УДК 811.161.2'367.4:004

Natalia Darchuk, Dr Hab., Doc,
Margaryta Langenbakh, Ph D
Taras Shevchenko National University of Kyiv, Kyiv

THE ELECTRONIC TARAS SHEVCHENKO'S LANGUAGE DICTIONARY: THE MODERN REPRESENTATION OF KNOWLEDGE

The article is dedicated to the theoretic and methodic basis of the Electronic Taras Shevchenko Language Dictionary. The internal structure, interface and linguistic parameters of dictionary is described. The article shows the advantages of electronic dictionaries, especially the Electronic Taras Shevchenko dictionary, and prospects of their usage in linguistic researches.

Keywords: *electronic dictionary, Taras Shevchenko, computer lexicography, linguistic database.*

The typical feature of modern science is active use of the computer tools that help not only to optimize the research process but to present the results in the more convenient and user-friendly way. Linguistics is not an exception in this trend. And one of the linguistics branches closely related to the information technologies is computer lexicography. Its tasks are:

– development of the systems that automatically convert paper dictionaries into digital formats;

– automatic language processing and dictionaries updating;

– creation of the electronic concordances, frequency dictionaries etc.;

– providing the multimedia and hypertext technologies into dictionaries units representation;

– creation of the translation dictionaries based on the parallel text corpuses;

– automatic dictionaries building;

– creation of the integrated lexicography systems etc.

On the ground of Ukrainian computer lexicography there is a set of various digital dictionaries: "The Ukrainian Affixes Dictionary", developed at the O. Potebnya Institute of Linguistics, NAS of Ukraine; electronic versions of "The Ukrainian Language Dictionary" (http://www.sum.in.ua) and "The Orthographic Dictionary", developed by the Ukrainian Language and Information Fund; The PolyDic – special tool for terminology databases and electonic dictionary creation,

developed by the Lviv Polytechnic University team; online concordance of the Grygory Skovoroda texts (http://www.artsrn.ualberta.ca/skovoroda/) etc. The Mova.info linguistic portal (http://mova.info) maintained by the Laboratory of Computational Linguistics, Kyiv National Taras Shevchenko University, also contains a collection of different electronic dictionaries. The latest their project is "The Electronic Taras Shevchenko's Language Dicitonary" (http://www.mova.info/cfqsh.aspx), dedicated to his 200th anniversary.

According to Y. Karpilovska the important advantage of the electronic dictionaries is the deep structuring of the information that allows users to work with the necessary data in the most flexible and convenient way. L. Belyaeva also adds to the list of digital dictionaries advantages such options as sufficient decreasing of the search time and systematical and well-balanced data collection. So long as the Taras Shevchenko is exceptionally important person for Ukrainian philology and his texts permanently attract scientists attention, there is an ultimate need to develop the advanced tool that helps to make the researches based on his texts.

The aim of the project was to create the electronic dictionary based on the modern lexicography technologies that represents different aspects of the Taras Shevchenko's language and makes the work with his texts more convenient and effective.

Such tasks were performed:

1) linguistic analysis of the Taras Shevchenko's texts;

2) creation of the database consisting of the detected language units with their grammatical and statistical characteristics;

3) development of the powerful and user-friendly interface that allows grammatical and statistical processing of the dictionary data according to the scientific needs of the researchers.

The project team was formed from the workers of the laboratory of computational linguistics lead by N. Darchuk (V. Sorokin, M. Langenbakh, Y. Khodakivska, O. Tiutenko) and students (M. Bilokon, A. Dryhina, K. Ksiondzyk, V. Miasnikova, M. Lysenko, A. Shkoda).

The object of research was the Taras Shevchenko's poetry language, the subject – its features on the different linguistic levels.

The project was divided into several stages that included linguistic and statistic processing, collecting of the data and development of the dictionary interface.

The dictionary entries were filled with the data that represented characteristics of the lexems, morphems and phrases used in the Taras Shevchenko's poems.

The linguistic processing was performed in two stages. The first stage included the automatic morphological and syntactical parsing. The words were attributed with the special tags representing the parts of the speech and grammatical forms, phrases were detected and marked with the appropriate types of syntactic links and grammatical models. The second stage provided the manual processing of the collected data in order to fix the errors and improve the results of the automatic processing.

Since syntactic and lexical characteristics are usually determined by the context within the sentence we defined its measures as the maximum context length.

