LINGUISTIC TECHNOLOGIES

APPLIED LEXICOGRAPHY AND SCIENTIFIC TEXT CORPORA

Larisa Beliaeva

Abstract: *Nowadays applied lexicography is a special domain of applied linguistics and language engineering in the framework of problem–oriented automated and automatic dictionaries and databases. Modern approach to dictionary creation assumes preliminary work with parallel or comparable text corpora to be considered as reference database for solving both research and practical lexicographic problems. Parallel text corpora are not always available. One of the options is to create a source lexicographical material as a text corpus with parallel presentation of initial texts, their machine translations and post-editing results. Analysis of comparable text corpus permits to reveal the set of terminological collocations (mostly noun phrases) on the translation level. The paper considers this process on the example of creating a dictionary on the Bologna process domain. The procedure permits to specify translations of lexical units in large text collections and to reveal the domain structure and its terminological system. This idea is shown on the examples of analysis for the collocations with component "higher education", being the most frequent in the Bologna process text corpora.*

Keywords: *applied lexicography, automatic dictionary, parallel text corpora, noun phrases, terminological system, machine translation, postediting.*

ACM Classification Keywords: *A.0 General Literature - Conference proceedings, H.2.5 Heterogeneous Databases, data translation, I.2.7 Natural Language Processing.*

Introduction

Dictionary creation is based on preliminary work with parallel or comparable text corpora, which can be considered as reference databases for solving both research and practical lexicographic problems. Parallel text corpora are the perfect source of lexicographical materials as these corpora are constructed on the basis of problem-oriented texts (articles, monographs, conference materials and their translations into other languages). Such corpora are to be sentence–by–sentence aligned, that permits to reveal and analyse terms and their translations, evaluate their level of standardization and translation conformity as well as prevalence of special variants. However, creation of such a corpus is not always possible. One of the options is to create a source lexicographical material as text corpora with parallel presentation of initial texts, their machine translations and post-editing results. These edited machine translations are to be agreed with experts in the proper knowledge domains. It is important, that the quality and potential of such corpus depends on cooperation with experts when selecting the source material and editing the machine translations. We ask authors to follow some simple guidelines.

Applied Lexicography and Automatic Dictionaries

Range of modern information technologies (IT), knowledge of which is now a significant part of any professional competence, makes it possible and really necessary to develop and implement different types of computerized tools for philological research work. These tools are used for developing new types of computer adaptive testing and tutorial systems, various systems of natural language processing (NLP), including technologies for

computational lexicography. Nowadays in the framework of open and multilingual communication a series of research and practical lexicographic problems are to be solved quickly and adequately. In general the main problems solving of which involve construction of modern lexicographic systems are as follows:

- information retrieval, information and knowledge mining when using various multimedia and multilingual sources;
- retrieval, presentation and dissemination of multilingual information;
- automatic mining of new facts from multimedia resources;
- using special knowledge sources for knowledge tagging and access (knowledge sources for various types of lexicons, thesauri, encyclopedia databases, etc);
- supporting human-computer natural language and interpersonal computer-based interaction;
- supporting distance learning in the open learning systems, including adaptive knowledge testing, electronic textbooks and computer-assistant tutorial systems development;
- creating intelligent tools for automatic bibliography, texts analysis and understanding;
- modeling and predicting of user needs and intentions on the basis of possible quests to different information systems;
- supporting human-computer oral interaction and speech analysis and generation.

These problem solutions define the necessity for creation and use of specialized systems for multilingual information processing in different domains. To solve these problems we need special lexicographic bases, thus all these problems relate to computational lexicography as creation of appropriate dictionaries determines translation and communication quality.

Correspondence with crucial research problems, relevancy and adequacy of lexicographic system spectrum define the level and relevance of knowledge and data mining from the texts of different nature, composition and function. Unfortunately modern translation dictionaries, both paper and automatic, do not correspond with the science and technology levels. The case is better for the pairs of the foreign language – Russian language, but is absolutely incredible with the pairs: Russian language – foreign language. This situation is not only the result of natural lagging of the lexicographical results but the result of traditional approach for creation of new dictionaries on the basis of dictionaries published. If we compile a new dictionary on the base of old lexicons and different glossaries with small part of the terms found by a lexicographer in course of his/her translation work and do it with the help of any IT means, the situation doesn't change. In this case any information technologies used have nothing to do with modern approach. In this context the information technologies only make the lexicographer work not so difficult and tiresome when comparing and compiling different dictionary sources and when editing the final word list. Thus finding a new way using computer tools for effective creation of the dictionaries that reflect the real terminology and domain structure is a special task [TKE, 2010].

