Information Technologies for a Language Worker*

Larisa Beliaeva [1]
lauranbel@gmail.com

Tatiana Gornostay [2]
gornostaja@tilde.com

Olga Kamshilova[1]
onkamshilova@gmail.com

[1]Herzen State Pedagogical University of Russia,
St. Petersburg, Russian Federation,
[2]Tilde Company, Riga, Latvia

Abstract

The paper considers the potential of information technologies and Web resources in practice of various language workers to support and ensure efficiency, accuracy and correctness of the information product they develop. It focuses on analysis of modern multilingual terminology databases, use of parallel and comparable corpora for term extraction and translation. As the modern industrial automation processes (Industry 4.0) determine new ways of processing, requesting and delivering information, the paper suggests a number of initiatives, considering optimization of a language worker's professional space and use of computer working practices.

Keywords: *automatic workstation; terminology data bases; text corpora; term extraction, termhood, terminology management; user translation dictionaries; professional linguistic education; language workers.*

1 Introduction

Modern development of science and industry in many ways is defined by the degree of information technologies adoption for Industry 4.0 actual demands. Naturally, the Industry 4.0 principles and methods depend on standardized methods of processing information on the project under development, production, operation and supply of materials.

Dramatic changes in science and technology, appearance of new research areas and, what is more, new knowledge domains result in sharp backlog of specialized language resources, supporting any language worker (terminologist, translator, lexicographer, technical writer, grammarian, teaching language specialist, etc.) studies and work. The available resources are mostly Web-based, which are not only stored but tagged according to ISO TC 37 standard (for more details see CLARIN project) [Broeder et al., 2010]. Specialized resources such as various

text corpora, grammars, dictionaries in paper form and/or embodied in various automated dictionary systems and information processing systems, unfortunately, satisfy in the majority neither the modern level of science and technology, nor the mainstream of knowledge domains development. This situation is determined not only by a natural backlog of such resources, connected with the necessity of terminological analysis of modern text files, but also with the conventional approach to dictionary creation and management on the basis of already published sources, with little account, if any, for currently published translated texts in a particular domain. Much is written and spoken today on the impact of the latest educational reforms in Russia. However, modern educational process has to face much stronger challenges than administrative restructuring. In the context of growing information volume, an evident conflict of "academic" approach to educating a language specialist and the need for competent language workers, pliable and prepared to change and develop their new qualities, we are placing an emphasis on a competence approach in training a language worker [Almazova et al., 2018; Chernyavskaya et al., 2018] and have proposed a possibility of training language workers on the basis of classical philological / linguistic education to justify the need for training new specialists [Beliaeva, Kamshilova, 2018].

The present paper suggests some initiatives that may be well applied both for training a language worker and for supporting a practicing specialist. It concerns issues of pragmatic nature, such as optimization of a language worker's professional space and use of computer technologies in their work.

The paper demonstrates the potential of information technologies and Web resources for language worker professional skills development, such as special text production (writing and translating) or practical lexicography. These resources, if appropriately applied and used, which implies the way they shape working space (as an automated workstation, for example), will support and ensure efficiency, accuracy and correctness of the information product various language workers may develop. The paper specially focuses on modern multilingual terminology databases analysis and parallel and comparable corpora use for terms extraction and translation.

2 Automated workstation for a language worker in a modern information environment

To support a language worker's practice of information extraction м.б. mining или retrieval?, translation, localization and dictionary creation a special complex of linguistic resources and software is required. Such complex can be arranged as automated workstation (AWS). Competent use of an automated workstation, which includes a set of resident dictionaries, thesauri, spell-checking systems, systems of information access due to data communication networks, becomes customary not only for a translator, but also for experts in various knowledge domains.

According to Russian GOST (State Standard) 34.003-90 automated workstation (AWS) is a program-technical complex of computer-aided systems for performing special automation activities. The core of AWS structure is a personal computer with standard and special software, as well as data and knowledge bases. However, this very State Standard does not mention any opportunity to create such AWS for humanitarian knowledge specialists. At the same time specifics of humanitarian approach to information and knowledge extraction gave rise to technological diffusion for the AWS, intended for operational processing of texts in

different natural languages and, first of all, in the field of automatic translation.

Actually, the AWS idea in humanitarian sciences began embodied in the 90s of the past century. Rapid development of specialized complexes for a professional translator interested both users and designers [Gordon, 1997].

Embodiment of the AWS idea is connected not only with comprehension of such tool necessity for professional activity organization, but also with accessible computer facilities for creation and implementation such AWS. At the same time, it should be borne in mind, that translator's AWS was not something absolutely new as the proposals to create its components (including translational memory) were made more than 50 years from this moment. AWS realization in its modern form has required computer system development, implementation of special technologies and devices, auxiliary aids and systems, but the necessity of effective use of computer devices and tools for translation was recognized long before. AWS for humanitarians constitutes an information system, elements of which are not material objects but one or another data (information) types. Under present-day conditions an information system is a set of hardware-in-the-loop tools and algorithmic procedures, intended for retrieval, input, storage, simulation and image presentation of data. Information system specifics is determined by the set of its functions [Belyaeva, 2009].

Changes in the situation with circuitry and electronic devices, namely personal computers propagation, Internet realization as the field of active work for different user categories have created entirely new conditions for specialists in humanitarian knowledge domain. These changes nowadays give rise to creation and gradual expansion of high-technology environments: industrial, research, educational, that, in turn, required a detailed understanding of new forms of information transfer, knowledge extraction, a new context of training included. Therefore, today one of the aspects of cultural and scientific interaction is creation of national knowledge funds in various fields of human activity, the funds which give opportunity to receive information on the items to be stored, to fill up and modify this information as necessary.

Unfortunately, exactly in this area of information technologies development a phenomenon, which is absolutely progressive, had a specific (sometimes destructive) impact: constant updating of computer facilities determines an inevitable transfer to the next generation computers. The fact is that for all information systems, connected with automatic information processing, this transition had been followed by a system creation all over again. This spiral development was characteristic for all information processing systems: systems of machine translation, information retrieval, annotating and abstracting, computer training etc. Following this development were the changes of principles and work methods of actual systems, and revision of their theoretical bases and realization principles. The modern condition of these systems is determined by the next computer revolution, which is connected with

• personal computers as well as advanced periphery and gadgets, hence, tools for mass user work,

• Internet and development of global Web as the mode of information transfer and its storage.

At the same time, many years' experience of using computer technologies for solving various types of problems has shown that the results produced depend on the results of natural language (NL) processing. As NL is the key tool to generate, store and transfer information, information technologies in NL processing field (linguistic technologies), that realize text automatic processing algorithms, are a necessary condition for solving the problems relevant to information technologies as a whole.

