Accentual Expansion of the Belarusian and Russian NooJ Dictionaries

Yury Hetsevich,
Sviatlana Hetsevich,
Boris Lobanov, Alena Skopinava
Yauheniya Yakubovich

Abstract

This paper focuses on ways of adding accentual information to dictionaries constructed in the NooJ format, as well as using this information by means of specially built algorithms.

Introduction

This article represents the continuation of the work begun last year by the group of researchers from the United Institute of Informatics Problems of the National Academy of Sciences of Belarus (Y. Hetsevich and S. Hetsevich 2011). Its subject covered the linguistic development environment NooJ in a context of the description of the Belarusian and Russian languages. So, in 2011 the first versions of NooJ modules for these Slavic languages were built. Further, in March 2012 the Belarusian NooJ module (already updated) was successfully published by Y. Hetsevich and S. Hetsevich. Now it includes twelve chapters from the first part of the novel by the famous Belarusian writer Uladzimir Karatkevich, *Spikes under your sickle*. In total, the texts have enriched the Belarusian dictionary with 52653 tokens and 20771 distinct annotations. The Russian module's first publication by Vincent Bennet took place in May 2012. The same year another Russian module was created by the authors of the article. As a basis, two of Anton Chekhov's narratives were used: "The House with the Mezzanine" with 7148 tokens (2365 different), and "A Hunting Drama" with 69197 tokens (14569 different).

However, none of these versions contains accentual information, though, obviously, the significance of these data can't be overestimated both for Belarusian and Russian linguistic resources. The two Slavic languages have much in common, and one of the similarities involves the preservation of free stress in words, so it can unpredictably fall on any vowel in a word. Subsequently, it cannot be described by a simple system of rules. According to the calculations performed in table 1, the Belarusian dictionary contains over 123 thousand lemmas (88% of the total number) with a constant position of an accented letter for each word form. In the Russian dictionary a fixed accent type is displayed by over 200 thousand lemmas (94% of the total number).

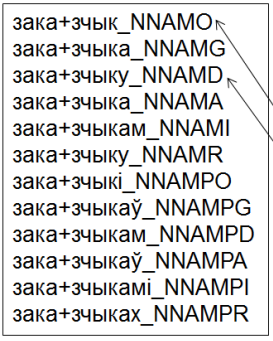
	Number of different accents in one lemma	**Number of lemmas**	**Examples**
BM	1	123529	м**а**ма, дыял**о**г
	2	11463	д**о**м-дам**а**мі, актыва**в**аць-актыв**у**ю
	3	1845	аб'**е**сціся-аб'я**с**і**с**я-аб'ядз**і**+мся
	4	574	злав**а**цца-злу**е**шся-злуяц**е**ся-злу**й**ся
	5	16	кл**а**сціся-клад**у**ся-кладз**е**шся-кладзяц**е**ся-кладуч**ы**ся
RM	1	201053	м**а**ма, д**о**м, **ё**кать, актив**и**ровать, арб**у**зный
	2	10297	бл**и**зок-близк**и**, **ё**рш-ерш**а**
	3	1543	борода-б**о**роды-бор**о**д
	4	90	добр**а**ться-добрал**ся**-добрал**и**сь-доберусь
	5	5	погн**а**ться-погон**ю**сь-погнал**о**сь-погнал**ся**-пог**о**нится

Table 1: Floating accent distribution within Belarusian (BM) and Russian (RM) NooJ modules

Thus, there are 12% Belarusian lemmas and 6% Russian lemmas with variability of accents within the framework of separate inflectional paradigms. The aim of our work is to improve NooJ methods of building dictionaries, namely to add accentual information and build special syntax grammars using this information.

Description of a fixed accent

When a lexeme preserves the same accent in all its inflectional word forms, the accent is fixed. As a model for its indication, we have a special format of the electronic dictionary base structure where accents are marked by a plus sign (+), while grammatical information is put in tags. Figure 1 illustrates an example from the Belarusian NooJ module, namely the paradigm of the noun *заказчык*. In all word forms (categories of gender and case are also specified), the accent can be observed on the fourth letter.



Tag	*Category*	…	*Gender*	*Case*
NNAMO	Noun		Masculine	Nominative
NNAMD	Noun		Masculine	Dative
NNAFPA	Noun		Feminine	Accusative
…	…	…	…	…

Figure 1: Fixed accent indication in the Belarusian NooJ module

Still, in order to make the computer retrieve this information, the need to create a special algorithm inevitably arises.