Using IT in the applied lexicography gives us an opportunity for

1. Supporting the lexicographer work at dictionary creation and maintenance:
   - solving the problems of lexical units (LU) selection, their lexicographic description, extraction of lexical unit information from the domain–oriented text files [Cerbach, Euzenat, 2001];
   - creation, editing and correction of the dictionary layout;
   - word lists creation and maintenance on the basis of LU selection from lexicographical databases according to the given criterion or a set of criteria;
   - creation and maintenance of terminological databases and ontologies.

2. Supporting the work of a specialist and/or interpreter when using different type of dictionaries in electronic or paper format:

- information extraction from various lexicographical sources (automatic, automated, resident dictionaries);
- research of lexical composition and lexical spectrum dynamics for a certain language/ sublanguage.

The task of effective terminology mining and description is solved when creating various types of automatic dictionaries to be used as the basic part of NLP systems and the quality of the NLP results depends on dictionary completeness and adequacy. Thus a sound approach to automatic dictionary (AD) creation is determined by the necessity to process a large volume of domain-oriented texts in order to access real terminology used. Special problem-domain orientation is one of the most important characteristics of a modern AD as it permits to solve lexical level ambiguity when parsing and translating separate words and collocations (terminological units).

Databases which are designed for various intellectual systems differ about their structure, composition, type of components, set of information and relation systems between the elements. But in spite of their differences all possible databases of NLP systems have common features and common problems which are to be solved when designing a database.

Dependency of the database structure on the knowledge domain and the main task of a natural language processing system and as a consequence, the necessity of the AD to be adjusted to the domain peculiarities are now mutually recognized. The same refers to the volume of a NLP-system database. It is now absolutely clear that creation of a practically usable expert system requires to design a huge database, items of which represent the main concepts and conventional terminology of the domain in question.

Not less than 95% of the source text items are to be distinguished and described with the help of a database if the NLP system is designed as a practical one. Naturally, particular volume of a NLP system database depends on the typology of the source language and the chosen procedure of morphological analysis, the aim of which is high-speed and accurate identification of the source text wordforms with the help of AD. The bottleneck of automatic machine phrases dictionaries lies in the necessity to establish for any database the following:

- the typology of machine phrases;
- the method of their recognition in course of text analysis;
- the method of storing the automatic machine phrases dictionaries.

The problem of automatic machine phrases dictionary corresponds with the fact that new and important notions in all contemporary languages are often expressed by means of phrases. Usage of a special automatic dictionary of phrases is absolutely necessary because of the different focal points of nomination: one and the same object in different languages has special descriptions and special features. In general, automatic dictionaries are created on the base of processing of huge samples of original texts (not less than 1 000 000 wordforms) translations, dictionaries and consultations with experts in the domain in question.

Thus, any linguistic database being a part of a machine translation system or a special entry to a knowledge or terminology database of any expert system shall include:

- source word dictionaries, which are organized both as dictionaries of words and dictionaries of stems,
- source phrase dictionaries and
- machine morphology for source and target languages.

In the most general sense selection of lexical units (words and collocations) for an AD shall be done on the basis of:

- statistical criterion that determines the necessity to include in the AD all the units for recognition of 95% wordforms of a text from the domain under consideration;
- criterion of syntactic independence that determines the necessity to include in the AD the units, structure of which is independent of the sentence structure and the nearest context structure;
- relevance criterion that determines the necessity to include in the AD the terminological units, which enter the terminology system, irrespective of their frequency in the learning text samples and their standardization level.

It is to be specially noted that in case of expanding information systems functions the machine translation and translational memory systems domain– and problem–oriented archives of such system are the optimum source for lexical items selection and description. The fact is that orientation on a specific data domain is very important characteristic of any AD, as it permits to solve the lexical unit ambiguity and to standardize the terminology translation on the lexical analysis level.

Modern approach to a translational dictionary creation assumes preliminary formation and use of parallel or comparable corpora of modern texts, which can be considered as a database for solving not only research tasks, but practical lexicographic tasks as well. Written text corpora, as a rule, include the texts as they are, as well as text tagging results: format boundaries and features, morphological characteristics of lexical units etc. These texts serve for creation of concordances, word and collocation lists in case of monolingual corpora, as well as for creation of multilingual lexicons and concordances if we have parallel corpora.