Development of hardware peripherals (scanners, modern information carriers, digitizers, etc.), that have generally changed text processing context by eliminating the need of labor-consuming text punching for subsequent computer processing has become an important stimulus for AWS idea realization.

Realization of AWS idea as a practical actual electronic product is connected not only with realization of a similar facility for text processing in a language worker's professional practice (freelance translator or a translators' group member, lexicographer, terminologist, technical writer, etc.), but also with coming and, that is equally important, availability of computer facilities for AWS creation and its subsequent implementation. Automated workstations can fulfill a complex of functions in order to organize a language worker's on-site activity, granting a huge range of integrated facilities for

- multilanguage text preprocessing,
- optical recognition symbols,
- information compression with information extraction,
- document transfer and obtaining over information networks
- spell and grammar checking,
- prepress and dummying,
- terminology management,
- production of concordances and other dictionary types,
- access to local or remote data banks or other linguistic resources,
- translation memory management,
- machine translation.

Pursuant to AWS tasks the structure of such complex includes the following program and linguistic modules:

- A linguistic automaton as a complex of automatic text processing tools [Belyaeva, Piotrowski, 2005]. This part of AWS is used for text translating, editing, annotating, information retrieval and analysis etc. Linguistic automaton is a hierarchical system of program modules, each of which fulfills a certain operation for text processing and can function both independently and in complex with other tools.

- A full-text database which supply text storage, modification and browsing. Besides the information directly stored this base should include references to national text corpora and net resources. This part of AWS can be used immediately for analysis of specific linguistic facts, textual and comparative analysis. Furthermore, such a base is an important source of information for building dictionaries of different composition and assignment.

- A terminological data base for storing, structuring, thesaurus and ontology building, as well as for term extraction and translation. The structure of this base includes electronic dictionaries, routinely used by translator, for example, in Russia the most large-sized multilingual dictionary (224 dictionaries for 19 languages) is Lingvo (www.abbyy.ru). Moreover, this data base can contain references to electronic dictionaries on the Russian Language Institute portal (www.slovari.ru).

- A base for reference and automatic dictionaries, specialized glossaries, automatic and educational dictionaries, united in an integrated complex, such that makes it possible to use any collected dictionary information for educational and scientific goals.

- A base of specialized linguistic software, among the other things the machine translation systems, translation memory, systems of terms extraction from parallel and comparable texts corpora.

Nowadays Internet network contains a set of similar AWS, for example one of early variants is a Canadian system Alis Translation Solutions (ATS), uniting series of tools for natural language processing and set of auxiliary services. ATS includes systems for supporting human and machine translation, on-line dictionaries, browsers, tools for operation control. The system can solve the tasks of dynamic information tracking, information publishing document circulation, information exchange on the basis of e-mail, information retrieval on the multilanguage sites and their translation [Coté, 1998].

Lexicographic resources to be included in the AWS structure can be oriented on terms and/or concepts choosing and processing, on working with a certain language pair, multi- or monolingual resources. The AWS for a language worker should provide an opportunity to use a content management system for storing the necessary data. Besides an AWS should have access to on-line tools for terminology management. In each individual AWS the language worker's own resources are to be used with on-line ones.

A system (a distributed network) which unites individual AWSs should assign who personally has access to dictionary databases and on which terms, since terminological units are extracted and translated by different users and on different phases of dictionary base formation. Creating common terminological resources implies a structured prescription for different users, whose work results in changing the term bases and translation dictionaries. The prescription should consider the potential situations [Großjean, 2009]:

- A terminological database user finds out a text term, absent in the dictionary base.

- A user-proposed pair "term – translation" can be entered into the resource of an individual AWS, besides, it is transferred to the terminological data manager, who delivers it for the expert evaluation, and hereinafter a decision on term introduction into common dictionary resource is to be made.

- A designer of actual terminological resource enters a new term and its translation.

- In this case a decision on introducing the term in question into the integrated base is required.

- A product designer enters a new term, which should be approved by experts.

Thus, the AWS resource base for a language worker, remaining an individual user tool, should be continuously evaluated from the standpoint of a common resource to be built.

Modern research in the field of linguistic resource creation assumes preliminary terminological work for extraction and description of terminology in different languages, followed by these descriptions harmonization and co-ordination of terminological systems of different languages. Furthermore, choosing the proper Web resources and means for automation of new terms extraction from texts in different languages is of special importance.

3 Web resources for technical translation and terminology management

Modern tools for a language worker are not only information technologies (IT), that are constantly developing, but IT-based language resources, oriented on storing both necessary information (lexicographic, expert, corpora) and machine translation and information processing systems. At present language resources comprise natural or artificial language descriptions and tools for their management and support to be used for presentation of appropriate language data (dictionaries, ontologies, thesauri, etc.), as well as for presentation of certain resources in various text processing systems and for solving the problem of empirical information extraction. Furthermore, language resources comprise text resources, collected in powerful databases and forming language knowledge sources.

A very important role in every terminologist's practice or any other language work is played by special lexicographic resources aimed at terminology management: a terminologist shall react quickly (and in a standard way) to satisfy the requirements of high-quality information processing and mark out terminological units that have not been earlier registered or are brand new. Results of the terminologist's activity are to be promptly entered into proper lexicographic resources.

As much as the terminologist's place and function in a certain production string, the place and function of technical translation and a technical translator need to be verified. To meet the demands of modern technological processes the technical translator's work today requires support of complex facilities, which include a chosen machine translation system, a complex of automated dictionaries, subject oriented text corpora and applied programs for text processing – on the whole, language resources to be organized properly.

Under development of information technologies language worker's activity (translators in the first place) is the base of information extraction and further analysis. Under such conditions the speed of performance and high translation quality acquire special importance as the translation, which is delayed or incorrect, can cause critical consequences (in seismic protection, nuclear engineering, medical science and other high-risk domains).

Evidently, both translator and terminologist today must be informed and proficient in relevant language resources that ensure efficiency, accuracy and correctness of the product they develop.

Terminological databases and terminology management It is to be noted, that terminology management systems were created long ago. As early as in the 70s of the previous century large companies and government institutions developed language machine funds: new terminology appeared simultaneously with economical and technical growth. Such funds were assigned for unification of terms within a special domain and for translation the domain specific texts. One of the largest funds was the TEAM data bank, developed by Siemens company for European

languages, it included about 700000 lexical units from various respectively grouped domains: natural sciences, business, engineering, etc. [Hutchins, 2001]. The fund materials were also used for creating specialized dictionaries. Later in the 80s national language machine funds began to be built (Russian language fund included). The purpose of such funds was creation of universal language databases.

