Anita Bartulović · Linda Mijić · Max Silberztein
Editors

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

17th International Conference, NooJ 2023
Zadar, Croatia, May 31 – June 2, 2023
Revised Selected Papers

Springer

Editors
Anita Bartulović
University of Zadar
Zadar, Croatia

Linda Mijić
University of Zadar
Zadar, Croatia

Max Silberztein
Université de Franche-Comté
Besançon, France

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-56645-5 ISBN 978-3-031-56646-2 (eBook)
https://doi.org/10.1007/978-3-031-56646-2

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

Paper in this product is recyclable.

Contents

Morphological and Lexical Resources

Syntactic and Semantic Resources

Recognition of Frozen Expressions in Belarusian NooJ Module

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

United Institute of Informatics Problems of the National Academy of Sciences of Belarus,
Minsk, Belarus
ssrlab221@gmail.com

Abstract. The article describes the resources for the automatic extraction of phraseological units in Belarusian within the research of syntagmatic delimitation of Belarusian prosody using NooJ. It comprises the dictionary of Belarusian phrasemes in NooJ format and 12 syntactic grammars for automatically searching different types of frozen expressions (phrasemes, nominal, adverbial, verbal, adjectival and mixed frozen expressions). Their implementation is essential for the computerized search of different types of syntagms for automatic speech delimitation to improve applications with voice accompaniment in the Belarusian language.

Keywords: Frozen Expression · Phraseological Unit · Syntactic Grammar · Intonation · Syntagma · Prosodic Delimitation · Segmentation

1 Introduction

In modern science, the question of the initial structural unit and perception of speech has no unambiguous solution because of various approaches and principles. However, speech is syntagmatic and comprises lexical units that form syntagms. The author's delimitation and proper intonation provide an adequate perception of speech. But synthesized speech, presented in various applications with voice accompaniment, is absorbed as unnatural, illegible and inexpressive. The way to solve this problem is to develop specific methods and algorithms for analyzing and processing intonation features of natural speech, its automatic syntagmatic separation and the implementation of all intonation constructions of a given language in NooJ [1]. It will lead to the automated reproduction of arbitrary text in the manner of human reading.

In the previous stages of the research, we composed syntactic grammars for extracting syntagms at the punctuational and lexical levels, highlighting the intonation boundaries [2–5]. The next task is to form a syntactic grammar for delimiting phraseological units or frozen expressions [6]. These phenomena are reproducible; at least two-component linguistic units that combine with words of free use and are integral in meaning. As a rule, they are stable in their composition and structure (that is, idioms). However, due to structural units, many frozen expressions in Belarusian are complicated to identify. For instance, the composition of phraseological expressions can be replaced by synonyms or other separate words. Or combinations, where one of the components is used in a

A. Bartulović et al. (Eds.): NooJ 2023, CCIS 1816, pp. 64–75, 2024.
https://doi.org/10.1007/978-3-031-56646-2_6

phraseologically related sense and the other in a free one. Another problem is the order of units: in the phraseology, it can be fixed, or, more often, it is used in the reverse order.

So, the core of the research is to develop resources for searching and extracting phraseological units using NooJ [1]. Firstly, it is necessary to compile a phraseological dictionary of the Belarusian language in NooJ format based on the etymological dictionary of the Belarusian phraseological units and annotate it [7]. The next step is to build syntactic grammars for searching the most typical groups of frozen expressions. And finally, to test them on the literary corpus of the Belarusian NooJ module [1]. This will contribute to studying the automatic processing of phraseological units for further extraction into separate syntagms and forming their intonation portraits.

2 Syntactic Grammars for Extracting Punctuation and Lexical Syntagms

Our research on Belarusian prosodic segmentation focuses on three main problems: (1) the absence of research in the field of Belarusian prosody and intonology (practically unexplored); (2) the lack of automatic prosodic segmentation into syntagms; (3) the lack of deep syntactic parsing for the automatic selection of syntagms. They lead to the search for new approaches to developing machine algorithms, methods and techniques by defining sequences of linguistic elements associated with certain semantic relationships.

There are no general rules for syntagma extraction in Belarusian speech. However, the statistical analysis results based on experimental data provide a basis for developing a general algorithm for syntagma delimitation. The system planned for finding the intonational boundaries of syntagms is based on a superficial parsing with emphasis on the grammatical features of the part of speech (POS) categories. The primary task of this research is to develop formal syntactic grammar rules and algorithms that divide sentences into syntagms. To implement the algorithm, it is necessary to consider all punctuation marks' phraseological units and to create a list of formal rules for splitting sentences into lexical syntax.

When dividing a text into syntagms, one should consider the following aspects: sentence structure, word order, the presence of members of the same kind, the nature of word combinations and other linguistic parameters. In addition, each language has its own rules regarding syntactic relations and their application. Most sentences can be read purely syntactically, based on a superficial syntactic structure fully indicated by punctuation in Belarusian texts. However, syntactic structure alone may not be sufficient for correct delimitation, especially when the context is ambiguous due to the diversity of writing styles and genres.

As was noted in previous papers [2–5], we have distinguished three groups of syntagms for this project: punctuation, grammatical and lexical.

A *punctuation syntagma* (PS) refers to a sentence or part of a sentence that is limited to punctuation marks. A *grammatical syntagma* (GS) marks stable word combinations (phraseological units and collocations). A *lexical syntagma* (LS) is a short sentence of two or three 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 [8]. The task of this study is the correct extraction of all syntagms (PS, GS, LS) by developing, testing and improving syntactic grammars based on NooJ. For example,

- Дзе бацька той (PS): | недзе круціцца ў віры на калу (GS), | што ніхто не знае (PS), | дзе і што (PS). ‘Where is the father (PS): | somewhere spinning in a whirlpool of the feces (GS), | that no one knows (PS), | where and what he is doing (PS)’.
- Ён нібы з неахвотаю (LS) | прыклаў да яе спіны далоні (PS). ‘He seemed reluctant (LS) | to put his hands on her back (PS)’.
- Рыцар беларускасці (LS) | ніколі не ганяўся за славай (LS) | па прынцыпе пеўня (PS) | калі не даганю (PS), | дык хоць сагрэюся (PS). ‘The knight of Belarus (LS) | never chased fame (LS) | on the principle of the rooster (PS) | (so) if I don’t catch up (PS), | at least I’ll warm up (PS)’.

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 significant blocks according to the definition of syntagms (punctuation, grammatical, lexical) (Fig. 1).