It is necessary to take into account, that lexicographical work even when using the whole set of IT means remains the work of art and can't be fully automated. At the same time, there is a vast potential for preparation of text files for automatic lexicographical analysis. Parallel text corpora are the perfect source of lexicographical materials [Lefever et al., 2009] as these corpora are to be constructed on the basis of problem-oriented texts (articles, monographs, conference materials and their translations into other languages). Such corpora are to be sentence–by–sentence aligned, that permits to reveal and analyze terms and their translations, to evaluate their level of standardization and translation conformity as well as prevalence of special variants. However, organization of such a parallel corpus is not always possible. One of the options to create a material for subsequent lexicographical analysis is formation of special text corpora that include parallel presentation of initial texts, their machine translations and the same translations after human postediting. These edited machine translations are to be agreed with experts in the proper knowledge domains. It is important, that the quality and potential of such corpus to a great extent depends on cooperation with experts when selecting the source material and editing the machine translations.

In case of lexicographical analysis sentence–by–sentence text alignment permits to compare initial sentence, its machine translation and the final sentence translation, thus we are able to reveal and describe the set terminological expressions (mostly noun phrases) on the translation level. In order to receive machine translation results it is expedient to use the dictionaries of a MT system, which contains the needed or comparable words and expressions. Let's consider this process on the example of creating a dictionary on the Bologna process domain (the examples presented below show in thick print those lexical units and their translations that require special attention and modification). Thus, initial sentence was as follows:

Student-centered learning *produces a focus on the* ***teaching-learning-assessment*** *relationships and the fundamental links between the design,* ***delivery****, assessment and* ***measurement of learning****.*

The sentence was translated using Word+ machine translation system with specialized AD for the linguistics domain:

Ориентированное на обучающегося обучение *производит фокус на отношениях* ***teaching-learning-assessment*** *и фундаментальных связях между проектом,* ***поставкой****, контролем знаний и* ***измерением обучения****.*

After human editing we had received:

При ***личностно-ориентированном обучении*** *основное внимание уделяется отношениям типа* ***преподавание-обучение-оценка*** *и фундаментальным связям между проектом,* ***подачей материала****, контролем знаний и* ***измерением качества обучения****.*

Thus, comparison of these three sentences permits to reveal the following units and their translations for dictionary registration:

student-centred learning	*личностно-ориентированное обучение*
teaching-learning-assessment relationships	*отношения типа преподавание-обучение-оценка*
measurement of learning	*измерение качества обучения.*

Besides, comparison of these three sentences permits to specify the translation of the word *delivery* as *подача материала*, this meaning corresponds with the terminology of the domain in question. All these expressions are terms and need specialized translation.

Using the sentence–by–sentence aligned texts gives us opportunity to specify translations of the lexical units in large text collections, domain structure and its terminological system. Let's consider this idea on the examples of analysis for the collocations with component *higher education*, being the most frequent in the Bologna process text corpus, a small research corpus (500 000 wordforms), the aim of this corpus was to verify the composition, structure and translations included in different Russian–English glossaries used for this very domain.

To show the potential of defining terminological system of the *higher education* field we analyze all lexical units in this corpus. There are 126 such elements in the corpus under study, for example:

higher education area, higher education assessments, higher education authority, higher education awarding bodies, national higher education frameworks of qualifications, national higher education qualifications.

The maximum length of noun phrases with the *higher education* component was 8 elements, the only noun group of this length is *New European Quality Assurance Network for Higher Education*, in which the collocation *network for higher education* is its head component. This last collocation has not been registered in this corpus as a separate lexical unit. At the following phase of analysis the whole set of noun phrases with the component *higher education* was used as the sample for receiving a frequency dictionary of lexemes. The function words were excluded from the word list, as a result we had received a key word list, which could be used as the base for terminological system structurization. The most frequent elements were lexical units *qualification, system, institution, program(mes), research, area, minister.* Convertible terms from the whole frequency list were united with the main (most frequent) key words, for example, a group with a key word *institution* consisted of the words *institute, school, university* from the word list.

On the basis of grouping the collocations in accordance with all the key words we had received the following subfields of the terminological system in question:

Programs of Higher Education, Systems of Higher Education, Institutions of Higher Education, Qualifications of Higher Education, Structure of Higher Education, Legislative Base of Higher Education, Types of Higher Education, Management of Higher Education, Degrees of Higher Education, Audit of Higher Education Quality.

These subfields correspond with the subfields for this domain which had been established on the basis of its structure analysis. Besides, the analysis of each of subfields permits to install the nomination peculiarities and variants. For example, the cluster *Management* of *Higher Education* consists of the following lexical units:

Department of Science and Higher Education, European Ministers in Charge of Higher Education, European Ministers of Higher Education, European Ministers Responsible for Higher Education, French Community Ministry

for Higher Education and Research, Minister for Higher Education, State Minister of Higher Education and Science.

