Magali Bigey · Annabel Richeton ·
Max Silberztein · Izabella Thomas (Eds.)

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

15th International Conference, NooJ 2021
Besançon, France, June 9–11, 2021
Revised Selected Papers

Springer

Editors
Magali Bigey
Université de Franche-Comté
Besançon, France

Annabel Richeton
Université de Franche-Comté
Besançon, France

Max Silberztein
Université de Franche-Comté
Besançon, France

Izabella Thomas
Université de Franche-Comté
Besançon, France

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-030-92860-5 ISBN 978-3-030-92861-2 (eBook)
https://doi.org/10.1007/978-3-030-92861-2

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

Approach to the Automatic Treatment of Gerunds in Spanish and Quechua: A Pedagogical Application 135
Andrea Rodrigo, Maximiliano Duran, and María Yanina Nalli

Using NooJ to Formalize French Cooking Expressions 147
Tong Yang

Natural Language Processing Applications

Paraphrasing Tool Using the NooJ Platform 163
Amine Alassir, Sondes Dardour, and Héla Fehri

The Recognition and the Automatic Translation of Dative Verbs. 174
Hajer Cheikhrouhou

From Laws and Decrees to a Legal Dictionary 187
Ismahane Kourtin, Samir Mbarki, and Abdelaaziz Mouloudi

Automatic Detection and Generation of Argument Structures Within the Medical Domain 198
Walter Koza and Constanza Suy

Meaning Extraction from *Strappare* Causatives in Italian 208
Ignazio Mauro Mirto and Mario Monteleone

Terms and Appositions: What Unstructured Texts Tell Us 219
Giulia Speranza, Maria Pia di Buono, and Johanna Monti

Automatic Generation of Intonation Marks and Prosodic Segmentation for Belarusian NooJ Module. 231
Yauheniya Zianouka, Yuras Hetsevich, David Latyshevich, and Zmicier Dzenisiuk

Author Index 243

Automatic generation of intonation marks and prosodic segmentation for Belarusian NooJ Module

Yauheniya Zianouka, Yuras Hetsevich, David Latyshevich and Zmicier Dzenisiuk

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Belarus
evgeniakacan@gmail.com, yuras.hetsevich@gmail.com, david.latyshevich@gmail.com, d.denissyuk@gmail.com

Abstract. The article depicts the problem of automated text segmentation into syntagmas at the punctuational and the semantic levels of Belarusian. It is aimed at generating prosodic transcription and delimiting long sentences. 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, Syntagma, Prosodic delimitation, Segmentation, Extraction, Text-to-speech systems

1 Introduction

The problem of localizing automatic intonation boundaries in a text is one of the main tasks of a prosodic processor, which is a mandatory unit in any speech and recognition system. The syntagmatic articulation of the speech flow allocates minimal semantic units and reflects the structural and semantic components of utterances [1]. The automatic selection of syntagmas is complicated by the lack of deep parsing, leading to the search for new approaches to the development of machine algorithms, methods and techniques by defining sequences of linguistic elements associated with certain semantic relationships.

To date, there are no general rules or mechanisms for an unambiguous definition of syntagmas in a written text or speech flow. The study of the 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 [2]. 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.

This work is a continuation of previous research where analyzed sentence parts were separated by punctuation [3, 4]. Most punctuation marks for sentences have been developed (up to 5 words) but the most frequent being three-word. Now the authors have expanded the study. The keystone is the number of syntagmas in a sentence that can significantly exceed the number of punctuation marks in the text. We applied a technique for automated phrase segmentation not only at the punctuational level but also at the semantic. It is also a system of marking intonation patterns in electronic Belarusian texts using NooJ [5, 6]. It widens the prosodic performance of the Belarusian text-to-

speech system and may also serve to improve the Belarusian NooJ module with so-called prosodic transcription at different levels.

2 Types of syntagmas in the Belarusian language

Intonation components are primarily related to syntax that represents a set of rules, theoretical systems and language processes. The delimitation of syntagmas is connected with the sentence structure, word order, the presence of homogeneous members, the nature of word combinations and other linguistic parameters. All the mentioned components should be taken into account and noted in separate syntagmas while developing new syntactic and morphological NooJ grammars [7].