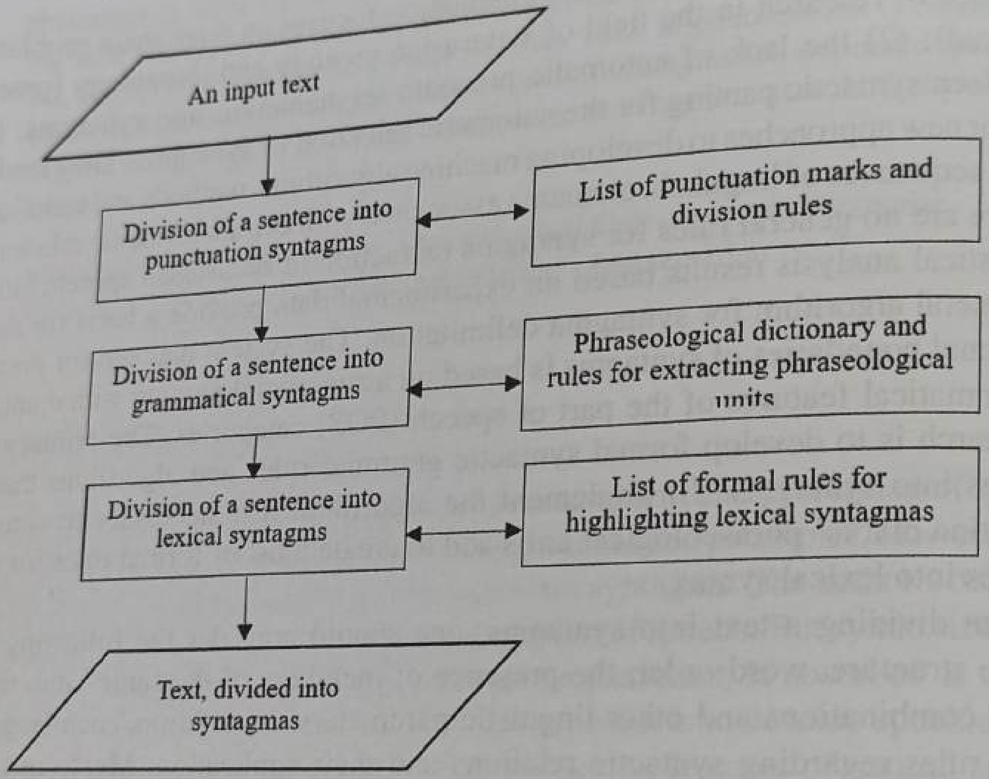

An input text
Division of a sentence into punctuation syntagms
List of punctuation marks and division rules
Division of a sentence into grammatical syntagms
Phraseological dictionary and rules for extracting phraseological units
Division of a sentence into lexical syntagms
List of formal rules for highlighting lexical syntagmas
Text, divided into syntagmas

Fig. 1. An algorithm for automatically searching syntagms in the Belarusian language.

So, at previous stages, we developed syntactic grammars for exporting punctuation and lexical syntagms. Syntactic grammar for extracting punctuation syntagms is an automatic phrase segmentation technique at the punctuation level that marks phrase intonation types in Belarusian electronic texts using NooJ [2, 3]. This tool separates syntagms, drawing on the syntactic structure of a sentence and punctuation. It may be helpful to improve the Belarusian NooJ module for so-called prosodic transcription. The main drawback of grammar is the selection of phrases or sentence fragments only by punctuation marks. Without punctuation marks, the system highlights long phrases that are not syntagms. Syntactic grammar for automatically extracting lexical syntagms solves this problem [2, 4, 5]. The central core of grammar is a morphological and syntactic principle that lies in the ability of a particular POS category to agree with other words and word forms and occupy a specific position in a sentence. Its concept is based on a superficial syntactic analysis of different texts (based on corpora in the literary and medical domains), emphasizing the grammatical features of the parts of speech that combine the accusative units. To complete the algorithm presented in Fig. 1, it is necessary to develop a methodology for extracting phraseological units in the text, which is what the research aims at.

3 Phraseological Units of the Belarusian Language

According to NooJ terminology, *Atomic Linguistic Units* (ALUs) are the smallest elements that make up the sentence, i.e., the non-analyzable units of the language (simple words, affixes, multiword units, frozen expressions) [6]. *Frozen expressions* (FE) are ALUs spelled as potentially discontinuous sequences of word forms. Frozen expressions or phraseological units are reproducible, at least two-component linguistic units that combine with words of free use and are integral in meaning.

For the Belarusian language, there is only one source of phraseological units: the *Etymological Dictionary of Phraseological Units* by Lepeshau in 2 volumes (editions of 1981 and 1993) [7]. It received a historical and etymological reference of more than 1,300 phraseological units. Many dictionary entries, especially in the 1981 edition, have been corrected, supplemented or shortened. The books combine the old and new etymologies, and reveal the origin of about 1,750 phraseological units of the modern Belarusian language.

Within the research, the most frequent types of FE were singled out. They are:

a. phrasemes,
b. FE with lexical variation,
c. FE with obligatory right context,
d. FE with limited meaning,
e. FE with mandatory left and right context,
f. FE similar to free-word combinations.

A phraseme is a semantically indivisible unit, the meaning of which is wholly inferred from the sum of the values of its components. Their semantic independence is entirely lost. They are idioms, collocations, clichés, pragmatemes, e.g.,

- З вышыні птушынага палёту ‘a bird’s eye view’
- Да мозгу касцей ‘to the core’
- З агню ды ў полымя ‘from fire to flame’
- Абое рабое ‘the kettle calling the pot black’

The group of phrasemes can be characterized by lexical variation. It means that one word can be replaced by its synonym. For example,

- Выбываць (*выбыць; выходзіць, выйсці*) са строю ‘drop out (exit) from the line’
- Выводзіць (*вывесці; спісваць, спісаць*) у расход каго ‘take out (write off) at the someone’s expense’
- Порах(-у) не выдумляць (*не выдумаць*). ‘Do not invent gunpowder.’

- Праз (*скрозь*) зубы ‘through (thru) the teeth’
- Заставацца пры сваіх інтарэсах (*пры сваім інтарэсе*) ‘remain in one’s interests’

Many phraseological units cannot be used without an obligatory object environment (“right context”). Their meaning is realized only in a strictly defined context. For example,

- Адальюцца слёзы *каму чые*. ‘Someone’s tears will go away.’
- Зуб за зуб зайшоў *у каго, з кім*. ‘Tooth for tooth went to someone.’
- Клюнуць на вудачку *чыю, каго, чаго, якую* ‘to peck with someone else’s fishing rod’

Some phraseological units have valently limited meanings. It is expressed only in a combination of phraseology with strictly defined words. For example, the expression куры не клююць ‘a lot of’ comes into contact only with the word грошы ‘money’.