When this corpus is expanded for lexicographic research this list of departments and ministers could be more exhaustive.

Research of the collocations revealed and organised on the basis of this procedure permits both to determine the collocations from the texts and there translations to be included in the dictionaries and glossaries and to define the potential collocations (see, for example, *network for higher education, higher education law, higher education program* which are not in the list but could be constructed from the longer ones).

If we use a full-text parallel corpus as a lexicographic base it is necessary to expand them with a corpora of machine translation results. Analysis and comparison of these text files will make it possible to allocate such lexical units, which should be considered as dictionary entries. The main problem is to establish the boundaries and structures of these lexical units – noun phrases.

Noun Phrases in a Scientific Text

During the translation process the text analysis is based on formal parsing and semantic analysis. Both of these processes are based on our possibility to understand the surface structure of a sentence and semantic relations between its components [Beliaeva, 2009].

Noun phrases are the objects of special research in both theoretical and applied aspects. Such phrases are functionally equivalent to a word, but at the same time they represent a convolution of a sentence, i.e. they are, rather, units of syntax, not lexicon. So we can assume that internal structure of a noun phrase correlate with internal dependencies structure of the sentence. The problem is to find a procedure to recognize this structure in a concise form of a noun phrase. The problem is related to the fact than when translating from English to any inflectional language we should know the relation structure between the noun phrase components. If NP is analyzed in scientific or technical texts its denotative or referential status is important, but the object definiteness is given by situation of a speech act both for the author, and for the recipient. This definiteness of an out-language object is the basis of scientific text understanding by the specialists in a domain, the necessary condition of such understanding. In machine translation such understanding can't be simulated which means necessity of postediting. The same problems arise as to whether some of these phrases should be included in the automatic or some other dictionary.

The main problem of special text postediting is to recognize information on noun phrase component relations in the domain in question. This information can be received on the basis of the whole text analysis. This approach seems expedient for as it is based on the formal indications of the author's intentions which are reflected both in the text structure and in the composition of different NP with the same constituents. Establishing the relation structure is only the first step for translation of word combinations as this translation is to be adjusted to the domain in question. In doing so we are confronted with the problem of context-sensitive terms, translation of which depends not only on the domain relations and meaning but on the nearest context.

In machine translation procedure the structure of each noun phrase and its boundaries are to be determined at the sentence analysis step, thus the task of noun phrase translation is to be performed in the framework of the following operations:

1. Establishing the head element of the English noun phrase;
2. Establishing semantic and syntactic structure of the English noun phrase;
3. Finding semantic, syntactic and lexical structure of the Russian noun phrase;
4. Translation.

Since a NP is a sentence convolution, a compression of this structure, such external simplification of both the structure and the form results in the noun phrase semantic complication. The markers of relations between actual components and types of relations between elements, which sentence shows with the help of different means, are absent in the English noun phrase. Absence of morphological markers of case and gender makes it impossible to establish the "host" for an attribute or a set of attributes in the preposition to a noun or a chain of nouns.

Basic noun phrases in English are two-element combinations with a head noun, frequency of which in scientific text three times exceeds the frequency of three-element combinations. However external simplicity of frequent English noun phrase structures is misleading. The fact is that this simplicity could be the result of initial noun phrase or sentence compression. Such compression, formal simplification of NP structure leads to its semantic complication. Pursuant to these, formation of noun phrases in a real text is based on either merging noun phrases and separate lexical units in a new, more complicated nominative construction, or on condensing multicomponent NPs at the expense of deletion of the units which are implicitly obvious.

Formation of a multi-component noun phrase in a text is realized in any of two ways depending on the type of nomination: either as a process of a step-by-step complication and specification of the nomination object (gradual complication of a noun phrase with addition of its head element characteristics), or as a process of sequential noun phrase convolution. This process is realized successively on several levels:

Level 1: transfer from a complex noun phrase to a simple one due to element inversion.

Level 2: elimination of component duplication in a new noun phrase.

Level 3: coordination of semes and elimination of components with duplicated semes.

Referential status of noun phrases in a scientific text permits us assume that author's attitude on information translation and its understanding requires explication of relations within the text. Analysis of texts in different subject domains had shown, that occurrence in the text a noun phrase with length more than 2 elements is followed by occurrence a 2-compound NP in the nearest context, within the limits of 2-3 sentences or combination of title, key words and abstract. Hence, at human translation we can use this situation as a key for structure diagnostic. At MT we need to create a special text translation memory.

