Languages with NooJ 2014. Revised Selected Papers, pp. 129–139. Cambridge Scholars Publishing, Newcastle (2015)
4. Hetsevich, Y., Okrut, T., Lobanov, B.: Grammars for sentence into phrase segmentation: punctuation level. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, Hanna (eds.) NooJ 2015. CCIS, vol. 607, pp. 74–82. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42471-2_7

5. Zianouka, Y., Hetsevich, Y., Latyshevich, D., Dzenisiuk, Z.: Automatic generation of intonation marks and prosodic segmentation for Belarusian NooJ module. In: Bigey, M., Richeton, A., Silberztein, M., Thomas, I. (eds.) NooJ 2021. CCIS, vol. 1520, pp. 231–242. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-92861-2_20
6. Lobanov, B.: Computer Synthesis and Cloning of Speech. Bielaruskaja Navuka, Minsk (2008)
7. Speech Synthesis and Recognition Laboratory. https://ssrlab.by/en/. Accessed 29 July 2022
8. Text-to-speech synthesizer. https://corpus.by/. Accessed 01 Aug 2022
9. NooJ: A Linguistic Development Environment. http://www.nooj4nlp.org/. Accessed 18 Feb 2021
10. Silberztein, M.: Formalizing Natural Languages: The NooJ Approach. Wiley, Hoboken (2016)