The first step is to define a constant accent position for all word forms in an inflectional paradigm. In order to mark the accent in a lemma, accent positions of the whole inflectional class (of each word form) should be taken into account. For instance, in the Belarusian noun *волат* the accent invariably falls on the second position, the second character of each lemma's word form. Obviously, the accent is fixed. Its accented letter marker is defined as *ap2*. In the dictionary file, each lemma is followed by an accented letter marker, apart from its category and title of a respective inflectional class: *волат,NOUN+FLX=АБАЛІЦЫЯНІСТ+AccentP=ap2* (*AccentP* denotes an accent letter position).

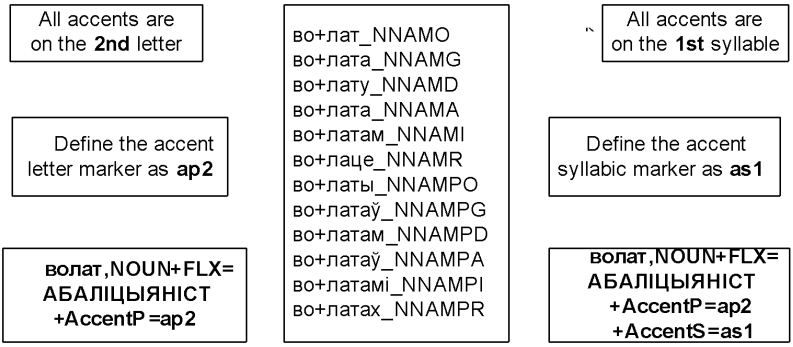


Figure 2: Adding an accent marker to the noun *волат*

In order to make the accent description more accurate (which is so important for the above-mentioned languages), we consider it necessary to mark not only an ordinal number of a letter where an accent occurs, but also an ordinal number of an accented syllable. The quantity of syllables is equal to the number of stressed vowels in a word.

For instance, the same Belarusian noun *волат* has the first stressed syllable (or the first vowel) in all its word forms. Thus, the accented vowel marker is defined as *as1*. Accordingly, for this lemma we have the following complete annotation:

волат,NOUN+FLX=АБАЛІЦЫЯНІСТ+AccentP=ap2+AccentS=as1

(*AccentS* stands for an accent vowel position or accent syllable position).

This means that the accent falls on the second letter and first syllable in all word forms of this lemma.

Description of a floating accent in regard to letters

Along with lemmas containing a fixed accent, there are lemmas with a floating accent both in Belarusian and Russian dictionaries (table 1). So the words with an accent that shifts within one inflectional paradigm require more sophisticated ways of annotation. Still, the basic principle in this case is the same as in the previous situation: each accent position should be considered.

In the paradigm of the Russian verb *идти*, the accent falls on three different letters: 2, 3 or 4 (fig. 3). Depending on a word form, the stressed syllable is either the first or the second one. In order to refer to the floating

accent letter position, the marker *apN* is used; to indicate the floating accent vowel (or syllable) position, the marker *asN* is applied.

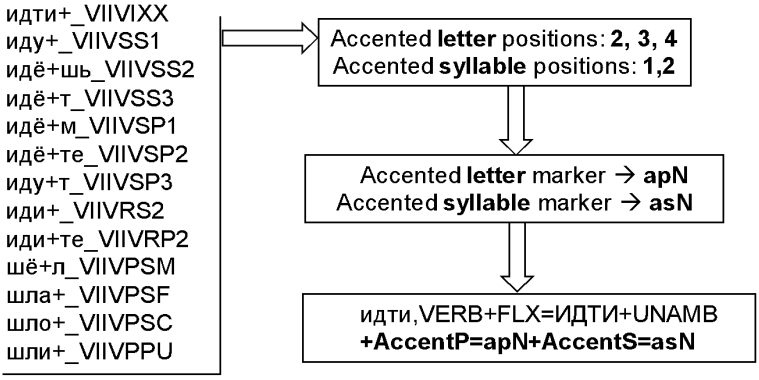


Figure 3: A marking process for a floating accent

In the dictionary one can find the following simplified description: *идти,VERB+FLX=ИДТИ +AccentP=apN+AccentS=asN.*

