Semi-automatic Part-of-Speech Annotating for Belarusian Dictionaries Enrichment in NooJ

Yu. Hetsevich[1], V. Varanovich[2], E. Kachan[1], I. Reentovich[1], S. Lysy[1]

[1] United Institute of Informatics Problems, Minsk, Belarus
[2] Belarusian State University, Minsk, Belarus

e-mails: yury.hetsevich@gmail.com, gamrat.vvv@gmail.com, evgeniakacan@gmail.com, mwshrewd@gmail.com, stanislau.lysy@gmail.com

Abstract. The paper describes the algorithm for the Belarusian main dictionaries enrichment in NooJ on the basis of the first one-million corpus for the Belarusian NooJ module. From the broad list of possible subject categories, the corpus focuses on literature of fiction, historical literature, medical literature, scientific literature, sociological literature and etc. The corpus is considered to be the finest source for searching unknown words of different domains.

So, for this purpose a specific algorithm for automatic word paradigms generation have been agreed to develop.

The authors have worked out a mechanism for further processing of all unknown (unique) words extracted from the corpus and adding them to the present dictionary on the basis of the Belarusian NooJ module. The algorithm is based on the required grammatical information of an entire word.

Keywords: corpora, Belarusian NooJ-module, part-of-speech tagging, countercheck, lexicology, dictionary, algorithm, online-service, paradigm

1 INTRODUCTION

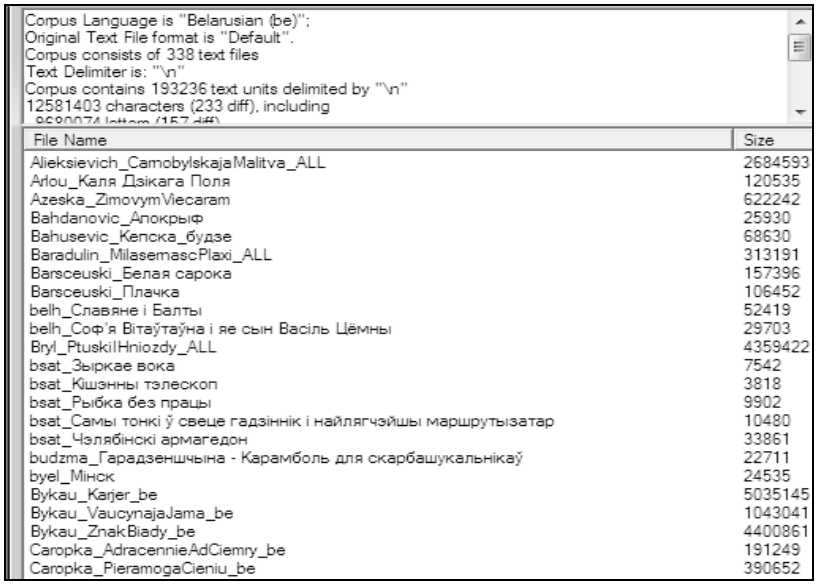
This research is a continuation of the overall work on the creation of The First One-million Corpus [1] for the Belarusian NooJ Module [2], which is applicable in a variety of thematic spheres and can be used in any linguistic research.

Today, the first one-million Belarusian corpus for the Belarusian NooJ module is used for solving the following fundamental tasks: optimizing and expanding the development of high-quality linguistic algorithms for the electronic text pre-processing in the TTS (Text-to-Speech) system [6].

The main task of the research is to work out a mechanism for further annotation of different categories and paradigms according to flexion classes of all unknown words extracted from Belarusian corpus and then to compose processed words to main Belarusian NooJ dictionary [3].

2 The Part-of-Speech Tagging Countercheck of unknown words

The corpus (see fig.1) was developed in an appropriate format for Nooj program last year. It composes 338 text files, where the total number of all word forms in the texts is more than 1 million, 197712 of which are unique well-known word forms (received by the <DIC> query, 1 occurrence per match) and 50186 – the unique unknown word forms (received by the <UNK> query, 1 occurrence per match) [1].