Terminological databases (TDB) represent computer-aided storages for terms of a special domain. In such storages the terms are supplied with additional information of both linguistic (combinatory power, frequency, semantic field, etc.), and extralinguistic (domain, definition, regularity, etc.) type. The main purpose of TDBs is information on separate words or collocations (descriptions, examples, translations), these banks were used as the base for compiling special text glossaries and new specialized translation dictionaries [Hutchins, 1988].

Many data bases were created as multilingual, nearly all of them had direct dialogue access, majority of these first databases provided extensive description of the data entry units, several of the first terminological bases were really huge. New terms were provided with examples from texts in other languages, definitions received from reliable sources, codes of data domains and references.

Modern multilingual lexicographic resources are Web-based, according to their multipurpose capability and availability they could be classified as governmental (for example, supported by European Commission) and initiative, developed by corporations or research groups. The survey of most popular governmental bases of terminological data is presented below.

Databank Eurodicautom [Johnson, Macphail, 2000] is an example of most powerful governmental terminological bases, covering all the languages of European Union and the Latin language. European Union issues documents in 24 languages and works with 552 language combinations. Up to 2008 the main dictionary base included 1 240 000 dictionary entries (5 million terms) and 325 000 abbreviations and acronyms. Data domain codes are based on the universal Lenoch classification. Completion of the dictionary database was based on the results of translation units (in Brussels and Luxemburg), the new terms and their proposed translations are systematized by Eurodicautom team. Besides, part of terminological information was imported by contracts with private companies and experts in separate knowledge domains.

The database updating took place every week [Rirdance, Vasiljevs, 2006]. Information input was organized in each translation unit pursuant to their own rules and approaches depending on various agreements about use and methods of co-operation of each language community and each country. Therefore, it became a necessity to merge all separate databases into one integrated coherent base with constant input of material for approximately 5000 translators from EC institutions.

In 2008 the European Parliament decided to create a special unit, which functions were co-ordination, tool support in terminology studies and result preservation in the IATE format (InterActive Terminology for Europe). This format is a terminological relational database. In addition, the function of this unit was collaboration with translation unit and other institutions under managing the new database which contained millions of terms, extracted from other bases and imported without any filtration. This database management provided for obsolete terms removal, as well as removal of recently added duplicate units, besides the database had to be filled up with terminology of new EC languages. The European Parliament organized a TermCoord department for terminology co-ordination, which provides access to EC terminology through public site and free tools, as well as through interinstitutional terminological

portal EurTerm [Maslias, 2014].

EuroTermBank represents another powerful Web governmental terminological base, covering all European Union languages and the Latin language. This linguistic resource unites 133 local resources, developed in various EC translation units, 2 650 976 terms (the number is permanently increasing), 710 705 dictionary entries, 221 512 definitions in 33 languages. Information Structure in the EuroTermBank assumes various options for choosing the source and target languages, data domain, information format presentation, etc. A special option provides the user with information on translation variants in different data domains and on terminological collocations. Dictionary information is free for all.

The Web resource of EuroTermBank can be considered as a proofed model for multilingual Web resource, creation of which is actual both for languages of Russia's national republics and for Customs Union languages, since it can supply proper terminological and lexicographic support for document translation in various areas of co-operation and knowledge.

The resource in its modern form permits language workers to search terms in various sources, to identify candidates for terms in the documents they process or develop, to extract terms automatically, to browse term translation variants in different data domains, to search for terms in several target languages simultaneously, to specify translations and to share information with other users. Access to the resource is possible from Microsoft Word directly.

At the same time, it should be borne in mind, that any Web lexicographic resource essentially includes terminology, recoverable as a result of standardization, and (in spite of their huge volumes) is not capable to cover all terminology in any domain, especially for press-forward knowledge domains. The main deficiencies of terminological resources today are their high cost and long term for their creation, insufficient scope and adequacy of term translation, especially for modern concepts nomination, non-sufficiency of terminological resources sharing and absence of devices involving a terminologist in working with Web resources.

Thus, Web lexicographic resources require constant "tracking" of new terms, that, in its turn, assumes development and use of both purely linguistic methods and statistic-linguistic analysis as well as metrics of automatic lexical unit extraction one-word terms and terminological collocations found in actual special domain texts. Lexical units to be automatically extracted from texts are called term candidates. Term extraction procedures and special metrics permit to evaluate:

- compatibility preferences of lexical units (unithood), their syntagmatic proximity;
- termhood (degree of terminological potential) of the extracted collocations considered as TC;
- charactiristics (salience) of a certain collocation in a specialized text corpus or in the terminology of a language for specific purposes (LSP).

Special methods for evaluating potential of certain lexical units to function as key words in the text (keyness) are now under development.

Text corpora for terminology extraction A comparatively new approach to language resources creation is based on formation and use of real text corpora, which can be considered as databases for solving not only research, but also practical lexicographic issues. As a rule, written text corpora include texts which are specialized according to a domain, author, function, etc. Corpora management systems (corpus managers) provide information retrieval in various linguistic aspects, which is possible due to annotating (tagging) texts and texts elements. The

tagged texts can be the basis for concordances, word and collocation dictionaries in case of monolingual corpus, as well as for creation multilingual lexicons and concordances for term extraction in case of parallel or comparable text corpora (see figure 1).

There are two main approaches to automatic extraction of source term – target term correspondence from multilingual text corpora. Within the first approach the procedure begins with parallel text alignment. The alignment is meant with heuristics for finding conformity points (or, rather, candidates to such points). Thus, numerals, abbreviations, dates, proper names, acronyms etc. can be used as conformity points for sentence alignment. For closely-related languages, for example, Spanish and Portuguese, words which identical roots can be positively considered as candidate pairs.

Under this approach, however, besides the complexities caused by terminological systems asymmetry in different languages, there is a specific problem of translated text choice for parallel text corpora, as the translation quality is frequently questionable. Even term glossaries issued by domain experts show term pairs which are not translation equivalents for each other.