There are phraseological units with double obligatory right and left context, e.g.,

- *Станавіцца/стаць* і *пад/на* роўную нагу з кім ‘to be on an equal footing with someone’
- Як *што/чаго* хоча (захоча) левая нага *каго, чыя*. ‘What does/will someone’s left leg want.’
- *<Адны>*скура ды косці засталіся *на кім, у каго, ад каго*. ‘<Only>skin and bones remained on someone, from someone.’

The last group can be confused with general word combinations. Their meaning may be interpreted only due to their semantics. Let us analyze the phrase божая кароўка’ ‘ladybug’ in the next sentences. Its general meaning is ‘red, yellow or white speckled bug’. Its FE meaning is ‘a quiet, harmless person who does not know how to stand up for themself’.

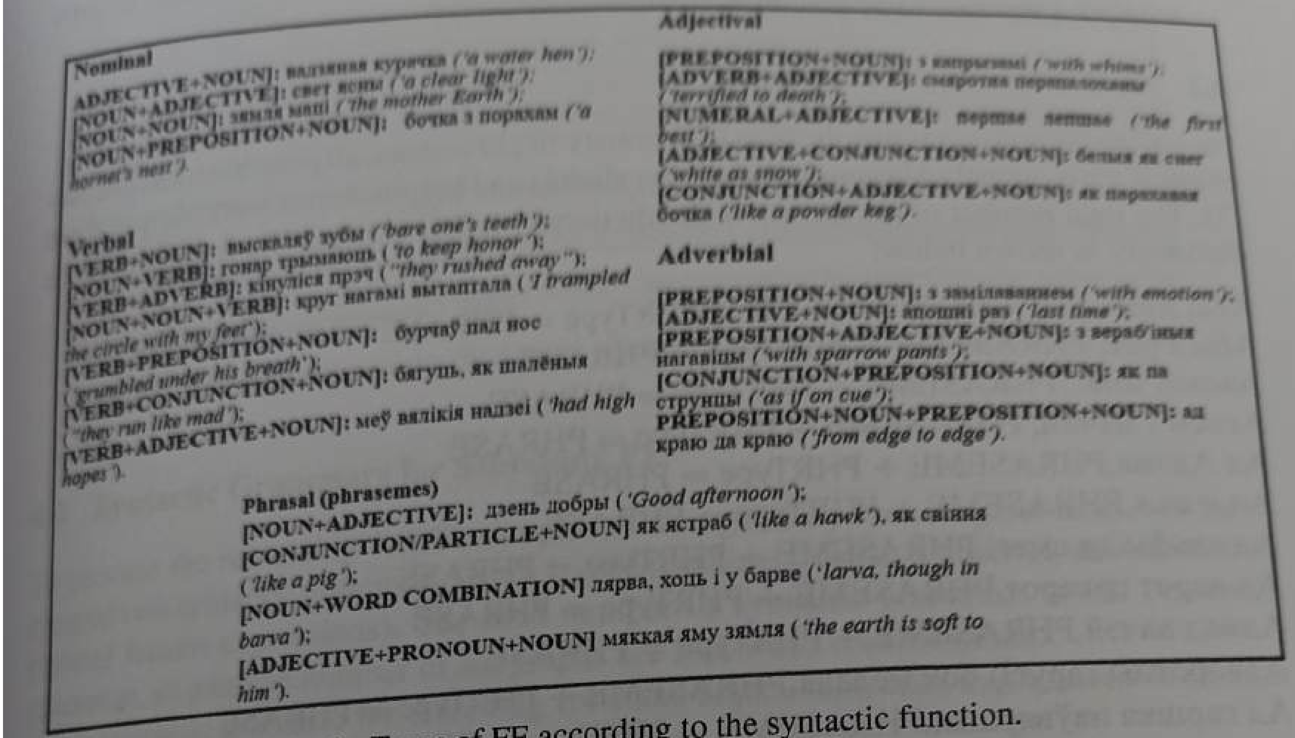
- Прыгрэтая на падаконніку божая кароўка раптам заварушылася і пацёпала сваімі чырвонымі падкрылкамі. ‘The **ladybug** warmed on the windowsill, suddenly stirred and fluttered its red wings’.
- Ня можа быць, каб ён быў шпіён, ён жа божая кароўка. ‘He can’t be a spy, he’s **a ladybug**.’

The first sentence is used in general meaning, and the second is FE. Only the context and the meaning of the phrase can indicate whether it is a FE of a free word combination. Other examples of this type are represented below:

- Агульнае месца ‘common place’
- Ад рукі пісаць ‘write by hand’
- Без гальштукаў ‘no ties’

As a part of the complex analysis of phraseological units, some specific features of phrasemes in Belarusian were identified. Firstly, the word order of some combinations can vary (гонар трымаць = трымаць гонар ‘to keep honor’). Secondly, some phrasemes admit lexical insertions (у той жа момант –>у той жа [самы] момант ‘at the same moment’). Finally, one or two lexical elements can often change their form. In the case of nominal phrasemes (multiword units) of the type [ADJECTIVE + NOUN], both elements are declined: вадзяная курачка, вадзяных курачак ‘water hen’.

Also, we emphasized the types of FE according to the syntactic function: nominal, verbal, adjectival, adverbial and phrasal frozen expressions. This classification depends on the main component of FE (Fig. 2).



Nominal
ADJECTIVE+NOUN]: вадзяная курачка (*'a water hen'*);
[NOUN+ADJECTIVE]: свет ясны (*'a clear light'*);
[NOUN+NOUN]: зямля маці (*'the mother Earth'*);
[NOUN+PREPOSITION+NOUN]: бочка з порахам (*'a hornet's nest'*).

Verbal
[VERB+NOUN]: выскаляў зубы (*'bare one's teeth'*);
[NOUN+VERB]: гонар трымаюць (*'to keep honor'*);
[VERB+ADVERB]: кінуліся прэч (*"they rushed away"*);
[NOUN+NOUN+VERB]: круг нагамі вытаптала (*'I trampled the circle with my feet'*);
[VERB+PREPOSITION+NOUN]: бурчаў пад нос (*'grumbled under his breath'*);
[VERB+CONJUNCTION+NOUN]: бягуць, як шалёныя (*"they run like mad'*);
[VERB+ADJECTIVE+NOUN]: меў вялікія надзеі (*'had high hopes'*).