The problematic field of our research involves next points:

– Each language has specific rules for syntactic relations and their application. It should be noted that not all the results obtained for one language are suitable for developing similar mechanisms in other languages. It is necessary to specify prosodic rules of Belarusian speech organization.

– Most of the sentences can be read purely syntactically based on the surface syntactic structure, which in the Belarusian written text is quite fully displayed by punctuation marks. Based on the knowledge of the rules operating in the language, the most and least probable boundaries are predicted both when encoding speech by the speaker and when decoding it by the listener. But sometimes the syntactic information is not enough for the correct delimitation especially for the ambiguity of the context.

– Not all the context can be rendered by syntax with the help of punctuation. First of all, this is due to the stylistic and genre diversity. For example, if a literary text is focused on evoking an emotional response, influencing the psycho-emotional sphere of the reader/listener, it is characterized by the use of visual and expressive means and various syntactic structures, separated by punctuation marks. On the contrary, medical texts are characterized by a strict, almost expressionless nature of scientific and journalistic content using special vocabulary, terminology, abbreviations and less punctuation. Also, analytical languages convey grammatical relations through individual parts of speech (prepositions, modal verbs, etc.), fixed word order, where context and/or intonation variations. It also can be reproduced by a various system of inflection using dependent morphemes (endings, suffixes, prefixes, etc.) but less by punctuation.

Therefore, within the framework of this study, three groups of syntagmas are defined: **punctuation, grammatical and lexical.**

A punctuation syntagma (PS) refers to a sentence or part of a sentence that is limited to punctuation marks. Belarusian punctuation includes next marks: “.”, “,”, “;”, “:”, “-”, ”. . . ”, ”!”, “?”, “?!”, “!!!”, “???”.

Examples: *Karboksihiemahlabin, (PS) jaki utvarajecca pry hetym, (PS) nie zdolny pieranosic kislarod (PS).*

Skoda, (PS) vyklikanaja nikacinam, (PS) zakranaje nie tolki samich kurcoŭ (PS).

Grammatical Syntagma (GS) marks stable word combinations (phraseological units and collocations).

Lexical Syntagma (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.

Examples: *Na zaniatkach školy prafsajuznaha aktyvu (LS) abmiarkoŭvalisia pytanni (LS) ab matyvacyi prafsajuznaha členstva (LS).*

Uračy pastajanna viaduc baracbu z kurenniem (GS) siarod moladzi (LS) i daroslaha nasielnictva (PS).

As the problem of extracting punctuation syntagmas is resolved in the previous works, the main and most difficult task of current research is to distinguish lexical syntagmas in a Belarusian written text.

3 An algorithm for dividing texts into syntagmas

Unfortunately, there are no general rules for the syntagma extraction of Belarusian speech. But the results of the statistical analysis of the experimental data fulfilled by the authors give grounds for developing a general algorithm for its delimitation. The system that is planned to be created to find the intonation boundaries of syntagmas is based on a superficial syntactic analysis with an emphasis on grammatical characteristics of speech parts. In programming, the syntax is determined by a set of rules applied in mathematical systems. Simplified versions of the original mathematical model of syntax and the method of its modeling on a computer are suitable for writing, on the basis of which these models are developed. The main task of this work is to develop rules and an algorithm of formal syntactic grammars that will divide a sentence into syntagmas. Types of syntagmas described above allow concluding that in order to develop an algorithm, it is necessary to take into account all punctuation marks, phraseological units and directly a list of formal rules dividing a sentence into lexical syntagmas.

An algorithm for determining syntagmas and intonation boundaries in sentences contains the next steps (see fig.1):

1. Dividing a text into sentences, according to punctuation marks emphasizing the end of the sentence. They are a period, a question mark, an exclamation mark, a question mark with an exclamation mark, three exclamation marks.
2. Delimiting a sentence into syntagmas:

2.1 Searching punctuation marks in a sentence that mark syntactic relations within a sentence: comma, semicolon, dash, colon, brackets, quotation marks. Arranging syntagma boundary and inserting a marker instead of a punctuation mark in combinatorial variants of intonation types.

