РАЗРАБОТКА БЕЛАРУСКОГО И РУССКОГО ЛИНГВИСТИЧЕСКИХ МОДУЛЕЙ ДЛЯ СИСТЕМЫ NOOJ В ПРИЛОЖЕНИИ К СИНТЕЗУ РЕЧИ ПО ТЕКСТУ

Гецевич Ю. С. (Yury.Hetsevich@gmail.com),
Объединенный институт проблем информатики
Национальной академии наук Беларуси
(ОИПИ НАН Беларуси), Минск, Беларусь

Гецевич С. А. (Novaeimya@gmail.com),
ОИПИ НАН Беларуси, Минск, Беларусь

Лобанов Б. М. (Lobanov@newman.bas-net.by),
ОИПИ НАН Беларуси, Минск, Беларусь

Ключевые слова: синтез речи по тексту, NooJ, беларуский язык, русский язык, лингвистические модули, словарь, морфологическая грамматика, синтаксическая грамматика, аннотирование текста, синтагмация, фонетические слова.

BELARUSIAN AND RUSSIAN LINGUISTIC PROCESSING MODULES FOR THE SYSTEM NOOJ AS APPLIED TO TEXT-TO-SPEECH SYNTHESIS

Hetsevich Yu. S. (Yury.Hetsevich@gmail.com)
United Institute of Informatics Problems of the National Academy of Sciences of Belarus (UIIP NAS Belarus), Minsk, Belarus

Hetsevich S. A. (Novaeimya@gmail.com)
UIIP NAS Belarus, Minsk, Belarus

Lobanov B. M. (Lobanov@newman.bas-net.by)
UIIP NAS Belarus, Minsk, Belarus

This paper describes the program NooJ, which provides a platform for the construction of modules for resolving linguistic problems in the area of text-to-speech synthesis. Belarusian and Russian are chosen as target languages. Basic and comprehensive electronic grammatical dictionaries for NooJ are described. We present the entire algorithm of converting dictionaries of the two languages into the form acceptable in NooJ, which retains all lexical, grammatical and accentual information. The dictionaries developed for NooJ help solve the problems of annotating words with lexical and grammatical categories and syllabic accents, as well as the problems of searching texts for a definite sequence of words according to their grammar and word forms. The grammars developed for NooJ are notable for their clarity and the speed at which they can be built. The structure and algorithm of the morphological grammar working on the text are given, including the localization of words written in Cyrillic and Roman letters. The structure and algorithm of the work of two syntactic grammars on a text are described: one aimed at searching for phonetic words and the other on searching for syntagms with a particular number of phonetic words. NooJ output produced after dictionaries or grammars have been applied to text are exported to text format Future work includes the completion of the transfer pf accentual information into Belarusian and Russian NooJ dictionaries,and construction of grammars for identifying accentual units in syntagms and grammars aimed at learning rhythmic structures of texts.

Key words: text-to-speech synthesis, NooJ, Belarusian language, Russian language, linguistic modules, dictionary, morphological grammar, syntactic grammar, text annotation, syntagmation, phonetic words

1. Introduction

The linguistic program NooJ (Silberztein 2003) contains numerous dictionaries, grammars and processed corpuses. The internal linguistic algorithms in NooJ permit the processing of separate files and also collections of files (corpora). For text processing, the following basic operations are developed: selection of tokens, bigrams, and unknown words, unambiguous and ambiguous words.

NooJ helps to create large lexical, morphological and syntactic grammars, which can be written both in text view and through an easy-to-use visual editor, greatly simplifying the process. Dictionaries and grammars can be applied to texts to find complex lexical, morphological and syntactic expressions. NooJ users are able to develop grammars for identifying semantic units such as proper names, dates, time, financial expressions[1], etc., in a large text.

In 2011 the authors had the opportunity to develop the first version of Belarusian and Russian NooJ modules. The discussion about the results took place at the international conference of the NooJ community in Dubrovnik, Croatia (Hetsevich 2012; Hetsevich 2012). This article will present an improved way of building Belarusian and Russian dictionaries for NooJ. Particularly for use in text-to-speech synthesis, more accentual information for words will be added from the basic electronic dictionaries. Grammars important for linguistic text processing in text-to-speech synthesis will also be examined.

2. Basic Belarusian and Russian electronic dictionaries for text-to-speech synthesis system Multiphone

The bilingual text-to-speech synthesis system Multiphone uses specially-prepared electronic dictionaries for Belarusian and Russian languages (Lobanov 2008). For the basic electronic dictionaries, a system of codifying lexical and grammatical categories was developed. The initial Belarusian electronic dictionary (BED) was based on the printed dictionary Biryla (Biryla 1987, Sovpel 2006). The Russian electronic dictionary (RED) was built in SSRL using the printed Zaliznyak dictionary (Zaliznyak 1980).

