Improvements of Dictionaries – Suggestions by Evroterm

Miran Željko
Secretariat General of the Republic of Slovenia, Translation Division,
Gregorčičeva 27A, SI-1000 Ljubljana, Slovenia
miran.zeljko@gmail.com

Summary

The paper presents some possibilities for improving electronic dictionaries from a translator's point of view. Dictionaries, glossaries, terminology databases and corpora are a translator's basic tools. The existing Slovenian electronic dictionaries are based on book dictionaries – data from the books were transformed into computer software using the path of least resistance. However, electronic dictionaries can provide more functions than books, e.g.: full-text search, fuzzy search, terminological analysis, corpus as a source of collocations, dynamically linked dictionary and corpus, and continuous improvement of a dictionary instead of new dictionary projects every few decades. The Evroterm terminology database is presented as a practical example of the proposed improvements.

Key words: terminology database, on-line dictionary, corpus, continuous improvement, full-text search, terminology analyser, Evroterm

1. Introduction

Dictionaries in Slovenia are generally first published as books and subsequently transformed into electronic form. The newest and the most extensive English-Slovene dictionary (Krek, 2005-2006) has not yet reached this stage at the time of writing this paper (http://slovarji.dzs.si/dokumenti/Dokument.asp?id=2).

In the future, it will be necessary to change this procedure completely: a book and software are two completely different products and should be designed and made separately. Dictionaries in books are useful for bibliophiles, while translators need electronic dictionaries and during the development of such dictionaries it is necessary to make full use of IT capabilities.

During the development of the Evroterm terminology database (term base) we added some features to it that make it much more useful than ordinary term bases. I believe that at least some of these features (if not all) could also be applied to dictionaries.

2. Improving dictionaries[1]

2.1. Full-text search

The first electronic dictionaries were books transformed into electronic form, e.g., the contents of *Veliki angleško-slovenski slovar* (Grad 1997) on CD is the same as the book under that title. The main advantage (together with some minor additions) is a faster search because the user does not need to turn pages. What is missing is the most useful improvement: instead of the search being limited to English headwords, a search of Slovene translations of words and collocations could be added.

2.2. Fuzzy search

We occasionally make errors when writing – sometimes because of mistyping and sometimes because we have the wrong spelling of a particular word in our mind. In a word processor, a spell checker gives a warning when it encounters an error. When an electronic dictionary does not find a word typed in, it would be user-friendly to show hits similar to the search word.

2.3. Corpora

A dictionary and a corpus seem two completely different products: words in a dictionary are sorted alphabetically, while a corpus is a disordered collection of data and the user gets some kind of a sorted output only when the search results are listed. However, a dictionary and a corpus are much more similar than they seem at the first glance. Suppose we have a glossary as the simplest form of dictionary and, in this glossary, one word in the source language (SL) corresponds to one word in the target language (TL) – then this is the simplest form of a corpus. On the other hand, in a large enough bilingual corpus we could find all the words from the glossary, the only obstacle being that we would have to find the mapping between the words (there is more on this topic in Vintar, 2003); a corpus can therefore be regarded as a sort of glossary with a large amount of noise.

Dictionaries and corpora are usually treated as two separate entities. On the web, e.g., there are *Slovar slovenskega knjižnega jezika* (SSKJ – Dictionary of Slovene Literary Language) and the *Nova beseda* (New Word) corpus, which incorporates Slovene literature. It seems natural that the software would list links to examples from Slovene literature when presenting a list of hits from SSKJ; the user would thus see how a particular word is used in the literary language. However, on-line and CD versions of SSKJ are the same as the book form.

The number of examples in a book is limited due to the nature of the medium: depending on paper size and thickness, it is possible to use a book if it contains

[1] In this paper I use the term “dictionary” for dictionaries and similar tools (e.g. terminology databases).

up to about 2000 pages. If a dictionary is too voluminous, it becomes too heavy. This problem can be overcome by publishing a dictionary in several volumes but in practice, we quickly hit a limit. In computer media, this limitation is higher by several orders of magnitude: e.g. (Grad 1997) contains almost 1400 pages with 5000 to 6000 characters on each page. By multiplying these numbers we get between 7 and 8.4 million of characters – which is just 1% of a CD capacity! And a CD is an old-fashioned medium by today's standards.

