Kinga Maślanko

A comparative study of terminology management tools in machine-assisted human translation.

Master's thesis
written under the supervision of
Prof. Włodzimierz Sobkowiak
School of English
Adam Mickiewicz University

Poznań, Poland 2004

ACKNOWLEDGEMENTS

I would like to express my gratitude to Marcin Feder Ph.D. from Adam Mickiewicz University for arousing my interest in terminology and computer-assisted translation, as well as for providing his illuminating comments and invaluable advice. I would also like to thank Peter Sandrini PhD from the University of Innsbruck for providing me with materials which greatly contributed to the development of my knowledge of terminology management and which pointed me in the right direction in my research. Last but not least, I would like to thank my supervisor for his encouragement and immense patience in supporting me at all stages of my work.

INTRODUCTION

In the global village we live in, the need to communicate seamlessly and effectively is a very significant one. The translation market in Poland is nowadays becoming highly competitive, demanding ever higher standards of performance and productivity both from experienced and novice translators. The requirements Polish translators and interpreters will have to meet are bound to soar even higher now that Poland has become a member of the European Union.

The Polish translation market can be characterized as highly fragmented, with very few medium-size companies. Many translators run single-person companies seated in their homes, sometimes not even owning legal software. The fact that the operating costs of these translation agencies are low, frequently paired with excellent customer service and high output quality can be viewed as an advantage. However, there are a number of serious limitations which act to the detriment of translation capacity and quality of the Polish translation agencies when it comes to large translation projects or technically demanding assignments (Argos 2002). Good management of translation projects, which involves, among others, efficient terminology management as well as the ability to use and benefit from the state-of-the-art language technology has become a necessity for those who wish to remain on the market.

Therefore, there is a great need for comprehensive writing on the tools that might help translators meet the ever-increasing expectations of their clients. This need includes not only a comprehensive presentation of the tools, their functionalities and advertising the different applications available, but also detailed and objective guidelines on how to evaluate such tools. There is a great abundance of sources presenting CAT tools. However, only testing the tools against objective and comprehensive criteria can give a real picture of the tools' applicability for a given user or a particular working environment. There is also an immense need to promote standards and new developments in the areas of language engineering and general computer technologies, in order to ensure more compatibility and exchangeability of terminology resources and translation memories among translators, technical writers, etc. (POINTER 1996)[1].

Bearing in mind the current situation on the Polish and global translation market, the author decided to devote this thesis to presenting a comprehensive study of terminology

[1] POINTER (Proposals for an Operational Infrastructure for Terminology in Europe) - a project carried out by terminology specialists in the years 1995 - 1996

management tools, which may address the above-mentioned needs, although in the limited way. The study will involve theoretical introduction into terminology management as a scientific discipline including a brief historical outline, followed by a suggested methodology of evaluation and an exemplary evaluation procedure comparing three terminology management tools. The author would like to emphasize that all the tools selected for presentation in the thesis were available for testing for free, and no software provider sponsored this project. The reasoning behind the selection of particular tools is given in the fourth chapter of the thesis.

In order to demonstrate the features of the tools, terminological databases were created in the programs selected. As a corpus for terminology extraction for the termbases, the author used a number of sources (see Appendix I). The selected programs were installed on two computers, both having Windows XP for the operating system. One had 256 MB RAM, the other 128 MB RAM. In both cases Office 2003 was used.

CHAPTER I
TERMINOLOGY – BASIC CONCEPTS

1. INTRODUCTION

This thesis contains four chapters, each devoted to a different aspect of the study of terminology management tools. This chapter acts as an introduction to terminology and terminology management in general. First the readers will be presented with the definitions of basic concepts of this discipline. The historical background of terminology management will introduce the most important developments in this branch of knowledge. The final part of this chapter will be devoted to illustrating the significance of terminology management and terminology management tools in translation.

2. DEFINITIONS

The notions central to terminology management, and thus to this thesis, are *term* and *concept*. As defined by Trippel "*term* is the language sign for a *concept*. This language sign does not necessarily have to be a single word, but it can also be a set of words - a fixed phrase - used only to denote a specific concept. Terms are not language independent while concepts are."(Trippel 1999).