2.2 Searching numbers, abbreviations, abbreviations and proper names. Splitting them off into a separate syntagma.

2.3. Searching phraseological units. Allocating them into a separate syntagma, arranging a syntagma boundary and inserting a marker according to the phraseological units.

2.4 Searching for conjunctions. Placing the marker of the syntagma 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).
2.5 Compiling a list of formal rules for dividing a text into semantic syntagmas.
3. The output of all sentences delimited by intonation boundaries with their formal markers.
Steps 1-2.1 of the algorithm described above can be fulfilled on the basis of the syntactic grammar developed in the previous research (see fig.2). It consists of 10 graphs depicting punctuation syntagmas of the Belarusian language. The authors applied a technique for automated phrase segmentation at the punctuational level and a system of marking types of phrase intonation in electronic Belarusian texts using NooJ.
Grammatical syntagmas can be searched for using formal markers with the help of the dictionary of phraseological units of I. Y. Lepeshev. For other steps, it is necessary to develop additional resources including a list of formal rules for delimiting grammatical and lexical syntagmas.

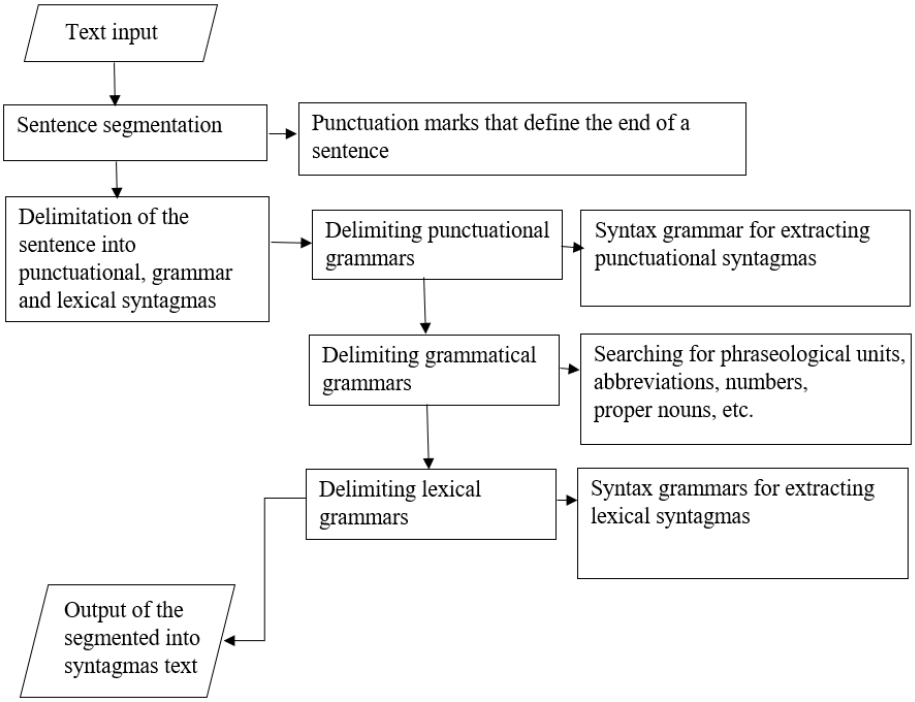

Fig. 1. An algorithm for extracting syntagmas and intonation boundaries.

A more difficult problem is separating lexical syntagmas, 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.