One of the basic arts of making a dictionary is therefore the selection of suitable examples of use (more on this in Drstvenšek 2003). Several problems can be encountered during this process:

- each author has limited knowledge, so some examples are missing in the dictionary (mostly newer collocations);
- the author may have wanted to prove some hypothesis and thus only selected examples that support his ideas;
- authors of dictionaries are usually people with several decades of experience – so dictionaries may contain words and collocations that are rarely used in modern texts.

These problems can be overcome if a corpus is compiled and a dictionary is designed so that the software searches the corpus directly. If the corpus data are not deliberately biased, the user should obtain the actual data on word use. There are errors in corpora due to the large volume of data but it is usually possible to find a rule from a larger set of hits.

2.4. Corpus linked with a dictionary

In book dictionaries, examples of use are static (there is no other possibility). In an electronic dictionary, it makes sense to create a dynamic link between a headword and examples of use; the link is established through search.

From a translator's point of view, a simplified expression of this is that a bilingual dictionary entry consists of three parts (Figure 1):

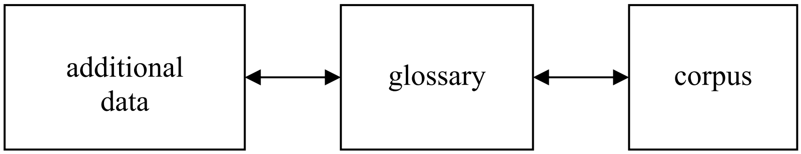
- glossary (headword in SL and TL);
- additional information about headword (depending on the part of speech, language, volume of a dictionary and target users); data that should be in a term base are listed in the ISO 12616 standard;
- examples of use – the simplest way is to use a corpus for this purpose; additional data may also be listed here, e.g., reliability of translation, subject, source, full text containing this word/term, etc.

Two search modes exist in such a system:

1. glossary → additional data → corpus: the user usually does not need all the data; so it makes sense for information to be displayed to him in layers: after making a search he first gets a list of headwords from the glossary. By clicking a particular item he obtains additional data, and with the next click he obtains examples of use from the corpus. The output de-

pends on the dictionary design (more complex operations can be performed in a dictionary that is installed on the user's PC than in an on-line dictionary), purpose of the dictionary, volume of data, etc.

Figure 1: dictionary elements and their relations



2. glossary → corpus → additional data: any dictionary contains a limited volume of words: a general dictionary contains too few technical terms and general terms are missing in a technical dictionary, so it always happens that some terms cannot be found in a dictionary. Some terms probably exist in a corpus of adequate size, so it makes sense to search in another direction. Again, the data are presented to the user in layers: the inputted word is first searched in the glossary and subsequently in the corpus. The user will find the meaning of the unknown word from its vicinity. If the search word exists in the glossary, the user should be able to obtain additional data from it.

From a translator's perspective, a dictionary linked with corpus has the following advantages:

- more search options;
- higher probability that the meaning of a word will be found;
- more data on the search word;
- faster and simpler search.

The disadvantage of this approach is the use of a corpus. A good corpus contains several tens of millions of words. It is not possible to check all of the corpus data because this would be too expensive, so corpora as a rule contain more errors than dictionaries. Users must be aware that if a particular hit deviates from the rest, it is possible that it is not an exception but an error and, in such a case, the data should be checked in another source.

An additional benefit can be obtained by using a multilingual corpus – if the meaning is not clear from a translation from one language into another, then an additional language may help in clarifying the term.

2.5. Terminological analysis of a text

A search by entering one word/term into the entry field originates from a search in a book. It is useful if the user wants to find meanings of just a few terms. The real life of a technical translator is entirely different: a translator often gets the following instructions: “When translating the text, use the terminology from our glossary”. If the glossary contains several hundred (or even thousand) terms that are continuously updated and supplied by various translators, it is impossible to know which terms are in the glossary. In such cases, computer-assisted translation tools offer basic help, (e.g., “Translate terms” in SDL Trados TWB). However, the translator needs more: the software should analyse the original text, mark the terms that exist in the term base and, by clicking them, the user should obtain corresponding terminology and corpus data.