The peculiarities of noun phrase formation in the text are to be analyzed as follows:

Connection of two two-element NPs into a new one which results in occurrence of

- four-component noun phrase, the structure of which depends on the structure of the merging NPs, for example, if two groups of *Adjective* + *Noun* type merger the noun phrase which plays the role of an attribute is embedded in the position of the head element of the first noun phrase attribute:

indirect method + *seismic analysis* $\Rightarrow$ *indirect seismic analysis method*

adult learner + *second language* $\Rightarrow$ *adult second language learner*

- three-component noun phrase in case, when one of the elements in two initial NP coincides, for example

mental processing + *processing operation* $\Rightarrow$

mental processing operation

with establishing direct relations between (in this case) the adjective *mental* and the noun *processing*,

- three-component noun phrase in case, when semantics of one of the NP elements is supported as a part of a new noun phrase by the semes of other noun phrase components, for example, merging the noun phrases

communicative method + *language learning*

results in occurrence of a new noun phrase

communicative language learning

- three-component noun phrase in case, when semantics of one of the elements is implied as a part of a new noun phrase at the expense of extralinguistic information of the domain, for example, merging the NPs

seismic stability + direct analysis

results in formation of a noun phrase

seismic stability direct analysis,

which in a text may be convoluted to a three-component noun phrase

seismic direct analysis

The cases of noun phrase transformation considered here under the condition of text coherence and cohesion do not show all possible variants of their development in a text, however, they give the basis for consideration of possible translation of a noun phrase with high degree of structure compression. Besides the research conducted permits to show that exactly two-element noun phrases present special difficulties at their analysis and translation.

To solve the problem of such noun phrase translation we can see only two approaches which can be used both in human and machine translation.

The first approach includes modelling the knowledge base of the domain in question (in the framework of a MT system) or appealing to such factual knowledge of a translator. In case of machine translation this approach is based on vast investigations of the possible relations between both the main concepts of the domain and the items of the linguistic data base. Creation of such a thesaurus or a semantic net is not only extremely laborious but space-consuming. But the most serious disadvantage of this approach is that an unambiguous solution of the problem sometimes can't be achieved. For example, for a noun phrase *constant amplitude deformation cycle* a semantic network would show relations between the nodes *constant* and *amplitude*, *constant* and *deformation*, *constant* and *cycle* and it is impossible to use this information to establish the dependencies structure of the noun phrase both in human and machine translation.

The second approach could be more formal: we can use the information, which can be received on the basis of the whole text analysis. This approach seems more expedient as it is based on the formal indications of the author's intentions which are reflected both in the text structure and in the composition of different noun phrases with the same constituents.

Conclusion

Investigations of text structure in terms of noun phrase composition in different subject domains (medicine, seismic isolation, space systems, power plants construction, language teaching etc) had shown that dependency structure of a noun phrase with three or more constituents can be obtained from the nearest context: a 2-component NPs would show the accurate relations relevant for this special text.

This means that for an English-Russian machine translation system we need a special tool for noun phrase analysis and translation within the text boundaries, something like text translation memory, which could store the history of noun phrase development. The same problem is to be solved with human text translation or dictionary creation.

Solving the problem of dictionary creation requires to distinguish between linguistic automata with which this problem could be solved automatically or semiautomatically and linguistic automata in work of which a specialist – lexicographer could participate both at he level of corpora tagging and alignment and text analysis, and at the level of lexicographic problem solving. In using the parallel corpora the second type is more expedient.

Bibliography

[Beliaeva, 2009] L.N.Beliaeva. Scientific Text Corpora as a Lexicographic Source // SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proc. from the Intern.Conference, November 25 – 27 2009, Smolenice, Slovakia. – pp. 19-25

[Cerbach, Euzenat, 2001] Cerbach F., Euzenat J., Using Terminology Extraction to Improve Tracebility from Formal Models to Textual Representations. // NLDB 2000, LNCS 1959. Berlin Heidelberg: Springer Verlag 2001. – pp. 115-126

[Lefever et al., 2009] Lefever, E., Macken, L. and Hoste, V., Language-independent bilingual terminology extraction from a multilingual parallel corpus. // Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, 2009. - pp. 496-504.

[TKE, 2010] TKE 2010: Presenting terminology and knowledge engineering resources online: models and challenges. – Dublin: Dublin City University, Ireland, 2010. – 102 p

Authors' Information

Larisa Beliaeva – *Herzen State Pedagogical University of Russia, Professor, Chief of Machine Translation Laboratory; e-mail: lauranbel@gmail.com*

Major Fields of Scientific Research: Human and Machine Translation, Applied Lexicography, Multi-dimensional information systems