New words from modern texts on literature, science and technology were added to the basic electronic dictionaries. To further facilitate the process, one of the authors has developed an application program, Expert Dictionary Editor (EDE) (Hetsevich 2011). EDE makes it possible to process texts, find new words for the dictionary, and allows the user to mark syllabic accents, lexical and grammatical categories (LGC) and priorities therein. Saved words (including their homographs) can be displayed together with all of the selected information (Fig. 1).

[1] http://www.nooj4nlp.net/pages/introduction.html

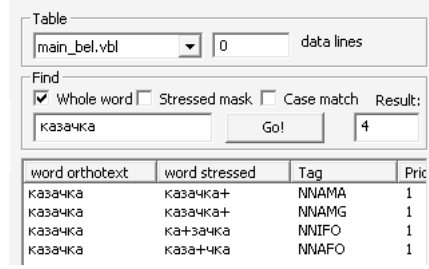


Fig. 1. The program EDE makes it possible to view, add and delete words in electronic dictionaries

Quantitative characteristics of the dictionaries (table 1) show that each dictionary has 14 lexical categories (or simply categories) of words. Category refers to part of speech (noun, verb, etc.), as well as several word groups (for instance, parenthesis, predicate). Each category is marked with the amount of corresponding word forms and tags.

Table 1. Content of Belarusian and Russian electronic dictionaries

	BED		**RED**	
Categories	**Word forms**	**Tags**	**Word forms**	**Tags**
Adjective	872 943	85	504 728	31
Adverb	6524	8	1366	33
Conjunction	45	2	563	49
Gerund	32 278	4	70 215	8
Interjection	8	1	184	1
Noun	592 245	244	540 684	168
Numeral	3571	96	1473	92
Parenthesis	45	1	65	1
Participle	189 609	116	1 897 033	246
Particle	57	3	348	75
Predicate	60	1	266	1
Preposition	121	15	277	32
Pronoun	1214	270	1064	307
Verb	419 734	68	429 587	78
Sum	**2.118.454**	**914**	**3.447.853**	**1122**
Sum(KB)	**41 267,2**	**50,5**	**92 774,4**	**80,6**

The total number of entries in the Belarusian electronic dictionary is more than 2,1 million word forms, and in Russian 3,4 million word forms. The total number of tags in each of the dictionaries is 914 and 1122 respectively. The total size of the files with word forms and tags for BED is 40,4 MB, and for RED 90,7 MB.

The dictionaries effectively manage the problem of identifying syllabic accent and LGC in a wide range of thematic texts for text-to-speech synthesis. Further, we shall try to realize the conversion of the aforementioned electronic dictionaries into NooJ so that they can be used by linguists more in processing texts in Belarusian and Russian.

3. Building Belarusian and Russian dictionaries for NooJ

For converting Belarusian and Russian electronic dictionaries into NooJ, various algorithms of conversion have been developed in Perl (Fig. 2) . The algorithms receive BED and RED in input, process them and put out dictionary files in NooJ according to the rules of its construction (Silberztein 2003). In place of XX is a language identifier: “be” for Belarusian and “ru” for Russian.

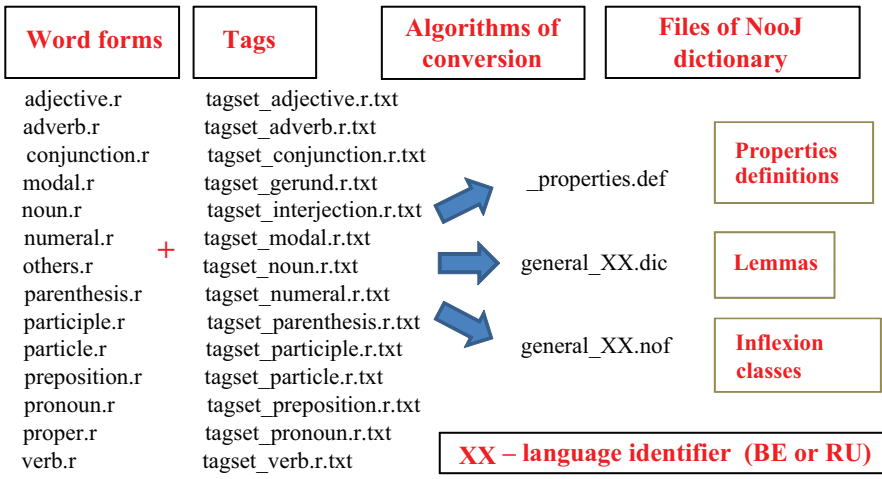


Fig. 2. Converting Belarusian and Russian electronic dictionaries into NooJ

Let us examine the steps of processing electronic dictionaries with developed algorithms.