Corpus Language is "Belarusian (be)";
Original Text File format is "Default".
Corpus consists of 338 text files
Text Delimiter is: "\n"
Corpus contains 193236 text units delimited by "\n"
12581403 characters (233 diff), including

File Name	Size
Alieksievich_CamobylskajaMalitva_ALL	2684593
Arlou_Каля Дзікага Поля	120535
Azeska_Zimovym\Viecaram	622242
Bahdanovic_Апокрыф	25930
Bahusevic_Кепска_будзе	68630
Baradulin_MilasemascPlaxi_ALL	313191
Barsceuski_Белая сарока	157396
Barsceuski_Плачка	106452
belh_Славяне і Балты	52419
belh_Соф'я Вітаўтаўна і яе сын Васіль Цёмны	29703
Bryl_PtuskiIHniozdy_ALL	4359422
bsat_Зыркае вока	7542
bsat_Кішэнны тэлескоп	3818
bsat_Рыбка без працы	9902
bsat_Самы тонкі ў свеце гадзіннік і найлягчэйшы маршрутызатар	10480
bsat_Чэлябінскі армагедон	33861
budzma_Гарадзеншчына - Карамболь для скарбашукальнікаў	22711
byel_Мінск	24535
Bykau_Karjer_be	5035145
Bykau_VaucynajaJama_be	1043041
Bykau_ZnakBiady_be	4400861
Caropka_AdracennieAdCiemry_be	191249
Caropka_PieramogaCieniu_be	390652

Fig. 1. The fragment of the first one-million corpus for the Belarusian NooJ module

Then a specialized dictionary of unknown words was composed for easy determination of categories for these words, firstly automatically, and then it was checked by linguists-experts (see fig.2).

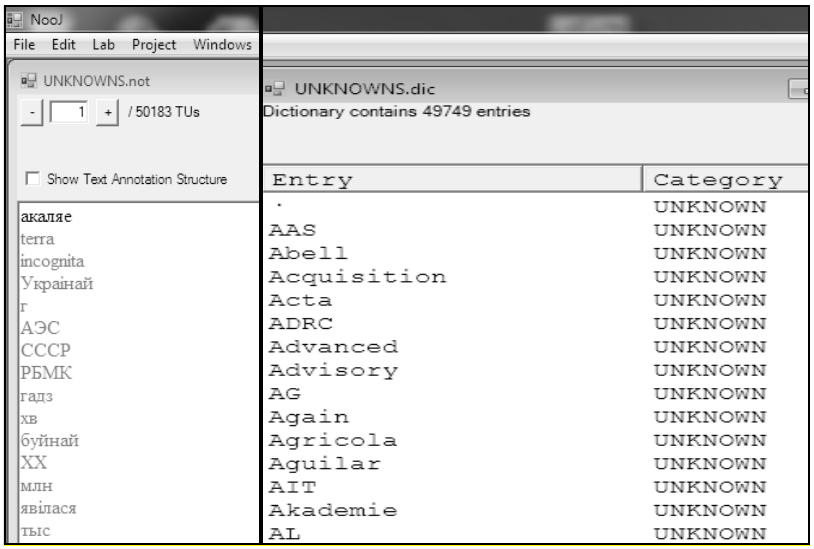


Fig. 2. The fragment of the specialized dictionary of unknown words

During the research of this year, new results to determine the unique categories of unknown words from the first million corpus for the Belarusian NooJ module were received. Statistical data are presented in table 1 under date of 25.05.2016.

- The total quantity of all unique unknown words after their lowercase conversion and spellcheck is 47206.
- The total quantity of annotated unique unknown words is 26983.
- The total quantity of unique unknown words annotated by the categories NOUN, ADJECTIVE and VERB is 21845.
- The total quantity of unique unknown words annotated by the remaining categories is 5138.
- The total quantity of unannotated unique unknown words is 18836.

Table 1. The statistics of **POS-annotated** and **POS-unannotated** unique unknown words in the first one-million corpus for the Belarusian NooJ module

MAIN INFORMATION ABOUT THE UNKNOWN WORDS	QUANTITY	QUANTITY (%)
All unique unknown words (according to NooJ results)	**50 183**	**100,00**
All unique unknown words (after their lowercase conversion and spellcheck)	**47 206**	**94,07**
The processed part	**2 977**	**5,93**