Figure 4 contains more examples of words with accentual information. It shows some excerpts of files with Belarusian (BN) and Russian (RN) lemmas both with fixed and floating accents.

```
...
абавязак,NOUN+FLX=АБАВЯЗАК+AccentP=ap5+AccentS=as3
адгалосак,NOUN+FLX=АБАВЯЗАК+AccentP=ap6+AccentS=as3
мама,NOUN+FLX=АБАТЫСА+AccentP=ap2+AccentS=as3
манастыр,NOUN+FLX=АБРУЧ+AccentP=apN+AccentS=asN
...
...
абстрактность,NOUN+FLX=АБСОЛЮТНОСТЬ+AccentP=ap6+AccentS=as3
аварийность,NOUN+FLX=АБСОЛЮТНОСТЬ+AccentP=ap5+AccentS=as3
адоптировать,VERB+FLX=АБЛАКТИРОВАТЬ+AccentP=ap6+AccentS=as3
мама,NOUN+FLX=АББАТИСА+AccentP=ap2+AccentS=as1
быстр,ADJECTIVE+FLX=БОДР+AccentP=apN+AccentS=asN
...
```

BN

RN

Figure 4: Examples of BN and RN with specified accentual information

When it comes to the problem of floating accent marking, the following algorithm is suggested. As an illustrative example, let's take the Russian noun *адресок* (fig. 5).

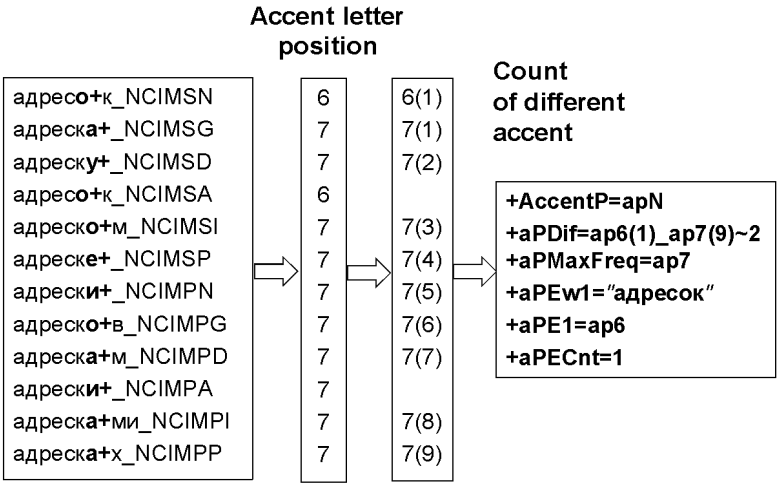


Figure 5: Floating accent marking of the noun *идти*

It is evident that there are two variable positions of the accented letter (*aPDif*): the 6th (in Nominative and Accusative) and the 7th (in all other cases). Besides definition of the stressed positions (*ap6* and *ap7*), the number of their occurrences in the paradigm should be counted: the 6th position occurs only in one case (for inequivalent word forms), while the 7th is stressed 9 times: *aPDif=ap6(1)_ap7(9)~*2. The quantity of different accents is also specially marked: *~2*.

The next step involves the detection of the most frequent accent position (*aPMaxFreq*). Since in the inflectional class *идти* word forms with the accent on the 7th character can be observed nine times more often – that is exactly the most frequent position: *+aPMaxFreq=ap7*. The less frequent positions cannot be ignored either. They are classified as exceptions (*aPE*) and can be described as follows: the word form *+aPEw1="адресок"* plus their respective ordinal numbers *+aPE1=ap6,* plus the quantity of its accentual exceptions *+aPECnt=1*.

The final step involves adding the resulting data to the lemma information: *адресок,NOUN+FLX=АВТОКРУЖОК*. Accordingly, the word *адресок* together with its accent marker takes the following form:

адресок,NOUN+FLX=АВТОКРУЖОК+AccentP=sN
+aPDif=ap6(1)_ap7(9)~2

+aPMaxFreq=ap7+aPEw1="адресок"
+aPE1=ap6+aPECnt=1 .

Description of a floating accent in regard to syllables

The algorithm for creating accent markers for syllables with floating accents is practically identical to the previous algorithm for letters. As an illustration, let's consider the Belarusian verb *лічыць* (fig. 6).