Adjectival
[PREPOSITION+NOUN]: з капрызамі (*'with whims'*);
[ADVERB+ADJECTIVE]: смяротна перапалоханы (*'terrified to death'*);
[NUMERAL+ADJECTIVE]: першая лепшая (*'the first best'*);
[ADJECTIVE+CONJUNCTION+NOUN]: белыя як снег (*'white as snow'*);
[CONJUNCTION+ADJECTIVE+NOUN]: як пароховая бочка (*'like a powder keg'*).

Adverbial
[PREPOSITION+NOUN]: з замілаваннем (*'with emotion'*);
[ADJECTIVE+NOUN]: апошні раз (*'last time'*);
[PREPOSITION+ADJECTIVE+NOUN]: з вераб'іныя нагавіцы (*'with sparrow pants'*);
[CONJUNCTION+PREPOSITION+NOUN]: як па струнцы (*'as if on cue'*);
PREPOSITION+NOUN+PREPOSITION+NOUN]: ад краю да краю (*'from edge to edge'*).

Phrasal (phrasemes)
[NOUN+ADJECTIVE]: дзень добры (*'Good afternoon'*);
[CONJUNCTION/PARTICLE+NOUN] як ястраб (*'like a hawk'*), як свіння (*'like a pig'*);
[NOUN+WORD COMBINATION] лярва, хоць і у барве (*'larva, though in barva'*);
[ADJECTIVE+PRONOUN+NOUN] мяккая яму зямля (*'the earth is soft to him'*).

Fig. 2. Types of FE according to the syntactic function.

Thus, the next step of the research after classifying the types of frozen expressions is an application of NooJ as a suitable software product for the automatic extraction of phraseological units.

4 Syntactic Grammars for Extracting FE in NooJ

To fulfill the automatic extraction of frozen expressions, NooJ offers two ways to format frozen expressions. NooJ's syntactic grammars can represent and recognize all possible contexts for a given expression. Also, NooJ allows linguists to link a dictionary describing the possible components of a frozen expression with a grammar describing its syntactic behavior. We combined two approaches according to the variety of FE types. There is a dictionary of phrasemes as well as a limited number of syntactic grammars for searching frozen expressions represented in the article. Two main strategies were realized within the research:

a. NooJ Dictionary of phrasemes (idioms, collocations, clichés, pragmatemes) was collected and compiled. The reason is that there are a lot of phrasemes that do not vary grammatically, and do not admit any insertion. For such phrasemes, one must construct a dictionary, and there is no need to elaborate a local grammar. For example,

Кожнаму свая шкура даражэй,PHRASEME + PHRType = PHRASE 'Everyone's skin is more expensive.'

b. We developed syntactic grammars for searching nominal, adverbial, verbal, adjectival and mixed frozen expressions. Some phrasemes have a similar structure (as ад краю да краю 'from stern to stern', ад цямна да цямна 'from dawn to dusk'), and can admit lexical insertions. They should be organized in groups, and each of these groups requires a separate local grammar.

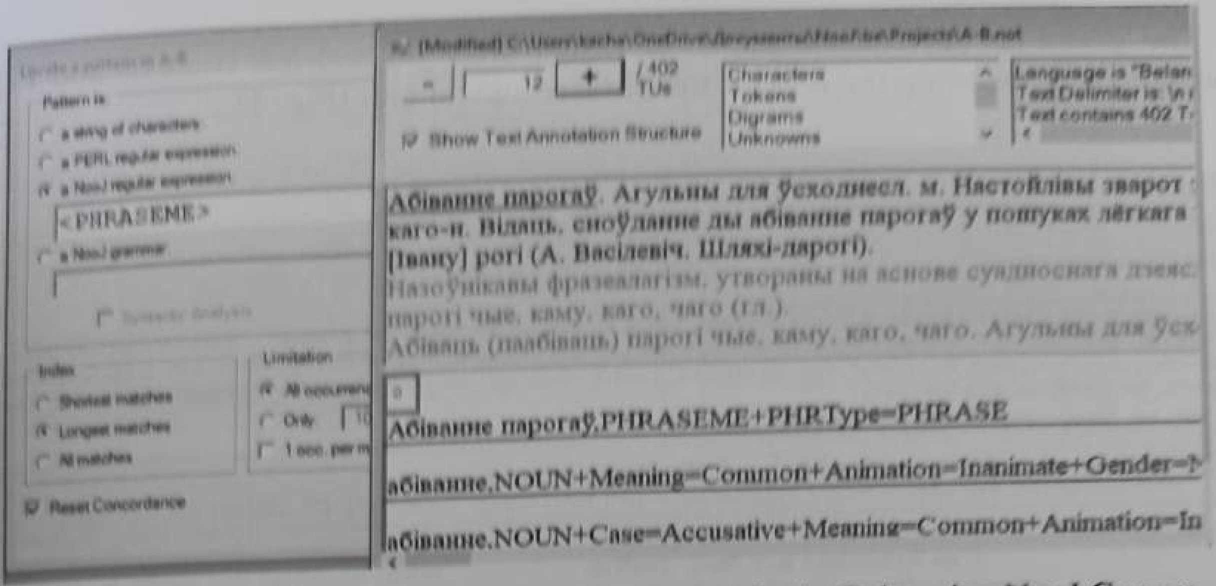
4.1 NooJ Dictionary of Phrasemes

For collecting and compiling the NooJ Dictionary of phrasemes, all possible fixed phraseological units, namely idioms, collocations, clichés and pragmatemes were chosen from [7]. The total number of phrasemes in the dictionary is 760 entries. The fragment of the dictionary is shown below:

Абы дзень давечара,PHRASEME + PHRType = PHRASE
Абы з рук, PHRASEME + PHRType = PHRASE
Авохці мне!, PHRASEME + PHRType = PHRASE
Агнём і мячом, PHRASEME + PHRType = PHRASE
Ад Адама,PHRASEME + PHRType = PHRASE
Ад а да я,PHRASEME + PHRType = PHRASE
Ад альфы да амегі,PHRASEME + PHRType = PHRASE
Ад варот паварот,PHRASEME + PHRType = PHRASE
Адвод вачэй,PHRASEME + PHRType = PHRASE
Адваротны (другі) бок медаля,PHRASEME + PHRType = PHRASE
Ад гаршка паўвяршка,PHRASEME + PHRType = PHRASE
Адданне чэсці,PHRASEME + PHRType = PHRASE
Ад дошкі да дошкі,PHRASEME + PHRType = PHRASE

We tested the dictionary of phrasemes by applying it to NooJ Belarusian Corpus "Kalasy 01_12". Locating the pattern "PHRASE" produces the list of frozen phrasemes as output. The function "Show Text Annotation Structure" confirms the use of the dictionary (Fig. 3).

