General and Professional Education
3/2013 pp. 21-27
ISSN 2084-1469

ON GETTING THE MOST OUT OF INTERNET RESOURCES TO RAISE TRANSLATION QUALITY OF PROFESSIONAL DOCUMENTATION

Svetlana Sheremetyeva
Department of Linguistics
South Ural State University
76, Lenin pr. Chelyabinsk
454080 Russia
e-mail: linklana@yahoo.com

Abstract: The paper is devoted to the problem of improving translation quality of scientific and technical documentation. It considers existing translation Internet-resources and raises the issues of usability and reliability of open online dictionaries, translation memories and machine translation systems for translating professional texts. A methodology for the efficient use of Internet-resources for finding correct cross-language equivalents of professional lexica is developed. Possible applications of the suggested methodology are discussed. A procedure for the development of a multipurpose reliable professional e-dictionary based on the Internet translation resources is presented. The approach is illustrated on the example of the Russian-English language pair in the domain of mathematical modelling for which an e-lexicon is implemented following the methodology described in the paper. The suggested techniques can be useful for a wide audience of scientists, technicians and professional translators. It can also be included in the course for training translation students.

Key words: Internet resources, translation, special texts, terminology acquisition, professional e-lexicon

Introduction

The wealth of knowledge contained in scientific and technical documentation cannot be rated high enough. Now, when exploding volume of professional publications demand an operative international exchange of scientific and technical information the problem of correct documentation translation is especially important. The quality of text translation in many respects depends on the correct translation of used lexical units, which, for highly specific texts, can be rather problematic. Accelerating technical progress leads to the continuous emergence of new terms and /or changes in use and meanings of the general-language lexical units.

A translator who, as a rule, does not possess enough of expert knowledge in the scientific or technological domain of a professional text spends about 75% of time for translating terms, which do not guarantee correctness of translation equivalents. In the professional context the percentage of mistakes in translating terminology, as well as, sometimes, in translating general-use lexical units reaches 40% [1]. The situation is even worse when scientists or engineers themselves attempt to translate their works relying on dictionaries where most often they either cannot find the term they need or they cannot decide on which of the suggested variants to select, let alone, on how to choose and combine components when translating multi-component terms.

Hard-copy dictionaries – until recently the main translation resource, are bulky, demand a lot of time to search for translation equivalents, as a rule, suffer low coverage and do not reflect recent changes in the linguistic inventory of professional language communication.

With the advent of Internet technologies, electronic dictionaries, translation memories, machine translation systems, etc., more and more people engaged in translation turn to the

Internet in an attempt to reduce the amount of translation time and effort while raising the quality of translation. This puts in focus the problem of the quality of electronic translation resources and the problem of their applicability for translating texts in specific domains. There is also a problem of the user awareness of the types and levels of reliability of such resources.

On usability of Internet resources for translating special texts

There is currently a huge number of different types of translation resources on the Internet. For example, for the query "Dictionary" alone the Google search engine yields about 8 million results. In general, the Internet resources, which can be useful for different kinds of audiences, such as translators, students trained to become translators, lexicographers, scientists, engineers, etc., can be classified into the following main categories:

- unilingual and multilingual electronic dictionaries and other reference sources (e.g., encyclopaedias);
- translation memory systems;
- machine translation systems;
- auxiliary resources.

Unilingual and multilingual *electronic dictionaries* and all kinds of other reference sources can be further divided into two main types:

- electronic copies of existing hard-copy reference sources and
- reference sources (dictionaries, encyclopaedias, etc) initially developed in the electronic format.

Electronic copies of paper dictionaries and other reference sources normally exist in the .pdf or .djvu file formats and are provided with a search service [2]. Such dictionaries do not differ in content from their hard-copy versions but are easily accessible and operative.

Electronic dictionaries of the second type (we will further call them *e-lexicons*) most often include a lexical database and a set of tools for data management, - visualises, editors, defaulters, etc. They are more flexible than e-versions of hard-copy resources and quite often together with translation equivalents output some morphological-syntactic information. For example, the user can be informed on a lexeme number or a paradigm of its word forms, textual contexts, as well as a certain amount of semantic information, most often in the form of comments or domain attribution (mathematics, mechanics, medicine, economy, etc.). Some of the e-lexicons are multilingual and can output equivalents in several languages which is, undoubtedly, their advantage.

