Wörterbuch und Übersetzung

4. Internationales Kolloquium zur
Lexikographie und Wörterbuchforschung
Universität Maribor
20. bis 22. Oktober 2006

Herausgegeben von
Vida Jesenšek
und Alja Lipavic Oštir

Georg Olms Verlag
Hildesheim · Zürich · New York
2008

MIRAN ŽELJKO

Integration of Terminology Database and Corpus of Translations

1 A vision

Professor Wolfgang Teubert finished his paper Korpuslinguistik und Lexikographie with the following vision:

> „Mittlerweile setzt sich die Überzeugung durch, dass die nächste Wörterbuchgeneration, die einsprachige ebenso wie die zweisprachige, zumindest korpusvalidiert, wenn nicht korpusbasiert sein muss. Doch letztlich will der eigentlich korpuslinguistische Ansatz mehr. Interaktive Verfahren sollen dem anspruchsvollen Benutzer den direkten Zugriff auf die Korpusevidenz ermöglichen und ihm die Interpretation der Sprachdaten überlassen, anstatt dass sie ihm, wie bisher üblich, durch die Brille der Lexikographen vermittelt werden." (TEUBERT 1999, 312).

In this paper an example of integrating a terminology database and a bilingual parallel corpus of translations will be presented – of course a dictionary is a much more complex product than a terminology database, but we have to start somewhere, and it seems wise to start with a simple task before we tackle a more complex one.

2 The past

Before Slovenia became a member of the EU, its translators had to translate an enormous volume of texts from English into Slovene (about 90,000 pages of the Official Journal). The time for this task was limited, so a large number of translators participated in this project. Expert-, legal- and language revisers revised their work. With so many participants it is important to ensure consistent translations (i.e.: the same term that appears in various legal acts must always be translated in the same way). In order to ensure this, the translators made

a glossary alongside each translation and those glossaries were afterwards revised in the same way as translations and finally stored in a terminology database. This database was regularly updated.

Translators used Trados software as their most important tool: MultiTerm was used for terminology management and Translator's Workbench for translation memory storage and retrieval. Trados software is rather expensive, and many of its features are useless for users who do not translate, so we had to provide a solution for revisers and external translators in order to ensure the use of consistent terminology.

Trados (now: SDL) offers web access to MultiTerm databases (even for users who do not have Trados software); the problem with this solution is that its server side is limited to the MS Windows operating system, while the Slovenian Government IT Centre has been using the Unix platform for its Internet-based applications. We did not want to change the whole web concept just because of a terminology database – but this meant that we had to develop our own solution. As far as translation memories are concerned, Trados does not provide access to them for users without Trados software.

In order to solve both of these problems we decided to develop a platform-independent solution: we exported the terminology and translation memory data into tagged text files, transformed them according to our needs, and stored them in a MySQL database on a Web server. Finally, we developed software for searching the databases (ZELJKO/KRSTIC 2002, 304–308).

Evroterm was developed and first presented in 2000 and Evrokorpus in 2002. Since then, there have been numerous and continuous improvements in this software. The number of languages in the terminology database has increased from 4 to 12, and the number of terms from 13,000 to 85,000. The volume of words in the bilingual corpus has increased from 4 million to 27 million.

3 The present

The terminology database is logically split into two parts:

- a glossary containing terms in various languages

- additional data containing the information supplied by the terminologist (e.g. the date when the term was first stored in the database and when it was last updated, the field in which the term is used, definition of a term, the source of translation, reliability of translation, notes (e.g. for synonyms), etc.).

Terminology database currently contains about 85,000 terms in English and Slovene, and far fewer terms in 10 other languages. On the Web, this database is known as Evroterm (http://www.gov.si/evroterm/index.php?jezik=angl).

Translation memory is transformed into a parallel bilingual (English-Slovene) corpus of translations. In addition to segments in both languages, the following data are provided: the field in which the segment was used, revision stage (as an indicator of quality) and the code of the document from which the segment was taken.

The English – Slovene corpus currently comprises more than 700,000 segments. On the Web, this database is known as Evrokorpus (http://www.gov.si/evrokor/index.php?jezik=angl).

Database design is such that the terminology database and the corpus can be developed independently of each other and a link between these databases is made only during search. There are two types of search possible in both databases: a simple and advanced search; in addition, a simplified version of terminology database for mobile phone users (or other small-display devices) has also been developed. Both databases are updated within the Translation, Interpreting and Revising Service of the Secretariat-General of the Government of the Republic of Slovenia. The software is developed in the same organisation.

As there are two databases, a search can be performed in two ways:

If the user conducts a terminology search (i.e. in Evroterm), the glossary is searched first, an alphabetically sorted list of hits is presented – each item corresponding to one entry in the database. This list is clickable: if the user selects a particular item, the details about this entry will be presented in the centre of the screen. Most terms in English and Slovene are clickable on this output page, too – by clicking them the user can see examples from the corpus (if there are any). In the corpus output page, a link to the full text of a document from which a particular sentence was taken can be selected – by doing this, the user can see a whole document in English, and upon request, the document can be aligned with any other EU language; by changing proper parameters in the URL of this

web page, the text can be seen aligned in any two EU languages (in case of EU legislation). Eur-Lex database is used for full-text versions of EU acts.

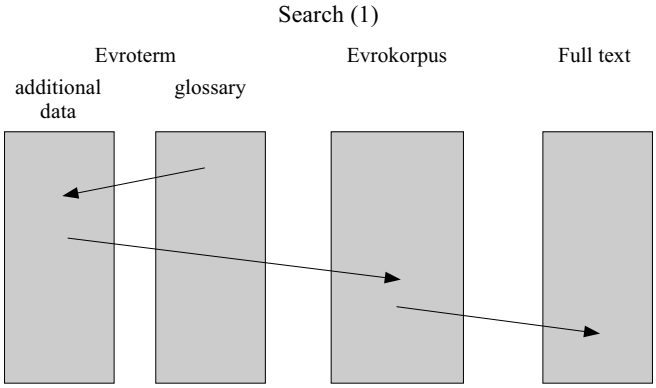

Fig. 1: Search path when the user starts searching in the terminology database.