NooJ cannot identify frozen expressions that contain one interchangeable component that depends on the author's usage, e.g., Ва ўсякім (у кожным) выпадку '**In any** (every) case'. The same is true for the use of different cases for nouns and tenses for the verb, e.g., Шмат (многа, нямала, колькі, столькі) вады сплыло (сплыве) '**a lot** of (**how much, so much**) water **has floated away (will float away)**'. A significant problem in using a dictionary is the insertion of an additional word inside a phraseme, e.g., Вось табе<бабка>і Юр'еў дзень! 'Here's to you<**grandma**>and St. George's Day!' The last one is the elimination of an object of action in phraseologism in the postposition of a transitive/intransitive verb with a controlling preposition, e.g., Волас з галавы не ўпадзе ў каго, з чыёй. 'A hair will not fall **off/from anyone's** head.' We solved this problem by constructing a separate syntactic grammar for similar phraseological units.



Pattern is:
○ a string of characters
○ a PERL regular expression
● a NooJ regular expression
<PHRASEME>
○ a NooJ grammar

Index
○ Shortest matches
● Longest matches
○ All matches

Limitation
● All occurrences
○ Only
☐ 1 occ. per m

☑ Reset Concordance

(Modified) C:\Users\kacha\OneDrive\Документы\NooJ\be\Projects\A-B.not

12 / 402 TUs

☑ Show Text Annotation Structure

Characters
Tokens
Digrams
Unknowns

Language is "Belar
Text Delimiter is: \n
Text contains 402 T

Абіванне парогаў. Агульны для ўсходнесл. м. Настойлівы зварот
каго-н. Віляць, сноўданне ды абіванне парогаў у пошуках лёгкага
[Івану] рогі (А. Васілевіч. Шляхі-дарогі).
Назоўнікавы фразеалагізм, утвораны на аснове суаднoснага дзеяс
парогі чые, каму, каго, чаго (гл.).
Абіваць (паабіваць) парогі чые, каму, каго, чаго. Агульны для ўс

Абіванне парогаў,PHRASEME+PHRType=PHRASE

абіванне,NOUN+Meaning=Common+Animation=Inanimate+Gender=

абіванне,NOUN+Case=Accusative+Meaning=Common+Animation=In

Fig. 3. An application of the dictionary of phrasemes in the Belarusian NooJ Corpus.

4.2 Syntactic Grammars for Searching FE

To process the remaining 1,000 examples of changeable phraseological units, we have created two syntactic grammars for each type (nominal, verbal, adjectival, adverbial and phrasal frozen expressions). The total number of syntactic grammars is 12. For each grammar, an average number of subgraphs are 5–8-word combinations.

A syntactic grammar for searching for some adjectival frozen expressions is shown in Fig. 4. It includes such phraseological units as:

•Заднім розумам моцны. ‘The hindsight is strong.’
•І жук і жаба ‘and the beetle and the frog’
•Лёд разбіты/паламаны ‘to break the ice’
•Мамчын/мамін сынок ‘mother’s/mum’s son’

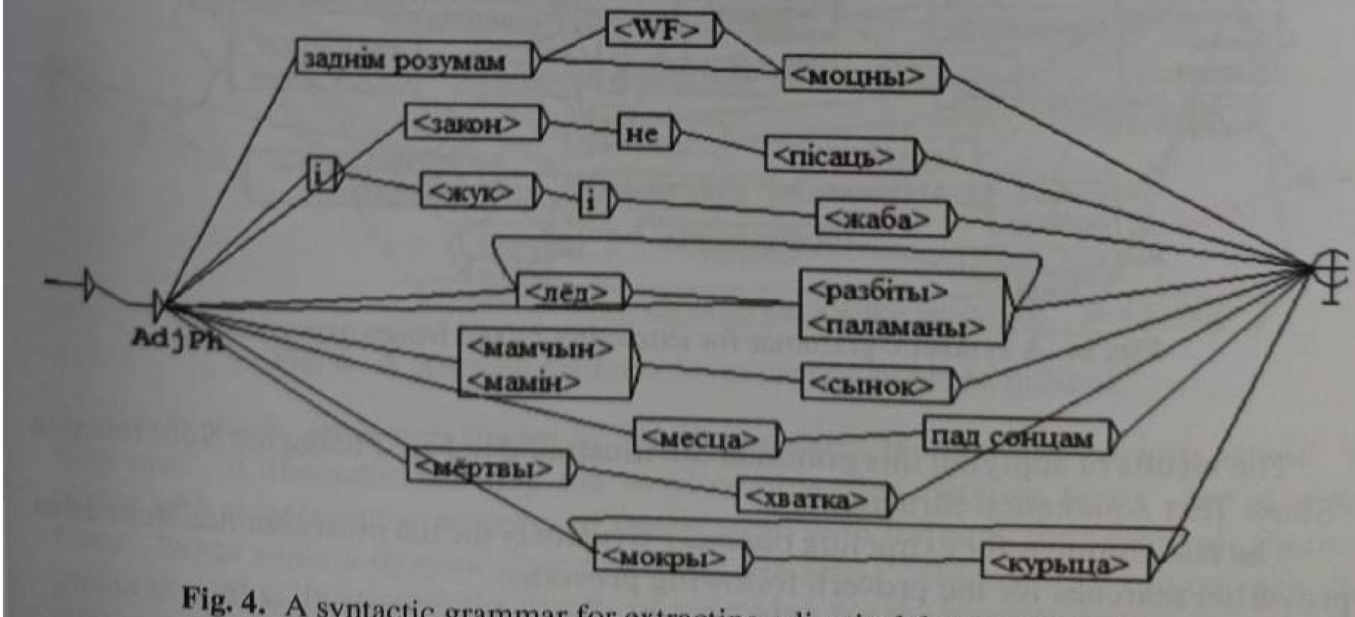
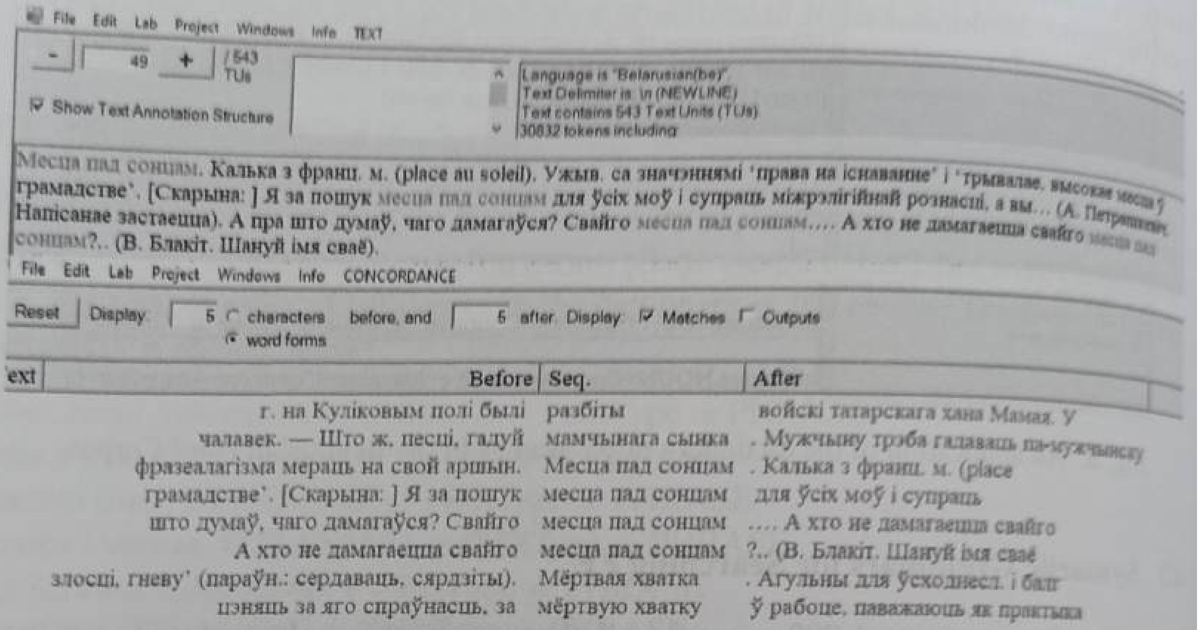