Minimum segment was the context in three positions 1-X +1 (where X is analyzed word form) – one word in the preposition and one word in the postposition to the analyzed form.

The syntactic information is represented in the subdictionary of the phrases models. The choose of such units in syntax description is motivated by the fact that it is both lexical and morpho-syntactic unit [Golovin : 190–193]. The phrase is the elementary syntactic construction and, what is important, it preserves its nominative character. The phrases form the more complicated structures, such as sentences, and make a platform for word-to-word links.

As the phrases we qualified the words complexes both with subordinating, coordinating and predicative links, because all of them have an important feature: they demonstrate the relations between the words and demand certain grammatical forms of them. Compatibility of words described by the following parameters:

1) the type of the phrase (based on the part of the speech of the main word): noun, adjectival, verbal, adverbial, numeral, pronoun;

2) the role of word forms in the phrase (for subordinative structures only, because in the coordinative and predicative phrases both words are equal):

a) core (the main word in the phrase);

b) subordinated (the dependent word);

3) the type of syntactic link: subordination, coordination, predication.

The statistical parameters of vocabulary items are presented by two options – absolute and average frequency. The absolute frequency is not

enough to make an objective picture of words usage because words often are used in the texts non-regularly. In order to solve this problem we added an average frequency and standard deviation of absolute frequencies – parameters which characterize regularity of words distribution in texts.

All data were fixed in the database (Fig. 1), which has the following structure: columns with the word, sentence and text id; phrase; each word with its form; their original forms (lemmas); type of syntactic link. The linguistic database helps to organize and structurize the information and significantly simplifies the process of data replenishment and editing [Langenbakh]. Additionally, the implementation of various search tasks requires the intervention of the researcher in the data representation in order to select only the necessary information, illustrate relationships between different types of data etc. All these tasks are easy to perform having the linguistic databases [Karpilovska : 36]. Thus, the databases are the formalized models of language objects and create a factual basis for solving of many linguistic problems [Darchuk 2008 : 205].

The dictionary interface consists of four pages. The first page contains information about the project, the rest represent various linguistic information. The structure of the pages is organized by a single principle (Fig. 2): the main field offers linguistic parameters for information selection and different search options: search for contexts by the word form or lexemes, morphemes or morphemic models of the words, grammatical categorical (gender, number, case, time, person, etc.), syntactic model of the phrase. The search results represent statistic data and contexts. The left sidebar menu contains the statistical search options and links to the other pages of the dictionary.

The parameters included in the dictionary interface make a basis for a wide range of the linguistic studies based on the works of Taras Shevchenko. For example, research of the valency models answers the following questions:

– whether there may be cases where a particular word form in a sentence is not a dependent element or has no dependents;

– words from which parts of speech can control the other words (parts of speech);

– in what constructions the word can appear;

– wher it is a main element and where is dependent;

– whether the word is able to form a predicate phrase and so on.