2.6. Presentation on a screen

A dictionary or corpus search usually produces a large volume of output. It is necessary to put these data in order and to present them so that the user quickly finds what he is interested in. In addition to a sensible layout, colours are very helpful: various data should be in various colours, less important data should be plain black. The aid of colours when using a dictionary can be seen from Amebis dictionaries.

Corpus output is more user friendly if individual units are clearly separated from each other and if the search word is coloured (if it is just bold (as in the SVEZ-IJS corpus) it is more difficult to find it on a screen). In a parallel bilingual corpus, it is easier to find the word in SL and TL if the two segments are parallel to each other. In a sequential output, visual aligning takes longer because the eye has to cover larger distances (this can be seen if the SVEZ-IJS corpus is compared with Evrokorpus). A coloured translation of the search word (if found in the glossary) is an additional help to the user.

2.7. Continuous improvement

In the past, a group of people made a dictionary. It was re-printed several times and it was on sale without changes for several years or even decades.

A different approach can be used in electronic dictionaries: the basic version of a dictionary is made as before. Every dictionary has errors and deficiencies, no matter how much effort has been put into its production. Large expenses are associated with changes of a book dictionary. In electronic dictionaries, the problems are solved by the very nature of the media: the cost of a CD is much lower than the book-printing cost; nevertheless, a problem remains because there are several versions of the dictionary on the market. Everything can be simplified by using the Internet: if a dictionary is on the web, the data only have to be updated at one location and each user has access to the newest version. People do not carry Internet-connected computers with them all the time, so it is good to make the dictionary accessible to mobile-phone users, too.

Manufacturing companies have been using the Deming principle of continuous improvements for several decades and this idea can also be applied to the development of dictionaries.

2.7.1. Improvement of contents

Updating consists of several tasks:

- correction of errors;
- adding new headwords;
- adding new meanings to existing headwords;
- marking or removing obsolete terms.

Users of dictionary will point out the most obvious errors. The question remains, which headwords to add to an existing dictionary.

One possibility is to use as large a corpus as possible, calculate word frequency and add the most frequent words. Lönneker 2004 suggests that a corpus of literary works should be used as a source of new terms with this approach. What about technical terms?

Another possibility is even simpler: those words are added that users did not find in the existing dictionary. Jakopin 2004 suggests that a web server's log file could be analysed for this purpose. This may be difficult to do on servers with heavy traffic, because log files grow very quickly. A better solution is for the search program to write unknown words into a special file. The most frequent words from this list are the first candidates for addition to the dictionary.

It rarely occurs to us that dictionary data should be reviewed in order to find and remove obsolete words; this subject is covered, e.g., in Brookes 2004.

2.7.2. Technical improvements

Most of the routine update procedures (conversion of data, transfer of data between servers, statistical processing) can be automated. Update frequency depends on the volume of new or changed data within a time unit; an update can be performed monthly, weekly or even daily. In addition to changes to the dictionary contents, software features can also be changed. New functions are available to all users from the moment they are implemented.

It is true that this imposes additional costs. However, the value of the dictionary is much higher because it always contains up-to-date information. High starting costs arise only with the initial preparation of the dictionary.

2.8. Copyright

Much more work is required to compile a new dictionary than to write a novel, so the price of the first is much higher. Because of this, the problem of unauthorised copying arises more often. Dictionaries on CDs are protected in such a way that they can only be installed on one disk, but this protection can often be overcome. And if we stick to the publisher's rules, we may encounter other types of problems:

- suppose I have a desktop PC and a notebook, but I use only one at a time: with this type of protection I need two licenses;

- suppose my hard disk (with the dictionary installed) breaks down and the data cannot be restored;
- or even worse: suppose my PC gets stolen.

In the latter two cases, it is probably possible to obtain another CD with proof of purchase, but some time is lost for this operation and the user will be without a dictionary for some days. All these problems occur because the license is attributed to a PC instead to a person.

If the publisher does not publish a dictionary on CD and stores everything on a web server instead, there appears to be even less protection (protection by username/password is not serious protection because people share passwords).

