ІНТЭРНЭТ-СІСТЭМА ГЕНЕРАЦЫІ ПАРАДЫГМАЎ СЛОВА ДЛЯ ПАПАЎНЕННЯ ЭЛЕКТРОННЫХ ГРАМАТЫЧНЫХ СЛОЎНІКАЎ

Ю. С. Гецэвіч[1], В. В. Варановіч[2], С. І. Лысы[1], І. В. Рэнтовіч[1], Я. С. Качан[1]

[1] *Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі*
[2] *Беларускі дзяржаўны ўніверсітэт*
Мінск, Беларусь
e-mail: *yury.hetsevich@gmail.com*

Разгледжана актуальнасць і акрэслены складанасці праблемы папаўнення беларускамоўных слоўнікаў праз аўтаматычныя сродкі апрацоўкі натуральнай мовы. Прыведзены алгарытм генерацыі парадыгмаў слова і апісаная інтэрнэт-сістэма, распрацаваная на дадзеным алгарытме.

Ключавыя словы: лексікалогія; слоўнік; алгарытм; анлайн-рэсурс; апрацоўка; парадыгма; генератар; папаўненне слоўнікаў.

ONLINE WORD PARADIGM GENERATOR FOR ELECTRONIC GRAMMAR DICTIONARIES COMPLETION

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

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

This article depicts the urgency and complexity of the problem indicated by the completion of the Belarusian dictionaries by means of electronic resources. It describes an algorithm of word paradigm generation and an online system for automatic dictionaries' enrichment on the basis of this algorithm.

Keywords: lexicology; dictionary; algorithm; online-service; processing; paradigm; generator; enrichment.

INTRODUCTION

Lexicology is the most flexible level of language, which is continuously changing, updating, improving. At the same time it responds to changes in the reality. Recently, in connection with the development of science, culture, technology and industry the vocabulary of the Belarusian language has changed, manifested in the appearance of new words and phrases denoting new objects, phenomena, concepts. New words reflect the current state of the vocabulary. Our research is aimed to develop the algorithm for the Belarusian dictionaries enrichment on the basis of a broad list of different text themes: fiction, historic, medical, scientific, sociological literature and etc., which are considered to be the finest source for searching unknown words of different domains.

MAIN PART

The main task of the research is to work out a mechanism for further processing (annotating different categories and paradigms according to flexion classes) of all unknown words extracted from different sources with adding annotated words to the present electronic laboratory's (the synthesis and recognition laboratory, UIPI NASB) dictionary with possibility to create user's own dictionaries.

The main concept is not only to get the category, but also the whole paradigm. The algorithm which was worked out by our team is the basis for the automatic generation of word paradigms. It consists of 11 consecutive interdependent steps, where each corresponds to a specific function. It results in output of one or several most suitable paradigms of a word. The algorithm searches the nearest paradigm(s) in matches of the last letters of a word you need to get the paradigm.

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

1. To search a word in the dictionary. If the word is there – to give the user a complete paradigm. If a word is a homograph – to give a few paradigms or ask the user to indicate an initial form. If not – step 2.

2. To ask the user to specify a part of speech.

3. Depending on the parts of speech to ask the user to specify the grammatical features (with the possibility to leave the fields empty if the user does not know).

4. To ask if it is changeable or unchangeable part of speech. If it is unchangeable – to give the user the word with annotation. If it is changeable – step 4a.

4a. If the user gives one form – Step 5. If the user gives more than one form – the algorithm processes the first form – step 5.

5. To search in the dictionary among words with marked grammatical features words without the first letter. If they are not found – without the first two letters and so on.

6. To take a sample of the first word in the list of found words, divided into «base» and "tail".

7. To select the "base" in the original word – the word without a "tail".

8. To separate the "tails" of other forms in the first found word.

9. If found words are more than 1 – the steps 5 and 6 for all words.

10. To compare the "tails". If they are the same in all forms, attach them to the base of the original word. If they are not the same – to highlight several types of changes, attach them all to the base of the original word.

10a. If in result there is only one paradigm – step 11. If the user has given one form – step 11. If the user has been given more than one form and after step 10 there are more than one paradigm – step 10b.

10b. To repeat step 5 for the second word form, which the user gives, but search in the list of generated paradigms, not in the dictionary.

10c. If the only match was found – step 11. If more than one – to pick them up and then – step 10b, and so on, until there is only one paradigm remains or forms, specified by the user.

11. On the basis of the chosen paradigm it is a hypothesis about stress arrangement in initial and indirect forms. The stress is indicated by "+" sign.

