УДК 81'322

Yu. I. V e r b i n e n k o
Ukrainian Lingua-Information Fund of NAS of Ukraine

Multilingual terminological dictionary in the context of a national terminology system formation

Вербиненко Ю. І. Багатомовний термінологічний словник у контексті формування системи національної термінології.
Розглянуто проблему створення національної термінологічної системи та пов'язані з цим аспекти та підходи до її розв'язання. Сформульовано загальні вимоги до багатомовного термінологічного словника, з урахуванням особливостей, що виникають при автоматизації процесу і створенні цифрової версії. Експериментальне моделювання проведено на корпусі тексту українсько-російсько-англійського словника зі зварювання. Представлено зовнішній інтерфейс ВЛЛ «Зварювання» та розглянуто основні завдання і функціональність системи. Розкрито потенціал віртуальної лабораторії і актуальність можливості оперативно взаємодіяти фахівцям різних галузей у режимі віддаленого доступу.
Ключові слова: Онтологія, національна терміносистема, ВЛЛ «Зварювання», багатомовний термінологічний словник.

Вербиненко Ю. И. Украинский лингво-информационный фонд НАН Украины. Многоязычный терминологический словарь в контексте формирования системы национальной терминологии. Рассмотрена проблема создания национальной терминологической системы и связанные с этим аспекты и подходы к ее решению. Сформулированы общие требования к многоязычному терминологическому словарю, с учетом особенностей, возникающих при автоматизации процесса и создании цифровой версии. Экспериментальное моделирование проведено на корпусе текста украинского-русско-английского словаря по сварке. Представлены внешний интерфейс ВЛЛ «Сварка» и рассмотрены основные задачи и функциональность системы. Раскрыт потенциал виртуальной лаборатории и актуальность возможности оперативно взаимодействовать специалистам различных отраслей в режиме удаленного доступа.
Ключевые слова: Онтология, национальная терминосистема, ВЛЛ «Сварка», многоязычный терминологический словарь.

Verbinenko Yu. I. Multilingual terminological dictionary in the context of a national terminology system formation. The problem of establishing a national system of terminology and related problems and methods of solution are considered in this article. The general requirements for multilingual terminological dictionary and specifications arising during the process of automation and digitizing were formulated. Experimental modeling was carried out on the body of the text of Ukrainian-Russian-English dictionary on welding. Interface of VLL "Welding" was presented and the basic tasks and system functionality were considered. Was unleashed the potential of the virtual laboratory and relevance of quick interaction between specialists of various fields.
Key words: Ontology, national system of terminology, VLL "Welding", multilingual terminological dictionary.

Major changes that occurring in modern lingual environment, primarily relate to terminology. According to researchers, at least 90% of new words belong to terminology of various subject areas. This is due to the continuous development of science, technology, economics and culture. Increased intensity of social communications and constant increasing of information volume had necessitated systematic research of terminological systems both new and already existing ones, traditional branches of knowledge.

The development of the information society leads to intense languages transfusion and their diffusion to the field of general language phenomena. An important side of these processes is multilingualism of the terminological systems and the necessity of their adaptation to modern information and communication realia. As there is no common language that could be used by experts, currently the primary tool for solving problem of multilingualism is harmonization and coordination of national terminologies. So, the necessity of creation multilingual terminological systems is constantly increasing. This task is particularly topical for the languages, which terminology is only formed, and Ukrainian language is among them.

At the present stage of science majority of new terms come from English, which is the language of international communication. Ukrainian terminology for a long time had been developing under the Russian influence. Therefore, it is advisable to develop

terminological dictionaries, combining Ukrainian, English and Russian terminology.

So, the topical task of terminological area in modern Ukrainian conditions is to create national terminology system, and to ensure its continuous and rapid improvement and adaptation to international standards, because of necessity to intensify international scientific, industrial and commercial contacts.

The development of such terminology system and corresponding dictionaries is possible only on condition of effective cooperation of specialists in different fields - science and technology (in corresponding subject areas), linguistic and information that identifies the necessity for designing and implementing effective mechanisms of intersectoral cultural and information integration based on modern information technologies.

