Experience from Translation of EU Documents

Gábor Prószéky
proszeky@morphologic.hu

MorphoLogic
Késmárki u. 8, 1118 Budapest, Hungary
http://www.morphologic.hu

1. Introduction

There are three main actors in the ideal translation workflow: the translator, the terminologist and the reviser. We describe all the three roles in terms of their input and output information in the translation workflow and developed a language technology toolkit to help them to solve their tasks as effective as possible. In the project "*Translation EU Legislation Texts into Hungarian"* translators and revisers belong to various agencies, and the agencies communicate to each other via the Central Program Office. This organisational issue helps to ignore technical difficulties in the workflow caused by the different technical education of the translators.

In the basic system called *MoBiDic* [Prószéky 1998] both the dictionary modules based on already published paper dictionaries of various publishers and the translators' own (common) glossaries are represented using existing standards (i.e. SGML, XML, HTML) enabling simultaneous access to multiple lexical resources performing queries through linguistic pre-processing (stemming, etc.). At the end of the project, the common glossary of the translators involved in the project will be published as the largest and most up-to-date "*English–Hungarian Dictionary of Legal Terms Used in EU Documents."*

In the project of translation of EU legal documents into Hungarian the above-introduced LT toolkit has been used since November 1998, supporting the translation of 40,000 pages by around 100 translators belonging to the Consortium. The Consortium consists of three translation agencies, a revision agency and a language technology partner, MorphoLogic. The toolkit based on the MoBiDic system has been written in standard C and C++, and is totally portable. Actually, the MoBiDic servers run on Windows NT and Unix (Linux, Solaris), and the MoBiDic clients are Windows 98/95 and Windows NT applications. Communication between the clients and servers is based on TCP/IP; consequently web-browsers (like Netscape or Explorer) can also be used as special clients of the *MoBiWeb* system. It also allows parallel lookup in a set of an unlimited number of MoBiDic-dictionaries and dictionaries on various web sites with the help of the embedded *MoBiGloss* subsystem.

2. Resources at the start of the project

The aim of the project is to provide the institutions preparing Hungary's join to the EC with precise translations of the European Community primary and secondary legislation and court decisions. The project concentrates mainly to translations into Hungarian and, in exceptional circumstances only, Hungarian legislation into EC languages. The project aims to deliver linguistically, legally and terminologically precise translations and – on an on-going basis – technical glossaries. The *Translation Co-ordination Unit* (TCU) in the Ministry of Justice supervises the program. At the start of the project the translators had various terminological dictionaries available from MoBiDic, the intelligent translation support system:

- the so-called *PHARE Glossary* (edited by Ministry of Justice) based on translations of Hungarian Rules of Law into English
- *English–Hungarian Dictionary of Law* [Adecom, Móra, 1990, MorphoLogic, 1994]
- *English-Hungarian Dictionary of Banking Terms* [Adecom,Véges, 1991, MorphoLogic, 1994]

- *English-Hungarian Dictionary of Foreign Trade* [Adecom, 1993, MorphoLogic, 1994]
- *English/American Economical Explanatory Dictionary* [Adecom, 1993, MorphoLogic, 1996]
- *English-Hungarian Manager Dictionary* [Novorg, 1993, MorphoLogic, 1996]
- *English–Hungarian Financial Dictionary* [Novorg, 1995, MorphoLogic, 1996)]
- *English-Hungarian Explanatory Business Dictionary* [Exorg, 1993, MorphoLogic, 1996]
- *English-Hungarian Explanatory Dictionary of Marketing* [Magyar Reklámszövetség, 1989, MorphoLogic, 1996]
- *Legislation Terminology Bank* distributed in 1998 by TCU to help the translation of EU legislation
- *Collection of Keywords* from the previous translations of EU legislation, collected by the member companies of the Consortium

3. Organisation of the translation–revision process

The scheme of the translation workflow we are describing here is not merely a theoretical frame. The three main actors in the translation workflow are the translator, the terminologist, and the reviser, but the terminologist's job can be done by the translator himself/herself. In the reality, all the translators and revisers belong to various agencies who communicate to each other via the *Central Project Office* (CPO). This organisational issue helps to ignore little technical problems in the workflow caused by different technical education of the translators. CPO receives the original text for translation and records it in the logbook, classifies the translation by subject area (26 possible classes) and allocates it to the translators. At the same, CPO makes a pre-allocation of the scheduled translation to revision. When the translation is ready, it is delivered to the CPO where it will be registered into the logbook and sent to the reviser. The revised text is delivered to the CPO where it is copied to the actual database (backup). The CPO delivers the completed and separately revised translation to the TCU periodically, or on request from the TCU when it is necessary.

