DESCRIBING SET AND FREE WORD COMBINATIONS IN BELARUSIAN AND RUSSIAN WITH NOOJ

YURY HETSEVICH, SVIATLANA HETSEVICH, ALENA SKOPINAVA, BORIS LOBANOV, YAUHENIYA YAKUBOVICH AND YURY KIM

Introduction

The goal of this research is to further improve the Belarusian and Russian dictionaries presented previously at the International NooJ Conferences (Hetsevich et al. 2013). A complete set of resources for a given language requires dictionaries and grammars that describe not only words but also word combinations. Hence it is planned to develop NooJ grammars and linguistic resources in order to identify and classify set word combinations (phrasemes) and syntax-free word combinations (including quantitative expressions with measurement units) in Belarusian and Russian texts.

Set word combinations can also be defined as *phrasemes* in the sense of the Meaning-Text Theory (Mel'čuk 2013), i.e. bound multiword phrases consisting of at least two lexemes, e.g.: *пойдзем сваёй дарогай* 'let's go our own way', *перайсці ўброд* 'to ford', *зоологический сад* 'zoological garden', *медовый месяц* 'honeymoon' etc. Phrasemes are indivisible multiword units and thus need to be processed as a whole by NooJ. To achieve this, we will describe them primarily in dictionaries and, if necessary, develop linked grammars.

Syntax-free word combinations are viewed as sequences of words bound together by means of standard language rules, e.g.: *новая кватэра* 'a new flat', *посматривал на часы* 'looked at the watch from time to time'. Words in syntax-free combinations can be easily replaced, e.g.: *новая (камфартабельная, сучасная…) кватэра* 'a new (comfortable, modern…) flat', *посматривал на часы (картину, девушку…)* 'looked at

the watch (painting, lady…) from time to time' etc. As these sequences are constructed ad hoc, they will be described primarily by means of syntactic grammars.

Special attention is given to processing of **quantitative expressions with measurement units (QEMUs)**, a specific subclass of syntax-free word combinations. For example, *Цягнік рухаўся з хуткасцю 200 км/г у Саарбрукен* 'The train was moving at a speed of 200 km/h to Saarbrücken'. The aim is to identify *200 км/г* '200 km/h', classify it as an expression with a SI-derived unit of speed, and turn it into orthographical words *дзвесце кіламетраў у гадзіну* 'two hundred kilometers per hour'. The problem of their processing is urgent due to the ubiquity of QEMUs, and at the same time it is not easy to solve because of their language-dependent character and the variety of ways in which QEMUs are expressed in writing. Texts containing QEMUs require resources for identification and processing in the following areas:

- Corpora and database management systems, libraries, information retrieval systems: to formulate extended search queries, locate specific expressions on the Internet, support automatic text annotation and summarization;
- Text-to-speech synthesis systems: to generate orthographically correct texts, their tonal and prosodic peculiarities;
- Publishing institutions: to automatically locate specified lists of expressions with measurement units and check quickly if the extended names of units are used correctly.

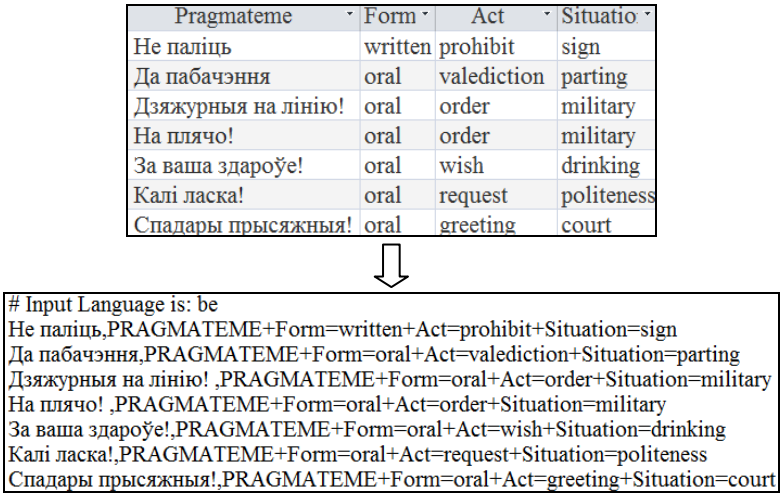
Building NooJ dictionaries and grammars for the above-mentioned problems will enable automatic recognition and annotation of these expressions in Belarusian and Russian texts. For instance, search engines should suggest the most common word combinations, and if they don't have sufficient statistical data, their suggestions can be based on resources describing structural types of word combinations. Such resources could be of great importance for text-to-speech synthesis. Speech synthesizers need to make pauses between free word combinations and set word combinations (as single units), but not inside of them (not between their components), which is impossible to accomplish if only basic dictionaries of separate tokens are applied. Such resources can be built in Belarusian and Russian NLP applications for many areas, e.g. syntactic parsing, prosody prediction, text-to-speech synthesis etc., and also as a didactic material for academic courses on computational linguistics and phraseology.