It is necessary to take into account, that when using comparable text corpora, the alignment problem turns specific. In case of parallel texts corpora, the main task is sentence-by-sentence alignment, which relies on formal indicators of boundaries and sentence parts, conformity of volumetric and pragmatic text structures. With all arising technical and linguistic complexities (see above) this process is quite realizable. In case of comparable text corpora only terminological alignment is possible, which relies on revealing one-word terminological units characteristic for both corpora and their comparison as translation equivalent

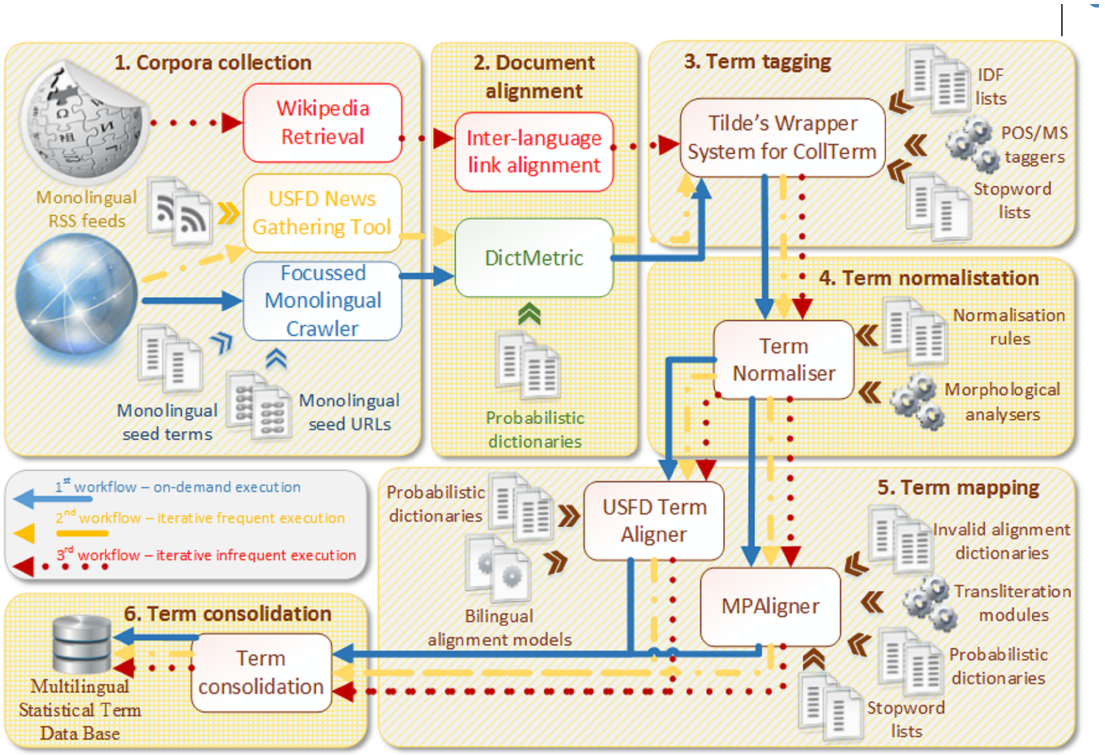


Figure 1: Technology of multilingual term extraction from the Web comparable corpora [see Vasiljevs et al., 2014]

candidates, as well as search of set expressions with these one-word terms [Hertog et al., 2010]. Further comparative analysis requires knowledge from automated translation dictionaries, enabling to verify the chosen term pair. Extraction of multicomponent terms can be produced on the basis of automatic syntactic analysis on functional segments level – noun phrases.

Thus, automatic term extraction (both one-word and multi-word lexical units / collocations) is based on preliminary alignment of different language texts, identification of terminological units in texts in each language and further establishing their translation equivalent or, rather, establishing feasible translation equivalent candidates.

Majority of Web term extraction systems apply either statistical or linguistic approach [Kageura, Umino, 1996]. They use lexical unit frequency, likelihood ratio for two-word terms, mutual information measures. For evaluation of longer collocations the only statistical parameter is term candidate frequency in the text corpora [Lefever et al., 2009]. Recently hybrid approaches have become the most popular ones as they represent the overcoming of unilateral approaches to solving the term extraction problem on the basis of both linguistic and statistical elements.

Dissimilarity of initial texts, levels of text specialization, end user purposes and profiles, and level of automation explain the absence of universal methods for solving the problem of term extraction from a natural language text.

The idea of creating automated systems for term extraction from parallel text corpora has been under study for more than 20 years and is partially realized in various terminological projects. As it was mentioned above, automated term extraction from parallel corpora faces two major problems: asymmetry of national terminological systems and inadequate quality of text translation (both machine and human). To face the first and avoid the second may help comparable text corpora which compile texts in different languages but of similar knowledge domain, register and format.

A specific source for comparable text corpora are conference proceedings, devoted to the same scientific problems, but organized in different countries and with different working languages.

Scientific text in a comparable corpus is to be considered as the result of information transfer and the source (starting point) of information mining and extraction. Thereafter the scientific text content, especially the part of its sense which is universal and can be extracted in case of minimum coincidence of the author's and recipient's thesauri, is mainly determined by information on the objects under consideration, linguistically described by their names – nouns and noun phrases (NP). Furthermore, the proper information extraction from a scientific text is determined by correct interpretation of terms (objects names). Thus, the adequacy of text perception on the lexical level is determined by text saturation with NPs, the degree of their compression and/or completeness of object nominations.

4 Comparable corpora method pitfalls: a linguistic guide for a language worker

A comparative analysis of such text corpora shows several additional pitfalls. "English" conference texts for the most part are written in global English, that means frequent infringement of syntactic sentence structures, caused by the authors' native language influence, and absence of terminology harmonization, due to which terms represent translations of appropriate lexical units of author native language, instead of using the standardized nominations. "Russian"

texts, in their turn, are "weighed down" by long Genitive chains in NPs, frequent use of syntactic structures with object in the first position of sentence and absence of explicit boundaries between terminological NPs that nominate the terms and may have different syntactic functions (see, for example, a sentence fragment *построение соответствующих различным конструктивным параметрам семейства силовых характеристик упругопластических демпферов, verbatim: creation of elastic plastic dampers force characteristics of which correspond with their design parameters).*

Under the condition of overdeveloped ambiguity caused by Genitive overuse, which is characteristic for Russian noun lexicon, it is very often impossible to determine term boundaries and structure.

NPs as terminology items 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 are results of a sentence compression, i.e. they are rather units of syntax, not lexicon. Thus, we can assume that internal structure of an NP correlates with internal dependencies structure of the appropriate sentence and reflects the peculiarities of the object nomination. The problem is to find a procedure or approach to recognize this structure in the compressed one.

One of the most serious problems of English scientific text analysis and machine or human translation into Russian is determination of dependency structure in NPs. The problem is related to the fact than when translating from English to any inflectional language we should know the relation structure between the NP components. In scientific texts simple NPs are multicomponent units with a large number of attributive elements in preposition to the NP head. Being dependent members of a sentence, these elements form one syntactic group with its head.

Since an NP is a sentence compression its external structure and form simplification causes the NP 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 NP. Basic NPs in English are two-element combinations with a head noun, their frequency in scientific texts is considerably higher, for example, in seismic construction domain texts their frequency is three-fold compared to frequency of three-component NPs.