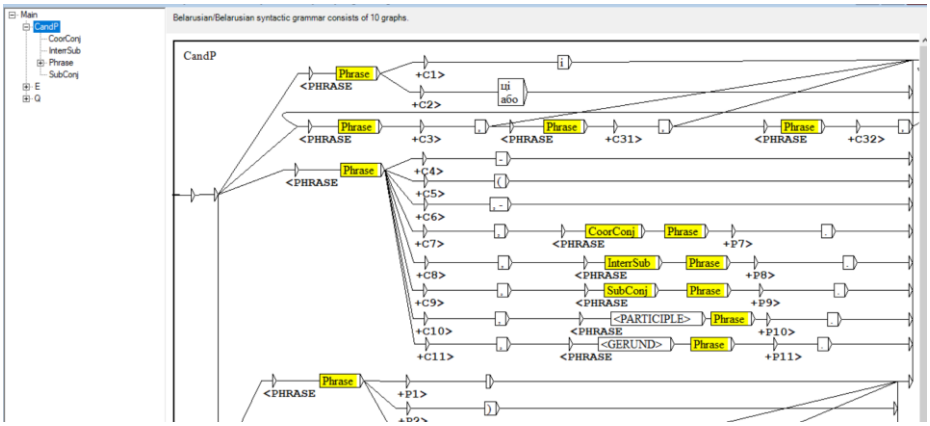

Fig. 2. Syntactic grammar for determining punctuational grammars.

4 Formal rules for extracting semantic grammars

Within previous research, the staff of speech synthesis and recognition laboratory created a text corpus of a medical domain (see fig.3). It was compiled on the basis of medical news published in next 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. At the current stage of development, the authors have added new medical texts and supplement them. Now, created corpus of medical domain consists of:

- 627 texts;
- 23623 text units;
- 24212 digits;
- 63017 other delimiters;
- 336146 tokens;
- 248917 wordforms;
- 55208 different annotations.

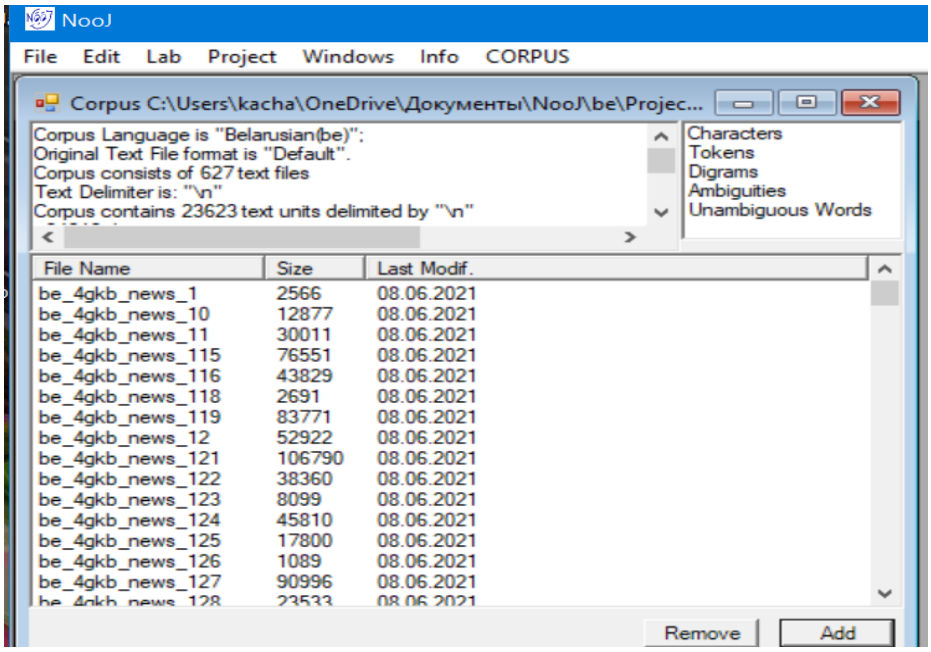

Fig. 3. Text corpus of medical domain.

The purpose of the corpus extension is to search and define syntactic constructions in the Belausian literary language, not separated by punctuation marks and a system of conjunctions. Their detailed analysis (mostly manual processing of each individual sentence) provides the basis for compiling a list of formal syntactic rules that will later be used by the expert system as a means of searching for identical structures in the input text and determining intonation boundaries within each sentence.

Fig. 4. Applying syntactic grammar for extracting punctuational syntagmas.

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

1. The boundary is drawn between a nominative group that includes a noun and several codependent or sub-dependent components: a noun and a group of words that convey the same related concepts; a noun and a complex name; a noun and a syntactically indivisible phrase. In these coordinated combinations of words, the main component is the noun.