A pull of e-lexicons can be further divided into *closed e-lexicons* and *open e-lexicons*. The knowledge bases for closed e-lexicons are normally developed by professional lexicographers; they are more reliable and can only be updated by their developers. The former is an obvious advantage, the latter, of course, limits the coverage of such lexicons. The user should also be aware of possible exaggerated claims of lexicon developers, e.g., about perfect translations, good coverage, the number of languages, etc. As shows our experience such promises should not be completely trusted.

To overcome the problem of coverage *open e-lexicons* are developed for online use and update. The main feature of an open e-lexicon is that it can be updated by the users. The most popular open e-lexicons for the Russian-English language pair are ABBYY Lingvo [3] and MultiTran [4]. The good thing about these lexicons is that they are capable of shallow parsing and in case a full translation of a multicomponent lexical unit is not found, component translations are output, which can still be a great help for the user. Ever rising lexicon coverage while more and more lexical units are being put in the lexicon by the users is obviously an advantage of open e-lexicons.

However, “the more, the better” does not always work. A huge number of user-supplied translation variants and their attribution to the user-selected domains can make e-lexicon navigation quite difficult and time consuming. Large open dictionaries can output too many translation variants for a lexeme and (or) its components, so that numerous accompanying contexts do not always help. The user thus faces the problem of selecting a correct equivalent among a lot of suggestions, the problem which she/he might not always be able to cope with.

For example, in the English-Russian scientific and technical e-lexicon which is a part of the electronic ABBYY Lingvo lexicon, the Russian term «*роевое представление частицы*» is absent, but given are the following equivalents of the term components: *«рой» – «swarm», «представление» – «conception», «expression», «representation», «частица» – «bit», «fraction», «particle».* The Multitran

dictionary offers the following variants for the same Russian term components: *«рой» – «swarm», «представление» – «performance», «configuration», «частица» – «shard», «corpuscle».* Based on these e-lexicons outputs even professional translators, let alone, other types of lexicon users will most probably have problems in deciding on a correct English translation for the Russian term.

And last, but not least, we have to mention here the main disadvantage of open e-lexicons, - their lower reliability as compared to hard-copy or professionally done closed e-lexicons. Well-intended users who are not necessarily lexicographers or translators can assign a wrong foreign equivalent or attribute a lexical unit to an inappropriate domain.

Translation Memory systems, for example, the well-known Trados [5] or Across [6], are databases of parallel "source language – target language" text segments. The segment can consist of a single word, group of words, sentence, or even a whole paragraph. When translation memory is run on a source language text it tries to automatically find text segments that match entries in the tool database. In case a match is found, a corresponding target language segment is offered to the user for approval. The user can accept, reject or edit the suggested translation. The user-edited translation variant can be put in the Translation Memory database for further use. In case of no match, the user is supposed to translate the segment her/himself and put in the database. Translation Memory systems are effective when translating similarly worded texts in very restricted domains, for example, maintenance instructions for a certain type of the equipment that contain a lot of repeated segments. In translating different text even in the same broader domain Translation Memory systems are ineffective and can hardly be recommended.

Machine Translation systems are getting more and more popular among lay-off people and translators. Though not perfect, their performance has greatly improved during the last decade. For the Russian-English pair of languages the system, which outputs the best translation results is a well known PROMT system [7].

However, the user should be warned against relying on free online machine translation systems when translating special texts, e.g., scientific or technical papers. The problem is that working quite satisfactory for the general use language such systems still cannot cope well enough with syntax and terminology of special domain texts. For example, PROMT translates the Russian term of cited earlier «*роевое представление частицы*» into English as *«royevy representation of a particle*». Such an output means that the Russian term "*роевое*" is absent from the system lexicon (not covered) and the word is simply transliterated, which can actually be confusing. Translation for another Russian term from the mathematical modelling domain "«далекие младшие разряды» is output by the same machine translation system as «far younger categories" that is completely wrong.

There exist machine translation systems that are initially developed for special domains but there are not so many of them and they still suffer the coverage problem. Even among paid systems it is impossible to find a system suitable for every particular type of professional texts.

The auxiliary resources include unilingual or multilingual (parallel or comparable) corpora and different tools to process them.

Parallel texts (corpora) are collections of texts of identical contents in different languages. Comparable or quasi-parallel texts are texts in different languages that have similar, but not completely identical contents. For example, comparable corpora are collections of scientific articles in different languages on one subject, that, however, are not translations of one another. Parallel or comparable corpora can be used by translators to compensate for the insufficient coverage of dictionaries. Unilingual special corpora in the source language can be used by a translator to correctly define the meaning of the language unit (in case of its ambiguity) by examining the context. A parallel (comparable) corpus can be used to find a correct foreign equivalent of the source lexeme to be translated, once its meaning is clear.