However external simplicity of the most frequent English NP structures is misleading. The fact is that this simplicity could be the result of initial NP or sentence compression. Such compression, formal simplification of NP structure leads to its semantic complication. NPs for terminological candidates: term extraction and termhood evaluation NP boundaries and their heads identification will allow to find terminological families in the text and to compare them with chosen term candidates in other languages on the basis of possible translations of head nouns. Most automated systems of term extraction use either statistical or linguistic approaches or their combination. These procedures are based on the frequencies of text lexical units, information on their compatibility preferences, likelihood ratio for two-component and multi-component terms, and other metrics. The hybrid approaches are of special interest, their use represents the attempt of overcoming the unilateral approaches constraints to terms extraction on the basis of both linguistic and statistical elements [Delpech, Daille, 2010].

Automatic extraction of multi-component terms provides a list of multi-component term candidates, which are then normalized according to their termhood degree, calculated by comparing summarized frequencies of collocation components, frequency of the collocation as a part of longer structures, and frequency of head word and collocation in national text corpora.

The list produced is evaluated by experts in actual data domain. As a rule, termhood rating approaches are based on a combination of linguistic and statistical information. Linguistic information is granted by grammatical tags in the text corpora, while a linguistic filter limits the type of the terms extracted and applies a list of stop-words.

- Linguistically, this method is supported by the following components:
- Information on the part of speech (PoS), provided by text corpora morphological tagging;
- Linguistic filter to exclude extraction of collocations prohibited (particularly, forbidden parts of speech combinations);
- List of stop-words.

Statistical part analyses collocation string frequencies of lexical units to define their status as term candidates. A hybrid approach supplements statistical information with special linguistic information. Selection of a specific linguistic filter depends on the way of balancing completeness and precision: preference of precision to completeness requires using a filter with a list of stop-words, while preference of completeness to precision demands using models.

Methods, used in statistical systems, vary from simple calculations of frequencies to calculations of complicated statistical indicators for measuring the connection force of collocation components in the term candidates chosen [TTC Project].

Building of a parallel or comparable text corpora for a certain language for special purposes and identification of complex of linguistic and statistical parameters represent correlated problems and find new terms and their translations.

5 Translation dictionary creation: text corpora as resource for user translation dictionaries

A specific language worker's professional competence is the competence in practical lexicography field, i.e. in the field of creation and management of user translation dictionaries. Today both language workers and specialists create a huge number of glossaries in various knowledge domains, these glossaries include collocations of different length and structure and in no way correlate with one another. With rapid growth of scientific and technical knowledge, accompanied by growth of new terminology, and dramatic backlog of specialized terminological and translation dictionaries practical lexicographic skills may seriously aid language work process.

Lexicographic research in the field of translation terminological dictionaries creation assumes preliminary work for screening and describing terms in different languages, these descriptions harmonization and co-ordination of terminological systems for various languages.

The process of translation dictionary creation includes the following main phases:

- creation of a term-pair list for the data domain or subdomain in question (terms extraction from parallel or comparable text corpora, their verification and description);
- ranking the term pairs in the framework of the term field under study (arrangement and analysis of the term system);
- normalization of the term pairs in relation to the target language (selection and approval, unification, optimization of the standard terms);

- term system codification (presentation as a standard dictionary, terminology standardization);
- terminology harmonization or cross-language ordering.

Unification of terms and term systems is one of the mainstreams for the applied terminology science, the basic aim of which is standardization, arrangement and harmonization of terminology on various levels of description and registration. The result of such unification is creation of terminological translation dictionaries, which are oriented on narrow data domains and languages for special purposes. Such dictionaries as a rule are bilingual, in Russia the main amount of translation dictionaries are those with English language is a source one, dictionaries with the source Russian language are not so wide-spread. On theoretical grounds bilingual translation dictionaries are concerned as one-way since the language lexical systems considered as a set of lexical units are asymmetric. Symmetry relation works only for the sets of nomens. In case of languages for specific purposes the variability of lexical unit translations sharply decreases and it is safe to say that introduction of local symmetry relation between sets of terms for two languages is justified. This local symmetry permits to develop terminological dictionary transformation methods. In this context the procedures for "turning a dictionary over" (in modern terminology for converting a dictionary) now become increasingly important [Egorova, 2015].

Consideration now will be given to terminology harmonization potential at creation of a source dictionary on the basis of specialized glossaries merging procedure, as well as to its conversion procedures and tools.

Terminology harmonization in the above-considered lexicographic work hierarchy is the final research phase, however when creating a translation dictionary due to information technology this phase coincides with the normalization phase. Discrepancy in term systems for different languages, among the other things, between the source language and the target language determines the necessity to determine and investigate the “source term - target equivalent” term pairs that permits to reveal the discrepancies in the term fields and systems for the data domain under consideration. Such discrepancies setting really permit to solve the problems of term translation and to sustain the efficiency cross-language communication [Belyaeva, 2000].

International standardization in special knowledge domains demands verification of term meanings, what's more, their accurate definition, which can be used in the dictionary structure taking into account specifics of the concepts nominated. A myriad of today glossaries, that reflect translators' and experts' ideas do not correlate with each other, the terminological collocations included have most different length and structure. Terminology management therefore becomes an important function of practical lexicographic and/or terminological work when creating problem-oriented translation dictionaries. It is especially important for terminology of high-risk areas (seismic protection, nuclear engineering, medical science and others). Hereafter is an example of terminology management based on lexicographic materials, published in seismic protection domain and a description of creating a problem-oriented translation dictionary for seismic safety texts.

In seismic safety, terminology harmonization is realized according to the International Standard requirements, namely Eurocode 8: Design of structures for earthquake resistance. Eurocode 8 is a European Standard for structural design in seismic zones, based on limit state design. The Standard was approved by the European Committee on standardization (CEN) on April, 23, 2004 and includes the major provisions and requirements of the previous standards (see table. 1). Table 1. Source-target terms nonconformity in seismic construction

domain glossary

In the process of terminology management the leading research and design institutes developed and published six dictionaries and glossaries. These glossaries, terminological dictionaries and building code represent the base for a research text corpus [Beliaeva, 2014], in which alignment is done on the lexical units, considered in these editions as terms, as well as on term definitions and translations. On the basis of such terminologically aligned corpus term candidates can be extracted and studied, and then, following special analysis, can be entered in a translation dictionary really harmonized by the described procedure.