Firstly the file with lexical property definitions entitled ***_properties.def*** is formed. For this all the tags and their decodings from the tag sets (***tagset_*.r.txt***) are read, but the only ones kept are those which are really used for marking LGC of words in the word file of a specific category (****.r***). For each property of the category all possible variants of their values are collected from the respective columns with tag decoding. They are written with “+” sign in the following format:

CATEGORY_propertyName = propertyVALUE1+propertyVALUE2 +…;

For example, fig. 4 presents excerpts of obtained files with property values for the category Noun.

Obtained from BED	**Obtained from RED**
… NOUN_Animation = Animate + Inanimate; NOUN_Case = Accusative + Dative + Genitive + Instrumental + Nominative + Prepositional; NOUN_Form = SubstanceAdjective; NOUN_Gender = Feminine + Masculine + Neuter; NOUN_Meaning = Common + Proper; NOUN_Flexion = ю + яў_аў; NOUN_Number = Plural; …	… NOUN_Animation = Animate + Inanimate; NOUN_Case = Accusative + Dative + Genitive + Instrumental + Nominative + Prepositional; NOUN_Gender = Feminine + Masculine + Neuter; NOUN_Number = Plural + Singular; NOUN_ProperCommon = Common; …

Fig. 4. Excerpts of the files with properties for the category Noun

Then the algorithm forms a file with word lemmas entitled general_XX.dic. For each paradigm of a specific category, lemma and inflection class with tags are formed. A lemma is defined as the initial form of a word. The inflection class for each lemma is of the form "inflection1_tag1, inflection2_tag2, …, inflectionN_tagN". Each inflection is obtained by deleting the minimal part of a word that remains invariable in all word forms of a word paradigm from the word form. Every obtained inflection class is checked according to the following rule:

Is there already such an inflection class with tags in the file ***general_XX.nof*** or not?

1. If not, the inflection class is registered in the file ***general_XX.nof*** with the resulting title being identical to the lemma. If there is such a class title, it is completed with a unique number. Couples of flexions and tags alternate with the sign "+". Commands to delete part of a lemma (number of characters) are inserted before the flexions, after which the NooJ parser adds flexions of its inflection class to the lemma. Decoding is taken from the file with decoding for each tag, the category title is deleted and the sign "+" is added between property values. The end of the inflection class is denoted by the sign ";".
2. If there is such an inflection class, it is already being used for some word, and a name has already been created for it.

Figure 5 contains excerpts of files with inflection classes for the category Noun.

Obtained from BED	**АБАВЯЗАК** = <B2>ак/Accusative+Common+Inanimate+Masculine+ <B2>ак/Common+Inanimate+Masculine+Nominative+ <B2>кі/Accusative+Common+Inanimate+Masculine+Plural+ <B2>кі/Common+Inanimate+Masculine+Nominative+Plural...;…
Obtained from RED	**АБСОЛЮТНОСТЬ** = <B1>ей/Common+Feminine+Genitive+Inanimate+Plural + <B1>и/Accusative+Common+Feminine+Inanimate+Plural + <B1>и/Common+Feminine+Inanimate+Nominative+Plural + <B1>и/Common+Dative+Feminine+Inanimate+Singular + <B1>и/Common+Feminine+Genitive+Inanimate+Singular...; …

Fig. 5. Excerpts of files with inflection classes for the category Noun

In order to mark accent, the algorithm collects all accent positions from each word worm of a lemma. If all accent positions are equal to the number A, a marker with the constant "+sA" is formed. If there is even one different accent position in any word worm of lemma, the marker of various accent in the paradigm is formed — "+sN".

The lemma is registered in the file ***general_XX.dic*** with the mark of the category, plus the command "+FLX=" and then the title of the corresponding inflection class in the following format:

LEMMA, CATEGORY + FLX = TITLE_OF_INFLECTION_CLASS + ACCENT_MARKER.

Figure 6 presents the resulting files with lemmas, marked categories, inflection classes and accents for BED and RED.

Obtained from BED	… абавязак,NOUN+FLX=АБАВЯЗАК+s5 адгалосак,NOUN+FLX=АБАВЯЗАК+s6 мама,NOUN+FLX=АБАТЫСА+s2 манастыр,NOUN+FLX=АБРУЧ+sN …
Obtained from RED	… абстрактность,NOUN+FLX=АБСОЛЮТНОСТЬ+s6 аварийность,NOUN+FLX=АБСОЛЮТНОСТЬ+s5 адоптировать,VERB+FLX=АБЛАКТИРОВАТЬ+s6 быстр,ADJECTIVE+FLX=БОДР+sN …

Fig. 6. Excerpts of files with lemmas

In the beginning of every resulting file of the NooJ dictionary, a corresponding heading with the mark of the NooJ version (V3), authors and other necessary information is added to save the correct format of the dictionary files.

