Addition of IPA Transcription to the Belarusian NooJ Module

Stanislau Lysy(✉), Hanna Stanislavenka(✉), and Yury Hetsevich

The United Institute of Informatics Problems, National Academy of Sciences of Belarus, Minsk, Belarus
stanislau.lysy@gmail.com, hanna.stanislavenka@gmail.com, yury.hetsevich@gmail.com

Abstract. This paper is based on earlier research works where the possibility was shown to represent a linguistic phonetic level in NooJ. The phonetic level was developed for the Belarusian NooJ Module and was embodied with the help of dictionaries, with transcriptions in different formats. However, while Cyrillic transcriptions traditionally used for Belarusian were generated and compiled correctly, the international format had inaccuracies. Among these, one is connected with stress positions in IPA transcription. The task of adding IPA transcription requires the solving of problems concerning splitting words into syllables. Thus, firstly we need to solve the problem of syllabification of words in Belarusian.

Keywords: NooJ · Belarusian language · Phonetics · IPA transcription

1 Introduction

NooJ as a linguistic development environment supplies tools to describe all levels of natural language. The main sphere of our interest in this research paper is phonetics. In earlier research papers, we have shown that it is possible to represent a phonetic level of language in NooJ [1].

We demonstrated this new yet undeveloped possibility of NooJ in the Belarusian module by creating the dictionary with transcriptions and by developing morphological NooJ grammars with the help of which one can create phonetic transcriptions of arbitrary orthographic words. Results that were presented in NooJ 2015 Proceedings are the following [1]:

1. The system of generating of dictionaries in NooJ format with four types of phonetic transcriptions was created.
2. NooJ-dictionary was successfully compiled and tested. It contained 46.384 first forms of nouns in 2015.
3. Development of morphological NooJ grammar for creating a phonetic transcription for orthographic words was launched.

However, there are still some issues in generating of transcriptions. One of them is connected with stress position in some types of transcriptions (in particular IPA

L. Barone et al. (Eds.): NooJ 2016, CCIS 667, pp. 14–22, 2016.
DOI: 10.1007/978-3-319-55002-2_2

transcription), where stresses (ˈ) are put before the stressed syllables in the phonetic word transcription. The solution of this issue requires multitasking work that we split into three main steps. Syllabification algorithm for generation of IPA transcriptions of orthographic words in Belarusian should be developed. Then it would be possible to advance high-quality tool for the generation of phonetic transcription of orthographic words in Belarusian [2]. With the help of the advanced tool, a dictionary in NooJ format for nouns and verbs can be created.

2 IPA-Transcription in Belarusian NooJ-Module

Above mentioned NooJ-dictionaries contains the following types of phonetic transcription:

- Cyrillic transcription (based on [3]);
- International Phonetic Alphabet (or IPA) [4];
- Simplified international format [5];
- Extended Speech Assessment Methods Phonetic Alphabet (or X-SAMPA) [6].

This paper is dedicated to the IPA as it is used worldwide by lexicographers, foreign language students and teachers, linguists, speech-language pathologists, other researchers, as well as singers, actors, anchors. Available NooJ dictionaries with international transcription format will contribute to the development and formalizing of the Belarusian language.

There are various traditions of marking stresses in a word transcription. Thus a stress mark in a word can be connected with a certain allocation of a stressed vowel sound (a stress mark above or after vowel, vowel written in uppercase, and so on), or with an allocation of a stressed syllable (a stress mark before syllable, after syllable, writing syllable in uppercase, and so on). The IPA-transcription marks stress not on the vowel (as in tradition for the Belarusian language) but before an accented syllable. The system of NooJ dictionaries generation puts a stress mark incorrectly for the IPA-format [2]:

смертнасць,NOUN+TranscriptionIPA=[sʲmʲᵒɛrtnasʲtsʲ]

The correct variant is as follows:

смертнасць,NOUN+TranscriptionIPA=[ᵒsʲmʲɛrtnasʲtsʲ]

Such seemingly irrelevant incorrectness can lead to misconstrued usage of the transcription of Belarusian words and correspondingly to the wrong results of research, and so on. A solution of the problem lies in developing of syllabification algorithm which is for the first time presented in the next part of paper.