The corpus output page lists the hits sorted according to the revision stage: translations that have been revised by multiple people are listed first. As usual, when using corpora, the user should use the page with caution: (s)he should decide whether some unusual result is a valid exception to the rule or is simply an error.

Some users need just the translation of the searched term while others are interested in or need additional data. In order to enable this, the software first produces small amount of output and if the user needs more data, (s)he can get it simply by clicking a link.

If the user is more interested in the use of a term in context and not so much in its terminology-related details, it makes more sense to search the corpus directly. Even in this case the software first makes a search in the terminology database's glossary and if the term is found in it, a link to additional output data is provided, which may be of interest to the user. The search term and its translation (if found) are coloured on the corpus search output page, so that the user can find the point of his/her interest more easily. As in the previous type of search, if the user wishes, (s)he can also see additional data related to terminology or the full text of the document.

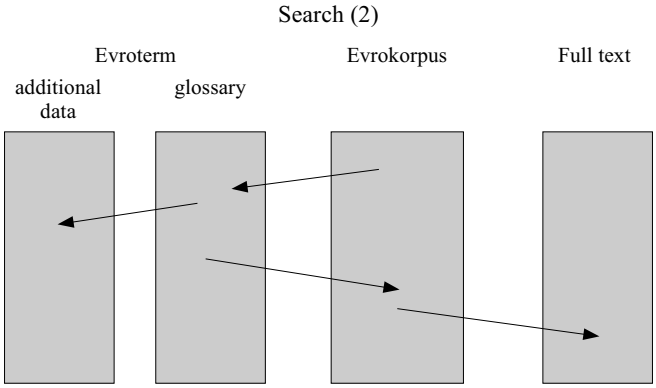

Fig. 2: Search path when the user starts searching in the bilingual corpus.

The corpus output can be either bilingual or monolingual – the latter is useful if a translator knows the meaning of a particular word or phrase but is interested in details about its use, which cannot be found in language handbooks.

The terminology database integrated with the corpus of translations exists only in electronic form – although it is possible to print the database it does not really make sense to do so.

User interfaces are currently available only in English and Slovene, but as these data are stored separately, a user interface in another language could be implemented easily.

There are many dictionaries, glossaries, terminology databases and corpora available. However, as a rule, the user can use only one tool at a time: either (s)he uses the dictionary (glossary, terminology database) or (s)he uses the corpus. We have seen from this paper that these two tools are logically connected – moreover, they are just two aspects of the same concept – the term. If we integrate these two tools, we get a system that has additional features:

- there are more data available
- this approach is more user-friendly
- there are more search options

- results are more relevant
- from the user's point of view, the search is faster.

4 The future

> These new dictionaries would (and could) not any more exist in printed form. Neither would they have a fixed electronic form. Rather, they would link a complex (but hopefully user-friendly) interactive query system designed to provide the users with the answers they are looking for to a continuously updated monitor corpus… Not every dictionary user is interested in the same depth of information. The interactive query system would initially offer more general information. More specific information would be offered only if prompted by the user. All the information required by the user would be generated in real time… (TEUBERT 2004, 17).

Professor Teubert had in mind a dictionary, rather than just a terminology database when he wrote the text cited above – and yet his words best describe the operation of Evroterm/Evrokorpus.

Owing to a careful database and software design, the terminology side can be extended, enriched, updated, maintained and further developed independently of the corpus side – and vice versa. Our terminologists update terminology data on the Web on a weekly basis. In addition to these routine tasks, in near future we will incorporate additional languages into the database. Corpus data are updated monthly. Currently, there is only a English-Slovene parallel corpus; we plan that we will soon add a German-Slovene and afterwards a French-Slovene corpus. These additional corpora will be much smaller, but they will further extend the basic principle of integrated terminology and corpus to multiple language pairs. However, the real equivalence to a multilingual terminology database would be a multilingual parallel corpus. With the publicly available European Commission Joint Research Centre results, such a solution does not seem to be in the all too distant future…

5 Summary

This paper has presented a web-based multilingual terminology database (the emphasis being on English and Slovene terms) and a bilingual parallel (English-Slovene) corpus of translations. These two tools can be used independently, however: by integrating them we have obtained new features that cannot be found in two separate tools. The terminology database and corpus can be extended and maintained independently; also independent is software development.

Literature

EUR-LEX. http://eur-lex.europa.eu/.

EVROKORPUS. http://www.sigov.si/evrokor/index.php?jezik=angl.

EVROTERM. http://www.sigov.si/evroterm/index.php?jezik=angl.

JRC-ACQUIS MULTILINGUAL PARALLEL CORPUS. Joint Research Centre. http://wt.jrc.it/lt/Acquis/

SDL: http://www.sdl.com/.

TEUBERT, WOLFGANG (1999): Korpuslinguistik und Lexikographie. In: Deutsche Sprache 4/99, 292–313.

TEUBERT, WOLFGANG (2004): Corpus Linguistic and Lexicography: The Beginning of a Beautiful Friendship. In: Lexicographica 20, 2004, 1–19.

ZELJKO, MIRAN / ADRIANA KRSTIC (2002): Web-based Trados databases – An Alternative Approach. In: Translation: New Ideas for a New Century, Proceedings of the XVI FIT Congress. Vancouver, 303–308.