Also it is worth mentioning that in the future it is planned to modify the algorithm for working with arbitrary flection classes' file and grammatical tags files.

The software prototype of this algorithm is Word Paradigm Generator service of the site www.Corpus.by (fig. 1). The user can specify multiple words of one paradigm by selecting a category with its grammatical attributes and clicking "Generate probable paradigms" button (fig. 2).

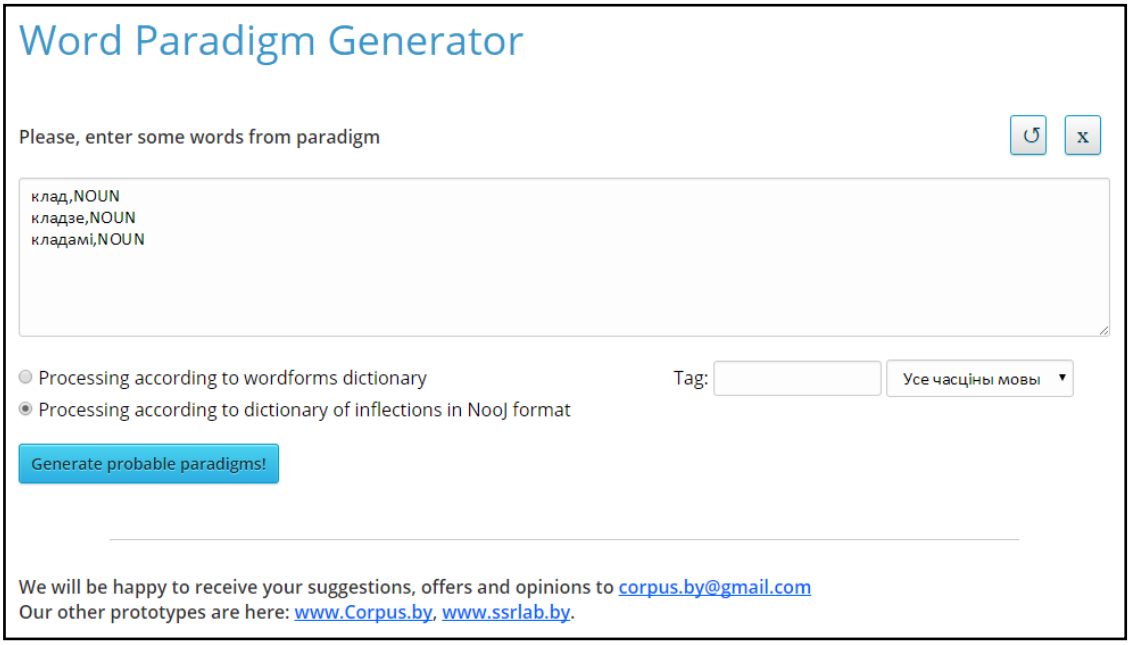

Fig. 1. The interface of Word Paradigm Generator

The resource outputs several variants of the words of the same grammatical categories with flection classes and their annotation. From the list of generated words the user himself chooses correct variant.