Identification of Set Word Combinations

Native speakers and those who seek to master a language often use phrasemes – bound multiword combinations. Therefore, a complete electronic dictionary should contain not only unigrams with necessary grammatical information but also syntactically and semantically combined multiword units. According to the NooJ terminology, we find two terms which are close to the specific term phraseme, namely frozen expressions and multiword units. The typology of phrasemes includes the classes of collocations, idioms, clichés and pragmatemes. These classes differ from each other mainly by the level and type of semantic fusion. However, phrasemes of all classes have been collected into the dictionaries regardless the degree of fusion.

Pragmatemes constitute a specific class of phrasemes and denote expressions which are formed according to grammar rules of a given language but with certain limitations. In a strictly defined situation, they convey a certain meaning, and only one of various grammatically and semantically possible expressions is used. For example, while riding a Belarusian or Russian bus (train etc.), one might hear *Асцярожна, дзверы зачыняюцца! Осторожно, двери закрываются!* (‘Attention, the doors are closing!’), but not *Увага, зараз я зачыняю дзверы! Внимание, закрываются двери!* (‘Attention, right now I’m closing the doors!’). Pragmatemes can be conveniently collected in a database and managed by means of MS Access. The pragmatemes databases are further converted into the dictionaries in the NooJ format (Figure 1).

Currently the dictionaries include over 300 Belarusian pragmatemes and over 170 Russian pragmatemes, subdivided into several categories according to the form of use (written or spoken); function (similar to the speech acts types: commands, prohibitions, advice, warnings, wishes etc.); situation (regarding temporal and spatial circumstances). For instance, the pragmateme *Асцярожна, злы сабака* ‘Beware of the dog’ is usually expressed in the written form, implies a warning, and is used to protect property. The pragmateme *Приятного аппетита!* ‘Bon appétit!’ is used as an oral expression of a wish at the table during the meal. In order to detect phrasemes in Belarusian and Russian NooJ texts, the first step was to collect them manually from the corpora. According to syntactic functions, which can be performed by phrasemes, we have subdivided them into nominal, verbal, adjectival, adverbial and phrasal (Table 1).



Pragmateme	Form	Act	Situatio
Не паліць	written	prohibit	sign
Да пабачэння	oral	valediction	parting
Дзяжурныя на лінію!	oral	order	military
На плячо!	oral	order	military
За ваша здароўе!	oral	wish	drinking
Калі ласка!	oral	request	politeness
Спадары прысяжныя!	oral	greeting	court

```
# Input Language is: be
Не паліць,PRAGMATEME+Form=written+Act=prohibit+Situation=sign
Да пабачэння,PRAGMATEME+Form=oral+Act=valediction+Situation=parting
Дзяжурныя на лінію! ,PRAGMATEME+Form=oral+Act=order+Situation=military
На плячо! ,PRAGMATEME+Form=oral+Act=order+Situation=military
За ваша здароўе!,PRAGMATEME+Form=oral+Act=wish+Situation=drinking
Калі ласка!,PRAGMATEME+Form=oral+Act=request+Situation=politeness
Спадары прысяжныя!,PRAGMATEME+Form=oral+Act=greeting+Situation=court
```