2. The boundary is placed between the verbal group, where the verb is the grammatical and semantic core of the sentence and enters into subordinate relations with a large number of subordinate members of the sentence.

3. The boundary is drawn between phrases with an adjective as the main component, which include an adjective and a noun in different cases. They can be an adjective, a preposition and a noun in different cases; an adjective in the forms of degrees of comparison with adverbs and particles; an adjective with an infinitive; an adjective with pronouns; word combinations of adjectives; an adjective and several co-dependent or sub-dependent components; an adjective and a group of words that convey the same related concepts; an adjective and a complex name; an adjective and syntactically indivisible phrase; an adjective with a noun in different cases and determinators; an adjective, a preposition with a noun in different cases and a determinator.

4. The boundary is placed between the adverbial group, where the adverb acts as the main component in the following combinations: adverb and qualitative/quantitative/qualitative-circumstantial adverb; qualitative/quantitative/qualitative-circumstantial adverb with a preposition and a noun in different cases; adverb and pronouns.

5. The boundary is placed between the prepositional group that separates long combinations of adjectives with nouns or the verbal group.

6. The boundary is placed before all types of conjunctions. With three or more homogeneous members, it combines the last two.

Using theoretical knowledge in the delimitation of Belarusian texts the authors have developed a list of formal rules for determining lexical syntagmas based on the medical text corpus (see fig.5). 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 syntagmas. The main principle is to take into account the right and the left context that separates syntagmas. 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 syntagmas (starts a new syntagma), forward slash "/" suggests possible variants of those parts of speech that begin the next syntagma. The"/ PUNKT /" symbol describes any of the punctuation marks that possibly separates punctuational syntagmas. 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 250 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 syntagmas.

104. P+N+R→C//PUNCT//V
105. P+I+P→V/C//PUNKT/
106. P+R+P→P+J+N+V
107. P+J+N+V→PUNKT//C
108. P+V+P+N→C//PUNKT/
109. D+J+J+N→I/C/V//PUNKT/L/PART2
110. J+J+J+N→I/C/V//PUNKT/L/PART2
111. R+N+N+J+N→I/C/V//PUNKT/L/PART2
112. R+IPL+N→I/C/V//PUNKT/L/PART2
113. R+I+J+J+N→I/C/V//PUNKT/L/PART2
114. R+P+INF→I/C/V//PUNKT/L/PART2
115. R+P+ЁСЦЬ→C//PUNKT//
116. R+I+N→I+N
117. R+MV+INF→N+I+N+N
118. R+V+I+N+N+N→I/C/V//PUNKT/
119. R+V+I+N+N→P

Fig. 5. The fragment of a list of formal rules for determining lexical syntagmas.

5 Application of formal grammars in NooJ

Based on the segmentation techniques proposed above, the authors have developed a NooJ syntactical grammar (see Fig. 6) representing the initial stage of prosodic processing for text-to-speech systems. It consists of 7 graphs: syntagmas which start a sentence, verbal, adverbialial, noun, pronoun, prepositional, conjunction groups. In their turn, some of them are also divided into subgraphs.