лічы+ць_VIC	2	2 (1)	**+AccentS=asN**
лічу+_VIIR1	2	2 (2)	**+aSDif=as1(6)_as2(7)~2**
лі+чыш_VIIR2	1	1 (1)	**+aSMaxFreq=as2**
лі+чыць_VIIR3	1	1 (2)	**+aSEw1="лічыць"+aSE1=as1**
лі+чым_VIIR1P	1	1 (3)	**+aSEw2="лічыце"+aSE2=as1**
лі+чыце_VIIR2P	1	1 (4)	**+aSEw3="лічаць"+aSE3=as1**
лі+чаць_VIIR3P	1	1 (5)	**+aSEw4="лічачы"+aSE4=as1**
лічы+_VIM2	2	2 (3)	**+aSEw5="лічыш"+aSE5=as1**
лічы+це_VIM2P	2	2 (4)	**+aSEw6="лічым"+aSE6=as1**
лічы+ў_VIIPM	2	2 (5)	**+aSEw7="лічыць"+aSE7=as2**
лічы+ла_VIIPF	2	2 (6)	**+aSEw8="лічыце"+aSE8=as2**
лічы+ла_VIIPN	2	2	**+aSECnt=8**
лічы+лі_VIIPP	2	2 (7)	
лі+чачы_VIB	1	1 (6)	

Figure 6: Floating syllabic accent marking of the verb *лічыць*

The marking procedure starts with the definition of all accented vowel positions (*aSDif*) and the subsequent counting of their occurrences within the inflectional paradigm of the verb *лічыць*. The variable ordinal number of stressed syllables can be either one (*as1*) or two (*as2*). The first position occurs six times while the second can be observed seven times: *+aSDif=as1(6)_as2(7)~2*, where the marker *~2* stands for the number of different accentual variants.

The second step is to detect the most frequently accented vowel position (*aSMaxFreq*). The 2nd syllable (7 times stressed) vs. the 1st syllable (6 times stressed): *+aSMaxFreq=as2*.

The less frequently accented positions are marked as exceptions, each one with its respective ordinal number:

+aSEw1="лічыць"+aSE1=as1
+aSEw2="лічыце"+aSE2=as1
+aSEw3="лічаць"+aSE3=as1
+aSEw4="лічачы"+aSE4=as1
+aSEw5="лічыш"+aSE5=as1

+aSEw6="лічым"+aSE6=as1.

Sometimes these exceptions can contain homographs, which also should be taken into consideration:

+aSEw1="лічыць"+aSE1=as2

+aSEw2="лічыце"+aSE2=as2.

Then the number of exceptions is added: *+aSECnt=8*. In the end all the obtained data are gathered as a marker, which is finally given to the analyzed lemma. As a result, we have floating markers in regard to both letters and syllables (table 2).

For letters	For syllables
аблічыць,VERB+FLX= БУРЫЦЬ +AccentP=apN +aPDif=ap6(8)_ap4(5)~2 +aPMaxFreq=ap6 +aPEw1="аблічыць"+aPE1=ap4 +aPEw2="аблічыце"+aPE2=ap4 +aPEw3="аблічаць"+aPE3=ap4 +aPEw4="аблічыш"+aPE4=ap4 +aPEw5="аблічым"+aPE5=ap4 +aPEw1="аблічыць"+aPE1=ap6 +aPEw2="аблічыце"+aPE2=ap6 +aPECnt=7	аблічыць,VERB+FLX= БУРЫЦЬ +AccentS=asN +aSDif=as3(8)_as2(5)~2 +aSMaxFreq=as3 +aSEw1="аблічыць"+aSE1=as2 +aSEw2="аблічыце"+aSE2=as2 +aSEw3="аблічаць"+aSE3=as2 +aSEw4="аблічыш"+aSE4=as2 +aSEw5="аблічым"+aSE5=as2 +aSEw1="аблічыць"+aSE1=as3 +aSEw2="аблічыце"+aSE2=as3 +aSECnt=7

Table 2: Excerpt with complete floating accent markers for both letters and syllables

Of course, floating markers for letters and syllables can be combined in one lemma, depending on the requirements for the dictionary module.

Annotating lemmas with accent markers in NooJ

Thanks to the NooJ text-annotating function, apart from grammatical information, one can observe how accent markers work (fig.7).