3 Syllabification Process

As it was mentioned above syllabification process is firstly demonstrated as a computational linguistic problem for the Belarusian language in this paper. For the algorithm

syllabification rules were developed according to the book "Phonetics of the Belarusian Literary Language" [3]. Developed rules have the following view:

'aa' => 'a|a'
'aka' => 'a|ka'
'ama' => 'a|ma'
'akka' => 'a|kka'
'akma' => 'a|kma'
'amka' => 'am|ka'
'amma' => 'am|ma'
'akkka' => 'a|kkka'
'akkma' => 'a|kkma'
'akmka' => 'akm|ka'
…

where "a" is an arbitrary vowel, "k" is an obstruent consonantal, "m" is a sonorant consonantal, and "|" stands for a syllable border.

For instance, the rule "'akmka' => 'akm|ka'" means that if between two vowels there are three consonant phonemes: one is obstruent, one is sonorant and one is again obstruent, first two phonemes will be in one syllable, and syllable border must be put before the second obstruent phoneme.

The algorithm works with certain compulsory data and has 5 main steps. Input data for the algorithm is:

– arbitrary word in Belarusian in an allophonic format W_a;
– set of syllabification rules $R_{syll} = \ll Pt_1, Pt_{syll1} > ,\ldots, < Pt_m, Pt_{syllm} \gg$, where Pt_i – sequence of allophonic class notations in the i^{th} rule, Pt_{sylli} – sequence of allophonic class notations with a syllable border in the i^{th} rule, $i = 1\ldots m$, m –number of syllabification rules;
– set of correspondences "allophone – class of allophone", $A_{class} = \ll A_1, Cl_1 > , \ldots, < A_n, Cl_n \gg$, where A_i – allophone code, Cl_i – allophone class notation, $i = 1 \ldots n$, n – number of allophones.

Note that in order to determine a position of a syllable border allophones were divided into three classes: vowel (notation "a"), obstruent consonantal (notation "k"), and sonorant consonantal (notation "m"). Scheme of the algorithm (Fig. 1) and description of its work steps are presented below:

Step 1. An input word W_a is divided into allophones, that are consecutively put into a list $L_a = < A_{w1},\ldots, A_{wn} >$, where A_{wi} – the i^{th} allophone in a word, $i = 1\ldots n$, n – number of allophones in a word.

Step 2. Sequence of allophone class notations is formed from the list L_a. For each allophone A_{wi} from the list L_a there is an allophone class notation Cl_i in the list A_{class}. Found notations are collected in a word pattern according to the classes of its allophones Pt_w.

Step 3. Consecutive browsing of Pt_i pattern from the set of syllabification rules R_{syll} is held. Algorithm searches each Pt_i pattern in a pattern of the input word template $Pt_{w.}$ When occurrence of the Pt_i pattern is found in the input word pattern $Pt_{w,}$ Pt_i pattern is

replaced with relative pattern but with a syllable border Pt_{sylli} from the set of rules R_{syll} in the input word pattern Pt_w.

Step 4. The algorithm is browsing symbols of the modified Pt_w pattern. If i^{th} symbol of the Pt_w pattern is a syllable border, then the i^{th} element with a syllable border is added to a list if allophones L_a of the input word $W_{a.}$

Step 5. The modified allophone list L_a of the input word W_a is elementwise assembled into one allophonic word with syllable borders W_{ares}.

Step 6. End of algorithm. The result of the algorithm is a syllabified allophonic word.