All the member companies of the Consortium operate according to DIN 2345 a Draft Standard for translation companies. The Draft was originally developed in Germany, but now the EU Translation Platform recommends the Draft with the Dutch extension (the certification procedure) for using in the EU countries. There are some paragraphs in DIN 2345, which are pretty important to the language technology tools supporting the translation process in this project:

Project function	**DIN Reference**
Terminology list for the client	Par. 4.2.1.5
External revision	Par. 4.2.1.8
Client's glossary and terminology list	Par. 4.3
Archiving project documents	Par. 4.4.2
Requirement against the target text	Par. 6
Checking the translation	Par. 7

4. MoBiDic H^+: An Intelligent Terminology Management System

The Consortium decided to have a well-defined language technoology support for the translation and revision work. Therefore, a system specification of a tailored software system, called *MoBiDic* H^+ was developed. It was also clear that the basic MoBiDic system (Prószéky 1998) in its original form was not enough for optimal solution of the project work. Due to the very special nature of the project, professional maintenance and solid user back-up were also a sound requirement for quality and assurance for work. About 100 distributed work places, the regular update of the Working Glossary was best carried out in a central service unit, called *Language Engineering Workshop* (LEW).

Translators translate source texts using partial translation units from dictionaries, new terminology from the terminologists, and translation proposals from the corpus of existing translations for the sentences to be translated. The first goal of developing a translation support tool has been, therefore, to create an intelligent dictionary system performing translation support to

some extent. Translators usually consult a variety of mono- and bilingual dictionaries and other reference works to find the precise meaning of a word in a given context. Traditionally, these were all printed materials, but today more and more such resources are available in electronic format (Chapter 2). At the same time, the variety of formats and proprietary interfaces of these offerings hinder the power of today's advanced search techniques. Even among the offerings of a single publisher, not all the electronic dictionaries have the same interface, and if they are designed to accommodate more than one dictionary, only one dictionary can be active at a time. Users are, therefore, required to deal with various search programs, cluttering their screens with redundant windows and toolbars. This lack of user-friendliness often leads translators to prefer traditional paper volumes, but they thereby lose the potential of the powerful search capabilities offered by desktop computers to improve and facilitate their work. For this reason, we have developed a software tool enabling simultaneous access to multiple lexical resources even from different publishers performing queries through linguistic stemming based on a multi-lingual morphological analyser.

The *multi-dictionary system* provides a uniform interface allowing parallel queries in multiple dictionaries regardless of the actual physical location of the resource (e.g. local network, wide area network, intranet or internet). Existing lexical resources have been semi-automatically converted into SGML/XML format with the help of a special subsystem we have also developed. Based on the dictionary management technology and other language technology modules, translators do not need to manually type the necessary words when working with a word-processor. Assuming that translators are networked: they are clients – with different access rights – to one or more dictionary servers over either LAN, or intranet/internet connections. For translators using intranet and internet, a HTML-interface – behaving as a special client to the users, called *MoBiWeb* – has been designed to access glossaries on the internet from the same user interface (see *http://www.mobidictionary.com*). Another special client – called *MoBiMouse* – of the client/server version is one that enables the user to see the translation of displayed text without a single mouse click. The user only has to move the mouse pointer over the appropriate word, then the program reads this text from the screen with a special character recognition technology, and displays its translation in another language. It is the fastest way in which translators can use dictionaries on a computer.

Nowadays, the most natural activity concerning electronic dictionaries is searching them for a single word. There is no problem if it can be found among the headwords of the actual dictionary. If the dictionary, however, does not contain the word in question, the translator must have a look for it in further dictionaries. No question, it is a time consuming task. MoBiDic, the intelligent multi-dictionary system, however, offers the ability of parallel search for the actual word or expression in an arbitrary set of dictionaries that is available for the translator.

4.1 Intelligent treatment of inflected words

The translator can easily start the look-up process by clicking onto an inflected word-form in the document to be translated. The inflected form usually cannot be found in the headword list. Typing the stem of the word, as it is supposed to be in the dictionary, is a boring and slow process. MoBiDic uses an integrated *stemming function* that provides the dictionary look-up module with the stem(s) of the input word automatically.