Figure 1: An excerpt of the Belarusian pragmatemes database, converted into the NooJ dictionary

Type and function	Model	Example
Nominal (subject or object)	[Adjective + Noun]	*вадзяная курачка* ‘a water hen’
	[Noun + Noun]	*зямлі маці* ‘Mother Earth’
	[Noun + Preposition + Noun]	*кража со взломом* ‘a break-in’
Verbal (predicate)	[Verb + Noun]	*выскаляў зубы* ‘bare one's teeth’
	[Verb/Imperative + Noun]	*пабойцеся Бога* ‘have a heart’
Adjectival (attribute)	[Preposition + Noun]	*з капрызамі* ‘with whims’
	[Adjective + Conjunction + Noun]	*белыя як снег* ‘white as snow’
Adverbial (adverbial modifier)	[Preposition + Noun]	*по обыкновению* ‘as usual’
	[Preposition + Noun + Preposition + Noun]	*ад краю да краю* ‘from edge to edge’
Phrasal units (sentences themselves)	Common models can hardly be stated. They are unique. Most often, such phrasemes turn out to be proverbs and pragmatemes.	*Лепей недаесці, як ястраб, чым пераесці, як свіння.* ‘It is better to eat like a bird than to overeat like a pig.’

Table 1: Types of phrasemes according to the syntactic functions they perform

The analysis of Belarusian and Russian phrasemes has led to the conclusion about some common features between these two languages. The word order is not always strict: *обратил внимание* = *внимание обратил* ‘drew attention’. Some phrasemes admit lexical insertions: *с благоговением* ‘with reverence’, *с тем же благоговением* ‘with the same reverence’. Often elements of set word combinations are declined: *судебный следователь* ‘an investigator’, *судебного следователя* ‘investigator’s’, *судебные следователи* ‘investigators’ etc. Almost every syntactic type of phraseme needs a local grammar. For example, nominal phrasemes of the type [ADJECTIVE+NOUN] can be found by means of the following simple graph (Figure 2):

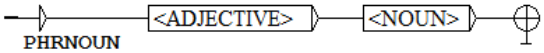


Figure 2: A graph for nominal phrasemes of the type [ADJECTIVE+NOUN]

Phrasemes which have a similar structure (e.g. *ад краю да краю* ‘from edge to edge’, *ад цямна да цямна* ‘from dawn to dusk’) should be organized into groups; each of these groups requires a separate local grammar (Figure 3):

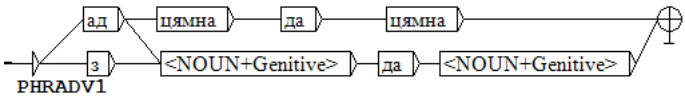


Figure 3: A graph for nominal phrasemes with similar structures

Those phrasemes which do not vary grammatically and do not admit any insertions must be included in NooJ dictionaries (Figure 4).

На назе бот рыпіць, а ў гаршку трасца кіпіць,
PHRASEME+PHRType=PHRASE+PHRSource=Kalasy04not

Figure 4: An excerpt of the Belarusian phrasemes dictionary in the format NooJ

Now in order to locate phrasemes in Belarusian or Russian texts, firstly, the dictionaries are applied, and secondly, the query PHRASEME (enclosed in angle brackets) is made through the “locate pattern” option (Figure 5).

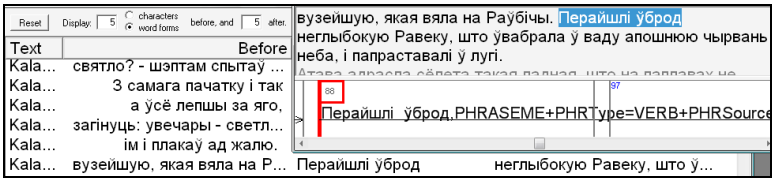


Figure 5: Locating phrasemes in Belarusian texts

To sum up, the practical results of the first part of the research include a dictionary of over 300 Belarusian and 174 Russian pragmatemes; a dictionary of over 420 Belarusian and 131 Russian phrasemes; graphs for 6 main types of free word combinations.