Words Annotated by Categories: general and additional	**QUANTITY**	**QUANTITY (%)**
NOUN	**12 303**	**45,60**
VERB	**4 843**	**17,95**
ADJECTIVE	**4 699**	**17,41**
PARTICIPLE	**1 495**	**5,54**
FOREIGN	**1387**	**5,14**
ADVERB	**981**	**3,64**
PRONOUN, NUMERAL, PREPOSITION, CONJUNCTION, PARTICLE, PARENTHESIS, INTERJECTION, PREDICATIVE	**553**	**2,05**
ABBREVIATION	**382**	**1,42**
GERUND	**340**	**1,26**
TOTAL ANNOTATED AND UNANNOTATED WORDS	**QUANTITY**	**QUANTITY (%)**
WORDS UNANNOTATED	**20 223**	**42,84**
WORDS ANNOTATED	**26 983**	**57,16**
Words annotated by the NOUN, ADJECTIVE, VERB categories	***21 845***	***80,96***
Words annotated by other categories	***5 138***	***19,04***

The Part-of-Speech Tagging Countercheck of unknown words was realized with the help of Levenshtein algorithm [4]. The algorithm revealed parts of speech of unknown words, picked up a possible correct form of the usage, and also gave an index of probability of correct forms. The stage of manual editing was carried out after computer-assisted Part-of-Speech detection: all parts of speech were checked by linguists-experts. In the case of the correct Part-of-Speech detection by the algorithm, this line of the table was marked as "true" (1). In an opposite case – "false" (0) (see fig.3).

ID	Seq -- specifies UNKNOWN words	DictSeq specifies words chosen by the algorithm	Similarity	PartOfSpeech	Ja&H
27073	рэфарміраванне	рэфармаванне	0,857142857	NOUN	1
27075	рэфарміраванню	фарміраванню	0,857142857	NOUN	1
27076	рэфарміравання	фарміравання	0,857142857	NOUN	1
27077	рэфектар	рэфлектар	0,888888889	NOUN	1
27078	рэфлэксаў	рэфлексаў	0,888888889	NOUN	1
27079	рэформ	рэформа	0,857142857	NOUN	1
27083	рэшаты	рэшата	0,833333333	NOUN	1
27086	рэштка	рэшткаў	0,857142857	NOUN	1
27089	рэшткай	рэштай	0,857142857	NOUN	1
27092	рэштку	рэшткі	0,833333333	NOUN	1
27094	рэюць	грэюць	0,833333333	VERB	1
27095	сілуэтам	сілуэтам	0,875	NOUN	1

Fig.3. Countercheck of Annotated Categories in unknown words by the three linguists-experts

The semi-automatic annotating of unknown words helped to form the version of the dictionary with annotated grammatical categories. It will be additional to the main dictionary, general_be.nod Dictionary of the Belarusian NooJ module (see fig.4). All parts of speech were tagged. The main difficulty was to find out the most effective way to generate all wordforms.

Fig. 4. A fragment of the latest additional dictionary for the Belarusian *NooJ* module

3 The algorithm for further annotating of all paradigms according to flexion classes in Nooj Format

The main concept is not only to get the category of a word but also the whole paradigm. The algorithm, which was worked out by the team, is the basis for the automatic generation of word paradigms. It consists of 16 consecutive interdependent steps. The algorithm outputs one or several most suitable paradigms of a word. It searches the nearest paradigm(s) in matches of the last letters of the word user needs to get the paradigm.

The algorithm for further annotating of all paradigms according to flexion classes is described below:

1. To search for a word in the dictionary of flexion classes. If the dictionary contains the word, then to display to the user a complete paradigm and go to step 16. If a word is a homograph, then do step 15. If the dictionary of flexion classes does not contain the word, then do step 2.
2. To propose the user to specify a part of speech of the word.
3. Depending on chosen part of speech to propose the user to specify the grammatical features (with the possibility to leave the fill-in-the-blank fields empty if the user does not know the features).
4. To define whether it is a changeable or un unchangeable word. If unchangeable, to display a word with annotation, then do step 16. If it is changeable, then step 5.
5. To take an unprocessed input word form for further processing, then step 6.
6. To search in the dictionary of flexion classes words (with marked grammatical features) that ends with the input word in current state. If the dictionary of flexion classes contains such words, then step 8, otherwise step 7.
7. To remove the first letter of the input word in current state. Then step 6.
8. To divide obtained words into "base" and "tail", where "tail" is a part of obtained word coinciding with input word in current state, and "base" – the rest part of the obtained word.
9. To select the "base" in the original input word by cutting the "tail".
10. To separate the "tails" in other word forms of the obtained words and to attach them all to the "base" of the original input word.
11. If there are more than one found words, then do steps 8-10 for all words.
12. To compare obtained paradigms. To delete all the duplicates, leaving only unique paradigms.
13. If the user has given only one form, then step 14. If there are more unprocessed word forms given by the user as an input, then step 5 for the other word forms, which the user has given, but search in the list of generated paradigms, not in the dictionary in step 6. If all word forms given by the user were processed, then step 14.
14. If in the result only one paradigm was found, then step 15. If more than one, to compare obtained paradigms. To delete all the duplicates, leaving only unique paradigms. Then step 15.
15. To give the user obtained word paradigms, to ask to specify the correct one.
16. The end of the algorithm.

Random 10 unknown words were selected for each of 14 categories to test the effectiveness of the algorithm. A total amount of words is 140 (see fig.5).

Short part-of-speech dictionary for Nooj			
word	**SK**	**category**	**COMMENTS**
абмакванне	,	NOUN	
болечка	,	NOUN	
вырай	,	NOUN	
нябытнасць	,	NOUN	
усіхнія	,	PRONOUN	
каторы	,	PRONOUN	
усенькае	,	PRONOUN	
чымсці	,	PRONOUN	
я	,	PRONOUN	
абароцісты	,	ADJECTIVE	

Fig. 5. A fragment of unknown words list for testing the algorithm

4 Word Paradigm Generator as the software prototype of the algorithm

The software prototype of the algorithm, which is described above, is Word Paradigm Generator service on the site www.Corpus.by (see fig.6) [5]. The user can specify multiple words of one paradigm by selecting a category with its grammatical attributes and clicking “Generate probable paradigms” button.

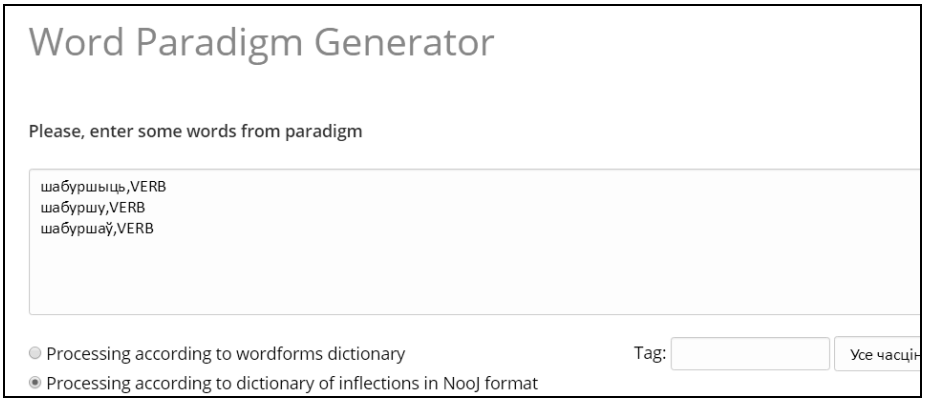


Fig. 6. The interface of “Word Paradigm Generator”

The resource outputs several variants of the words of the same grammatical categories with flexion classes and their annotation. From the list of generated words the user himself chooses the correct variant (see fig.7).