However, the user engaged in translation should be aware that, firstly, it is often impossible to find online a ready-to-use parallel (comparable) corpora of a particular scientific/technical domain, especially for the Russian-English language pair and, secondly, the process of effective search of translation equivalents in parallel (comparable) texts is far from being trivial and demands both a good knowledge of foreign languages, and special training. Effective application of corpus mining for

interlingual equivalents often requires special auxiliary tools.

Auxiliary tools, such as n-gram calculators, frequency sorters, concordances, extractors of typed lexical units, interlingual aligners, etc., to mention just a few, are, as a rule, intended for lexicographers and trained developers of electronic dictionaries, translation memories, and machine translation systems. They are used, in particular, for the automated lexical acquisition for the databases of uni- and multilingual lexicons. If not used properly and not checked by a human translator, lexical cross-language equivalents can be erroneous.

What follows from the discussion above is that, the current translation Internet resources though having a great advantage of availability and operativeness are still far from being completely reliable and require special skills to get the most out of what they can offer. We have developed and tested a methodology for the efficient use of the Internet resources for lexical acquisition in special domains, which is described in the next section.

A methodology for the efficient use of Internet in translating special texts

Nowadays a translator not using computer tools and Internet resources to increase the efficiency of his/her work cannot be competitive on a labour market. However, it is not everybody, who realizes that, on the one hand, a human sill plays the major role in the translation process, and, on the other hand, the Internet resources should be used with a certain caution, especially when translating highly specialized texts or lexica.

As was shown in the previous section the following should be kept in mind when searching for the cross-lingual terminological equivalents.

- It is not recommended to use free on-line machine translation systems as they are normally tuned to the general-use language and cover neither terminology, nor complex syntactic structures of special domain texts.
- On-line dictionaries of open type do not guarantee correct translation equivalents as those dictionaries entries might have been created by the users who are not professional translators or lexicographers.
- Translation should start with first checking electronic copies of domain-specific hard copy specialized lexicons created by professional lexicographers. In case you can find the term or expression you need (which in many cases can unfortunately be problematic) it should be preferred over those from open online lexicons.
- The most reliable sources of up-to-date correct interlingual equivalents are on-line corpora of parallel and comparable texts. It is important to remember that foreign-language parallel (comparable) texts have to be written by native speakers or should be printed in magazines where translation is proofread by native speaker translators and professional editors. Translated texts that do not meet these requirements cannot guarantee complete correctness.

When selecting a translation variant among several translation suggestions for a term as a whole or for its individual components it is necessary to make sure that the meaning of a lexical unit is understood correctly. The best way to do it is to check the context of the term in question in the source language document and monolingual e-reference books (e.g., a corresponding domain encyclopaedia that contains term definitions).

The term meaning understood, in case it is not included in any lexicon and a parallel text is not available a decision can be made to use a comparable corpus (that can consist of just one text). It is important to make sure that a comparable foreign-language text (corpus) really corresponds to the domain in question. It is recommended to find on the Internet native speaker papers listed in the references of the source document.

If no reliable full term translation can be fount in lexicons and neither parallel, nor comparable bilingual corpora are available, the following algorithm of actions can be recommended.

- Translate term components using any available dictionaries/lexicons.
- Make translation hypothesis for the whole term translation using different combinations of component translations.
- Use your translation hypothesis as keywords in an Internet search engine in the target language. The search results will help to determine the correct wording of the term translation.

To illustrate this procedure let us apply it to our example Russian term «*роевое представление частицы*». In the previous section we have shown the inefficiency of the direct use of on-