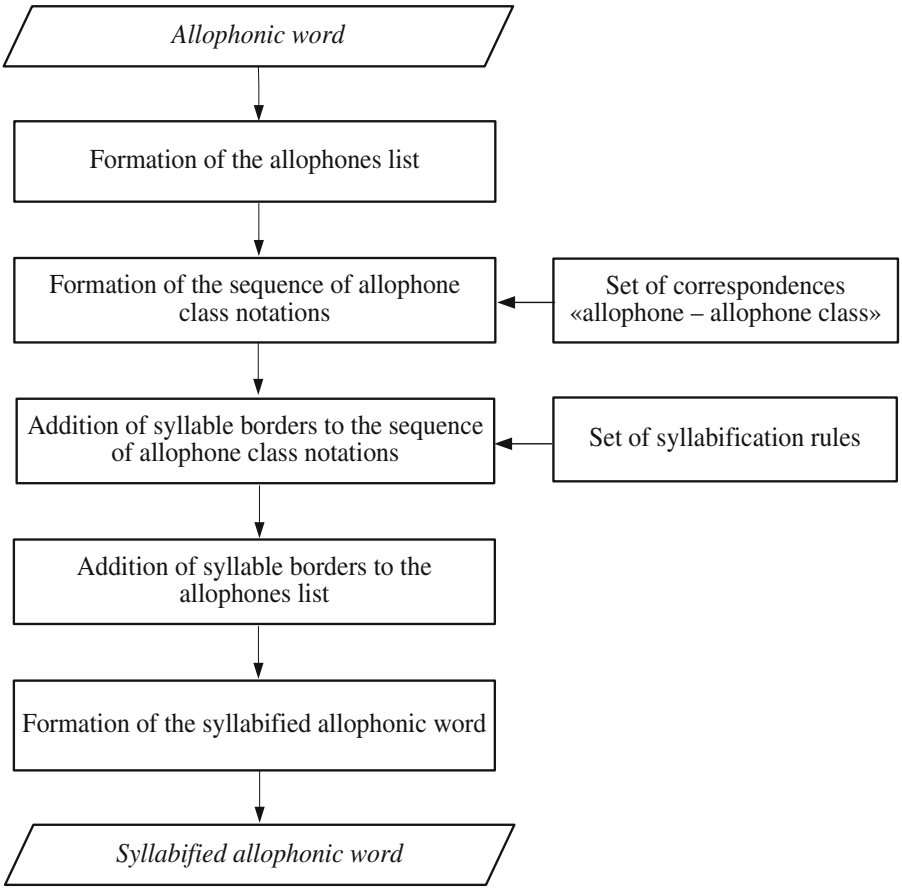

Fig. 1. General scheme of allophonic word syllabification algorithm

4 Software Prototype "Syllabifier"

For testing above described algorithm, a web-service "Syllabifier" was created. "Syllabifier" is a program prototype that embodies the allophonic word syllabification algorithm [7]. Its interface is presented in Fig. 2.

Fig. 2. Interface of the web-service "Syllabifier"

The web-service "Syllabifier" can take as input data both word lists and texts in Belarusian. Input data is sent to the text processor of the Internet-version of text-to-speech synthesizer, where words are extracted from an input list or text, normalized and united into syntagmas [8].

Then each syntagma is sent to the phonetic processor of the synthesizer and there allophonic view of syntagmas is formed with the help of "grapheme-to-phoneme" and "phoneme-to-allophone" algorithms. The allophonic view of syntagmas is processed by the syllabification algorithm.

As a result of the work of the web-service "Syllabifier", one gets the input text in allophonic format, as well as the input text in an allophonic format with syllable border marks ">" (Fig. 3).

Fig. 3. Interface of web-service "Syllabifier" with output data

5 NooJ-Dictionaries Compilation

The described above algorithm was inbuilt to online service "Orthoepic Dictionary Generator" in which there is an option – "Headwords processing in NooJ format" [2]. It means that the first word in every line of the input text is processed and transcription of the word written in the NooJ format appears. After processing we get material with

which we create Dictionary for NooJ. Here the fragment of generated NooJ dictionary is shown:

смертнасць,NOUN+TranscriptionCyr=[с'м'э́ртнас'ц']+TranscriptionIPA=['sʲmʲɛrtnasʲʦʲ]
смертнік,NOUN+TranscriptionCyr=[с'м'э́ртн'ік]+TranscriptionIPA=['sʲmʲɛrtnʲik]
смертніца,NOUN+TranscriptionCyr=[с'м'э́ртн'іца]+TranscriptionIPA=['sʲmʲɛrtnʲiʦa]
смертухна,NOUN+TranscriptionCyr=[с'м'э́ртухна]+TranscriptionIPA=['sʲmʲɛrtuxna]
смерць,NOUN+TranscriptionCyr=[с'м'э́рц']+TranscriptionIPA=['sʲmʲɛrʦʲ]
смерч,NOUN+TranscriptionCyr=[с'м'э́рч]+TranscriptionIPA=['sʲmʲɛrʧ]
сметнік,NOUN+TranscriptionCyr=[с'м'э́тн'ік]+TranscriptionIPA=['sʲmʲɛtnʲik]
сметніца,NOUN+TranscriptionCyr=[с'м'э́тн'іца]+TranscriptionIPA=['sʲmʲɛtnʲiʦa]
сметнішча,NOUN+TranscriptionCyr=[с'м'э́тн'ішча]+TranscriptionIPA=['sʲmʲɛtnʲiɕʧa]
смех,NOUN+TranscriptionCyr=[с'м'э́х]+TranscriptionIPA=['sʲmʲɛx]
смехата,NOUN+TranscriptionCyr=[с'м'эхата́]+TranscriptionIPA=[sʲmʲɛxa'ta]
смешкі,NOUN+TranscriptionCyr=[с'м'э́шк'і]+TranscriptionIPA=['sʲmʲɛɕkʲi]
смірна,NOUN+TranscriptionCyr=[с'м'і́рна]+TranscriptionIPA=['sʲmʲirna]
смог,NOUN+TranscriptionCyr=[смо́х]+TranscriptionIPA=['smɔx]
смок,NOUN+TranscriptionCyr=[смо́к]+TranscriptionIPA=['smɔk]
смоква,NOUN+TranscriptionCyr=[смо́ква]+TranscriptionIPA=['smɔkva]