Fig. 4. A syntactic grammar for extracting adjectival frozen expressions.

The grammar depicts subgraphs in which the right and left contexts and FE are represented, preserving the core of the phraseological unit (Fig. 5).


File Edit Lab Project Windows Info TEXT
- 49 + / 543 TUs
☑ Show Text Annotation Structure

Language is "Belarusian(be)"
Text Delimiter is: \n (NEWLINE)
Text contains 543 Text Units (TUs)
30832 tokens including

Месца пад сонцам. Калька з франц. м. (place au soleil). Ужыв. са значэннямі 'права на існаванне' і 'трывалае, высокае месца ў грамадстве'. [Скарына:] Я за пошук месца пад сонцам для ўсіх моў і супраць міжрэлігійнай рознасці, а вы… (А. Петрашкевіч. Напісанае застаецца). А пра што думаў, чаго дамагаўся? Свайго месца пад сонцам,… А хто не дамагаецца свайго месца пад сонцам?.. (В. Блакіт. Шануй імя сваё).

Display 5 ○ characters ● word forms before, and 5 after. Display: ☑ Matches ☐ Outputs

ext	Before	Seq.	After
	г. на Куліковым полі былі	разбіты	войскі татарскага хана Мамая. У
	чалавек. — Што ж, песці, гадуй	мамчынага сынка	. Мужчыну трэба гадаваць па-мужчынску
	фразеалагізма мераць на свой аршын.	Месца пад сонцам	. Калька з франц. м. (place
	грамадстве'. [Скарына:] Я за пошук	месца пад сонцам	для ўсіх моў і супраць
	што думаў, чаго дамагаўся? Свайго	месца пад сонцам	…. А хто не дамагаецца свайго
	А хто не дамагаецца свайго	месца пад сонцам	?.. (В. Блакіт. Шануй імя сваё
	злосці, гневу' (параўн.: сердаваць, сярдзіты).	Мёртвая хватка	. Агульны для ўсходнесл. і балт
	цэняць за яго спраўнасць, за	мёртвую хватку	ў рабоце, паважаюць як практыка

Fig. 5. An output of syntactic grammar for extracting adjectival frozen expressions.

The following syntactic grammar searches mixed verbal and adjectival frozen expressions with the main component вочы 'eyes' (Fig. 6). For example,

- Вочы вялікія адкрыць/раскрыць/адчыняць/рабіць **(verbal FE)** 'eyes wide open'
- Зрабіць вялікія вочы (**verbal FE**) 'make big eyes'
- Вочы на мокрым месцы (**adjectival FE**) 'eyes on a wet place'

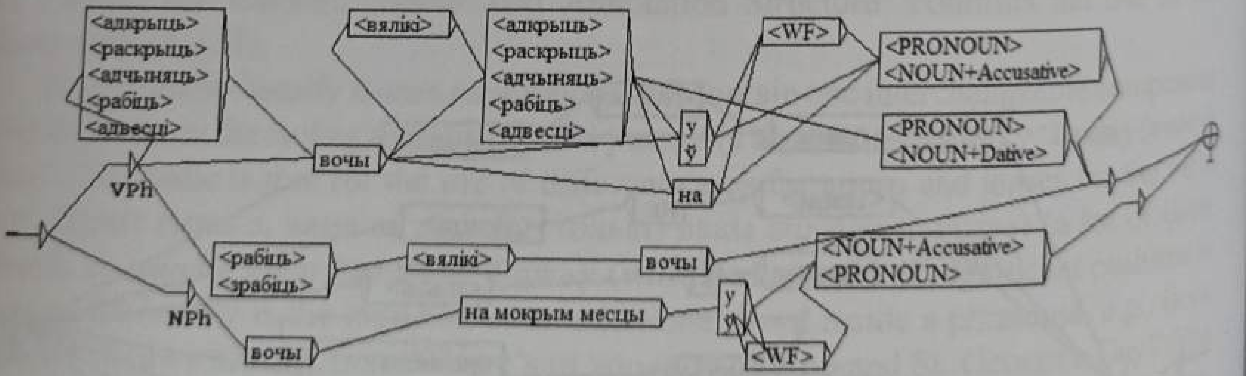
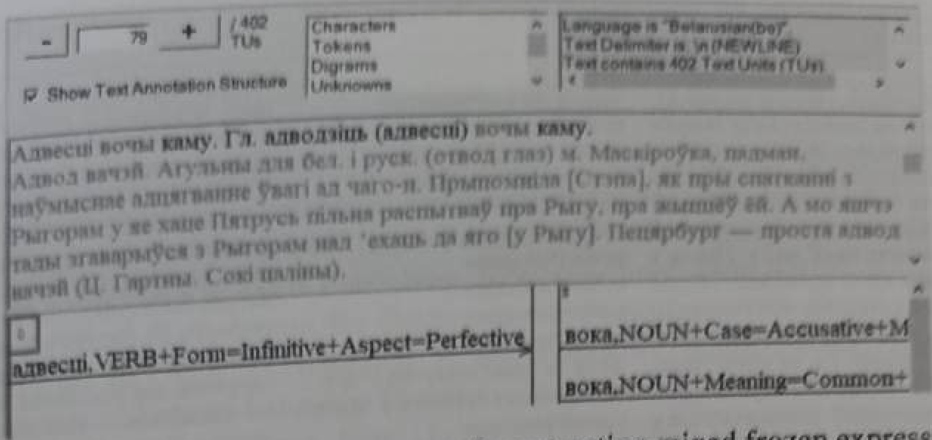


Fig. 6. A syntactic grammar for extracting mixed frozen expressions.