line resources to get a correct English translation of this term. Machine translation suffered a coverage problem and was useless. The English-Russian scientific and technical e-lexicons ABBYY Lingvo and Multitran lexicon, do not contain a full translation of the Russian term «*роевое представление частицы*» but they give the following English equivalents of the Russian term components:
«рой» – «swarm»
«представление»–
«conception»,«expression»,«representation», «performance», «configuration»
«частица» – «bit», «fraction», «particle», «shard», «corpuscle».
If you create a translation hypothesis by using the first suggested translation for every component of the Russian term you will get: *«swarm conception of a bit ».* If we use this hypothesis as key words in Google, the search results will not contain these words combined in a term. This means that this translation hypothesis should be rejected. Another hypothesis *«swarm representation of a particle»* used as key words in Google immediately gives, for example, the English term *«***Particle Swarm** *Optimization and Priority* **Representation***»* from the paper by Philip Brooks. This shows that the term is used in an English native speaker paper related to mathematical modelling and its wording can be trusted. By analogy it is safe to consider «***particle swarm representation***» as a correct English translation of the Russian term «*роевое представление частицы*»**.**
Acquisition of reliable cross-linguistic equivalents on a large scale requires professional linguistic/lexicographic knowledge and acquaintance with modern approaches and techniques of text mining. This work is very effort and time consuming to be done on a regular basis by a translator, let alone, other categories of translation users. Therefore specialised e-lexicons where the users (e.g., translators, researchers, engineers, etc.,) could easily find required lexical inventory of the domain of their activity are very much in demand. In what follows we try to contribute to the problem by showing how the methodology of using Internet resources described above can be applied to the development of real-life professional e-lexicons.

Development of reliable e-lexicons for special domains

In this section we suggest a methodology to develop contemporary electronic multilingual dictionaries for highly specialized texts based on the Internet resources. We describe its realization on the example of the development of a bilingual Russian-English electronic lexicon for the domain of mathematical modelling.
At the first stage of lexicon development the purpose, domain and users of the lexicon should be defined. Our purpose was to create an electronic resource to support translation of scientific papers on mathematical modelling from Russian into English. Our intention was to implement the lexicon as a multipurpose tool that can easily be used directly by humans of the different levels of English proficiency and as a lexicographic module for automated text processing systems. In this paper we concentrate on the human-oriented features of the e-lexicon.
Quite a number of translation mistakes can be made in selecting grammar forms of foreign verbs, especially by people with restricted foreign language proficiency. We therefore designed the e-lexicon so as to output verb translations in the text grammar form (keeping the voice, time, number, person grammar features) rather than a base form. English equivalents of other lexical units are output their base form. We further augmented the human oriented functionalities of the e-lexicon with a special feature that allows the user (in addition to getting translations of query terms in a usual way) processing whole paper texts and instantly getting English equivalents of all single- and multi-component lexical units used in the paper. This feature saves a lot of translation time that would otherwise be wasted on the search of every individual term.
At the second stage the lexicon knowledge base model that could provide for the desired lexicon features should be defined. It is necessary to decide on the types and amount of linguistic information for a lexeme to be included in the lexicon, as well as the formalism to represent this information in the lexical entry. To meet the requirement described above our bilingual Russian-English lexicon is built as a set of cross-referenced set of monolingual entries. The structure of the Russian entries and their

English equivalents are identical and to provide for the human oriented output as specified above they contain the following zones:
ZONE 1: part of speech
ZONE 2: morphological features, such as number, gender, etc.
ZONE 3: explicitly listed domain specific wordforms.
The verb lexicon entries contain additional zones with syntactic and semantic information, but they are only needed when the lexicon is used as a module in a larger text processing system (e.g., machine translation) and therefore the description of these zones is beyond the scope of the current paper.

The third stage of the e-lexicon development consists in deciding on the content, source and the methodology of the acquisition of the source language (Russian, in our case) vocabulary. Our translation experience shows that it is not only the terminology that might cause problems in translation. Professional texts contain a lot of commonly used lexical units, introductory or other phrases that require special attention. We therefore together with the terminology cover other types of lexica used in mathematical modelling domain. Following our purpose to reduce translation problems as much as possible, not neglecting single word lexemes we put in focus multicomponent lexical units (up to 8 components). In creating our e-lexicon we used the most contemporary corpus-based approach. The Russian vocabulary was created in two steps.

First, an initial corpus of Russian scientific papers on mathematical modelling of approximately 80 000 wordforms was acquired on Internet. From this corpus domain specific typed lexical units (NPs, VPs, ADJs, etc) consisting of 1 up to 4 components were extracted automatically with the help of a lexical extractor. For this purpose we ported the English extraction tool LanA-Key [8] to the Russian language. The automatically extracted lists of lexemes were further checked by human acquirers and used as a seed lexicon.