However, there is professional protection available; banks use it for on-line access by their customers, and government uses it for communication with citizens when transferring sensitive data (e.g., tax data): a digital certificate.

A simplified description of digital certificate protection: a company that wants to limit access to its dictionary must obtain a digital certificate for its server, while a user of the dictionary (client) must obtain a digital certificate for his browser. The client pays a yearly subscription, which is much cheaper than the cost of a dictionary, and he then has on-line access to the dictionary for a specified period.

It is possible that users would share digital certificates but this possibility is rather theoretical, because a user of a borrowed bank certificate would have access to all on-line bank services, and the user of a borrowed government certificate would have access to all personal government-related data. I believe that the volume of false-identity-dictionary-access frauds would be much lower than the volume of unauthorised copied CDs.

There are several advantages when transferring a dictionary to the web and using digital certificates for access protection:

- the publisher of the dictionary maintains data and software in one location;
- all users have access to the most recent dictionary version;
- there is no more production and distribution of CDs;
- dictionary-use license is limited to a person, not to a PC. If the user has several devices and uses only one at a time (e.g., at home, in job, notebook, mobile phone), he can legally access the dictionary from any of them. If he has a copy of the digital certificate, there is no problem if he has to change the PC;
- the user has to pay a much lower initial charge than when buying a CD, so there are more potential customers.

The disadvantage of this approach is that the dictionary is accessible only online; but more and more people have full-time Internet access today, which makes this solution ever more applicable.

3. A practical example: Evroterm

A term base that uses the improvements mentioned in section 2 (with the exception of limited access) is Evroterm combined with Evrokorpus and Terminator (terminology analyser).

The term base contains terms in 15 languages (the emphasis is on English and Slovene terms – there are more than 100,000 terms in these languages).

The corpus side of the database consists of several corpora: there are five bilingual corpora (English, French, German, Italian and Spanish paired with Slovene as the second language) and one 22-lingual corpus.

The Eur-Lex database is used for access to full-text data.

Modern software has many functions (and many of them are never used), so some functions are hidden deeply in the menu system. Google made an important improvement in this field: with the exception of some lines above and below the entry field the screen is practically empty. On the other hand, terminology experts need additional functions in order to limit the volume of output. It is therefore possible to use either simple or advanced search in Evroterm and Evrokorpus.

In simple search, the user enters the search term (a word, part of a word or several words that can be combined with wildcards) into the entry field and clicks the search button. As a result, he obtains a list of hits in all languages. If there are no hits, the program writes a warning and switches to fuzzy search. If there are no hits even in this case, the program searches Evrokorpus directly. If there are no hits even in the corpus, the program makes a search in IATE (EU term base).

If the user does get a list of hits then he gets details about the first term on the list. Details about other terms on the list are available just by clicking them. If terms on this detailed output exist in the corpus, they are clickable and lead to corpus output. If the corpus segment has been published in the Eur-Lex database or in a database of international treaties that Slovenia has concluded with other countries, a link is provided to the full text of this document. By clicking a Celex number, the user gets a monolingual output, and if he clicks another language at the top of the page, he gets a bilingual output.

If he uses the mobile-phone version, he gets only the basic data.

In advanced search, the user can define:

- SL and one or more TLs,
- one or more fields,
- word-match pattern,
- the type of output.

Corpus search is similar: the program first checks whether the search term exists in the glossary. If it does, its translation is listed with a link to additional data. Afterwards, hits from the corpus are shown. The search terms that were found in the corpus are coloured blue on the output screen. If the program finds a translation of the search term in the corpus output, the translation is also coloured blue. Every corpus unit also has revision stage data appended to it. The output is sorted so that the user first gets the hits of the highest quality. As before, the user can define specific search criteria in the advanced search.

In Termacor (multilingual terminology combined with multilingual corpus) there is just one user interface for both terminology and corpus search and the user selects the type of output (terminology, corpus or both). The user selects one SL and any number of TLs. If there are up to five TLs, the output is parallel, otherwise the output is sequential.