Thereby, the resulting files ***_properties.def***, ***general_XX.dic*** and ***general_XX.nof*** correspond to the rules of construction for files in the NooJ dictionaries.

Quantitative characteristics of the resulting Belarusian (BN) and Russian (RN) NooJ dictionaries are presented in Table 2. There are nearly 137 thousand lemmas in BN 123 565 (88 %) of which are lemmas with precise accent position. There are nearly 214 thousand lemmas in RN, 199 638 (94 %) of which are lemmas with precise accent position. The total number of inflection classes in BN is 3137 and 1851 in RN. The number of property values in BN and in RN is 61 and 56 respectively. These numbers of property values are almost identical. This can be explained by the fact that Belarusian and Russian are kindred languages, having similar grammatical and categorical structure.

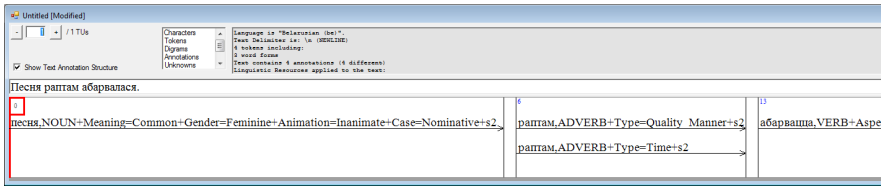
Table 2. Quantitative characteristics of Belarusian and Russian electronic dictionaries for NooJ

	Lemmas		**Inflection classes**		**Property values**	
Categories	**BE**	**RU**	**BE**	**RU**	**BE**	**RU**
Adjective	32 350	31 583	33	61	8	5
Adverb	6524	1366	8	30	2	4
Conjunction	45	563	2	49	1	4
Gerund	1836	32 977	282	160	2	4
Interjection	8	184	1	1	0	0
Noun	53 054	44 396	1427	309	8	5
Numeral	169	99	38	27	8	5
Parenthesis	45	65	1	1	1	0
Participle	12 293	68 006	13	57	8	9
Particle	57	348	3	73	2	5
Predicate	60	266	1	1	1	0
Preposition	105	277	17	27	2	4
Pronoun	66	1064	56	307	9	5
Verb	30 718	33 127	1255	748	9	6
Sum	**137 330**	**214 321**	**3137**	**1851**	**61**	**56**
Sum (KB)	**6064**	**10 446**	**1809**	**1113**	**5**	**5**

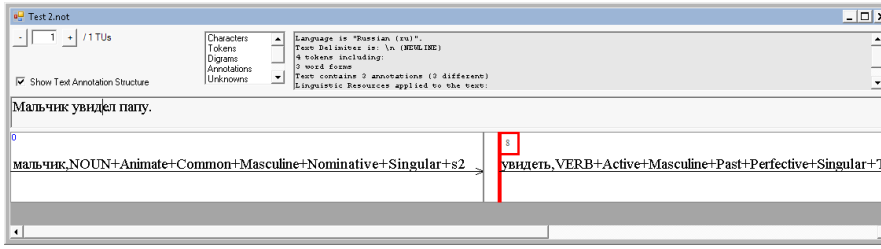
The total size of BN (the file with lemmas, inflection classes and property values) is 7,7 MB, and that of RN is 11,3 MB. If these files are processed with the NooJ parser to create a dictionary in one file, the result will be a file of the type ***general_XX.nod***. For Belarusian it will take up 8,2 MB, and for Russian 5 MB.

4. Linguistic processing of Belarusian and Russian texts with developed dictionaries for text-to-speech synthesis

When Belarusian and Russian dictionaries for modules of the program NooJ are developed, they can be used for text annotating. For instance, for each known word, all variants of accents and grammatical categories present in the dictionary are marked (Fig. 7).



(a)



(b)

Fig. 7. Examples of annotating words with accents and grammatical categories for Belarusian (a) and Russian (b) languages

It is possible to receive an XML-export of the annotated sentence (Figure 8). For each word, the export contains its lemma, category, inflection class name, expanded grammatical marks and accent marker. As such, this information can be processed further using the linguistic processor of the text-to-speech synthesizer.

BN	<LU LEMMA="**песня**" CAT="**NOUN**" FLX="ПЕСНЯ" Meaning="Common" Gender="Feminine" Animation="Inanimate" Case="Nominative" TYPE="**s2**">**Песня**</LU> <LU LEMMA="**раптам**" CAT="**ADVERB**" FLX="АБАВЯЗАЦЕЛЬНА" Type="Quality_Manner" TYPE="**s2**"> <LU LEMMA="**раптам**" CAT="**ADVERB**" FLX="АДВЕКУ" Type="Time" TYPE="s2">**раптам**</LU></LU> <LU LEMMA="**абарвацца**" CAT="**VERB**" FLX="АБАРВАЦЦА" Aspect="Perfective" TYPE="**sN**">**абарвалася**</LU>.