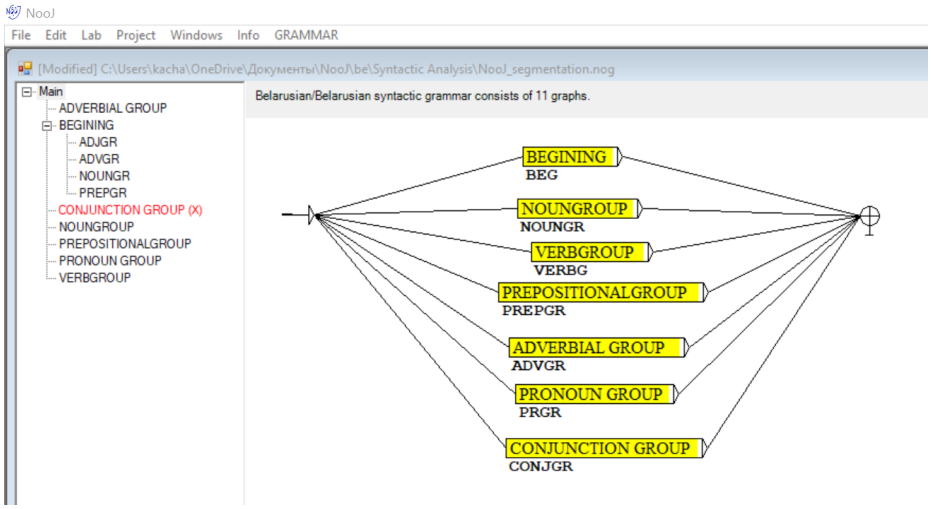

Fig. 6. Syntactic grammar for extracting lexical syntagmas in NooJ.

Firstly, the system automatically defines syntagmas that start the sentence. For this, it specifies punctuation marks that finish the previous sentence and uses a ”Beginning” graph that is the combination of different speech parts (see fig. 7, 8). It also subdivided into four subgraphs according to the main component of accentual unit. After finding the correct subgraph that corresponds to a certain rule from the list, the system analyzes the right context: formal markers that indicate the boundary between syntagmas (main/auxiliary parts of speech or punctuation). This marker is an indicator of a new syntagma. Thus, a boundary is drawn between the word combination of one subgraph and defined markers that begin a new syntagma. According to this principle, the syntactic grammar should work.

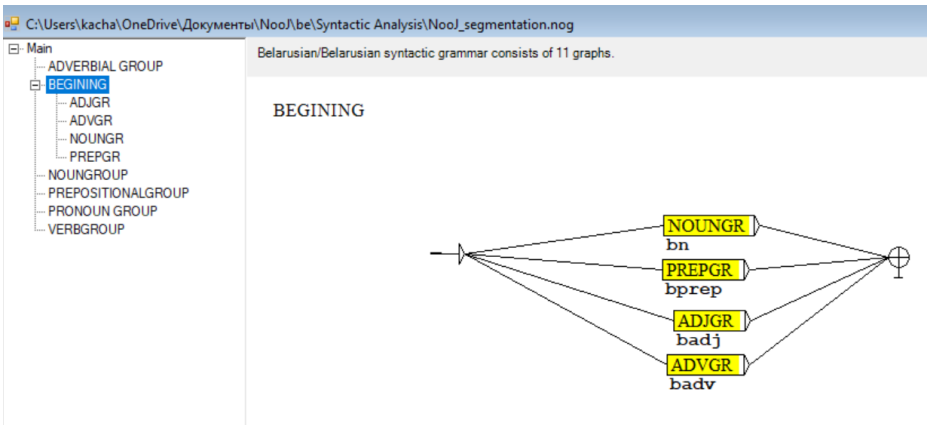

Fig. 7. A “Beginning” graph of the syntactic grammar for extracting lexical syntagmas in NooJ.

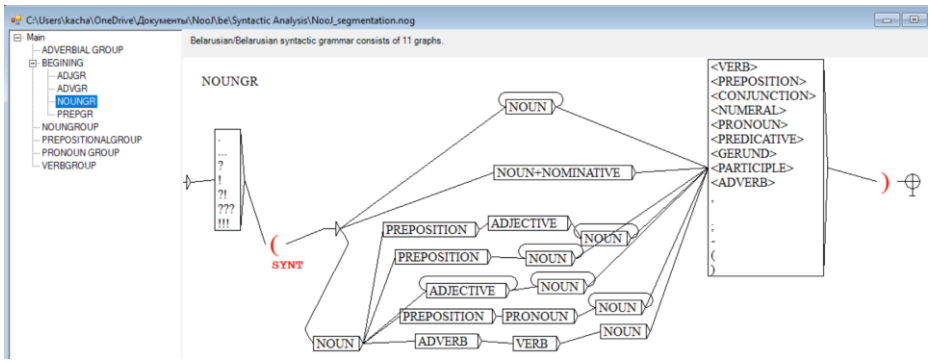

Fig. 8. Syntactic grammar for extracting noun groups at the beginning of the sentence.

