The Enlargement of Electronic Lexical Database by Computational On-line Free System

Zanouka E.
United Institute of Informatics Problems
National Academy of Sciences,
Minsk, Republic of Belarus
Email: evgeniakacan@gmail.com

***Abstract*—The article describes the urgency of electronic vocabularies' update. It also depicts the principle of electronic Belarusian grammar dictionaries composing by means of computational program "Word Paradigm Generator" which is the main source of Belarusian lexical database enlargement for natural language processing in computational and linguistic platform "corpus.by" [2].**

***Keywords*—computational linguistics, lexical databases, text and speech platform.**

I. Introduction

Lexicography is a branch of linguistics which studies a process of dictionary compiling and dictionaries themselves. The lexicography accomplishes different functions: it describes normalized language, teach the language, provides interlanguage communication. But the main function of the brunch is the scientific study of vocabulary. A vocabulary is in a certain way organized collection of words, usually with the addition of notes, which provides information about the features of their structure and / or their function. Today many scholars pay attention to the new scientific field – computational lexicography for many reasons: it's a new way of compiling and collecting dictionaries. Electronic dictionary is a dictionary stored as machine (computer) data instead of being printed on paper. Many researches have pointed out that these dictionaries have a number of distinctive peculiarities In contrast with printed. Electronic format:

1) accelerates the process of finding a required vocabulary unit;
2) extends the capabilities of a unit description;
3) electronic format makes it possible;
4) allows adding extra information: additional illustrations, diagrams, pictures, use recording or live-sounding speech, etc;
5) allows adding extra information: additional illustrations, diagrams, pictures, use recording or live-sounding speech, etc;
6) gives an opportunity of quick renewal of new lexicology

In this paper the author represents the program for creating new electronic grammar dictionaries for the Belarusian language and demonstrates the mechanism of its functioning on the example of her own dictionary which composes 1500 lexemes.

II. Word paradigm generator

Service Ẅord Paradigm Generator[1] is a computational program which processes unknown words. It analises the structure of a word, defines its possible grammatical features and outputs the whole paradigm and a tag of this word (see figure 1). The tag is an annotation of a word which describes all grammatical features of a word depending of a part of speech to which it belongs to. The necessity of the tags is the next: morphological information is widely used in scientific and normative grammars, as they simplify the description of the language and reflect its systemic nature. Paradigms of one part of speech in most cases have the same features, the same structure and a set of endings for the same type of declension or conjugation, combine sets of similar words stems, and often characterized by similar stresses and / or morphonological phenomena.

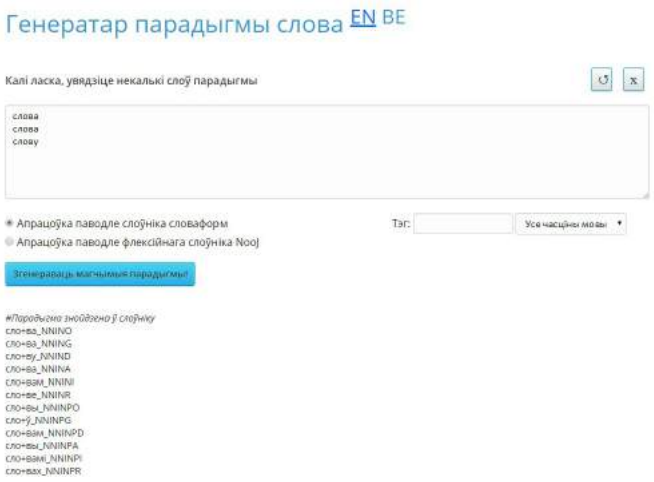

Figure 1. The lexeme which is found in the dictionary of the system

The system works on the principle of formal pattern: for a necessary word it chooses words from the database with the same grammatical characteristics and structure and outputs a list of the variants for this word. The user by himself chooses the right paradigm for this word. There are two main dictionaries which are built in the system: the dictionary of word forms and a dictionary of NooJ-format [4]. If any of these dictionaries contain a testing word, the system outputs the whole paradigm with tags (see figure 1). Otherwise it searches words corresponding to the query and outputs all

possible variants (see figure 2). In other words, the algorithm searches for the nearest paradigm(s) matching the last letters of the word, to which the user needs to get the paradigm. As it is the program for grammar dictionaries enrichment, the service processes only changeable parts of speech (noun, verb, adjective, participle, pronoun, numeral). For auxiliary parts of speech the system outputs their annotation [3].