RN	<LU LEMMA="**мальчик**" CAT="**NOUN**" FLX="АБРЕК" TYPE="Animate" TYPE="Common" TYPE="Masculine" TYPE="Nominative" TYPE="Singular" TYPE="**s2**">**Мальчик**</LU> <LU LEMMA="**увидеть**" CAT="**VERB**" FLX="ВОЗНЕНАВИДЕТЬ" TYPE="Active" TYPE="Masculine" TYPE="Past" TYPE="Perfective" TYPE="Singular" TYPE="Transitive" TYPE="**s3**">**увидел**</LU> <LU LEMMA="**папа**" CAT="**NOUN**" FLX="БОНЗА" TYPE="Accusative" TYPE="Animate" TYPE="Common" TYPE="Masculine" TYPE="Singular" TYPE="**s2**">**папу**</LU>.

Fig. 8. Examples of XML-export of the data annotated from NooJ

The resulting dictionaries have been tested in NooJ using two types of queries for Belarusian and Russian texts:

- To find words and their contexts so that they correspond to the given categories with indicated property values (Figure 6);
- To find words and their contexts so that they correspond to word forms of the given paradigm (Figure 7).

(The paradigm is set through any word form of this paradigm).

query for BN: <NOUN><NOUN+Genetive>

Before	Seq.	After
трапятаў	агеньчык каганца	. Каля
ліловым	адценнем пліткі	жалезняку
ведаць	асновы свету	, што
старажытныя	белыя званіцы	. А
песню.	Бог салдата	свайго

query for RN: <VERB><ADVERB>

Before	Seq.	After
человека, который	вставал очень	рано
части, и	было немножко	страшно
позднего вечера	бродил где-нибудь	. Однажды
кончились; я	прошел мимо	белого
слушать ее	было интересно	, хотя

Fig. 9. Locating words and their contexts by the given categories and property values

query for BN: <горкі>

Before	Seq.	After
На	горцы	лясок
Каля	горкі	сям
Чаму	горкія	асіны

query for RN: <жить>

Before	Seq.	After
помнил,	жила	моя
редактор.	Живет	в
он	живет	. Выяснилось

Fig. 10. Locating words and their contexts, where words correspond to word forms of the given paradigm

As a result, it is possible to study the forming correlations between words in sentences according to grammatical features of words.

In summary, the Belarusian and Russian dictionaries for NooJ help to solve basic problems with the linguistic aspect of the text-to-speech synthesizer and to study the

correlations between words in texts for the building of syntactical grammars in order to define phonetic words and accentual units in text.

5. Linguistic processing of Belarusian and Russian texts with morphological and syntactical grammars for text-to-speech synthesis

NooJ makes it possible to develop graphic grammars for solving common problems with processing specific words and their sequences in text-to-speech synthesis. This program does not require knowledge of a programming language; it has a special visual editor for building grammars.

Input texts can contain many words unknown to dictionaries, which causes the words to be read incorrectly in the text-to-speech synthesis system. This can be the case with words formed from mixed alphabets (Cyrillic and Roman, for example). And the most difficult is the case when two similar symbols have different codes for Belarusian/Russian and English languages (/i/, /i/), (/c/, /c/). For the localization of such words, a special morphological grammar has been developed with the visual editor NooJ (Figure 11). This grammar performs the following actions:

- marks foreign words with the category EN (searches words containing all letters from the set ENGLISH_LETTERS and marks them);
- marks words with mixed alphabets with the category MIXED and marks Roman symbols with brackets (Figure 12);