The methodology applied to build the corpus included:

1. Building a parallel corpus of glossaries and terminological dictionaries materials. In the case described the corpus core was the dictionary developed by Intergovernmental Scientific and Technical Commission on Standardization, Technical Regulation and Conformity Evaluation in the construction domain. The dictionary material was supplemented with information from all other dictionaries of the domain and thus the basic aligned parallel corpus was built.

Even in the small fragment (table 1) there is a lexical unit *accidental eccentricity of the mass of one storey from its nominal location. Extracting term candidates at this stage may find collocations of the type accidental eccentricity of the mass of one storey from its nominal location* which refers to an important seismic safety aspect, but entering it in the translation dictionary as a term is absolutely inadequate. Analysis of this aligned parallel corpus shows, that such lexical units present large majority, and they require further analysis and detection of term candidates. For this purpose we built one more corpus.

Table 1. Source-target terms nonconformity in seismic construction domain glossary

English construction	Parallel Russian construction
Correlation factors to derive the pile resistance from ground investigation results, not being pile load tests	поправочные коэффициенты для оценки результатов испытаний свай статической нагрузкой
accidental eccentricity of the mass of one storey from its nominal location	случайный эксцентриситет массы одного этажа от его номинального положения
diameter of confined core in a circular column	диаметр бетонного сечения, замкнутого поперечной арматурой в круглой колонне
section factor [m-1] of the part i of the steel cross-section (non –protected member)	площадь обогреваемой поверхности на единицу длины
SPT blow count value normalized for overburden effects and for energy ratio	количество ударов при стандартных испытаниях на погружение, нормированное по эффектам слагающих пород и по коэффициенту использования энергии
endurance under stress	выносливость
geometric bow imperfection	искривление
	прогиб
key value of the stress	приведенное напряжение
leg members	пояса ствола мачты или башни
number of stress range cycles	количество циклов
upper shelf region	область разрушения образцов при ударном изгибе при температуре выше порога хладноломкости

2. Compiling a joint corpus on the basis of Russian and English lexical units and their machine translations comparison. Creation of integrated English-Russian dictionary index based on comparison of English and Russian lexical units and their machine translation make it possible to reveal English terms candidates by computerized extraction of simple NPs in the English terms and definitions.

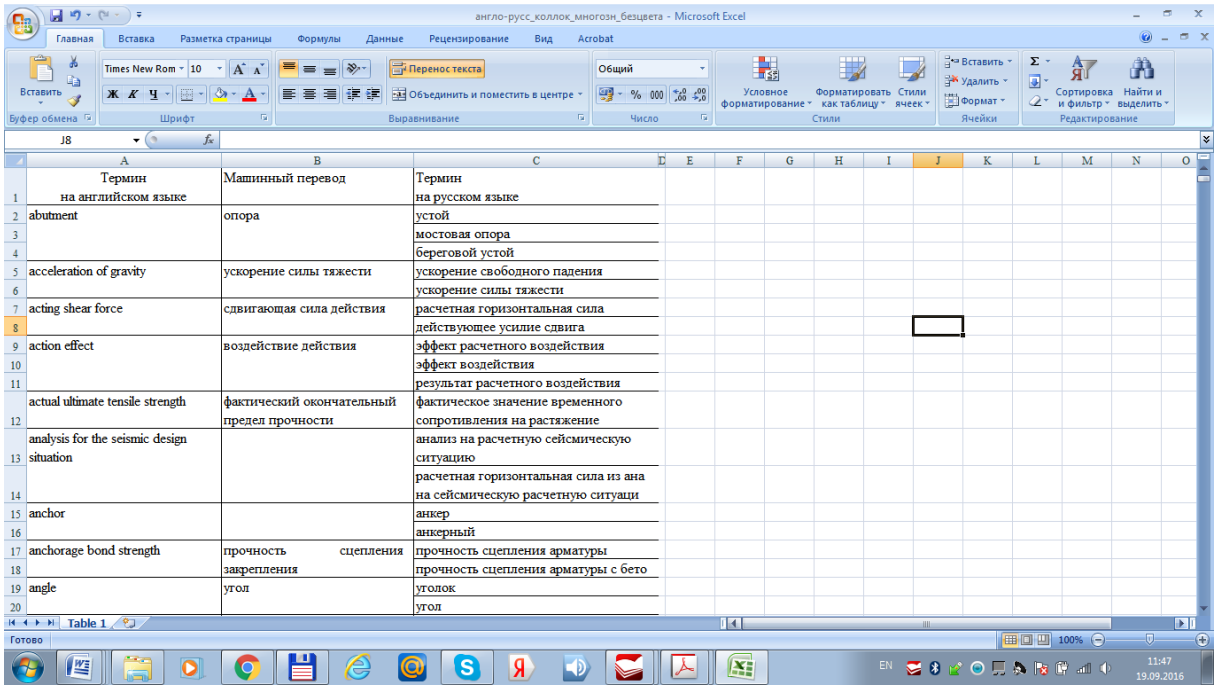
3. Compiling an English-Russian dictionary on the basis of the joint corpus (figure 2).

4. Creating a Russian-English dictionary on the basis of English-Russian dictionary conversion, with term candidates added in the form of simple NPs (figure 3).

5. Working with experts and verification of initial terms and their translations list, recording synonyms revealed and normalization of initial terms and translations.