Describing Free Word Combinations

Usually sentences contain at least one free word combination. For instance, in the sentence *Разам з лазняй сплыла чорная гісторыя* 'The dark story together with the bathhouse disappeared' two free word combinations can be found: *разам з лазняй* 'together with the bathhouse' [Adverb + Preposition + Noun], *чорная гісторыя* 'the dark story' [Adjective + Noun]. Free word combinations are easier to be described with graphs or regular expressions. Depending on the part of speech of the first word, free word combinations are subdivided into several groups (Figure 6).

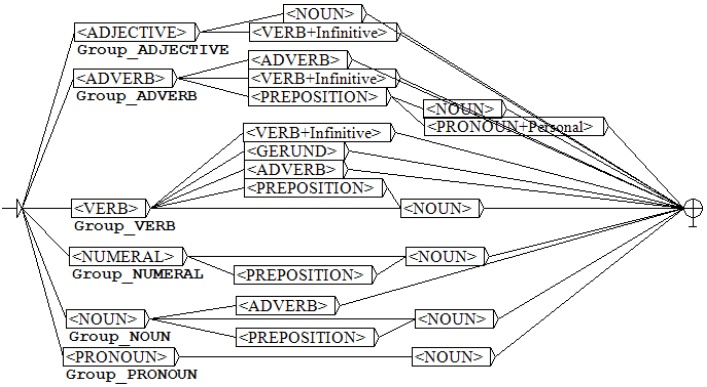


Figure 6: A graph for identifying Belarusian or Russian free word combinations

Table 2 gives examples of Belarusian and Russian free word combinations found by the constructed grammar.

	Free word combinations	Types	Sources
Bel.	*абыякавыя галасы* 'indifferent voices'	Group_ADJECTIVE	Kalasy_12.not
	адразу зразумець 'at once to understand'	Group_ADVERB	Kalasy_03.not
Rus.	*потом подобрать* 'to pick up later'	Group_ADVERB	Dom s mezoninom.not
	горе с счастьем 'grief with happiness'	Group_NOUN	Drama na ohote.not

Table 2: Some results of identification of Belarusian (Bel.) and Russian (Rus.) free word combinations

Processing of Quantitative Expressions with Measurement Units

To start with, under the term "a quantitative expression with a measurement unit QEMU", we mean a character-literal expression, combining two elements: a numeral quantifier (a number or a numeral) and a symbol or a word denoting a metrological unit, e.g. *123 мА* '123 mA', *пяць кілаграмаў* 'five kilograms', *200 кДж* '200kJ', *36°C* '36°C', etc. When dealing with units of measurement, many difficulties arise. Numeral quantifiers and names of units are found in a great variety, both in writing and formation (Skopinava et al. 2013). Another difficulty lies in variable agreement of a unit with a certain number/numeral (e.g. *25 метраў* '25 meters', *21 метр* '21 meter', *23 метры* '23 meters'), synonymous written forms (e.g. *2000 метраў* '2000 meters' = *2000 м* '2000 m' = *2×10³ м* '2×10³ m' = *2 кіламетры* '2 kilometers' = *2 км* '2 km' = *два км* 'two km'…). Therefore, creating localization rules for all cases is practically impossible, and it is extremely important to use tools that allow users to easily modify previously-developed rules and add new ones. QEMUs are language-dependent: *гадзі́на* (Belarusian) = *час* (Russian) = *hour* (English) = *Stunde* (German) = etc. Thus, it is essential to make accurate provisions for each language.

Significant results have already been achieved by European researchers and developers of the Quantalyze semantic annotation and search service[1],

[1] http://www.stn-international.com/numeric_property_searching.html

and Numeric Property Searching service in Derwent World Patents Index on STN[2] (Hetsevich et al. 2013). However, language orientation and limited thematic coverage are the reasons why theoretical or practical results cannot be completely suitable for Belarusian or Russian.

Our first goal was to create **resources for identification and classification of QEMUs for Belarusian and Russian according to the International Bureau of Weights and Measures**[3] (Figure 7).