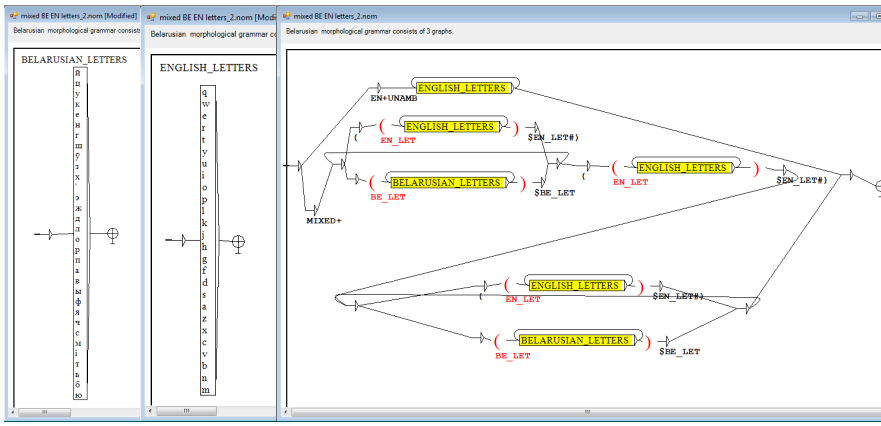
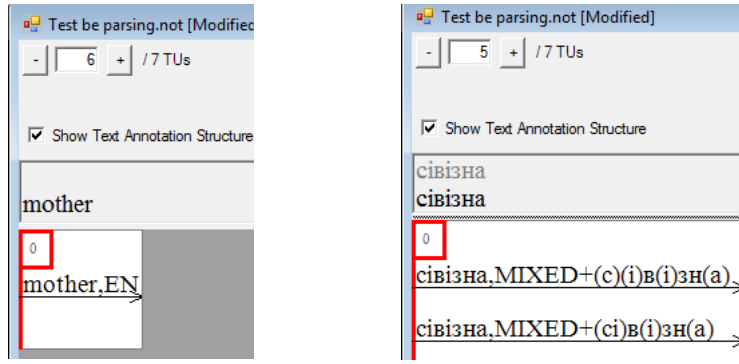


Fig. 11. Example of building morphological grammar for finding words written in mixed alphabets

For convenience we shall understand ***complex phonetic words*** as words formed from two or more words, and ***phonetic words*** as those formed from only one word.

In order to mark complex phonetic words, a syntactical grammar has been developed (Figure 13). The grammar works in a way that left and right context of a word is checked, and according to its meaning, the word sequence can be marked as a complex phonetic word in the following ways (Figure 14). If there are prepositions (one or more) to the right of the word, such a word sequence is marked as a complex phonetic word (PHONETIC_WORD) formed with prepositions (+PRE), for example, /**у кавалкі**/, /**на куски**/. If there are particles (one or more) to the left of the word, such a word sequence is marked as a complex phonetic word (PHONETIC_WORD) formed with particles (+PAR), for example, /**што ж**/, /**что же**/. If the word has both prepositions (one or more) and particles (one or more) as right and left context, such a word sequence is marked as a complex phonetic word (PHONETIC_WORD) formed both with prepositions and particles (+PRE+PAR), for example, /**за што ж**/, /**за что же**/.



(a) (b)

Fig. 12. Examples of work of the morpholgical grammar for finding words written in Roman (a) and mixed (b) alphabets

NooJ makes it possible to import a grammar into another one and to use sub-grammars in grammar. In order to define possible syntagmas containing four phonetic words, a corresponding grammar has been developed (Figure 15). It marks word sequence with the category **SYNTAGMA+4**, if four phonetic words follow each other (Figure 15, a). In order to identify four phonetic words (including complex words), a sub-grammar **WF_or_PHW** is invoked. This sub-grammar is employed if a simple word is found (**WF**) or if a complex phonetic word is found (**phonetic_word**) according to the sub-grammar on the Figure 15, b. An example of work of the grammar on a specific text is shown in Figure 16.