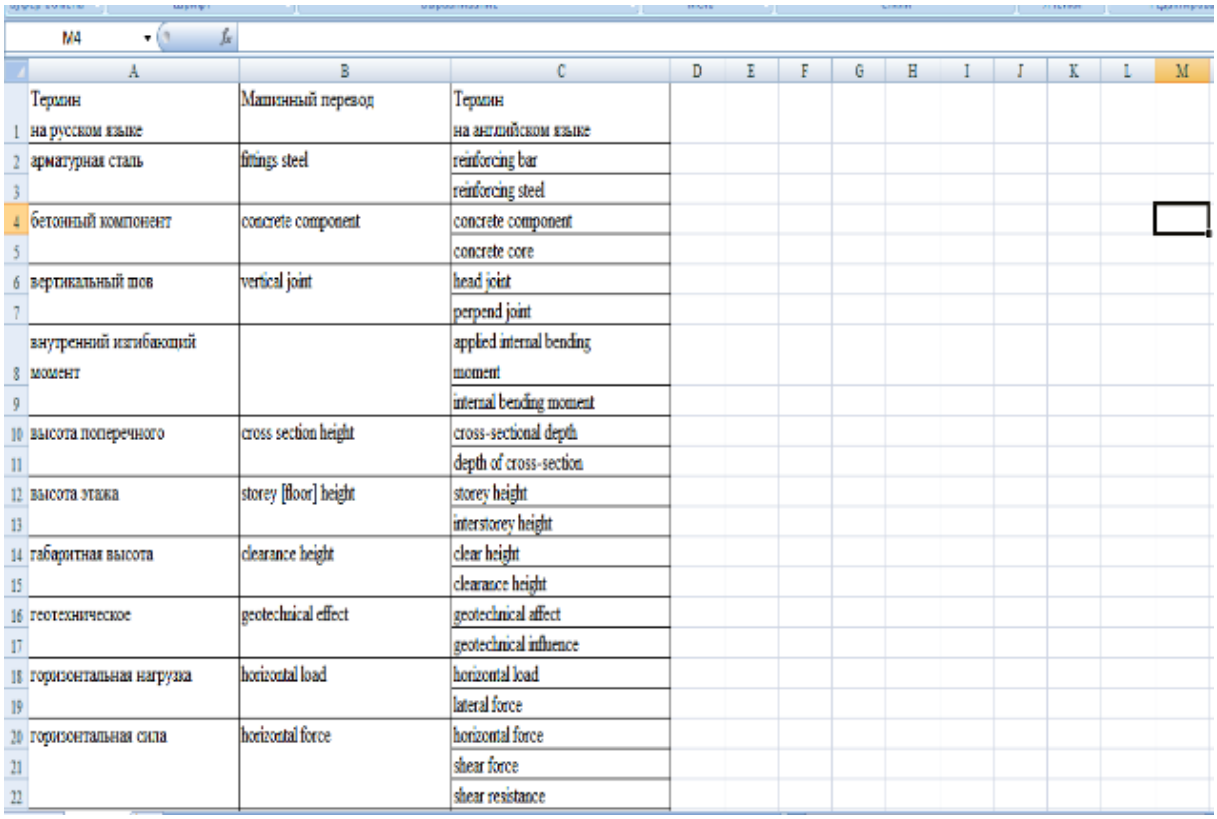
As the result of this terminology harmonization several leading research and design institututions in Russia (Intergovernmental Scientific and Technical Commission on standardization, technical regulation and conformity evaluation in construction, Ministry of Regional Development of Russian Federation, Building Construction Research and Development Institute) developed and published in 2011-2014 period the following dictionaries, glossaries and codes:

- Design and construction. Conceptual and terminological dictionary to Eurocodes EN 1992 EN 1996, EN 1998, EN 1999. Recommendations of National Constructors Association. (2014, 102 P.)
- English-Russian dictionary on building construction design (ranked according to numbers of Eurocodes). (2011, 35 P.)
- English-Russian dictionary on building construction design (in alphabetic order). (2011, 29 p.)



	A	B	C
1	Термин на английском языке	Машинный перевод	Термин на русском языке
2	abutment	опора	устой
3			мостовая опора
4			береговой устой
5	acceleration of gravity	ускорение силы тяжести	ускорение свободного падения
6			ускорение силы тяжести
7	acting shear force	сдвигающая сила действия	расчетная горизонтальная сила
8			действующее усилие сдвига
9	action effect	воздействие действия	эффект расчетного воздействия
10			эффект воздействия
11			результат расчетного воздействия
12	actual ultimate tensile strength	фактический окончательный предел прочности	фактическое значение временного сопротивления на растяжение
13	analysis for the seismic design situation		анализ на расчетную сейсмическую ситуацию
14			расчетная горизонтальная сила из ана на сейсмическую расчетную ситуаци
15	anchor		анкер
16			анкерный
17	anchorage bond strength	прочность сцепления закрепления	прочность сцепления арматуры
18			прочность сцепления арматуры с бето
19	angle	угол	уголок
20			угол

Figure 2: English-Russian joint dictionary (fragment)



	A	B	C
1	Термин на русском языке	Машинный перевод	Термин на английском языке
2	арматурная сталь	fittings steel	reinforcing bar
3			reinforcing steel
4	бетонный компонент	concrete component	concrete component
5			concrete core
6	вертикальный шов	vertical joint	head joint
7			perpend joint
8	внутренний изгибающий момент		applied internal bending moment
9			internal bending moment
10	высота поперечного	cross section height	cross-sectional depth
11			depth of cross-section
12	высота этажа	storey [floor] height	storey height
13			interstorey height
14	габаритная высота	clearance height	clear height
15			clearance height
16	геотехническое	geotechnical effect	geotechnical affect
17			geotechnical influence
18	горизонтальная нагрузка	horizontal load	horizontal load
19			lateral force
20	горизонтальная сила	horizontal force	horizontal force
21			shear force
22			shear resistance

Figure 3: Russian- English joint dictionary (fragment)

- Code of Practice. Antiseismic and Seismic Isolated Buildings. Design rules. Official Edition. (2014. 51 p.)
- Code of Practice. Construction in Seismic Areas. Design rules. Official Edition. (2014. 85 p.)
- Terminological dictionary for national norms and standards in the framework of Eurocodes. (2014. 199 p.)

The compiled parallel corpus aligned by lexical units considered as terms in the abovementioned editions and by term definitions and their translations provides the ground for term candidates extraction. Following a subsequent analysis, the candidates then can be entered in a really harmonized translation dictionary.

But for all that it shall be taken into account that even the most refined system of terms extraction does not give the final version of a translation dictionary and grants only conveniently organized and "on-the-fly" resource for a language worker. Translation dictionaries received following the suggested methodology with its merging and converting procedures are to be carefully edited by language workers and experts, but the volume of this work is not comparable with laboriousness of a translation dictionary manual creation.

6 Conclusions

In this paper we have suggests a number of initiatives, considering optimization of a language worker's professional space and use of computer working practices.

A promising way to organize and optimize a language worker's working space is an automated workstation (AWS). A regular automated workstation shall include a set of resident dictionaries, thesauri, spell-checking systems, systems of information access due to data communication networks. The idea of AWS for humanitarian issues introduced yet in the 90s for machine translation can be effectively applied for a general language worker professional space. We have considered an AWS for a language worker as a structured complex including relevant program and linguistic modules. An AWS can perform a complex of functions that organize a language worker's on-site activity, granting a huge range of integrated facilities for operations with multilanguage text preprocessing, symbols optical recognition, information compression and information extraction, document transfer, spell and grammar checking, terminology management, machine translation and translation memory management, production of user dictionaries, and access to local or remote data banks or other linguistic resources.

We have recommended a number of information technologies and Web resources, which potential considered through the lens of various language workers' practice may support and ensure efficiency, accuracy and correctness of the information product they develop. Among those are modern multilingual terminology databases, such as Databank Eurodicautom and EuroTermBank.