The results of applying this grammar are illustrated in Fig. 7 using the NooJ function "Show Text Annotation Structure".

The last grammar for extracting phrasal FE (namely the ten most common Belarusian proverbs) searches for the proverb following proverbs:

- Да Абрама на піва трапіць. 'Go to Abram for a beer.'
- Дзяліць скуру незабітага мядзведзя. 'Split the skin of an unkilled bear.'



Language is "Belarusian(be)"
Text Delimiter is: \n (NEWLINE)
Text contains 402 Text Units (TUs)

Адвесці вочы каму. Гл. адводзіць (адвесці) вочы каму.
Адвод вачэй. Агульны для бел. і рус. (отвод глаз) м. Маскіроўка, падман.
наўмыснае адцягванне ўвагі ад чаго-н. Прыпомніла [Стэпа], як пры спатканні з
Рыгорам у яе хаце Пятрусь пільна распытваў пра Рыгу, пра жыццё ёй. А мо яго
тады згаварыўся з Рыгорам над 'ехаць да яго [у Рыгу]. Пецярбург — проста адвод
вачэй (Ц. Гартны. Сокі цаліны).

адвесці,VERB+Form=Infinitive+Aspect=Perfective
вока,NOUN+Case=Accusative+M
вока,NOUN+Meaning=Common+

Fig. 7. An output of syntactic grammar for extracting mixed frozen expressions.

- Забіць двух зайцаў. 'Kill two birds with one stone.'
- Узяцца за гуж 'to take the buzz'
- Ухваціцца за саломінку 'grasping at straws'
- З'есці пуд солі 'to eat a pinch of salt'
- З мухі зрабіць слана 'to make an elephant out of a fly'
- Есці з сямі печаў хлеб 'to eat bread from seven ovens'
- Калоць вочы 'to sting eyes'
- Лезці са сваім статутам у чужы манастыр 'to go with your charter to someone else's monastery'

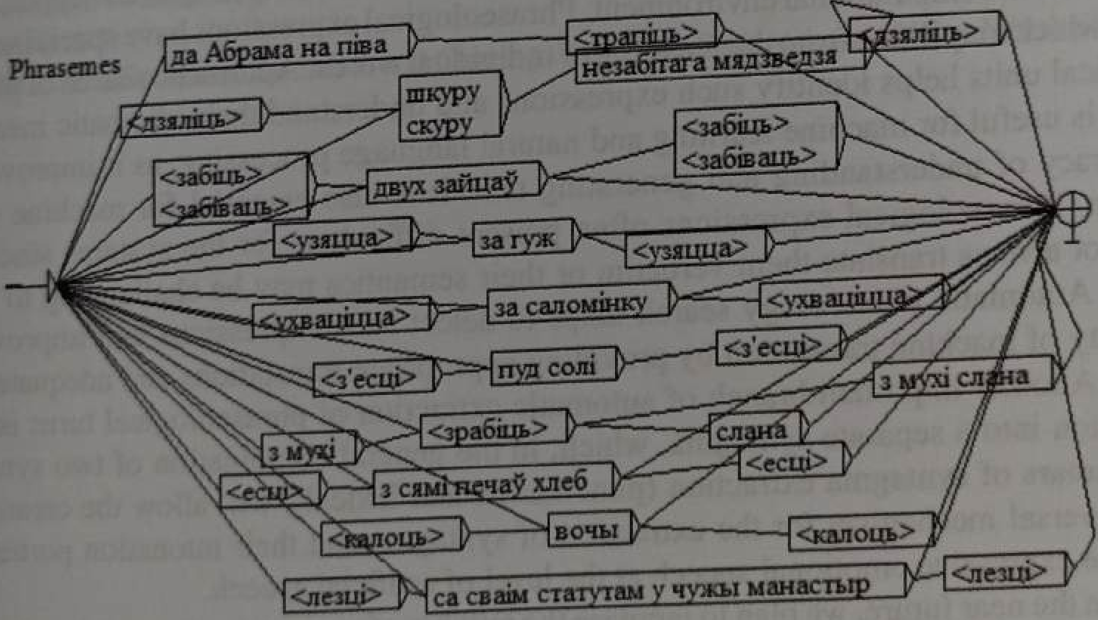
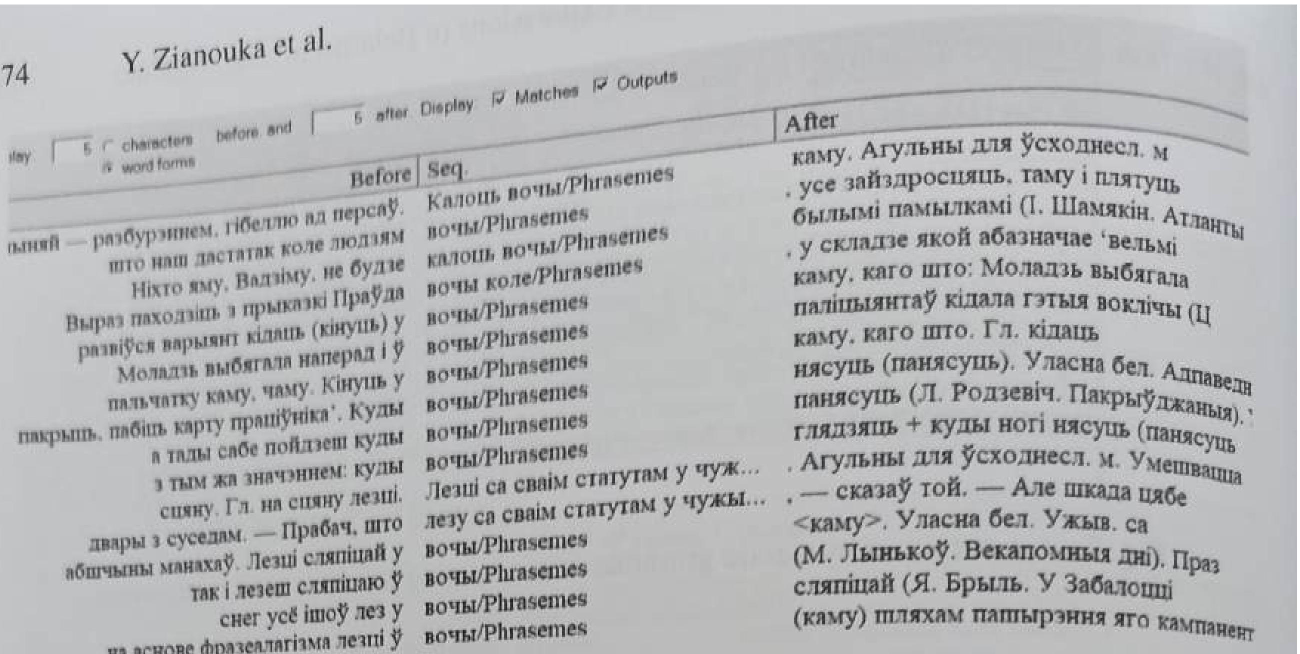