The E.O. Paton Electric Welding Institute (EWI) of the NAS of Ukraine in cooperation with the Ukrainian Lingua-Information Fund (ULIF) of NAS of Ukraine on the initiative of Boris Paton are carrying out a project, which aim is to create modern electronic terminology system of the subject area "Welding". The work is perfomed through the introduction of the theory and practice of constructing lexicographical systems [8]. In order to create and implement mentioned terminology system the Virtual Terminographic Laboratory "Welding" was developed and put into operation; and this Laboratory provides creative interaction between welding specialists and linguists that work in order to create final terminological product. It would be a dynamic, open, based on ontological approach [4] electronic terminology dictionary on welding, which in its content and structure is correspond to the current trends in the field of subject knowledge representation and adapted to constant updating, extension, integration with other terminology systems, development in quantitative and qualitative terms, etc.

According to Gruber, we understand ontology as a system of explicit conceptualization of a subject area, i.e. its formalized representation [10].

It should be noted that there are different interpretations of such general definition. But all formalizations have common thigs, such as selection of objects' set (concepts, notions), relations between them, and rules for establishing relations and axioms, which determine inference rules on the relation set. [11; 12; 13]

Considering the use of ontologies for automatic text processing tasks there are two approaches to establish relation between subject area ontology and subject area language, particularly its vocabulary.

On the one hand, at first a system of concepts is built, and then sets of language equivalents (words, terms, phrases) are assigned to these concepts. Presence of such objects in the text allows to initiate relevant concepts and rules related to them [10].

On the other hand, it was noted that existing linguistic and lexicographical resources (dictionaries, glossaries, thesauruses) also contribute to a conceptualization of a subject area [6].

As a result, term "ontology" corresponds to a wide range of structures that represent knowledge about particular subject area. Examples of ontologies, formalized in various ways are following [14]:

1. Dictionaries with definitions;
2. Simple taxonomy;
3. Thesaurus (taxonomy with terms);
4. Model with a free set of relations;
5. Taxonomy and random set of relations;
6. Fully axiomatized theory [2].

Note that structures mentioned above may be considered as a particular case of a lexicographical environment, arranged in some manner or other. [8]

Despite the fact that nowadays such information resources as thesauri and ontologies are considered the most popular for the construction of structured descriptions of knowledge about world and language, in practice, their application in information systems, based on automated information processing, is not very common. This situation is associated with a range of circumstances.

First, if it is proposed to use a certain linguistic resource, it must contain a description of tens of thousands of words and phrases. The amount of recourse errors should be so low to not to spoil the possible improvements expected from its use. In addition, it should be understood that the introduction of any linguistic resource will always be behind of the development of the subject area.

Secondly, the use of thesauri and ontologies for information search requires a high accuracy in solving linguistic ambiguity problems. Appropriate tests conducted at the SemEval [https://www.cs.york.ac.uk/semeval-2012/], and Senseval [http://www.senseval.org//] conferences, had showed that existing means of ambiguity elimination has not yet reached a level sufficient for the effective application of thesauri and ontologies, particularly during information search.

Third, the use of thesauri and ontologies relationships for requests expansion may face the problem of inaccurate description of relations or relations irrelevant to context request. Usage of such relationships often leads to a significant decreasing of the search accuracy. Thus, recently, global search engines Yandex and Google have begun to use requests extension by cognate words that can be consider as a minimum thesaurus, but in many cases even a minimal query expansion is also appears irrelevant.

Finally, there is a thought that used statistical methods are implicitly taking into account linguistic information, that text is just a set of features, which considered by statistical models very well. Helen Vorhes gives following examples of linguistic approaches modeling by statistical methods: morphological analysis can be approached by

stemming, phrases identifying – identifying pairs of words that frequently occur, ambiguity settlement procedures can be simulated by means of similarity contexts [3].

In practice, creating a linguistic ontology associated with solving a number of problems. First one is complexity of automatization of the process of ontology concepts forming. Each concept in the ontology must correspond to a certain element of subject knowledge. Distinguishing such elements is complicated due to unclear semantic boundaries between objects and real world processes. Even trained experts of the subject area have difficulties [5]. In addition, different experts may distinguish different sets of concepts or different semantic identifiers of the same concept. The second problem is that not only concepts themselves are important elements of the ontology, but semantic relations between them are also important. Sometimes the connection between concepts so fuzzy or hidden that identifying its presence or qualify its semantic nature (type of relationship) is almost impossible, which often leads to the necessity for special research within the corresponding subject area. In addition, different experts have very different views on defining quantitative characteristics of this connection regarding its intensity (weight) [6].