Much attention has been given to practical lexicography as a basic skill of terminologists, translators, technical writers and language workers in general. We have focused on a comparatively new approach to language resources creation which is based on corpus linguistics methods, namely formation and use of real text corpora, which can be considered as databases for solving not only research, but also practical lexicographic issues. The paper demonstrated the use of parallel and comparable corpora for terms extraction and translation dictionaries creation.

We have described briefly the technology of multilingual term extraction from comparable corpora and analyzed possible pitfalls of operating with comparable corpora, giving a linguistic commentary to assist a novice or practicing language worker in avoiding such pitfalls.

We insist that term extraction must be performed in accord with actual demands for termhood evaluation and harmonization and demonstrate a way to this in the process of translation dictionary creation, either for the user (personal resource) or general purposes (official resources). We marked the main phases of translation dictionary creation process and characterized a corpus-based methodology applied to build a translation dictionary.

We sincerely hope that the suggested pragmatic initiatives, such as optimization of a language worker's professional space and use of computer technologies in their work, alongside with linguistic analysis of interpreting linguistic facts relevant for text processing tasks (NPs length and structure, translated text quality and others) may be well applied both for training a language worker and for supporting a practicing specialist.

References

[Broeder et al., 2010] Broeder, D., Kemps-Snijders M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C. (2010) A Data Category Registry- and Component-based Metadata Framework. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Malta, 19-21 May, 2010, 43-47.

[Almazova et al. 2018] Almazova, N. I., Beliaeva, L. N., Kamshilova, O.N. (2018) Towards Textproductive Competences of a Language Worker and Novice Researcher. 18th PCSF 2018 - Professional Culture of the Specialist of the Future: The European Proceedings of Social Behavioural Sciences EpSBS. Volume LI, 103-109.

[Chernyavskaya et al., 2018] Chernyavskaya, V., Beliaeva, L., Nefedov, S. (2018) Language Worker in the Framework of Information 4.0. 18th PCSF 2018 – Professional Culture of the Specialist of the Future: The European Proceedings of Social Behavioural Sciences EpSBS. Volume LI, 608-614.

[Beliaeva, Kamshilova, 2018] Beliaeva, L. N., Kamshilova, O. N. (2018) Problems and Perspectives of Language Worker Professional Training. International Journal of Open Information Technologies. 6(12), 35-42. (In Rus.) = Problemy i perspektivy professional'noy podgotovki lingvotekhnologa. Avaible at http://injoit.org/index.php/j1/article/view/641/629

[Gordon 1997] Gordon I. (1997) Translation memory systems. Proceedings of the Eleventh Conference of the Institute of Translation and Interpreting. Ed. by C. Greensmith and M. Vandamme. 8-10 May 1997, Harrogate, ITI, London, 15-20.

[Belyaeva , 2009] Belyaeva, L. (2009) Scientific Text Corpora as a Lexicographic Source. SLOVKO 2009. NLP, Corpus Linguistics, Corpus Based Grammar Research, Proceedings from the International conference, 25 – 27, November 2009, Smolenice, Slovakia, 19-25.

[Belyaeva, Piotrowski, 2005] Belyaeva, L., Piotrowski, R. (2005) Linguistic Automaton. Quantitative Linguistics. An International Handbook. Berlin. N.Y.: Walter de Gruyter, 922-931.

[Coté, 1998] Coté, N. (1998) System Description. Demo of Alis Translation Solutions. Overview. Machine Translation and the Information Soup. Proceedings of Third Conference of the Association for Machine Translation in the Americas AMTA'98 Langhorne, PA, USA, 28–31 October, 1998, Volume 1529 of the series Lecture Notes in Computer Science. Ed. By Dr. Farwell. Berlin Heidelberg: Springer-Verlag, 494-497.

[Großjean, 2009] Großjean, A. (2009) Corporate Terminology Management: An approach in theory and practice. VDM Publishing. – 100 pp.

[Hutchins, 2001] Hutchins, J. (2001) Machine Translation and Human Translation : in Competition are in Complementation? Machine Translation: Theory Practice. New Delhi, 5-20.

[Hutchins, 1988] Hutchins, J. (1988) The Origins of the Translator's Workstation. Machine Translation, 13, 287-307.

[Johnson, Macphail, 2000] Johnson, I., Macphail, A. (2000) IATE – Inter-Agency Terminology Exchange: Development of a Single Central Terminology Database for the Institutions and Agencies of the European Union. Avaible at http://www.mt-archive.info/LREC-2000-Johnson.pdf

[Rirdance, Vasiljevs, 2006] Rirdance, S., Vasiljevs, A. (2006) Towards Consolidation of European Terminology Resources. Experience Recommendations from EuroTermBank Project. Riga, Tilde – 123 p.

[Maslias, 2014] Maslias, R. (2014) Combining EU Terminology with Communication and Ontology Research. Terminology and Knowledge Engineering. Berlin,19-21 June 2014, 48-56.

[Vasiljevs et al., 2014] Vasiljevs, A., Pinnis, M., Gornostay, T. (2014) Service model for semi-automatic generation of multilingual terminology resources Terminology and Knowledge Engineering. Berlin,19-21 June 2014, 67-76.

[Hertog et al., 2010] Hertog, de D., Heylen, K., Speelman, D., Kockaert, H. (2010) A Variational Linguistic Approach to Term Extraction. TKE 2010: Presenting terminology and knowledge engineering resources online: models and challenges, Dublin: Dublin City University, Ireland, 226249.

[Kageura, Umino, 1996] Kageura, K., Umino, B. 1996. Methods for Automatic Term Recognition: A Review. Terminology, 3(2), 259–289.

[Lefever et al., 2009] Lefever, E., Macken, L. and Hoste, V. (2009) 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, 496-504.

[Delpech, Daille, 2010] Delpech E., Daille B. (2010) Dealing with lexicon acquired from comparable corpora: validation and exchange. Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE), Fiontar, Dublin City University, 229-223.

[TTC Project] TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora. Avaible at http://www.ttc-project.eu/about-ttc/concept-and-objectives.

[Egorova, 2015] Egorova K. (2015) Editing an automatically-generated index with K Index Editing Tool. Proceedings of the fourth biennial conference on electronic lexicography, eLex 2015: Linking lexical data in the digital age, 11-13 August 2015, Sussex, United Kingdom, 268-280.

[Belyaeva, 2000] Belyaeva, L. (2000) Machine Translation Methods and Text Structure as a Source for Translation Competence Study. Across Languages and Cultures, 1(I), 85-96.

[Beliaeva, 2014] Beliaeva, L. (2014) Applied Lexicography and Scientific Text Corpora. Transcations on Business and Engineering Intelligent Applications. Ed. By Galina Setlak, Kassimir Markov. Rzeshov, Poland: ITHEA, 55-63.