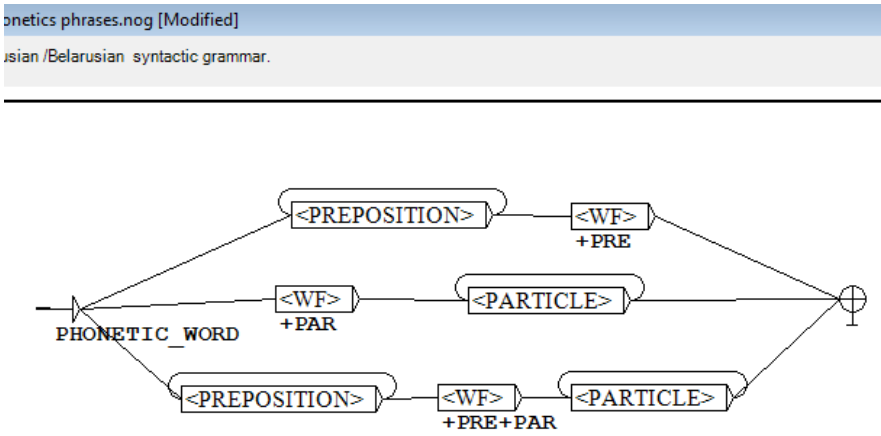
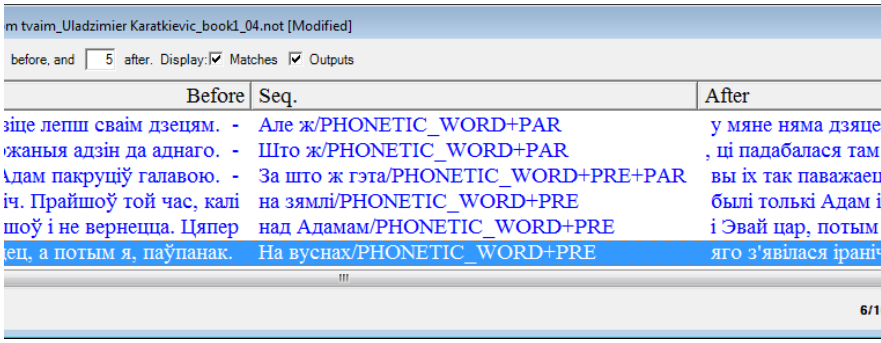
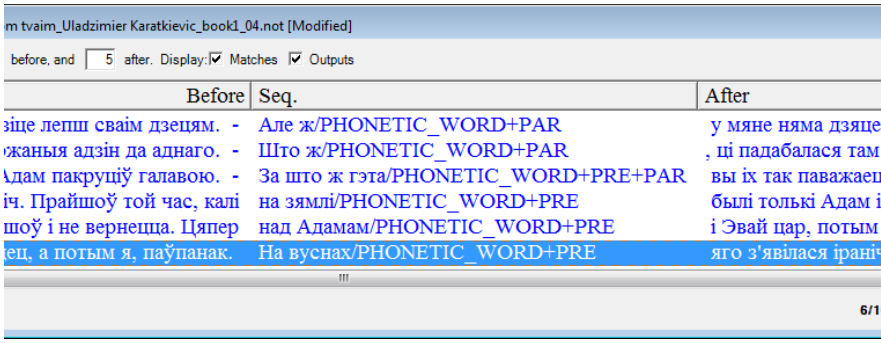


Fig. 13. Syntactical grammar for defining complex phonetic words

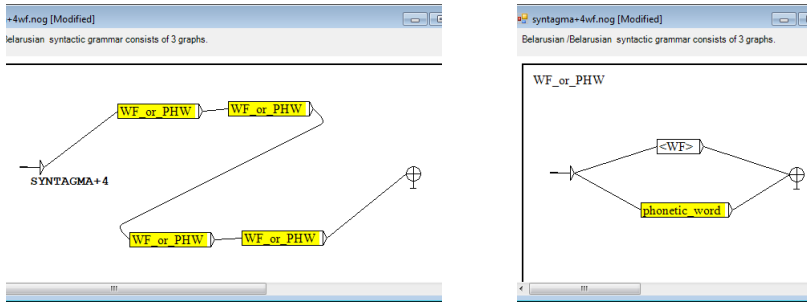


(a)



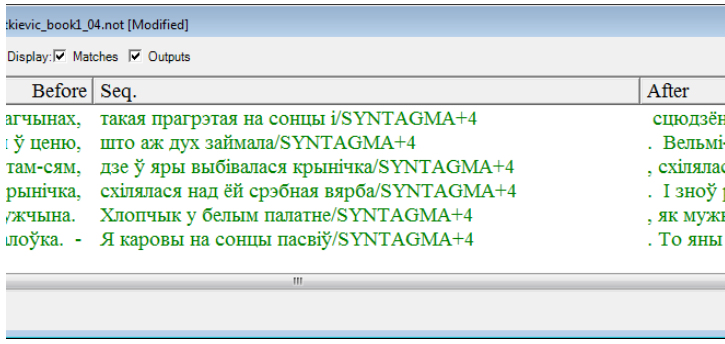
(b)

Fig. 14. Results of work of the syntactical grammar for defining phonetic words in texts in Belarusian (a) and Russian (b) languages

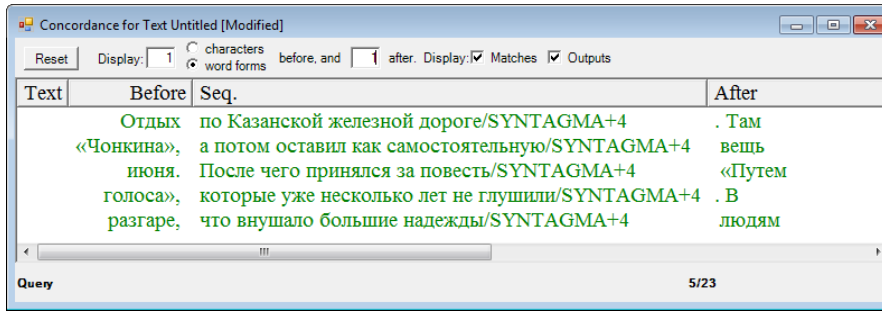


(a) (b)

Fig. 15. Syntactical grammar (a) for separating possible syntagmas into four phonetic words according to the sub-grammar (b)



(a)



(b)

Fig. 16. Examples of work of the syntactical grammar for separating possible syntagmas into four phonetic words for Belarusian (a) and Russian (b) texts

It should be noted that a grammar for separating syntagmas into four phonetic words can be extended to any number of phonetic words. This research gives an example of building grammars according the most frequent number of phonetic words in a syntagma (Lobanov, 2011).