The following requirements for terminological dictionary were formulated in the work [6]:

1. Adequate coverage of special vocabulary of chosen subject area;
2. Availability of the necessary information about special lexical units;
3. Rejection of unnecessary, redundant data, that increases dictionary volume and complicates relevant information searching;
4. Unification of composition and apparatus of links of similar dictionaries in order to facilitate for users moving from one dictionary to another;
5. The maximum harmony (coherence) between all elements of methodological guideline and dictionary composition.

Note that point 3 is significant only for paper dictionaries, because modern technology can increase the speed of information searching, and operate with almost unlimited amounts of it. However, modern computer lexicography does not devoid drawbacks of paper dictionaries yet:

1. Large number of possible translations of terms without interpretations and comments on the scope of their use and differences [1];
2. Lack of encyclopedic information that explains one or another term, because a translator usually needs more information for adequate translation than specialist of the subject area [8];
3. Lack of illustrative examples that would clearly show the peculiarities of a particular term in different contexts;
4. Targeting only one stage of translation [1].

There are three main problems that arising during creating a multilingual terminological dictionary. First one is heterogeneity repertoire of language units for special purposes. The dictionary cannot be restricted only by terms of specific subject area, there should be present also terms of close branches and other linguistic units, close to terms or essential for special discourse. A problem of forming and maintenance of text corpus, that should be a source of illustrative material and basis for terminographic constructions verification, is very important.

The conceptual model of a terminological system "Welding", that considers the above, became as a result of the joint project of EWI and ULIF. The internal structure of terms representation is designed with the respect of following conditions:

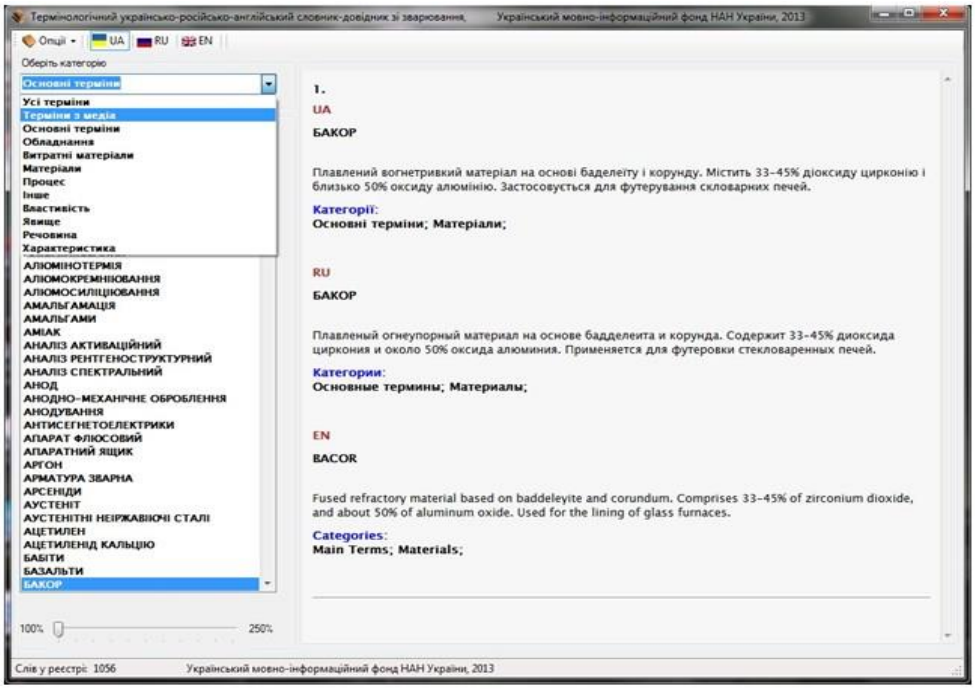
- Possibility of a random order of translation language in dictionary entry;
- Independence of the entry structure from the word's source language;
- Possibility to increase the number of translation languages.

Experimental simulation was conducted on the corpus of electronic Ukrainian-Russian-English dictionary of Welding [9] with the register volume about 12 thousand of terms. From this registry were selected about 1000 nuclear words, and then they were divided into the following categories: equipment, supplies, materials, processes, properties, phenomena, substances, characteristics etc.

A terminology unit was chosen as a main structural element of the dictionary entry – a structural part of a dictionary entry that contains term or terminology phrase, its grammatical parameters, possible synonyms, phonetic or morphological variants that correspond to a certain terminological concept in one of the languages.