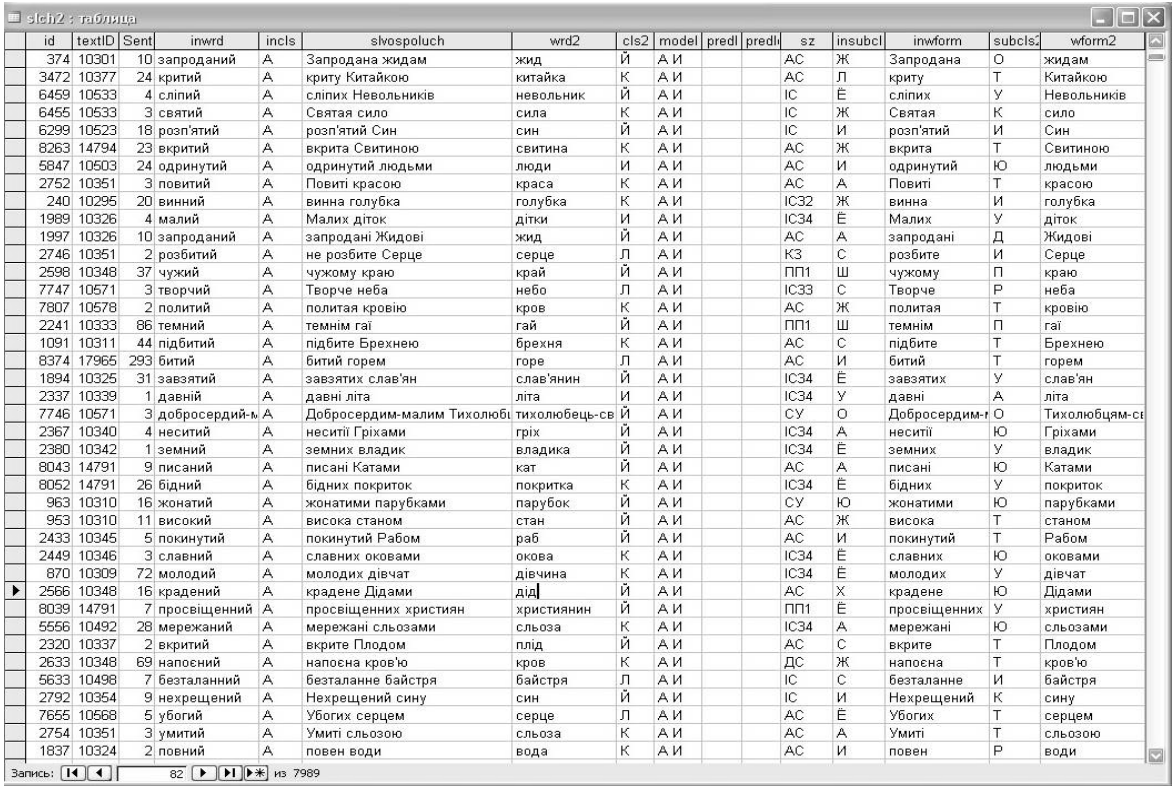
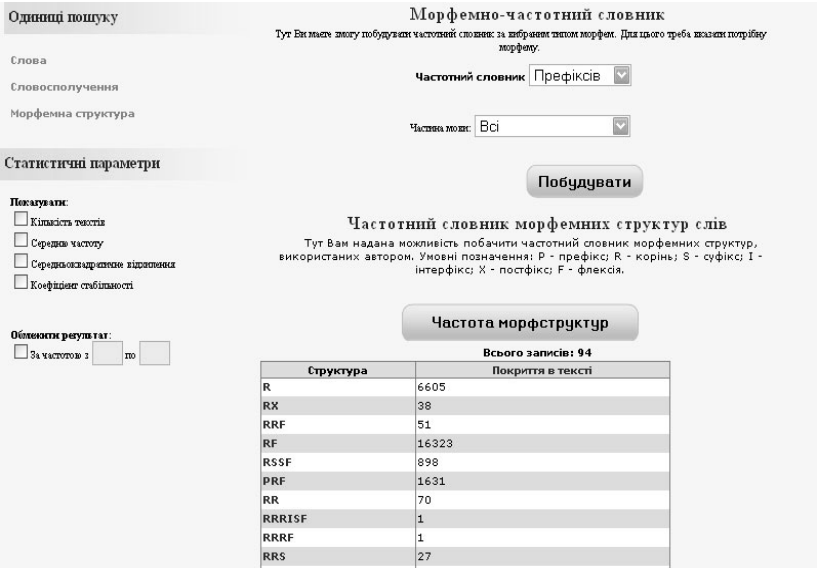


Fig. 1. The database structure



Структура	Покриття в тексті
R	6605
RX	38
RRF	51
RF	16323
RSSF	898
PRF	1631
RR	70
RRRISF	1
RRRF	1
RRS	27

Fig. 2. The search page

As an illustration, let's look at the list of the valency models obtained from the Taras Shevchenko's texts study:

I. Models verbs

1. Core verb models:

– verb + noun: ***благословить*** *дітей* 'to bless the children';

– verb + adjective: ***ставати*** *зеленим* 'to become green';

– verb + preposition + noun: ***дивлюсь*** *на тебе* 'look at you';

– verb + verb: *жити* ***хочу*** 'want to live'.

2. Dependent verb model:

– verb + verb: *ліг* ***одпочить*** 'lay to rest';

3. Predicative verb model:

– noun + verb: *Вітер віє* 'the wind blows'.

4. coordinated verb model

– verb + verb: *гралися, хвалили* 'played, sang praises'.

II. Noun models

1. Core noun models:

– adjective + noun: вольнії ***села*** 'free villages';

– pronominal adjective + noun: *мій* ***квіте*** ‘my flower’;

– noun + noun: ***день*** радості ‘a day of joy’;

2. Predicate noun models:

– noun + verb: *верба похилилась* ‘the willow bowed’.