Second, the seed lexicon was used to acquire more and longer lexemes both from the pre-constructed corpus, and the Internet, which is in fact an unlimited corpus. This procedure of extending the seed vocabulary with the help of the Internet is one of the specific features and know-how of our acquisition methodology. The seed list of the Russian terminological units extracted from the initial corpus was used as keywords in the Internet search engines. New terms were selected from the two first pages of the search results. For example, for the seed (key) term *«псевдообращение»* the following multi-component terms of mathematical modeling domain were found on the Internet: *«псевдообращение сопряженной системы», «псевдообращение матриц с вырожденными весами», «псевдообращение Мура-Пенроуза»,* etc. As a result, the Russian vocabulary was extended up to 60 000 single- and multi-component units up to seven-eight words long.

At the fourth stage of lexical acquisition the sources of English equivalents for the created Russian vocabulary are to be decided upon and the English equivalents proper for each unit of the Russian vocabulary are to be found. For this purpose we used both specialized hard copy dictionaries, and the Internet translation resources following the methodology described in the previous chapter.

Special attention was paid to facilitating the problem of selecting a correct translation equivalent from several possible ones. It was achieved, first of all, by tuning the e-lexicon to a rather specific domain of mathematical modeling and by focusing on the multi-component lexical units that cover the main part of the e-lexicon vocabulary. Such units are mostly unambiguous. In case a Russian lexical unit has several synonymous domain-relevant English translations only one, the most frequent equivalent, was included in the lexicon. The Russian lexemes that have several meanings in the domain are included in the lexicon (and output to the user) in an unambiguous context. For example, the ambiguous word *«решение» («solution», «decision»)* is included in such entries as *«принимать решение»* (*«make decision»*), *«находить решение»* (*«find solution»*), etc.

The fifth stage includes the specification and programme implementation of the lexicon. The electronic shell of our lexicon is created by the reuse and adaptation of the TransDict [9] programme. It includes the module of the Russian-English linguistic knowledge base as specified above and the interfaces for the end user and the lexicon acquirer/developer. Interfaces are supplied with effective services to facilitate the user and developer activities. The developer interface, or example, includes morphological generators both for Russian [10], and for English (ported from TransDict) to

automatically fill out Zone 3 in the lexicon knowledge base.

The sixth stage – filling out the e-lexicon knowledge base (Russian-English entries) in the required format by means of the developer interface. At the time of the preparation of the current article the Russian-English electronic lexicon described in this section contains 60 000 entries.

Conclusions

In this article the translation potential of modern Internet resources is analyzed and a methodology to efficiently use these resources is suggested. The main emphasis is placed on the problem of identifying correct translation equivalents for professional lexica with the help of the Internet search engines. A technique to use internet resources for creating professional e-lexicons and a development procedure for such lexicons are described on the example of the Russian-English e-lexicon for the domain of mathematical modeling. The suggested techniques can be useful for a wide audience of scientists, technicians and professional translators. It can also be included in the course for training translation students. This research and development has been financed by the Russian Federation grant for scientific research in the frame of state project 01201262684.

References

1. Кудашев, И. С. Проектирование переводческих словарей специальной лексики [Текст] / И.С. Кудашев. – Helsinki University Print, 2007. – 445 с.
2. Мюллер, В.К. Англо-русский словарь, [Электронный ресурс], http://www.twirpx.com/file/486488/ (Дата обращения 4.10.2013).
3. Онлайн-словарь ABBYY Lingvo. [Электронный ресурс], Pro http://lingvopro.abbyyonline.com/ru (Дата обращения 24.09.2013).
4. Словарь Мультитран [Электронный ресурс], http://www.multitran.ru (Дата обращения 4.10.2013).
5. Trados [Электронный ресурс], http://www.translationzone.com/trados.html (Дата обращения 24.09.2013).
6. Across Personal Edition [Электронный ресурс], http://www.my-across.net/ (Дата обращения 25.09.2013).
7. Онлайн-переводчик [Электронный ресурс], PROMT http://www.translate.ru (Дата обращения 25.09.2013).
8. Sheremetyeva, S. Automatic Extraction of Linguistic Resources in Multiple Languages. Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, Wroclaw, Poland, 2012, pp.44-52.
9. Sheremetyeva, S. Application Adaptive Electronic Dictionary with Intelligent Interface. Proceedings of the workshop on Enhancing and using electronic dictionaries in conjunction with the 20th International Conference on Computational Linguistics. COLING 2004, Geneva, Switzerland, August, pp 23-28.
10. Babina, O. Modes of Automating Lexicon Compilation as a Component of Practical Studies for Linguists // General and Professional Education. 2012. No. 2, pp. 3-12.