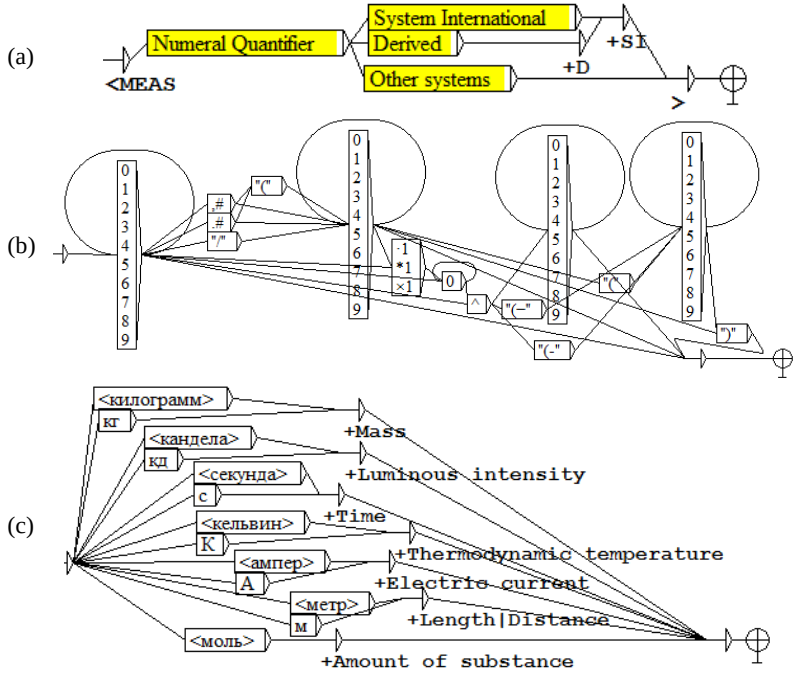


Figure 7: Graphs for identifying and classifying QEMUs according to the SI

The main graph (Figure 7a) consists of four graphs. Any text fragment is initially checked in the 1st subgraph (Figure 7b) if it has a numeral descriptor. Very often it is recorded not only with prime, decimal and fractional numbers, but also as compound expressions with exponential parts, periods etc. This subgraph is language-independent. If the graph detects a numeral quantifier, it keeps moving along one of the tree

[2] https://www.quantalyze.com/en/
[3] http://www.bipm.org/en/si/si_brochure/general.html

branches in accordance with the SI-classification: SI-basic (Figure 7c), SI-derived and extra-systemic units. Inside of each subgraph, units are differentiated, e.g.: *секунда* ‘a second’ is a unit of time. After applying the graph to the text, different search requests are possible via “Locate Pattern” (Figure 8).

Before	Seq.	After	
з разрашэннем да 1–	5 м	. З улікам камерцыйных	5 m
інфармацыі на ўзроўні 1–	10 м	. Такая дэталёвасць неабход	10 m
стартавая маса перавышае	3600 кг	. Разліковы тэрмін актыўнага	3600 kg
масай меней за	10 кг	, а праз 10–20 гадоў	10 kg
гадоў – масай парадку	1 кг	, якія змогуць устанаўліваць	1 kg
этым складала парадку 150–	500 метраў	. У 70–80-х гг	500 metres

Figure 8: Search results for the request <MEAS+SI-D> (quantitative expressions with SI-basic MUs) for Belarusian (English translations provided)

According to the values of precision, recall and their average harmonic mean, the graphs achieve 72% accuracy.

Our next goal was to develop **resources for identifying and classifying QEMUs according to word formation peculiarities**. The solution is based on successive application of three blocks (Figure 9). The first block includes resources with basic stems, metrological prefixes and prefix symbols. At this stage, morphemic and lexical analysis is performed. Metrological word stems and prefixes are identified. Stems can be either full (*Ампер* ‘ampere’) or shortened (*А* ‘amp’), while prefixes or prefix symbols can be either multiple (*кіла**ампер* ‘kiloampere’) or submultiple (*м**А* ‘mA’). Measurement units with several sequences of prefixes (e.g. *мікрамегафарад* ‘micromegafarad’) are uncommon and need to be extracted as mistakes. All these variants are considered by the second block, which consists of four morphological, language-independent grammars. Next morphological analysis is performed. Words denoting units of measurement are identified. They also receive certain markers: *Mump* (with multiple prefixes), *Musp* (with submultiple prefixes), *Muhp* (with several prefixes), *Mub* (basic units without prefixes). Identified measurement units with prefixes inherit all grammatical and inflectional characteristics of initial words. For example, *дэкалітрамі* ‘deciliters (in the Instrumental case)’ correlates with the dictionary word *літр* ‘liter’. The graph lets *дэкалітрамі* remain the noun with all its inflectional endings and grammatical features, though it is not included in the resource dictionary. The third block is represented by two syntactic grammars which accumulate morphological data and, finally, identify QEMUs and give them the marker <MUEXPR> (Figure 10).