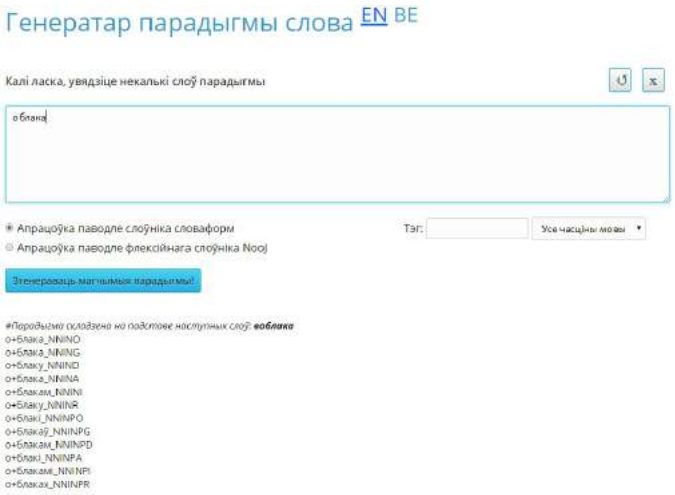

Рис. 2. The output of processing unknown word «Облака»

III. The composition of lexical database

Any dictionary is built according the next principles:

1) the nature of the dictionary;
2) the information they contain;
3) the language of the explanations;
4) an intended user.

While preparing our own Belarusian grammar dictionary, we have taken into account some other criteria for its compiling:

1) The meaning of lexical units upon which the lexicographer concentrates his attention presents very special difficulties of its own.
2) A lexicographer searches new words for the users whose needs are more practical.
3) Each entry must be self-contained, as we consult a single entry in a dictionary.
4) an intended user.

To correspond to all these points it was decided to analyze the vocabulary of the authors of 20th century both Belarusian and foreign (translations of foreign authors), such as Uladzimir Karatkevich (Dzikae paliavanne karalia Stakha), Vasil Bukaw (Zhurawliny kryk, Kar'er, Vawchynaia iama), Anton Adamovich (Tryvoha), Hohal (Vechary na khutary bliz Dzikanki), Daniel Defo (Zhytstsio i dziwnyia pryhody marakhoda Rabinzona Kruza) and so on. As a result, it was examined nearly 600 pages of different books, where nearly 5 thousand of words are unknown for the system [2].

The author has chosen 1600 lexemes in a haphazard way and divided all these words into the next groups:

1) Non-obscene words (the dictionary contains these words).
2) The words that are absent in the dictionary of the system.

The main attention is paid to the last group as all these words are unknown, therefore they are the core of a new dictionary. They are also subdivided into:

1) Proper Nouns (Анеліна, Менск, Яніна, etc.);
2) Dialectisms (дамо, йшла, etc.);
3) Russisms (начаць, дзяйстві+цельны, палучацца etc.);
4) Authorial words (кашчуннічайце, коця, дрыкганты, бомкнуцца etc.);
5) Words written in Tarashkevitsa - unofficial variant of Belarusian orthography (разьбованымі, купальле, губэрнатар, госьць, дысэртацыя).

To fall under these criteria, we have created a Belarusian grammar dictionary which contains the paradigms of a lexeme with its annotation. The dictionary composes 1500 separate lexemes, including changeable and unchangeable parts of speech (see figure 3).