Results of the work of the syntactical grammar on a text can be exported into a text file marked by the beginning of the given syntagma and its length. Such results can be processed with another grammar in order to form accentual units.

Conclusion

This article presents qualitative features of Belarusian and Russian electronic dictionaries which are used in the text-to-speech synthesis system Multiphone. Algorithms for converting basic dictionaries for Belarusian and Russian modules of linguistic resources of the program NooJ have been developed.

The total number of lemmas for the Belarusian dictionary is nearly 137 thousand, and for the Russian approximately 214 thousand. Information about precise accent position in lemmas' paradigms has been transferred to the Belarusian dictionary at 88% and to the Russian — at 94%. The resulting dictionaries are able to annotate words with lexical and grammatical categories and accents and also to help with studying the syntax of sentences.

NooJ tools for building visual morphological and syntactical grammars allow linguists to structure quickly and graphically important linguistic algorithms for text-to-speech synthesis: searching for words written in Cyrillic and Roman letters, searching for phonetic words, searching for syntagmas with a specific number of phonetic words.

The developed Belarusian and Russian NooJ modules will be improved. In addition, all accentual information will be transferred to dictionaries; syntactical grammars for defining accentual units in syntagmas will be developed; and the problems of learning rhythmic structure of texts will be solved.

The modules for dictionaries in NooJ can be also used by linguists and philologists in order to learn Belarusian or Russian languages.

Acknowledgments

Many thanks to the linguist from USA Adam Morrison for his help in revising this paper.

References

1. *Biryla M. V.* (1987), Slounik belaruskaj movy: Arfagrafija. Arfaepija. Aktsentuatsyja. Slovazmjanenne [Belarusian dictionary: Orthography. Orthoepy. Accentuation. Inflection]. Minsk: BelSE
2. *Hetsevich Y., Hetsevich S.* Overview of Belarusian and Russian dictionaries and their adaptation for NooJ, Selected Papers from the NooJ 2011 International Conference "Automatic Processing of Various Levels of Linguistic Phenomena", Dubrovnik, Croatia. Newcastle: Cambridge Scholars Publishing, 2012, pp. 29–40.

3. *Hetsevich Y., Hetsevich S., Lobanov B., Yakubovich Ya.* (2012) “Belarusian module for NooJ”. Available at: http://www.nooj4nlp.net/pages/belarusian.html.
4. *Hetsevich Y. S., Vyaltsev V. N.* Editing and completion system for the dictionaries of audio interface of question-and-answer system for Belarusian and Russian [Sistema redaktirovanija I popolnenija slovarej rechevogo interfejsa voprosno-otvetnoj sistemy dlja belaruskogo i russkogo jazykov]. Open Semantic Technologies for Intelligent Systems: Proceedings of the International Scientific and Technical Conference. Minsk, 2011, pp. 413–424.
5. *Lobanov B. M., Hetsevich Y. S.* Statistical Characteristics Of Sentence Syntagmatic Segmentation In Applying To Expressive Text-To-Speech Synthesis [Statisticheskie harakteristiki sintagmaticheskogo chlenenija predlozhenija v prilozhenii k sintezu vyrazitel'noj rechi po tekstu], Komp’iuternaja Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2011” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2011”]. Moscow, 2011, pp. 434–447.
6. *Lobanov B. M., Tsirulnik L. I.* (2008), Kompjuternyj sintez i klonirovanie rechi [Computer synthesis and cloning of speech], Belaruskaja Navuka, Minsk.
7. *Silberztein, M.* (2003), “NooJ Manual”. Available at: www.nooj4nlp.net
8. *Sovpel, I. V.* Computer corpus of Belarusian language and its applications [Kompjuternyj fond belorusskogo jazyka i ego prilozhenija]. Materialy 3 mezhdunarodnoj konferentsii “Informatsionnye sistemy i tehnologii” IST’2006, chast’ 2 [Proceedings of the 3 International Conference “Information systems and technologies” IST’2006, part 2]. Minsk, 2006, pp. 71–76.
9. *Zaliznjak, A. A.* (1980), Grammaticheskij slovar’ russkogo jazyka: Slovoizmenenie. Ok. 100 000 slov [Russian grammatical dictionary: Inflection. Near 100 000 words], Stereotip, Moscow.