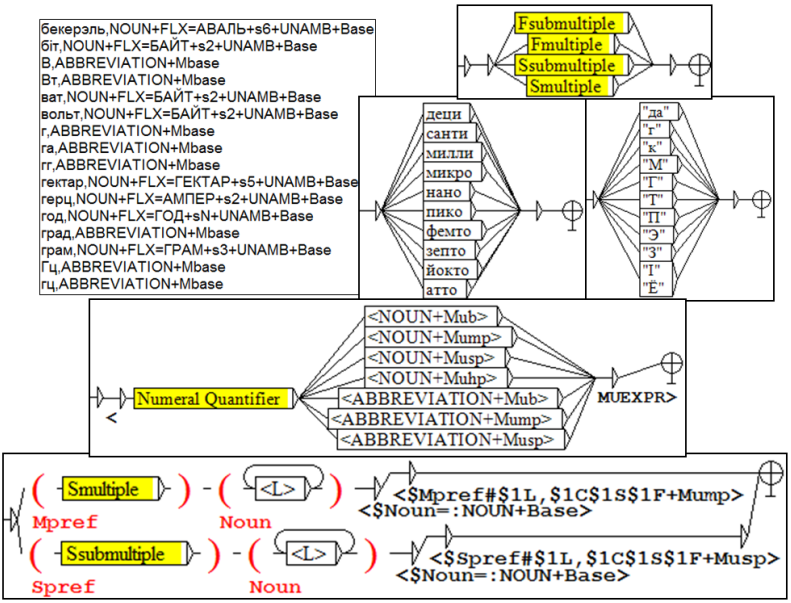


Figure 9: Resources for identifying and classifying QEMUs according to word formation peculiarities

Before	Seq.	After	
ток не превышал	1 мА	. На человека токи ст	1mA
сказать «файл в	100 килобайт	»). При обозначении	100 kilobytes
личиной от 1 до	100 МОм	, чтобы протекающий	100 Mohm
д пикотеравольт	13 йоттайоктограммов	Каждая строка соде	13 yottayoktograms
до 64 Мбит/с) и	137,4 МГц	(метровый диапазон	137,4 MHz
зумруд» массой	1383,95 каратов	. Изумруды выраще	1383,95 carats
мюонов - около	2.2 мкс	- осложняет задачу	2.2 μs
казалась равной	22 фемтограммам	(1 фг = 1•10^(-15) г).	22 femtograms
ались равными:	8.1·10^21 Дж	(уменьшение массы	8.1·10^21 J
км - наклонение	98,00 град	. Срок активного сущ	98,00 deg

Figure 10: Search results for the request <MUEXPR> for Russian (English translations provided)

Finally, we proceed to the third complex of resources, which is aimed **at generating QEMUs or turning them into orthographical word sequences**. Such tokens as *7000 м* '7000 m' are supposed to turn into *сем тысяч метраў* 'seven thousand meters'. The graph (Figure 11) contains subgraphs of two types: for generating numeral quantifiers (numbers from 0 to 999999999999) and for generating words denoting units of

measurement. QEMUs pass from input to output by means of one of seven ramifications, depending on the inflection of Belarusian/Russian nouns after numbers. The graph which processes numbers resembles Russian dolls: the subgraph for numbers of one triad (0-999) includes the subgraph *ThMlnMlr* for numbers of two triads (1000-999999) with the in-built subgraphs *MlnMlr* (three triads) and *Mlr* (four triads). After generating numbers the graph proceeds to processing nouns which denote MUs. For the present, the graph can generate basic SI units and some frequently used ones. Thanks to the visibility of finite-state automata, the graph can be easily and rapidly improved by adding more MUs. Figure 12 demonstrates some results using the graph.