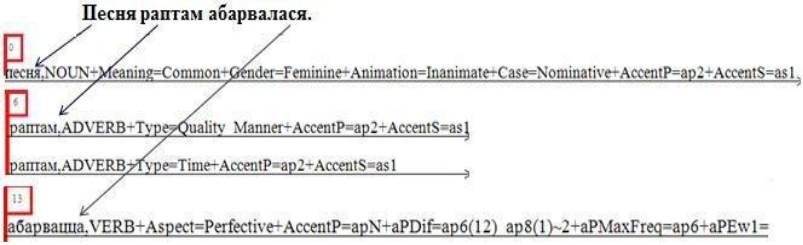


Figure 7: Grammatically and accentually annotated Belarusian phrase

Moreover, *Locate Pattern* gives an opportunity to get specific bigrams both for the Belarusian and Russian texts. Figure 8 illustrates search request specification in order to obtain all phrases consisting of a noun with an adjective, both of which have an accent on the first syllable.

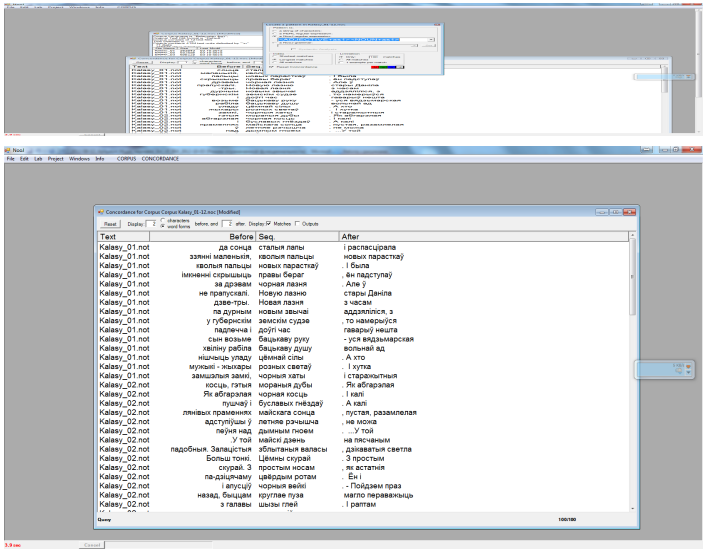


Figure 8: A specified search request for necessary accentual data and the results of its application

The next step involved the creation of a syntactic grammar for defining accurate accent vowel positions in any word form (fig. 9):

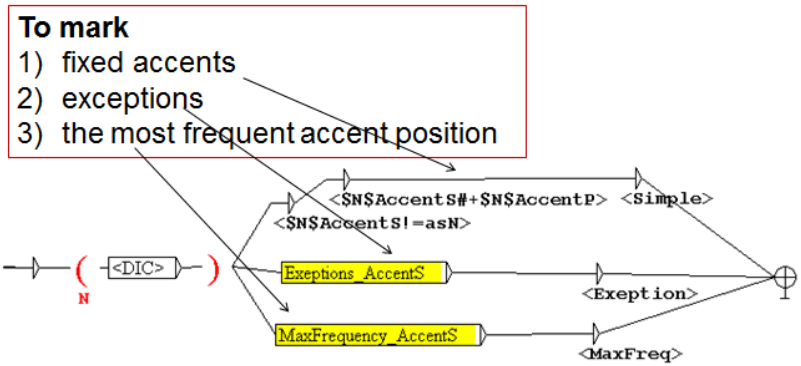


Figure 9: A syntactic grammar based on accentual data

The grammar consists of three parts. The first part defines a fixed accent (when *AccentS* is not *asN*). Part two is designed as a subgraph depicting the situation when a text contains word forms with accent exceptions (fig. 10).