The term *terminology* has two possible interpretations. The first one says that it is a specialist vocabulary, used in a particular subject field, also referred to as *technical jargon*. The other reference of this term is the theory or science dealing with the relations between terms and concepts (Trippel 1999). Another definition states that it is 'a structured set of concepts and their designations (graphical symbols, terms, phraseological units, etc.) in a specific subject field.' (POINTER 1996). On the whole, it is an interdisciplinary branch which involves both theoretical and practical aspects of creation, introduction, interpretation, usage, validation, evaluation, correction and classification of terms. There are a number of applications of terminology, among which the most significant are: standardization, research and development, marketing communications, consumer information, language engineering applications, knowledge engineering, computer-aided language learning (CALL), distance learning, computer-aided instruction (CAI), technical writing, corporate information

systems, information retrieval, term databanks (TDB), computer-aided translation (CAT), machine translation (MT), human translation, and nomenclature (POINTER 1996).

It is instructive to draw a distinction between the seemingly similar disciplines of terminology and lexicology, as well as terminography and lexicography. While the methodology of the disciplines in question may be in some cases similar, their focus is different. Lexicology is a linguistic specialty dealing with general language vocabulary, while terminology deals exclusively with special language lexis (POINTER 1996). Similarly, the general language dictionaries, compiled as a result of lexicographical work, contain some specialist terms as part of the general vocabulary, however usually embedded in the general language entries. Terminography in turn, deals with compiling special language vocabulary collections solely (POINTER 1996). Another difference is manifested in the direction of work. Terminology collection, usually restricted to a specialist domain, begins with concepts, not terms themselves and proceeds with the mapping of the domain with the concept delimitations i.e. terms, whereas lexicography work starts with vocabulary collection. However, there are linguists who claim that the distinction between the two disciplines may soon be no longer valid due to the imminent convergence of their methodologies (Campenhoudt 2001).

In this thesis the author will focus on the practical aspects of terminology, and its application in machine-assisted human translation (MAHT), therefore only selected issues connected with *terminology management* will be discussed.

Terminology management involves a number of activities, ranging from terminology collection or extraction, to terminology creation and validation, to classification, storage, retrieval and exchange. For the purposes of this thesis, we will focus only on the following aspects of terminology management: terminology extraction, organization, storage, retrieval and exchange. Some aspects of validation will also be mentioned.

This thesis is devoted to discussing terminology management tools which are often referred to as *terminology management systems (TMS)*. They are software systems which help to create and store terminological data in the form which allows for a controlled use of the data. Terminology management systems have nowadays become indispensable tools for translation agencies and translation project managers. Thus, at least rudimentary knowledge of these systems is required of translators who seek employment with such agencies.

Another central notion is that of *terminological database*, or *termbase*:

> 'Termbase: Short form of *Terminology database*. A termbase is the collection of information on a term or concept in a structured, electronically readable way combined with a *terminology management system*. It is mostly used synonymously with *termbank*, though some terminologists distinguish them. If they are distinguished, *terminology databases* do not include the organizational environment but *termbanks* do.' (Trippel 1999) (cf. Galinski 1998).

In this thesis the terms *termbank* and *termbase* will be used interchangeably.

Terminology management tools are part of a larger group of software tools referred to as *computer-assisted translation (CAT)* tools. CAT is defined as 'direct translation by humans with the help of a computer interface which makes translational expertise accessible through "translation-intelligent" software'. (Neubert 1991:56). In other words, CAT applications are a group of software tools assisting translators, where the human knowledge and linguistic competence are the key factors, and it is the human translator who plays the dominant role and makes the final decisions concerning terminology and phraseology choices.

Modern CAT tools, referred to as workbenches, consist of a number of modules or components, terminology management systems being part of them. The module which is considered the central one though, is the *translation memory* module.

> 'There are different TM programs currently available on the market, but they share similar features, albeit with some differences in speed and data management. Normally, the core of TM is *the memory*, a complex database where source text sentences are aligned side by side with the corresponding target text sentences. The ways in which *the memory* can be accessed and managed vary from one TM program to the other, but the philosophy behind the tool is basically the same: reusing previous work.'(Rico Pérez 2001).