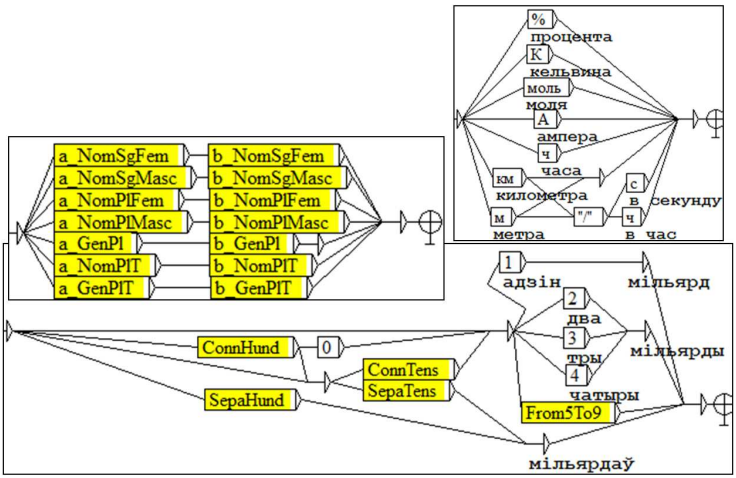


Figure 11: Resources for expanding QEMUs into orthographical words

Seq.
15 т/пятнаццаць тонаў
123 кг/сто дваццаць тры кілаграмы
5678904 К/пяць мільёнаў шэсцьсот семдзесят восем тысяч дзевяцьсот чатыры кельвіны
123 А/сто дваццаць тры амперы
6785672 м/шэсць мільёнаў семсот восемдзесят пяць тысяч шэсцьсот семдзесят два метры
787879 сут/семсот восемдзесят сем тысяч восемсот семдзесят дзевяць сутак
6761 сут/шэсць тысяч семсот шэсцьдзесят адны суткі

Figure 12: After applying the resources for expanding QEMUs into orthographical words to the Belarusian text corpus

Conclusion

It can be concluded that the goal of developing resources which find set and free word combinations in Belarusian and Russian text corpora has been achieved. Particular attention is paid to analyzing quantitative expressions with measurement units as a specific subclass of free word combinations. Three complexes of NooJ visual morphological and syntactic grammars have been developed. They identify, classify (with two approaches) and expand QEMUs into orthographical word sequences. In future it is planned to build local grammars for phrasemes, expand the resources for describing wider groups of free word combinations, and disambiguate cases when the graphs "confuse" some units (e.g. the same initial letter г for год 'year', грам 'gram', гадзі́на 'hour').

Acknowledgements

We would like to thank the linguist Adam Morrison for his help in revising the language of this paper.

References

Hetsevich, Yuras, Sviatlana Hetsevich, Boris Lobanov, Alena Skopinava and Yauheniya Yakubovich. 2013. "Accentual expansion of the Belarusian and Russian NooJ dictionaries". In: *Formalising Natural Languages with NooJ: Selected Papers from the NooJ 2012 International Conference*, edited by Anaïd Donabédian, Victoria Khurshudian and Max Silberztein, 24–36. Newcastle: Cambridge Scholars Publishing.

Mel'čuk, Igor. 2013. "Tout ce que nous voulions savoir sur les phrasèmes, mais". In: *Cahiers de lexicologie. Revue internationale de lexicologie et de lexicographie*, 129-149.

Skopinava, Alena, Yuras Hetsevich and Boris Lobanov. 2013. "Processing of quantitative expressions with units of measurement in scientific texts as applied to Belarusian and Russian text-to-speech synthesis". In: *Компьютерная лингвистика и интеллектуальные технологии: По материалам Международной конференции «Диалог»*, 634–651. Moscow: Russian State University for the Humanities Publishing.