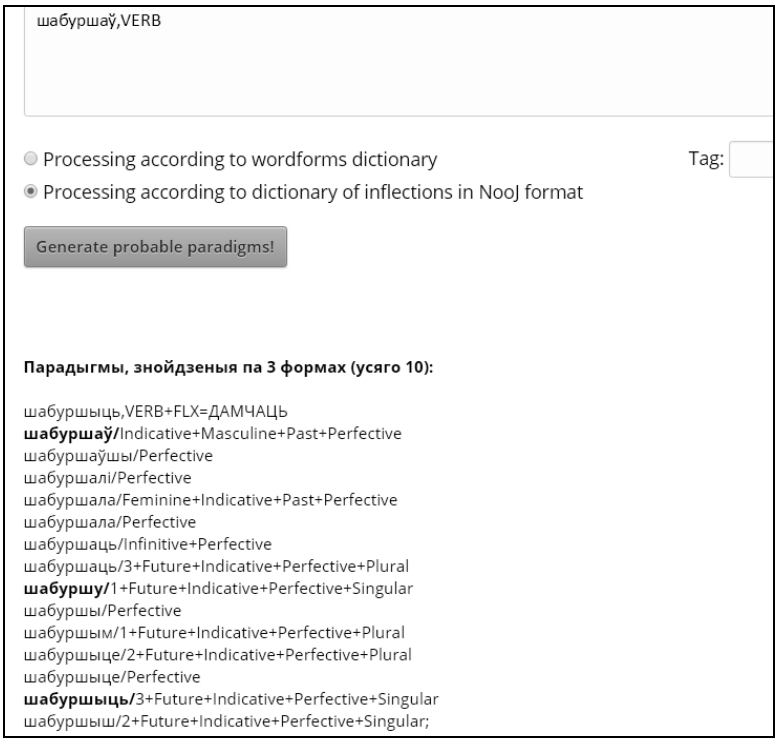


Fig. 7. An output example of “Word Paradigm Generator” operation

It should be noted that only changeable parts of speech (noun, verb, adjective, participle, pronoun, numeral) can be processed as they have a paradigm. The user can get annotation (tag) with stress arrangement of unchangeable parts of speech (adverb, preposition, conjunction, particle, parenthesis, interjection, predicative, gerund) only if a word is found in the dictionary (see fig.8). Otherwise, he needs to choose the right variant among proposed. It would be better if the user could also indicate the tag of a word (see fig.9).

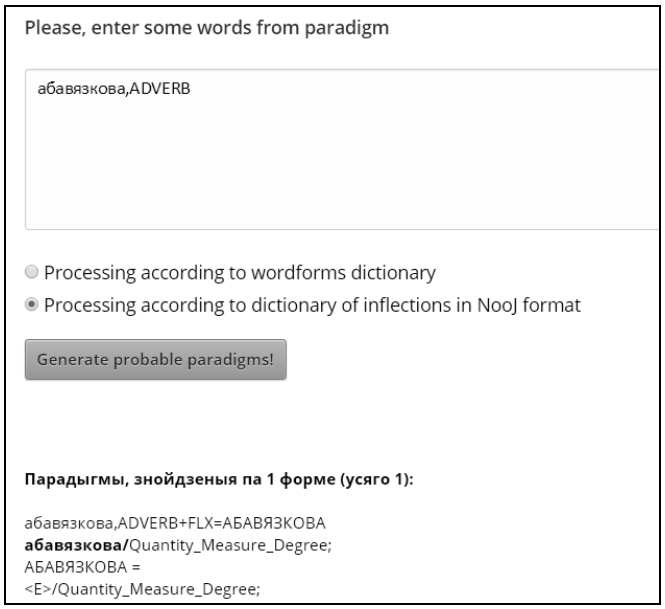


Fig. 8. Unchangeable adverb found in Nooj dictionary

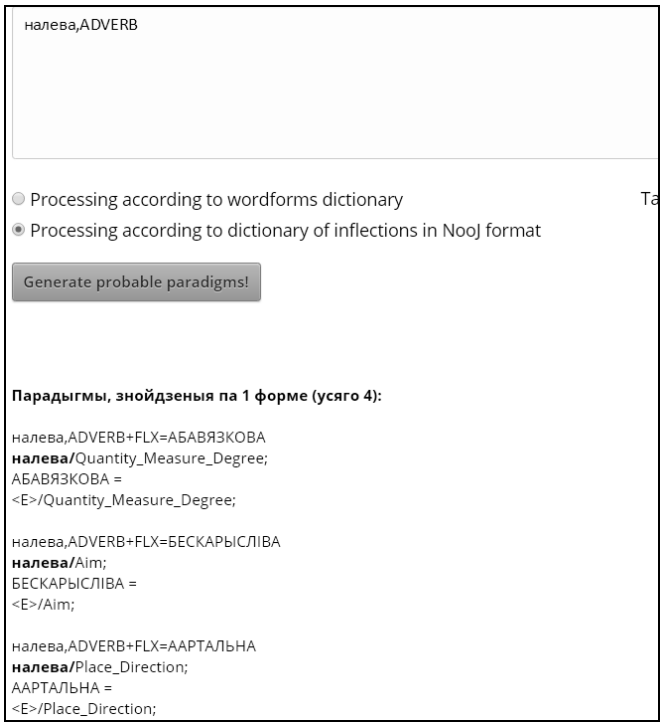


Fig. 9. Unchangeable unknown adverb processed by “Word Paradigm Generator”