As it is seen on the fragment above, stresses in IPA transcriptions are put in correct places.

With the help of the web-service "Orphoepic Dictionary Generator" we can compile dictionaries for NooJ with many entries. Thus we made a dictionary with nouns and with verbs [2]. A dictionary for nouns contains more than 49 000 entries An extract from this NooJ dictionary one can see on Fig. 4.

Entry	Category	TranscriptionCyr	TranscriptionIPA
смертнасць	NOUN	[с'м'э́ртнас'ц']	['sʲmʲɛrtnasʲʦʲ]
смертнік	NOUN	[с'м'э́ртн'ік]	['sʲmʲɛrtnʲik]
смертніца	NOUN	[с'м'э́ртн'іца]	['sʲmʲɛrtnʲiʦa]
смертухна	NOUN	[с'м'э́ртухна]	['sʲmʲɛrtuxna]
смерць	NOUN	[с'м'э́рц']	['sʲmʲɛrʦʲ]
смерч	NOUN	[с'м'э́рч]	['sʲmʲɛrʧ]
сметнік	NOUN	[с'м'э́тн'ік]	['sʲmʲɛtnʲik]
сметніца	NOUN	[с'м'э́тн'іца]	['sʲmʲɛtnʲiʦa]
сметнішча	NOUN	[с'м'э́тн'ішча]	['sʲmʲɛtnʲiɕʧa]
смех	NOUN	[с'м'э́х]	['sʲmʲɛx]
смехата	NOUN	[с'м'эхата́]	[sʲmʲɛxa'ta]
смешкі	NOUN	[с'м'э́шк'і]	['sʲmʲɛɕkʲi]
смірна	NOUN	[с'м'і́рна]	['sʲmʲirna]
смог	NOUN	[смо́х]	['smɔx]
смок	NOUN	[смо́к]	['smɔk]
смоква	NOUN	[смо́ква]	['smɔkva]

Fig. 4. Extract from the NooJ dictionary for nouns

Moreover, a dictionary with transcriptions for verbs was compiled. It contains more than 30 000 entries. An extract from this NooJ dictionary one can see in Fig. 5. With the help of the IPA transcription, anyone can read a word in Belarusian.