III. Adjective models

1. Core adjective models:

– adverb + adjective: *дуже* ***цікаве*** ‘very interesting’;

– adjective + preposition + noun: ***великая*** *в женах* ‘the great among the women’;

– adjective + conjunction + noun: ***червоних*** *як калина* ‘red like a guelder rose’.

2. Dependent adjective models:

– verb + adjective: ***був*** *дужий* ‘was strong’;

– adjective + noun: *вольнії* ***села*** ‘free villages’.

3. Predicative adjective models

– noun + adjective: ***ангелом*** *святим* ‘by the saint angel’.

4. coordinate adjective models:

– adjective + conjunction + adjective: *зелений і синій* ‘green and blue’.

IV. Adverbial models:

1. Core adverbial models. The adverb is considered to be core if it can be omitted without infringing content, that is lexically bound

– noun + adverb: *гріха* ***менше*** ‘less sin’;

– adverb + adverb: *досі* ***нудно*** ‘still bored’.

2. Dependent adverbial models:

– adverb + adverb: *досі* ***нудно*** ‘still bored’;

– adverb + adjective: *надто* ***молодую*** ‘too young’.

3. Predicative adverbial models:

– noun + adverb: *билини кругом* ‘the grass is around’.

4. Coordinative adverbial models:

– adverb + adverb conjunction: *любенько та тихо* ‘lovely and quietly’.

The range of potential studies using the Dictionary includes the morphological, lexical, syntactic and stylistic researches. Such wide potential becomes possible thanks to the linguistic diversity and flexibility markup and powerful user-friendly interface that demonstrate the high efficiency of the electronic dictionaries.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *L. Belyaeva.* The potential of the automatized lexicography and the applied linguistics / L. Belyaeva // The News of The A. I. Hertzen RSPU. – № 134. – SPb, 2010. – P. 70–79.

2. *B. Golovin.* The introduction into linguistics / B. Golovin. – M. : Highest School, 1966. – 328 p.

3. *N. Darchuk.* The computational linguistics (automatic text processing): textbook / N. Darchuk – K. : Printing Center "The Kyiv University". – 351 p.

4. *T. Gryaznukhina.* The frequency Dictionary of the modern ukrainian publicistics / N. Darchuk, T. Gryaznukhina // Movoznavstvo. – 1996. – № 4–5. – P. 15–18.

5. *Y. Karpilovska.* The introduction to the applied linguistics: the computational linguistics: textbook / Y. Karpilovska. – Donetsk, 2006. –188 p.

6. *M. Langenbakh.* The electronic database of the semantic and syntax noun valency in the Ukrainian language / M. Langenbakh // The Scientific Bulletin of Lesya Ukrainka Volyn National University. – Lutsk, 2008. – p. 249–252.

7. *R. Mysak.* The electronic dictionaries: classification and buildind principles / R. Mysak // The Problems of the Ukrainian terminology. – Lviv, 2008. – P. 52–55.

8. *V. Perebyinis.* The Traditional and computational lexicography / V. Perebyinis, V. Sorokin. – K., 2009. – 218 p.

Стаття надійшла до редакції 14.04.14

Дарчук Наталія, д-р філол. наук, доц.,
Лангенбах Маргарита, канд. філол. наук,
КНУ імені Тараса Шевченка, Київ

Електронний словник мови Тараса Шевченка: сучасне представлення знань

Стаття присвячена теоретико-методичним аспектам Електронного словника мови Тараса Шевченка. Розглядається внутрішня структура та інтерфейс словника, лінгвістична параметризація текстового матеріалу. Оцінюються переваги електронних словників у цілому та даного проекту зокрема, перспективи його використання у лінгвістичних дослідженнях.

Ключові слова: *електронний словник, Тарас Шевченко, комп'ютерна лексикографія, лінгвістична база даних.*

Дарчук Наталия, д-р филол. наук, доц.,
Лангенбах Маргарита, канд. филол. наук,
КНУ имени Тараса Шевченко, Киев

Электронный словарь языка Тараса Шевченко: современное представление знаний

Статья посвящена теоретико-методическим аспектам Электронного словаря языка Тараса Шевченко. Рассматривается внутренняя структура и интерфейс словаря, лингвистическая параметризация текстового материала. Оцениваются преимущества электронных словарей в целом и данного проекта в частности, перспективы его использования в лингвистических исследованиях.

Ключевые слова: *электронный словарь, Тарас Шевченко, компьютерная лексикография, лингвистическая база данных.*