Figure 9 shows the Pronoun graph which consists of 6 subgraphs. The main principle of this grammar is the combination of pronouns as main words of syntagmas and their subordinate components. It's very important to consider the right and the left context (words/expressions) which surround this syntagma for delimiting its boundaries. For greater clarity let's consider the first subgraph. The left context can be a conjunction, a numeral adverb, or some punctuation marks represented in this figure. The syntagma is a combination of a pronoun, an adjective and a noun or some nouns that close it in case if the next sign is a verb, a preposition, a conjunction, a numeral, an adjective, a pronoun, a predicative, a gerund, a particle, an adverb, or any punctuation mark. The other 6 graphs have the same structure and principle of working out.

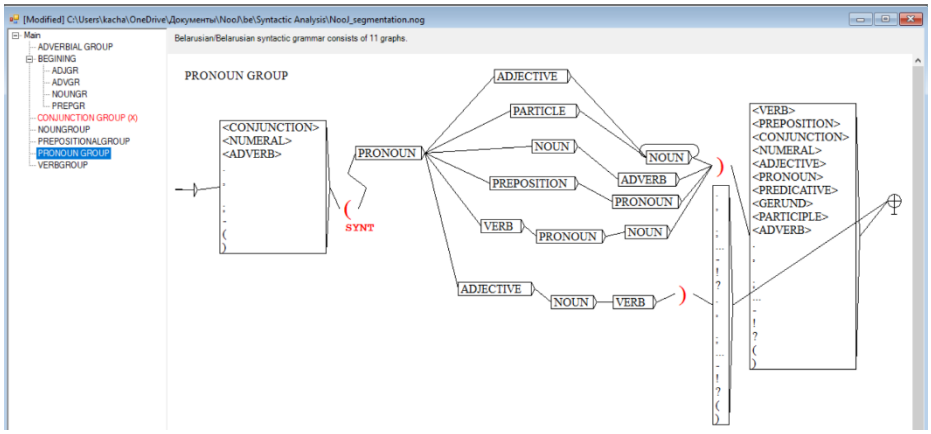

Fig. 9. The graph for extracting pronoun group of lexical syntagmas.

The results of applying syntactic grammar (in particular, noun group and prepositional group) are demonstrated in figure 10. It separates the left context, a syntagma and the right context. 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. Except this, 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 on a medical domain for searching new word combinations into syntagmas, their adding into graphsand correcting mistakes and bugs.