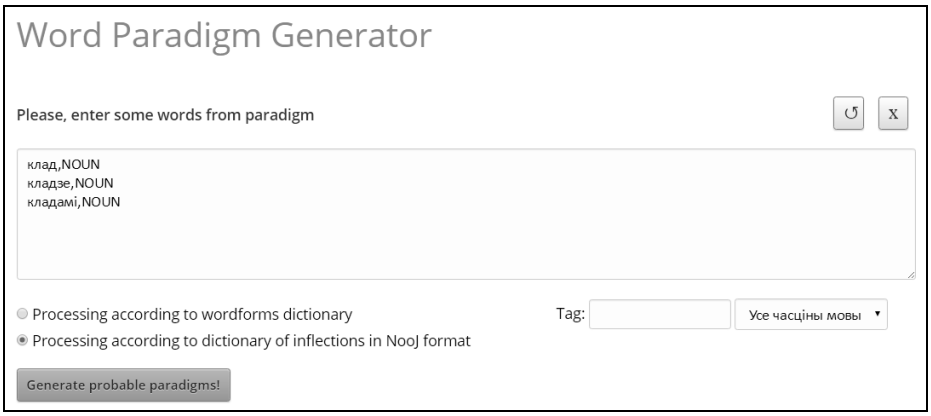


Fig. 10. Searching for the paradigm of the word *"клад"* according to dictionary of inflections in NooJ format

The most probable paradigm of the word "клад" (see fig.10) chosen by the expert after the word paradigm generation process in the service is shown in figure 11.

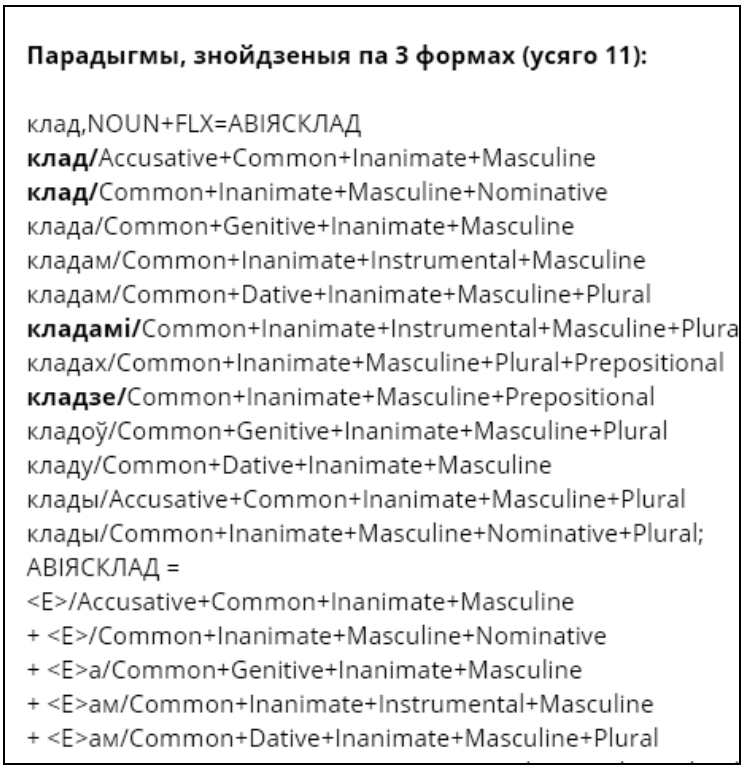


Fig. 11. A fragment of most probable paradigm of the word *"клад"*

5 Additional NooJ dictionary (*general_be(add).dic*) on the basis of annotated unknown words

As a result, an additional NooJ dictionary (general_be(add).dic) for the Belarusian module was composed. The dictionary of 365 words was generated by "Word Paradigm Generator" in NooJ format. Every line provides the information about an unknown word, its part of speech, and a word from the dictionary "main, general_be.nod", which has the same paradigm. These two words of one line belong to the same flexion class (see fig.12).