Figure 3. An electronic database of 1500 separate lexemes

According to developed dictionary the author has prepared the algorithm of lexical database enlargement which consists of 5 steps:

1) Vocabulary selection.
2) Words processing by Word Paradigm generator.
3) Manual editing of output data.
4) Implementation it the system.
5) Vocabulary approbation.

IV. The specification of the program

While working with Word Paradigm Generator, the user should take into account the next points:

1) It processes one word at once (the user needs to input only single word).
2) It partially arranges stresses (some words have stress arrangement, some – no).

3) A user needs a manual editing for verification.

In the process of composing new grammar electronic dictionary the author has used Word Paradigm Generator as the main system which processes unknown words and on their basis generates grammar dictionaries. It was marked some more peculiarities of the program functioning which is necessary to correct. They are presented below. The system:

1) gives an incomplete verb paradigm (the future or present tense form is absent, depending on the verbal aspect + not all forms of participle) (see figure 4);
2) gives an incomplete paradigm of the adjective (without short and comparative forms) (see figure 4);
3) outputs participle II paradigms (in Belarusian "дзеепрыслоўе", which is unchangeable part of speech and has only aspect);
4) outputs a paradigm for unchangeable nouns (for example, са+ні, ка+ва, etc.);
5) outputs an incomplete paradigm of female nouns with endings –сць, -сьць);
6) outputs a full paradigm for complex words, where the first word is unchangeable and the second is changeable, and doesn't output a paradigm for both changeable parts of a complex word;
7) doesn't output a paradigm for changeable parts of speech in Tarashkevitsa.

There are also some mistakes with tags:

1) It outputs tags of common nouns for proper nouns.
2) Very often it outputs tags of participle I (in Belarusian "дзеепрыметнік") for adjectives and vice versa.
3) It can confuse verb aspects, therefore output incorrect tag for verbs.

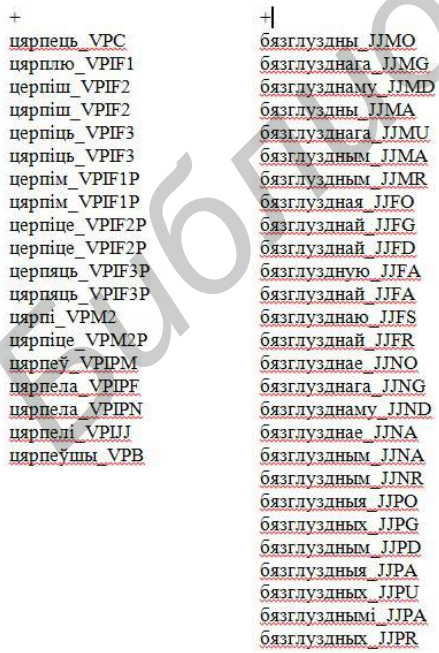

Рис. 4. The output of verb "Цярпець"(on the left) and the adjective "Бязглуздны" (on the right)

Among this list of program disadvantages we should distinguish two main points: the system problems, connecting with program incompleteness, and problems, connecting with unfilled database of the whole system which works only on the basis of one dictionary of word forms. All in all, the system works on 86% and this is a rather high level [3]. In future it is planned both to improve the capability of the program on the basis of this research and enlarge the lexical database of corpus.by.

V. Conclusion

The article covers the problem of electronic Belarusian grammar dictionaries composing by means of electronic program "Word Paradigm Generator" on the example of electronic lexical database consisting of 1500 lexemes. It is a little dictionary which contains different lexical groups including unofficial vocabulary. It provides two types of information: grammatical (the whole word paradigm) and morphological (a word annotation). Word Paradigm Generator is described as the means of electronic grammar dictionaries enrichment. The mechanism and principles of its functioning are also pointed out as well as all it deficiencies which need to be improved.

References

[1] Word Paradigm Generator [Электронны рэсурс]. — 2016. Рэжым доступу: http://corpus.by/WordParadigmGenerator/. — Дата доступу: 17/07/2016.