In a nutshell, TM tools play the role of a perfect memory that can be accessed anytime during the translation process. It is a memory that never fails to retrieve the requested information and prevents the translator from struggling with the same translation problem twice. The fact that translation memory stores aligned sentence pairs in source language (SL) and target language (TL) makes the tool extremely useful in translating

repetitive texts e.g. technical manuals. When a new document is being translated in a workbench environment, the program automatically searches the translation memory for identical or similar segments, and whenever a match is returned (exact or fuzzy) it will be displayed in a special pane or grid or directly in the space where the target segment should be entered.

There are, however, technical texts that are very dense in terms of specialist terminology, but do not contain as much as two identical sentences. In this case the terminology management component of a workbench comes in as the right solution (Benis 1998). Thanks to terminology management modules even in the case of non-repetitive texts we can still benefit from the workbench packages in terms of speed and quality of translation, even though translation memory is not applicable.

Other components which are normally part of workbench applications are alignment tools (applications used for building translation memories from the corresponding SL and TL documents), analysis modules performing word frequency and repeatability calculations, sometimes also database and project maintenance modules.

At this stage it is necessary to draw the distinction between two terms which are frequently confused, i.e. computer-assisted translation (CAT), also referred to as machine-assisted human translation (MAHT), and machine translation (MT). While it clearly transpires from the very term that MAHT is the type of translation where the human translator plays the crucial role (cf. Feder 2001:51, Neubert 1991:57) it should be noted that 'MT aims at assembling all the information necessary for translation in one program so that a text can be translated without human intervention' (Craciunescu *et al.* 2004). The difference between MAHT and MT applications is also in the output quality. In the case of MAHT tools, the translations are usually of publishing quality. The up-to-date MT systems, on the other hand, deliver translations of unacceptable quality or requiring much post-editing. However, the advent of new MT systems applying neural networks and artificial intelligence technology is only a matter of time and we may expect the quality of their output to improve (Champollion: 2001).

Finally, we should bear in mind that computer-assisted translation is a complex process consisting of a several stages. Managing large translation projects involves a number of phases and tasks which can be broken down into translation and non-translation tasks or pre-translation, translation and post-translation tasks[2]. In such classifications, only the actual

[2] http://www.ad-ex.net/process.pdf

building of the TL equivalent of a SL text is regarded as a translation task, while all the remaining tasks, i.e. terminology management, desktop publishing (format conversions), text extraction[3], proofreading and customer's review, are considered to be non-translation tasks.

3. HISTORICAL BACKGROUND

The first efforts in terminology began in 1960's, probably as a result of the publication of the (in)famous ALPAC report which advocated developing software tools to aid translators instead of carrying out machine translation research (Feder 2001:15, Hutchins 1996, Palacz 2003:8). Terminology, however, had not been perceived as a discipline distinct from lexicology and other linguistic disciplines until the publication of the *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie*[4] by Eugen Wüster in 1979 (Wüster 1979) (cf. Campenhoudt 2001).

The first terminology projects were only available for large organizations because the terminology management software required mainframe computers (Rico Pérez 2001). This situation led to the development of large-scale termbanks, e.g. Termium, Eurodicautom, Banque de terminologie du Quebec (now Le Grande dictionnaire terminologique). The termbanks developed at this stage are still in use nowadays, although the systems underwent general overhauls, e.g. Eurodicautom now runs on entirely new platforms (Oracle and Fulcrum)[5].

The 1980's saw the first electronic dictionaries and terminology management software developed for personal computers and available for individual translators, following the development of translation memory software. However, these tools had many limitations. First of all, they were not networkable i.e. it was impossible to share terminology collections over local area networks. The first generation of terminology management tools often offered only unidirectional searching e.g. EN-GR but not GR-EN. There were also restrictions on the number and type of data fields, as well as of storage capacity.

The new generation of terminology management tools followed the publication of the concepts of a three-level integrated translator's workstation (Melby 1992) (cf. Feder 2001).