4.2 Look-up for multi-word expressions

Translators frequently want to find the word as a part of multi-word expressions, mainly in texts of law, economics, agriculture, etc. If the user does not know whether the actual word is part of some phrasal compound or idiom, the use of traditional printed dictionaries are pretty difficult. It is a time-consuming task to try to find another word in the actual sentence being the headword of an entry containing a multi-word expression with our original word.

4.3 Symmetric treatment of the two languages of the dictionaries

'Bi' is somehow misleading in the name MoBiDic (MorphoLogic Bilingual Dictionaries). Bilingual in this sense means that the source and the target language are not the same types of object for the program. In MoBiDic, source language means the language of known morphology to provide the user the adequate output. The output is expected to be in the target language, and MoBiDic has the knowledge about the characters, the alphabetic order, etc. of the target language to display the hits on the screen in correct form. The role of source and target languages (that might be the same, e.g. in explanatory dictionaries, synonym dictionaries) can be changed, that is, the same set of MoBiDic dictionaries supports translations both from and to Hungarian. Several publishing houses supply the lexicographic basis for MoBiDic. MorphoLogic actually licensed dictionaries that have already published and users were familiar with the paper version. When translators use the electronic version of these dictionaries they can decide which dictionaries should be activated. Currently, if all the available dictionaries are open, MoBiDic manages about one million lexical entries.

4.4 Building terminological databases on-line

Translators usually like to use their own terminological collections, vocabularies, glossaries, and the various client's special terminology. These user dictionaries are rarely published but translators want to use them via the same tool that is used for lookup in published dictionaries. MoBiDic supports to build user dictionaries, glossaries and terminological databases, and to use them as its "own" dictionaries. With the help of this module, consistent building, reviewing, and annotating terminology databases is guaranteed. In case of a networked system, terminologists can work simultaneously on the terminology database. With the help of this feature, at the end of the project a comprehensive dictionary will also be available, as a special by-product of the system.

4.5 Direct access from word-processors

Translators, using their favourite word processor (e.g. Microsoft Word) can reach directly from the terminology management environment. In MoBiDic, there are special frames for the input text to be translated, and another one for the construction of the perfect translation. That is, translators do not need to go back to the word-processor after each single dictionary look-up.

4.6 Pre-translator system

The pre-translator system helps the user to analyse the source text by going through the sentences, and checks whether all the words and expressions of the actual sentence can be found in the active dictionaries. If the pre-translator finds the equivalent, makes an annotation to the word in the source file, but if the entry is missing from every open dictionary, it automatically adds to the list of unknown terms.

4.7 Document-specific glossary builder

This application looks up all words and expressions of the source document in a given set of MoBiDic dictionaries. The output can either be a word-processor file containing the glossary (text file or formatted document) or a document-specific MoBiDic-compatible dictionary module. The document specific dictionary modules can be merged. By merging several dictionaries of this kind, dictionaries of a given area are possible to build.

4.8 Unknown words' glossary creator

This application finds all unknown words of the source document(s), according to a given set of MoBiDic dictionaries. The output can be either a text file (one sided) glossary or a special MoBiDic dictionary module, which can be completed later by the terminologist who receives words and expressions that have no – or not generally accepted – equivalents in the target language. To support his/her job, a special subsystem shows the term to be translated in a simple

terminology-editing window. In case of existing proposals for the target equivalent, the information who and when created the actual translation is also shown.

4.9 Consistency Checking System

First, an aligner module combines sentences or sentence sequences of the source and the target files to construct special couples called aligned bi-text. It gives error messages if the number of paragraphs is different in the source and the target texts, or the number of sentences is different in any of the paragraphs. The consistency checker module based on the MoBiDic dictionary system and it can be used for checking the (terminological) consistency of translations. It uses the output of the aligner and all the available dictionaries. If the source side of an aligned sentence pair contains a word or expression the translation of which cannot be found in the target equivalent, the consistency checking system marks the place of inconsistency. Later, the list produced by this system is processed by the translator who happened not to translate the term in question according to the common terminology database.

5. Working Glossary: An Increasing Terminology Database

An important part of the expected output is the glossaries to be given at the end of each and every translation. It is the request of the Contracting Authority to use these terms in future translations and revisions with the aim to reduce the error rate on each successive translation and revision, and to ensure consistency of similar content. In the present project environment, the very nature of terminology requires flexible and decentralised working facilities allowing timely follow up of the evolution of terms, the creation of new ones, and the offer of normalised terminology for each participant (translators and revisers) during the execution and revision exercises. The overall result is a running facility, the *Working Glossary* (WG), showing how translators and revisers interact to achieve their respective goals in terminology production chain. In addition, the WG will facilitate the TCU to update existing terms and to realise consensus about new terms. With the use of WG, the Consortium will ensure cost-effectiveness in terms of timeliness of output and improved quality of the services (translation and revision) provided, justifying the investment in advanced technology and the process-re-engineering effort.