[2] Камп'ютарная платформа для апрацоўкі тэксту і маўлення // http://corpus.by/ [Электронны рэсурс]. — 2016. Рэжым доступу: http://corpus.by/index.php?lang=be. — Дата доступу: 16/12/2016.

[3] Гецэвіч, Ю.С. Інтэрнэт-сістэма генерацыі парадыгмаў слова для папаўнення электронных граматычных слоўнікаў / Ю.С. Гецэвіч, В.В. Варановіч, С.І. Лысы, І.В. Рэентовіч, Я.С. Качан // Международный конгресс по информатике: информационные системы и технологии=International Congress on computer science: Information systems and technologies / БГУ; под ред. С.В. Абламейко. — Минск, 2016. — С. 584-588.

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

ПОПОЛНЕНИЕ ЭЛЕКТРОННОЙ ЛЕКСИЧЕСКОЙ БАЗЫ ДАННЫХ ЧЕРЕЗ КОМПЬЮТЕРНУЮ СИСТЕМУ “ГЕНЕРАТОР ПАРАДИГМЫ СЛОВА”

Зеновко Е.С.

Данная статья описывает проблему создания и пополнения электронных словарей. Она также отражает принцип пополнения электронных белорусских грамматических словарей с помощью компьютерного сервиса в открытом доступе "Генератор парадигмы слова". Программа является основным источником расширения белорусской лексической базы данных для обработки естественного языка в области компьютерной и лингвистической платформы "corpus.by"[2].

Введение

Лексикография является разделом лингвистики, изучающим процесс составления словарей. Лексикография выполняет различные функции: она описывает нормированный язык, обучает языку, обеспечивает процесс коммуникации. Но основная ее функция - научное изучение лексики. Сегодня многие ученые обращают внимание на новое научное направление - компьютерную лексикографию по многим причинам: это новый способ сбора и пополнения словарей. Многие исследователи отмечают, что электронный формат словарей имеет ряд отличительных особенностей, таких как ускорение процесса поиска необходимой единицы словаря, расширение возможности описания единиц словаря, добавление дополнительной информации по словарных единицах, быстрое пополнение словарей и т.д. В данной статье автор представляет программу для создания новых электронных белорусскоязычных грамматических словарей и демонстрирует механизм его функционирования на примере своего собственного словаря, который составляет 1500 лексем.

Описание программы «Генератар парадигмы слова»

В данной главе описывается сервис "Генератор Парадигмы Слово"[1], который представляет собой компьютерную онлайн программу, которая обрабатывает неизвестные слова. Она анализирует структуру слова, определяет его возможные грамматические особенности и выводит всю парадигму и тэг этого слова. Система работает по принципу формального шаблона: для необходимого слова он выбирает из базы данных слова с одинаковыми грамматическими характеристиками и структурой и выводит список вариантов для этого слова.

Принцип и механизм создания лексической базы данных

В данном разделе описывается процесс подбора лексики для белорусскоязычного грамматического словаря, демонстрируются его особенности и механизм пополнения через Генератор Парадигмы Слова.

Особенности функционирования «Генератара парадигмы слова»

Третья падглава отражает особенности функционирования онлайн ресурса "Генератор Парадигмы Слова описывает все недостатки программы, которые должны быть исправлены, и пути их исправления.

Заключение

В статье рассматривается проблема составления электронных белорусских грамматических словарей, которые могут пополняться с помощью электронной программы "Генератор Парадигмы Слова». Данный механизм демонстрируется на примере электронной лексической базы данных, которая состоит из 1500 лексем. Это небольшой словарь, содержащий различные лексические группы слов, в том числе ненормативную лексику. Он дает два типа информации: грамматическую (вся парадигма слова) и морфологическую (аннотацию слова). Генератор Парадигмы слово описывается как средство электронного обогащения грамматических словарей. Отмечены механизм и принципы его функционирования, а также все недостатки, которые должны быть улучшены.