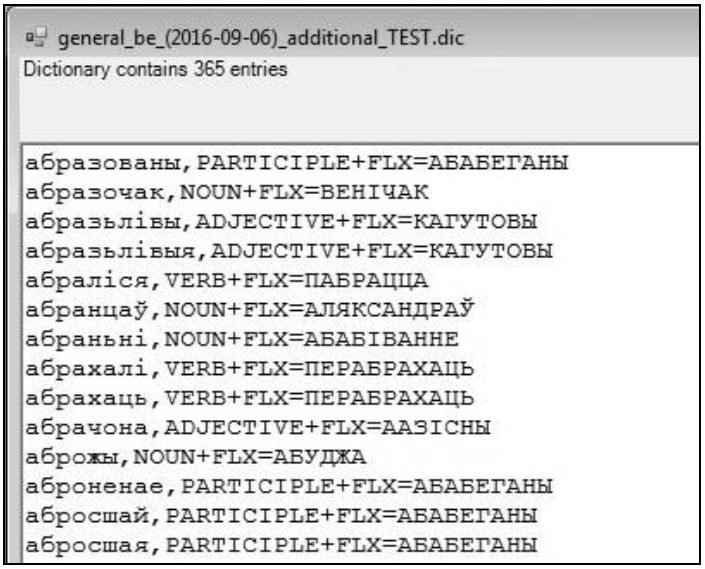


Fig. 12. NooJ dictionary (general_be(add).dic) for the Belarusian module

It takes approximately 0.035 of an hour (2.1 min) to process one word in "Word Paradigm Generator" by one linguists-experts. It means that we need nearly 945 hours to annotate 27 thousand words. More detailed statistics is represented in table 2.

Table 2. The Word Paradigm Generator level of efficiency according to its process by the user

Name	Quantity of words	Time consumed (h)
10 random words from 14 taken categories	140	4,9
First 365 words taken from an additional dictionary	365	12,775
An additional dictionary to be completely annotated	26 983	944,405

The corresponding table provides information about service efficiency. It should be noted that only changeable parts of speech due to NooJ system flection class (+ FLX), were tested. Unchangeable parts of speech were represented as follows: - (+ UNCH – "unchangeable").

6 Conclusion

Today, the algorithm for annotating different categories and paradigms according to flexion classes was worked out. It was realized in the online prototype – "Word Paradigm Generator" (http://corpus.by/WordParadigmGenerator/). A list of unknown words extracted from Belarusian corpus was examined (365) and added to annotated words to the present dictionary on the basis of the Belarusian NooJ module.

The next task is being planned: to develop the mechanism of automatic stress arrangement for all forms of an entire word on the basis of the Belarusian NooJ module.

7 REFERENCE LIST

1. Reentovich, Ivan The First One-Million Corpus for the Belarusian NooJ Module / Ivan Reentovich, Yuras Hetsevich, Valery Voronovich, Evgenia Kachan, Hanna Kozlovskaya, Angelina Tretyak, Uladzimir Koshchanka // Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015, Minsk, Belarus, June 11-13, 2015, Revised Selected Papers / Springer; ed. T. Okrut, Y. Hetsevich, M. Silberztein, H. Stanislavenka. — Springer International Publishing, 2016. — P. 3-15.
2. NooJ: A Linguistic Development Environment [Electronic resource]. — 2015. Mode of access: http://www.NooJ4nlp.net/. — Date of access: 08.05.2015.
3. 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.
4. The Levenshtein-Algorithm [Electronic resource]. — 2015. Mode of access: http://www.levenshtein.net/. — Date of access: 24.09.2015.
5. Word Paradigm Generator [Electronic resource]. — 2016. Рэжым доступу: http://corpus.by/WordParadigmGenerator/. — Date of access: 17.07.2016.
6. Hetsevich, Yu. Semi-automatic part-of-speech annotating for Belarusian dictionaries enrichment in / Yu. Hetsevich, V. Varanovich, E. Kachan [et al.] // NOOJ 2016 International Conference - Book of Abstracts (6-9 June, 2016, Czech Republic) / University of South Bohemia; ed. Jan Radimsky. — Ceske Budejovice, 2016. — P. 47.