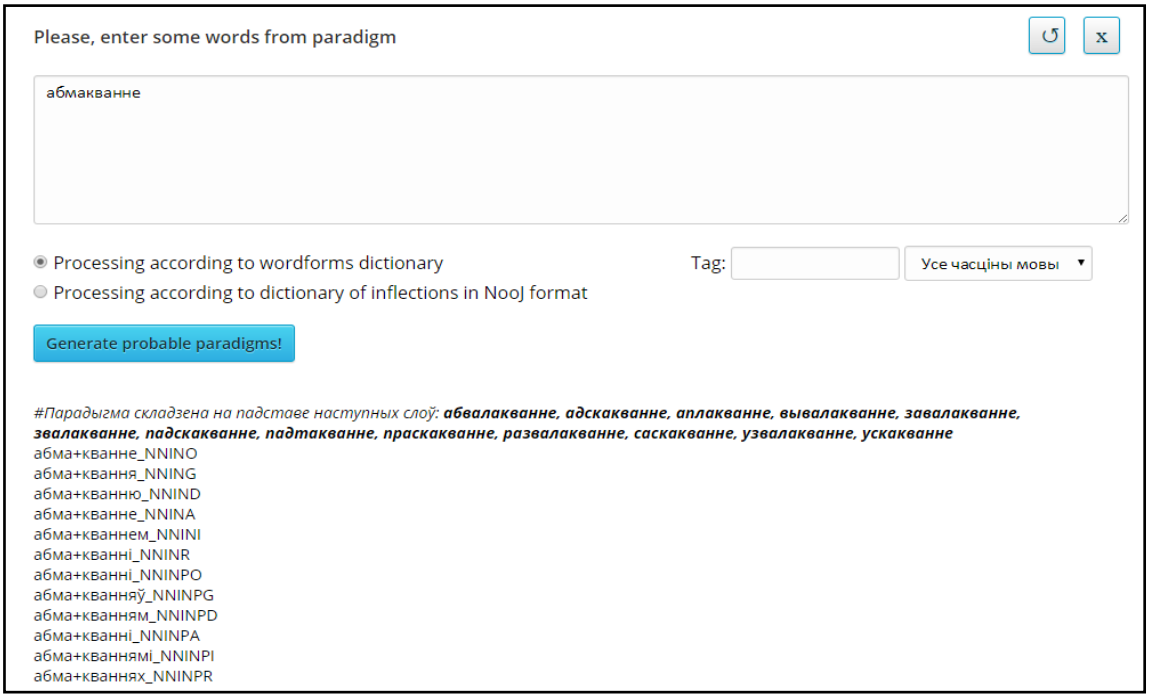

Fig. 2. An output example of Word Paradigm Generator operation

It should be noted that only changeable parts of speech (NOUN, VERB, ADJECTIVE, PARTICIPLE, PRONOUN, NUMERAL) can be processed as they have 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 the word is found in the dictionary (fig. 3). Otherwise he needs to choose the right variant among proposed. It would be better if the user could also indicate the tag of word (fig. 4).

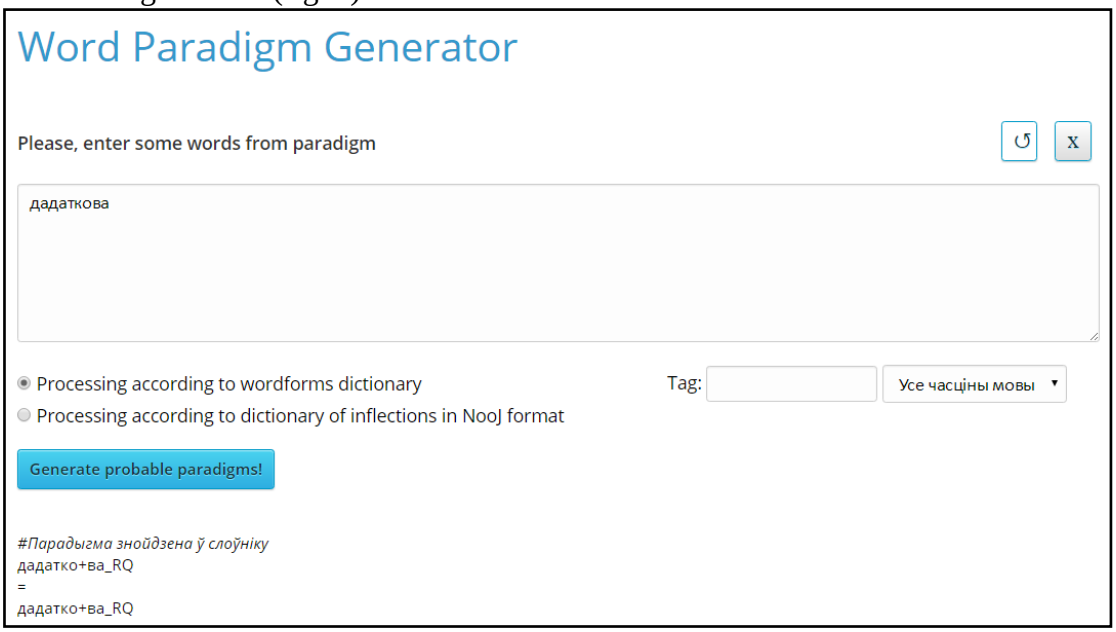

Fig. 3. Unchangeable part of speech (adverb) processing found in the dictionary by Word Paradigm Generator