Fig. 8. A syntactic grammar for extracting proverbs in NooJ.

The syntactic grammar, shown in Fig. 8, indicates the proverb 'калоць вочы' 'to sting eyes'. It illustrates not only the word combination 'калоць вочы', but also the phrase with an opposite word order of this proverb with different time intervals ('калоў вочы', 'вочы калоць будуць', 'не будзе калоць вочы') (Fig. 9). As you can see, the grammar works, illustrating the results of searching for the proverb 'калоць вочы' 'to sting eyes'.



lay 5 ○ characters ● word forms before and 5 after Display ☑ Matches ☑ Outputs

Before	Seq.	After
ынай — разбурэннем, гібеллю ад персаў.	Калоць вочы/Phrasemes	каму. Агульны для ўсходнесл. м
што наш дастатак коле людзям	вочы/Phrasemes	. усе зайздросцяць, таму і плятуць
Ніхто яму, Вадзіму, не будзе	калоць вочы/Phrasemes	былымі памылкамі (І. Шамякін. Атланты
Выраз паходзіць з прыказкі Праўда	вочы коле/Phrasemes	. у складзе якой абазначае ‘вельмі
развіўся варыянт кідаць (кінуць) у	вочы/Phrasemes	каму, каго што: Моладзь выбягала
Моладзь выбягала наперад і ў	вочы/Phrasemes	паліцыянтаў кідала гэтыя воклічы (Ц
пальчатку каму, чаму. Кінуць у	вочы/Phrasemes	каму, каго што. Гл. кідаць
пакрыць, пабіць карту праціўніка’. Куды	вочы/Phrasemes	нясуць (панясуць). Уласна бел. Адпаведн
а тады сабе пойдзеш куды	вочы/Phrasemes	панясуць (Л. Родзевіч. Пакрыўджаныя). ’
з тым жа значэннем: куды	вочы/Phrasemes	глядзяць + куды ногі нясуць (панясуць
сцяну. Гл. на сцяну лезці.	Лезці са сваім статутам у чуж...	. Агульны для ўсходнесл. м. Умешваца
двары з суседам. — Прабач, што	лезу са сваім статутам у чужы...	. — сказаў той. — Але шкада цябе
абшчыны манахаў. Лезці сляпіцай у	вочы/Phrasemes	<каму>. Уласна бел. Ужыв. са
так і лезеш сляпіцаю ў	вочы/Phrasemes	(М. Лынькоў. Векапомныя дні). Праз
снег усё ішоў лез у	вочы/Phrasemes	сляпіцай (Я. Брыль. У Забалоцці
на аснове фразеалагізма лезці ў	вочы/Phrasemes	(каму) шляхам пашырэння яго кампанент

Fig. 9. An output of syntactic grammar for searching the proverb ‘калоць вочы’.

5 Conclusion

Within the current research, we performed several essential tasks: we created the NooJ dictionary of Belarusian phrasemes (nearly 760 entries), and developed 12 complex syntactic grammars for searching nominal, adverbial, verbal, adjectival and mixed frozen expressions. We verified them using NooJ corpus “Kalasy 01-12.noc”.

An automatic search of phraseological units has several advantages and applications in a modern computational environment. Phraseological expressions have special semantics, which may differ from the meaning of individual words. Automatic search of phraseological units helps identify such expressions and understand their idiomatic meaning. This is useful for machine learning and natural language processing, as it improves the accuracy of understanding and generating text. It is also essential for machine translation. Phraseological expressions often present difficulties for the systems since one cannot always translate them verbatim or their semantics may be challenging to interpret. Automatic phraseology search helps to detect such expressions and improve the quality of machine translation by providing appropriate translations and adequate context. Another important branch of automatic extraction of phraseological turns is their division into a separate syntagma, which, in the general combination of two syntactic grammars of syntagma extraction (punctuation and lexical), will allow the creating of a universal mechanism for the extraction of syntagms and their intonation portraits to create expressive emotional speech at the level of artificial speech.

In the near future, we plan to increase the number of syntactic grammars by choosing the most popular and typical FE for Belarusians. Another one is to unify syntactic grammars for extracting punctuation, grammatical and lexical syntagms in one syntactic grammar as a single algorithm for automatic syntagma extraction. The results obtained will be used for further research in the automatic processing of Belarusian prosodic structure, in particular for computer systems with voice accompaniment.

References

1. NooJ: A Linguistic Development Environment, http://www.nooj4nlp.org/. Last accessed 21 July 2022
2. 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, pp. 129–139. Cambridge Scholars Publishing, Newcastle upon Tyne (2015)
3. Hetsevich, Y., Okrut, T., Lobanov, B.: Grammars for the sentence into phrase segmentation: punctuation level. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H. (eds.), Automatic Processing of Natural-Language Electronic Texts with NooJ. 9th International Conference, NooJ 2015, Minsk, Belarus, June 11–13, 2015, Revised Selected Papers, pp. 74–82. Springer, Cham (2016)
4. 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
5. Zianouka, Y., Hetsevich, Y., Suprunchuk, M., Latyshevich D.: Prosodic segmentation of belarusian texts in NooJ. In: González, M., Reyes, S.S., Rodrigo, A., Silberztein, M. (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, pp. 50–62. Springer, Cham (2022)
6. Silberztein, M.: Formalizing Natural Languages: The NooJ Approach. Wiley-ISTE, London (2016)
7. Lepeshau, I.Y.: Etymological Dictionary of Phraseological Units. Bielaruskaja encyklapiedyja, Minsk (1993)
8. Lobanov, B.: Computer Synthesis and Cloning of Speech. Bielaruskaja navuka, Minsk (2008)