To create a digital version of the terminological dictionary, in contradiction to paper dictionary, the following rules were adopted: separate dictionary entries for terminological phrases; separate entries for the reflexive verbs; there are no reference entries in dictionary; dictionary entry has multiple terminological concepts.

External interface of VLL "Welding" is presented below. There are three registers of terms and terminological phrases on the left side of the interface, where the alphabetical order is automatically induced by system for each of the selected languages – Ukrainian, Russian or English. A search box that located above the register is used for terms searching. On the right side there is a dictionary entry that was built dynamically from the elements of lexicographical database. User has possibility to print a dictionary entry he (she) needs, after reviewing it.



One of the main tasks of this work was to create opportunities for the detailed information presentation of terminological concepts. For this purpose, the structure of VTL "Welding" and corresponding lexicographical database were developed with the possibility of adding to the individual terminology units of dictionary entries media in any modern format. Digital dictionary is equipped with editing tools (terminological unit is an editing unit). The user is able to introduce new concepts to digital terminology dictionary. Editing entries are made at the terminological concepts level.

Such system functionality is necessitates to divide users' powers, from user level which rights are restricted to the possibility only to view dictionary entries and comment units of terminology concepts from particular part of the register, to editorial and administrative levels of access.

Virtual Lab enables operative interaction between specialists in various branches in remote access.

Custom version of the dictionary was released on CD-ROM in amount of 1000 copies. Also it is available on-line on the website [7].

References

1. Гринев-Гриневич С. В. Введение в терминографию: Как просто и легко составить словарь. Учебное пособие / Гринев-Гриневич С. В. – [Изд. 3-е, доп.] – М. : Книжный дом «ЛИБРОКОМ», 2009. – 224 с.
2. Добров Б.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного-поиска. Ученые записки Казанского Государственного Университета. Серия Физико-математические науки / Добров Б. В., Лукашевич Н. В. – Казань, 2007 – Т.149. Книга 2. – С. 49–72.
3. Лукашевич Н. В. Тезаурусы в задачах информационного поиска / Лукашевич Н. В. – М. : Издательство МГУ, 2011. – 512 с.
4. Мазур А. А., Формування національної терміносистеми в галузі зварювання на основі віртуальних лексикографічних лабораторій / А. А. Мазур, В. С. Петрук, С. В. Пустовойт, М. В. Надутенко, И. В. Остапова, Н. Н. Сидорчук, В. А. Широков // Вісн. НАН України. – 2014. – № 4.
5. Палагин А. В. Системно-онтологический анализ предметной области [Текст] / А. В. Палагин, Н. Г. Петренко // УСиМ. – 2009. – № 4.
6. Потапова Е. В. Лингвистическая онтология "Физика магнитных явлений" и ее приложения: дис. ... канд. техн. наук : 10.02.21 / Потапова Елена Владимировна; Нац. акад. наук Украины, Нац. б-ка Украины им. В. И. Вернадского. – К., 2013. – 180 с.
7. Українсько-російсько-англійський словник зі зварювання [Електронний ресурс] : [Веб-сайт]. – Режим доступу: – http://lcorp.ulif.org.ua/
8. Широков В. А. Комп'ютерна лексикографія / Широков В. А. – К. : Наукова думка, 2012. – 352 с.
9. Широков В. А. та ін. Українсько-російсько-англійський словник зі зварювання. ISBN 978-966-507-276-8 – 2010. – Ел. вид.
10. Gruber T. R. A translation approach to portable ontologies / Gruber T.R. // Knowledge acquisition. – 1993. – № 5 (2). – P. 199-220.
11. Guarino N. Formal Ontology and Information Systems // Proc. of the 1st International Conference on Formal

Ontologies in Information Systems FOIS'98 / N. Guarino, (ed). – Trento, Italy : IOS Press, 1998.

12. Hovy E. H. Combining and standardizing large-scale, practical ontologies for machine translation and other uses // Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). – Granada, Spain, 1998.

13. Stumme G., Using ontologies and formal concept analysis for organizing business knowledge. // Proc. Referenzmodellierung 2001.

14. Welty C., Ontologies: Expert Systems all over again / Welty C., McGuinness D., Uschold M., Gruninger M., and Lehmann F. // AAAI-1999: National Conference on Artificial Intelligence, Austin, Texas. – 1999.