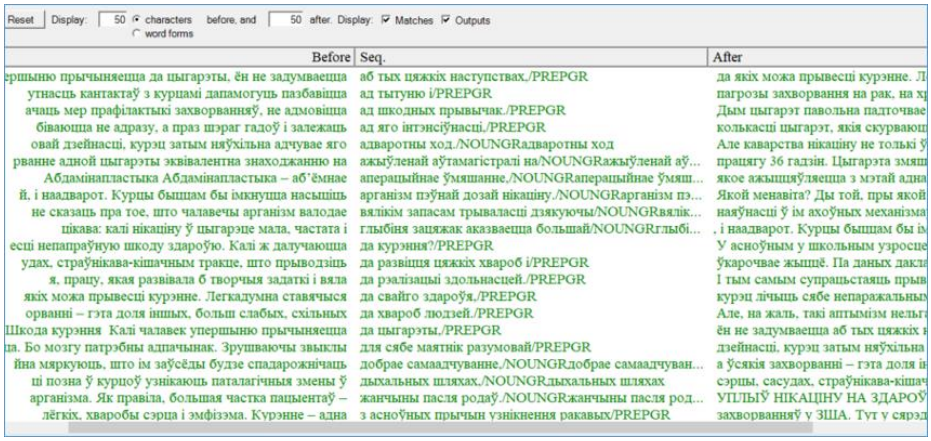

Fig. 10. Applying lexical grammars of noun and prepositional groups in the corpus of a medical domain.

```
Corpus contains 627 texts.

Corpus contains 1970166 characters.

Rank    Term    Frequency
1       юрыдычная дапамога;/NOUNGRюрыдычная дапамога    2
2       сацыяльнае суправаджэнне;/NOUNGRсацыяльнае суправаджэнне        2
3       эмацыянальная падтрымка;/NOUNGRэмацыянальная падтрымка  2
4       для асоб,/PREPGR        1
5       за дапамогай у/PREPGR   1
6       тэрытарыяльных цэнтраў сацыяльнага/NOUNGRтэрытарыяльных цэнтраў 1
7       сацыяльнага абслугоўвання насельніцтва вы/NOUNGRсацыяльнага абслугоўвання насельніцтва  1
8       часовы прытулак./NOUNGRчасовы прытулак  1
9       за дапамогай у рэлігійныя арганізацыі,/PREPGR   1
10      правядзення першаснай кансультацыі і/NOUNGRправядзення першаснай кансультацыі   1
11      тэлефона для асоб,/NOUNGRтэлефона для асоб      1
12      цэнтраў сацыяльнага абслугоўвання насельніцтва вы/NOUNGRцэнтраў сацыяльнага абслугоўвання насельніцтва  1
13      тлушчавых адкладаў,/NOUNGRтлушчавых адкладаў    1
14      грамадскіх арганізацый,/NOUNGRграмадскіх арганізацый    1
15      са стацыянарнага тэлефона /PREPGR       1
16      у рэлігійныя арганізацыі,/PREPGR        1
17      пакарання ў выглядзе штрафу альбо/NOUNGRпакарання ў выглядзе штрафу     1
18      Грамадскія арганізацыі./NOUNGRГрамадскія арганізацыі    1
19      па сацыяльнай рабоце для правядзення першаснай/PREPGR   1
20      стацыянарнага тэлефона /NOUNGRстацыянарнага тэлефона    1
21      арганізавана работа крызісных/NOUNGRарганізавана работа 1
22      па сацыяльнай рабоце./PREPGR    1
```

Fig. 11. Statistical analysis of applying lexical grammars in the corpus of a medical domain.

It is also planned to develop a syntactic grammar for extracting collocations and combine three types of grammars (punctuation, grammatical and lexical).

6 Conclusion

This paper represents an algorithm for extracting grammars and the grammar for automatic highlighting the intonation boundaries between syntagmas at the syntax level. The main core is the morphological and syntactic principle. The approach is confined in the ability of a particular speech part to match with words of other lexical and grammatical classes and occupy certain syntactic position. The concept is grounded in a superficial syntactic analysis of a text with the emphasis on grammatical characteristics of speech parts that combine accentual units. The results can be used for further research in phrase delimitation of Belarusian. Identified prosodic rules for dividing speech flow not only at punctuation level estimate value of intonation peculiarities of a certain language. This grammar is also wholesome to create an algorithm for segmenting textual information in the Belarusian speech synthesis systems.

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 5nd International Title Suppressed Due to Excessive Length 9 Workshop Speech and Computer SPECOM'2000. Revised selected papers, pp. 75- 78. St.-Petersburg, (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 3-rd International conference, Dresden, Germany, May 2–5, V. 2. Revised selected papers, pp.553-556. Dresden (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: Johanna Monti, Max Silberztein, Mario Monteleone and Maria Pia di Buono (eds.). Formalising Natural Languages with NooJ 2014. Revised selected papers, pp. 129-139. Newcastle : Cambridge Scholars Publishing (2015).
4. Hetsevich, Y., Okrut, T., Lobanov, B. Grammars for the Sentence into Phrase Segmentation: Punctuation Level. In: T. Okrut, Y. Hetsevich, M. Silberztein, H. Stanislavenka (eds.). International Scientific Conference on the Automatic Processing of Natural-Language Electronic Texts "NooJ'2015": Abstracts (11–13 June, 2015, Minsk, Belarus), Revised selected papers, pp.74-82. Springer International Publishing (2016).
5. NooJ: A Linguistic Development Environment. http://www.nooj-association.org. Last accessed 2021/02/18.
6. Silberztein, Max: Formalizing Natural Languages: the NooJ Approach. Wiley Eds. Hoboken, NJ, USA (2016)
7. Silberztein Max, 2003. NooJ Manual. Available for download at: http://www.nooj4nlp.org