If the user wants to use the terminology analyser (Terminator), he just copies the text into the text box (word(s), sentence(s) or complete text), selects SL and starts processing. On the output page all terms that exist in the term base are converted to Evroterm links. If input text is bilingual ("Trados segmented"), each sentence will be presented in a separate cell table; the idea is to simplify the search of new terms. The terminology analyser has several functions:

- translators use it to check and use existing terminology;
- terminologists use it to check the glossaries supplied by translators and to find new terms in existing translations.

More than 50,000 searches are performed every workday in the term base.

Wolfgang Teubert finished his paper (Teubert 1999) with the idea that the user of a dictionary should be able to check the corpus data himself – instead of obtaining filtered data by lexicographers. A dictionary is much more complex system than a term base – but it is necessary to start somewhere and it can be said that Evroterm in combination with Evrokorpus is a step in this direction.

Conclusion

On the basis of the development of the Evroterm term base and Evrokorpus bilingual corpus, the paper presents some possibilities of how to design electronic dictionaries to overcome book limitations:

- searching in both languages in a bi-lingual dictionary;
- full-text search;
- fuzzy search;
- division of dictionary into three parts: glossary, additional data and examples of use;
- independent development of these three parts;
- use of a corpus for retrieving examples of use;
- a corpus as a supplement to dictionary data and a glossary as a supplement to corpus data;
- terminological analysis of a text to be translated;
- continuous improvements of software and data;
- dictionary copyright protection with digital certificates.

References

Amebis dictionaries: http://www.amebis.si (access date: 10 August 2009)

Brookes, Ian. Painting the Fort Bridge: Coping with Obsolescence in a Monolingual English Dictionary. // *Proceedings of the Eleventh Euralex International Congress*. Lorient: France: Euralex 2004, pp. 221-231.

Deming cycle of continuous improvements: http://www.hci.com.au/hcisite3/toolkit/pdcacycl.htm (access date: 10 August 2009)

Drstvenšek, Nina. Vloga besedilnega korpusa pri postavitvi geselskega članka v enojezičnem slovarju. // *Jezik in slovstvo*, 48/5, 2003, pp. 65-81.

ELAN, SVEZ-IJS and TRANS corpora: http://nl2.ijs.si/index-bi.html (access date: 10 August 2009)

Eur-Lex: http://eur-lex.europa.eu/ (access date: 10 August 2009)

Evrokorpus: http://evrokorpus.gov.si/index.php?jezik=angl (access date: 10 August 2009)

Evroterm: http://evroterm.gov.si/index.php?jezik=angl (access date: 10 August 2009)

Grad, Anton, Škerlj, Ružena, Vitorovič, Nada. Veliki angleško-slovenski slovar. Ljubljana: DZS. 1997

IATE: http://iate.europa.eu/ (access date: 10 August 2009)

ISO 12616. Translation-oriented terminography. ISO, Geneva. 2002

Jakopin, Primož, Lönneker, Birte. Query-driven Dictionary Enhancement. // *Proceedings of the Eleventh Euralex International Congress*. Lorient: France: Euralex 2004, pp. 273-284.

Krek, Simon (ed.). Veliki angleško-slovenski slovar Oxford. Ljubljana: DZS. 2005-2006

Lönneker, Birte, Rozman, Katarina. Online SLO-DE-SLO: spletni slovensko-nemški in nemško-slovenski slovar. // *Zbornik 7. mednarodne multikonference Informacijska družba IS 2004*, book B: Language technologies. T. Erjavec, J. Gros (ed.). 2004, pp. 56-63.

Nova beseda corpus: http://bos.zrc-sazu.si/s_beseda.html (access date: 10 August 2009)

Slovar slovenskega knjižnega jezika: http://bos.zrc-sazu.si/sskj.html (access date: 10 August 2009)

Termacor: http://evrokorpus.gov.si/k2/index.php?jezik=angl (access date: 10 August 2009)

Terminator: http://evroterm.gov.si/x/indexe.html (access date: 10 August 2009)

Teubert, Wolfgang, 1999: Korpuslinguistik und Lexikographie. Deutsche Sprache 4/99. 1999.

Vintar, Špela. Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov: doktorska disertacija. Ljubljana. [COBISS.SI-ID 21981538]. 2003