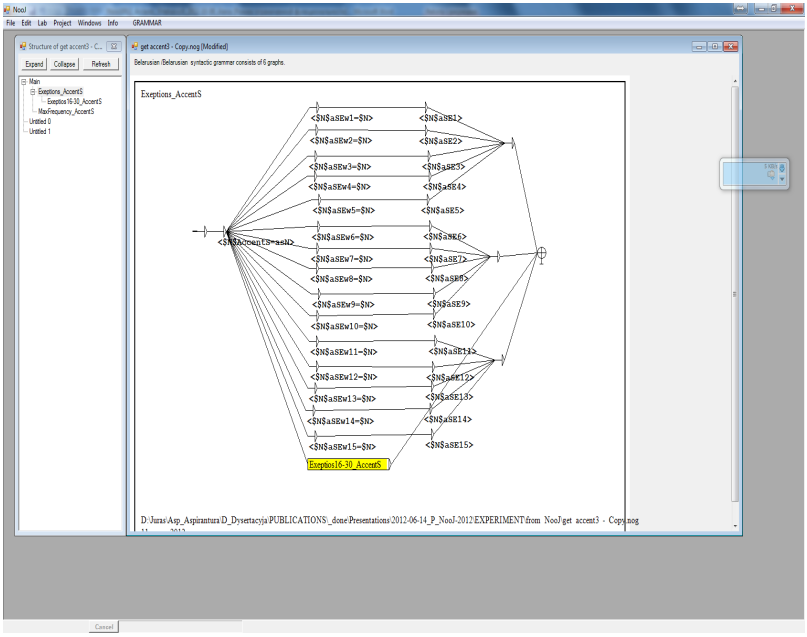


Figure 10: Subgraph 2 for marking accent vowel exceptions

The third part (the second subgraph) is generated by NooJ when the previous ones don't occur, that is, the case with the most frequent accent vowel positions (fig. 11).

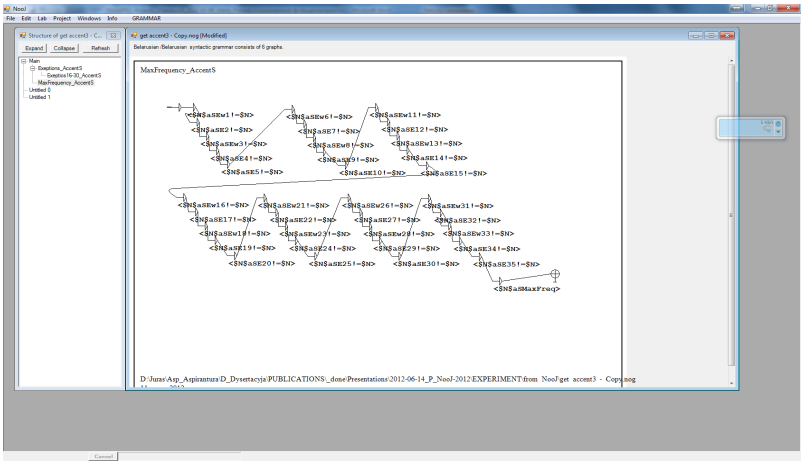


Figure 11: Subgraph 3 for marking the most frequent vowel positions in words

Table 3 gives illustrative examples of annotations given to lemmas after applying the constructed syntactic grammar.

BM	RM
быў/<as1><MaxFreq>	привыкший/<as2+ap5><Simple>
непадобны/<as3+ap6><Simple>	тут/<as1+ap2><Simple>
на/<as1+ap2><Simple>	тоже/<as1+ap2><Simple>
другі/<as2+ap5><Simple>	стал/<as1+ap3><Simple>
і/<as1+ap1><Simple>	интересоваться/<as5><Exeption>
ўсё/<as2><MaxFreq>	новыми/<as1+ap2><Simple>
няўлоўна/<as2+ap5><Simple>	лицами/<as1><MaxFreq>
падобны/<as2+ap4><Simple>	Сидя/<as1+ap2><Simple>
Гэта/<as1+ap2><Simple>	павильоне/<as3+ap7><Simple>

Table 3: Words annotated by means of the obtained syntactic grammar for both Belarusian (BM) and Russian (RM) modules

Conclusion

The first versions of the Belarusian and Russian NooJ modules with accentual information have been completed, which is extremely useful in the field of text processing for human perception, as well as in the area of learning Belarusian and Russian.

Thanks to text annotation operations (by means of the created syntactic grammar), it is possible to locate phonetic words and specific phases for reading and also to check understanding of phrases, sentences, and texts. Accordingly, the modules can be used by journalists, copyrighters, and foreigners who are studying the Belarusian or Russian language, etc.

Acknowledgements

We would like to thank Xavier Blanco-Escoda for his help in preparing the theoretical part of this paper. Many thanks to the linguist Adam Morrison for his help in revising the language of this paper.

References

Hetsevich, Y. Overview of Belarusian And Russian dictionaries and their adaptation for NooJ / Y. Hetsevich, S. Hetsevich // Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 Intern. Conf. / eds. Vučković Kristina, Bekavac Božo, Silberztein Max. – Newcastle : Cambridge Scholars Publishing, 2012. – P. 29–40.