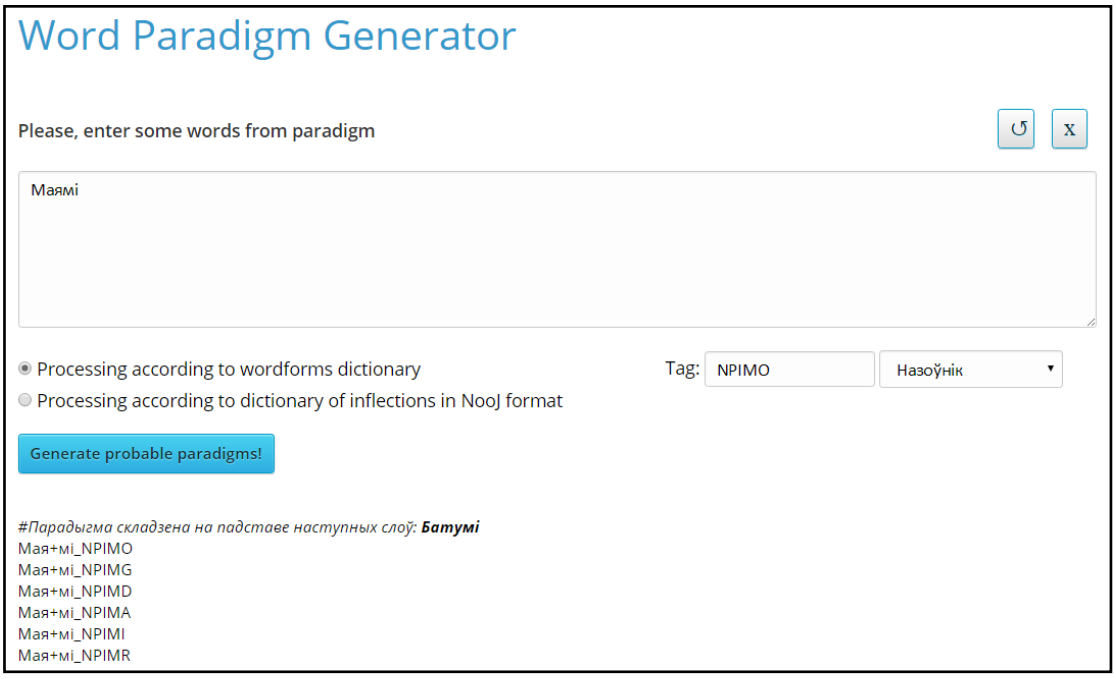

Fig.4. Unchangeable unknown part of speech (adverb) processing by Word Paradigm Generator

For testing the effectiveness of the algorithm, it was decided to process alphabetical subject index of Belarusian Universal Decimal Classification (UDC-2015). The total amount of unknown words is 3049. This figure includes basic (evident) changeable words (2584 words), unchangeable words (265 words) and problematic (inexplicit) words (200 words). To be clear, many of these words were absent in lexical database of the Corpus.by site. The results are represented in table below.

General statistical data about the results of unknown words processing (from the electronic UDC ASI) with the Word Paradigm Generator Service

Entity Description	The "status" of entity and its quantity of words			Total (Pr + Part-ly Pr)	Grand-Total	Percentage Grand-Total
	Processed (Pr) (with generated paradigm and stress)	Unprocessed (UnPr) (without generated paradigm and stress)	Partially processed (Part-ly Pr) (either with generated paradigm or stress)			
Basic changeable* words (BDWs)	994	229	1 361		2 584	84,75%
Unchangeable** words (IDWs)	0	265	0		265	8,69%
Problematic words (PWs)	0	180	20		200	6,56%
Grand-Total	**994**	**674**	**1 381**	**2 375**	**3 049**	**100,00%**
Percentage Grand-Total	**32,60%**	**22,11%**	**45,29%**			

The main idea was to get words' paradigms and stress arrangement. These words were divided into three groups: processed (getting paradigm and stress – 994 words or 32,6 %), partially processed (getting either paradigm or stress – 1381 words or 45,29 %) and unprocessed at all (674 words or 22,11 %). The overall percentage of processed and partially processed words is 77,89 % (84,75 % of them are basic changeable words, 8,69 % are unchangeable words, 6,56 % problematic words. The statistics shows that the system works on the sustainably high level. It effectively processes changeable words, arranges stresses. On the other side, while working with problematic words users should apply their own knowledge as it concerns semantics as well as homographs, a group of words which need special attention.

CONCLUSION

So, the main task is fulfilled: the mechanism for annotating different categories and paradigms according to flexion classes was worked out. This is Word Paradigm Generator. This program has its own novelty: it is the first Belarusian online open-free service with such functions, it means that a user doesn't need to check in and can independently process any data with the possibility of choosing the right variant. Moreover everyone can create their own electronic dictionaries without any obstructions. What is more important that the system all the time is being specified and improved. For further research we are planning to develop automatic stress arrangement for all forms of entire word.

LITERATURE

1. SEMI-AUTOMATIC PART-OF- SPEECH ANNOTATING FOR BELARUSIAN DICTIONARIES ENRICHMENT IN NOOJ / Yu. Hetsevich [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.

2. Word Paradigm Generator [Electronic resource]. URL: http://corpus.by/WordParadigmGenerator/ (date of access: 17.07.2016).