The Working Glossary is

- a customised product not available in the market,
- an advanced product with basic language processing technology and features such as terminology database, procedures and system management facilities,
- a tested product for a cost-effective, high quality solutions to the translation process.

The Working Glossary together with the established procedures and rules for updating, revising and consolidating this facility will allow basic functions such as

- rapid collection and dissemination of new terms documented in individual piece of legal texts,
- improved quality of translation through centrally-controlled dissemination of updated WG,
- effective and efficient management of workflow,
- common platform for consistency of documents (revised translations) produced,
- operating workbench with items of one-word (keyword), expression, whole sentence, bunch of sentences.

The concept of the Working Glossary is the key element of providing the same tool for each translator to ensure consistency of translations. In order to assure the quality of translations and keep control of decentralised implementation of project activities, in particular collecting new terms entered into the WG by individual translators, there is a need for well-defined procedure mandatory for each translator and reviser. The TCU will also receive regularly copy of revised (updated) WG in order to make "super revision" as it will consider appropriate.

6. Related works

It is rather difficult to find fully automated translation workflow management software, we try to compare all the modules one by one to some existing applications. Multi-dictionary lookup systems are not typical on the electronic dictionary world. *CompLex* – among others – uses a bookshelf where the dictionaries take place, but they cannot be used simultaneously, as in case of our toolkit. Web interfaces for dictionaries are usually quite simple: in their development the potential competition with non-web dictionary interfaces have never been an important issue. Besides many mono-dictionary web-sites, *OneLook* offers parallel access to dictionaries – as our system does – but OneLook does not use stemming modules, and does not have non-HTML user interface, as our system does. In the literature we have not found real client/server implementation of dictionary software, because user dictionary management usually belongs to the terminological area (e.g. *MultiTerm*), and those tools do not offer access traditional dictionary, as we do. The only client/server dictionary system we have found is *DicoPro*, but it is still under development and does not have linguistically motivated stemming modules, as we do. In handling of multi-word expressions Xerox's *Locolex* and *Compass* systems are rather sophisticated (using finite-state representation for the multi-word unit) but they are not able to cope with multi-dictionary access. Among fast dictionary access systems based on screen-OCR we found *Babylon*, but it has a single source language (English), does not work without clicking, and contains canned dictionary content only, in spite of ours being a special client of the client/server application.

7. Further Plans

Additional translation support features are to be implemented in the near future by means of the parser module (context-based disambiguation, phrase detection etc.). Another important enhancement is the linguistically sound translation memory. Recent translation memories either attempt only literal matching, i.e. can only retrieve the exact match of a sentence, or employ fuzzy matching algorithms to retrieve similar target language strings, flagging differences. The new module will work with linguistic structures given by linguistic analyses rather than strings.

8. References

Breidt. E., F. Segond and G. Valetto (1994), Local Grammars for the Description of Multi-Word Lexemes and Their Automatic Recognition in Texts. *Proceedings of Complex-94 – Papers in Computational Lexicography*, pp. 19–28.

Calzolari, N. (1994), Issues for Lexicon Building. In: A. Zampolli, N. Calzolari & M. Palmer (eds.) *Current Issues in Computational Linguistics: In Honour of Don Walker.* Kluwer / Giardini Editori, Pisa, pp. 267–281.

Feldweg, H. and E. Breidt (1996), COMPASS – An Intelligent Dictionary System for Reading Text in a Foreign Language. *Proceedings of Complex-96 – Papers in Computational Lexicography*, pp. 53–62.

Hutchins, J. (1996), Introduction. *Proceedings of the EAMT Machine Translation Workshop* Vienna, pp. 7-8.

Kingscott, G. (1993), Applications of Machine Translation. In: Kohn, J. (ed.) *Transferre necesse est… (Current Issues of Translation Theory)*, pp. 239–248.

Nerbonne, L. Karttunen, E. Paskaleva, G. Prószéky and T. Roosmaa (1997), Reading More into Foreign Languages. *Proceedings of the Fifth Conference on Applied Natural Language Processing,* Washington, pp. 135-139.

Prószéky, G. (1998), An Intelligent Multi-Dictionary Environment. *Proceedings of 17th International Conference on Computational Linguistics (COLING 98),* Montreal, pp. 1067–1071.