Entry	Category	TranscriptionCyr	TranscriptionIPA
дзынкаць	VERB	[z'ы́нкац']	['d͡zʲɨnkatsʲ]
дзынкнуць	VERB	[z'ы́нкнуц']	['d͡zʲɨnknutsʲ]
дзьмухаць	VERB	[z'му́хац']	['d͡zʲmuxatsʲ]
дзьмухнуць	VERB	[z'му́хнуц']	['d͡zʲmuxnutsʲ]
дзьмуцца	VERB	[z'му́ц:а]	['d͡zʲmutstsa]
дзьмуць	VERB	[z'му́ц']	['d͡zʲmutsʲ]
дзюбануць	VERB	[z'убану́ц']	[d͡zʲuba'nutsʲ]
дзюбацца	VERB	[z'у́бац:а]	['d͡zʲubatstsa]
дзюбаць	VERB	[z'у́бац']	['d͡zʲubatsʲ]
дзюбнуцца	VERB	[z'у́бнуц:а]	['d͡zʲubnutstsa]
дзюбнуць	VERB	[z'у́бнуц']	['d͡zʲubnutsʲ]
дзюрчаць	VERB	[z'урчаи́ц']	[d͡zʲur'ʧatsʲ]
дзюрчэць	VERB	[z'урчэи́ц']	[d͡zʲur'ʧɛtsʲ]
дзяваццa	VERB	[z'аваи́ц:а]	[d͡zʲa'vatstsa]
дзяваць	VERB	[z'аваи́ц']	[d͡zʲa'vatsʲ]
дзяжурыць	VERB	[z'ажуры́ць']	[d͡zʲa'ʐurɨtsʲ]
дзякаваць	VERB	[z'ака́ваць']	['d͡zʲakavatsʲ]
дзяліцца	VERB	[z'ал'іи́ц:а]	[d͡zʲa'lʲitstsa]
дзяліць	VERB	[z'ал'іи́ц']	[d͡zʲa'lʲitsʲ]
дзяржацца	VERB	[z'аржаи́ц:а]	[d͡zʲar'ʐatstsa]
дзяржаць	VERB	[z'аржаи́ц']	[d͡zʲar'ʐatsʲ]
дзяўбацца	VERB	[z'аўбаи́ц:а]	[d͡zʲaw'batstsa]
дзяўбаць	VERB	[z'аўбаи́ц']	[d͡zʲaw'batsʲ]
дзяўбці	VERB	[z'аўпц'іи́]	[d͡zʲaw'ptsʲi]
дзяўбціся	VERB	[z'аўпц'іс́'а]	[d͡zʲaw'ptsʲisʲa]

Fig. 5. Extract from the NooJ dictionary for verbs

6 Conclusion and Further Steps to Take

Summarizing our research the following results should be underlined. Firstly, syllabification algorithm for generation of IPA transcriptions of orthographic words in Belarusian was developed and tested on a special tool. Secondly, high-quality tool for generation of phonetic transcription of orthographic words in Belarusian that was presented in the results of the previous research was advanced with the help of the developed syllabification algorithm. Thirdly, dictionaries in NooJ format that include correct phonetic transcriptions in IPA format for nouns and verbs were created.

The results presented in the paper are a part of the ongoing research. Provided that plans are to examine correctness of the IPA transcriptions for NOUN and VERB; to

build NooJ morphology grammar for letter-to-phoneme conversion with right syllable positions; to add IPA-transcription for Adjectives and Adverbs NooJ dictionaries.

The results of this research can be used in introducing and learning the norms of the literary pronunciation of the Belarusian language. The process of syllabification can also be further implemented into other computational linguistic programs and into NooJ modules of other languages.

References

1. Zahariev, V., Lysy, S., Hiuntar, A., Hetsevich, Y.: Grapheme-to-phoneme and phoneme-to-grapheme conversion in belarusian with NooJ for TTS and STT systems. In: Okrut, T., Hetsevich, Y., Silberztein, M., Stanislavenka, H. (eds.) NooJ 2015. CCIS, vol. 607, pp. 137–150. Springer, Heidelberg (2016). doi:10.1007/978-3-319-42471-2_12
2. Orthoepic Dictionary Generator. Available at http://corpus.by/OrthoepicDictionaryGenerator (2016)
3. Падлужны, А.І.: Фанетыка беларускай літаратурнай мовы. Навука і тэхніка, Мінск (1989)
4. The International Phonetic Alphabet and the IPA Chart. Available at http://www.internationalphoneticassociation.org/content/ipa-chart (2016)
5. Кошчанка, У.А.: Беларуска-англійскі размоўнік. Артыя Груп, Мінск (2010)
6. Computer-coding the IPA: a proposed extension of SAMPA (2016). http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm
7. Syllabifier (2016). http://corpus.by/Syllabifier
8. Text-to-Speech Synthesizer (2016). http://corpus.by/TextToSpeechSynthesizer

http://www.springer.com/978-3-319-55001-5

Automatic Processing of Natural-Language Electronic Texts with NooJ
10th International Conference, NooJ 2016, České Budějovice, Czech Republic, June 9-11, 2016, Revised Selected Papers
Barone, L.; Monteleone, M.; Silberztein, M. (Eds.)
2016, XII, 259 p. 155 illus., Softcover
ISBN: 978-3-319-55001-5