[3] If a project manager does not receive a source file in a text format, e.g. the source file is in a protected PDF format, the text that is to be translated will often be extracted from the file and delivered to the translator in a text format, in order to enable the translator to use CAT software. If the source text is delivered in hard copy, text extraction will refer to using optical character recognition (OCR) in order to receive the source text in the electronic form.
[4] General Theory of Terminology and Terminological Lexicography – An Introduction
[5] http://europa.eu.int/eurodicautom/Controller?ACTION=about

The first release of MultiTerm for Windows and DOS Translator's Workbench package was in 1992. Another CAT tool which is now one of the market leaders – Déjà Vu, was first released in 1993. These tools offered more possibilities than the earlier generation and gained significance especially due to the fact that the developers e.g. Trados, recognized the opportunities lying in the growing popularity of local area networks (Brace&Joscelyne 1994).

Since that time, many new tools and new versions of the first CAT tools have been released, catching up with the developments in language engineering and computer technology. Currently, there are two main tendencies in CAT development. On the one hand, software developers tend to isolate the functionalities which used to be part of terminology modules into separate tools, e.g. term extraction module was part of MultiTerm 5.0 but is no longer in a package with Multiterm *i*X. On the other hand, there is the tendency for integration of the typical MAHT tools, including terminology management tools, with machine translation systems and localization tools (Melby&Wright 1999), resulting in the so-called hybrid systems (Feder 2001:32). The application of hybrid systems and highly integrated translation environment is usually most advanced in large institutions, e.g. European Commission (Blatt 1998, Hutchins 1989).

4. SIGNIFICANCE OF TERMINOLOGY MANAGEMENT FOR TRANSLATION:

After the introduction of the basic notions related to terminology we should now focus on how terminology management tools can assist translators in their work. The best answer to this question is provided by translators themselves:

> 'Why do translators need to consult dictionaries, databases and/or experts when they work? The answer is so obvious that we tend to forget how important it is: Translators are not experts. This fact colors our whole approach to our work, particularly in areas where we are less than confident of our mastery of the subject matter.' (Titchen&Fraser 1996)

As we can see, applications designed to create and facilitate the use of specialist reference sources, tailored to the needs of translators cannot be replaced by any other tools. For a translator who has no knowledge of a particular subject area and needs a number of technical

terms which cannot be found in general-language dictionaries there is no alternative, but to create reliable terminology collections themselves. Another linguist mentions the following argument in support of termbases in translation work:

> '(…) in his daily work routine even the experienced translator encounters countless "new" problems having to do with the almost unlimited influx of words, terms and phrases that are not part of his average or even specialist lexical knowledge. The way he successfully copes with these gaps can most efficiently be modeled by term banks on the computer.' (Neubert 1991:58)

One more reason for the emphasis placed nowadays on efficient terminology management employing state-of-the-art technologies is the impact of terminology used, on the localization market, i.e. one of the most important areas of technological development.

> 'Efficient terminology management is crucial for publishers and manufacturers when translating and localizing their products. Translation vendors and translators may change over time – but the quality of the localized product should always adhere to the highest possible standard. Consistent terminology is necessary for ensuring continued familiarity with a product, and it is essential for functional compatibility between different versions of a product on one or multiple platforms.' (Project Review 2000)

Another advantage of using terminology management tools in translation is that a search for a given term is more time-efficient when compared to searching in printed dictionaries and other sources. It is not only faster however, but also more reliable as it was pointed out by the specialists working on the POINTER project in 1995 and 1996:

> 'Analysis of various dictionary entries demonstrates that the extraction of terminological data from currently-available LGP[6] dictionaries (both monolingual and bilingual) is problematic from a number of different points of view, including the inconsistent and imprecise use of subject-field labels,

[6] Language for General Purposes

> the absence of adequate pragmatic information, and varying definitional practices. Terms are also often deeply nested in entries, even as sub-senses of polysemous headwords. The unsatisfactory use of subject-field labels is of particular importance for the automatic extraction of data.' (POINTER 1996)

Terminological databases are designed to avoid problems of inconsistency and imprecision of LGP dictionaries. Terminological records provided in termbases are prepared by translators on the basis of sources they trust, with usage contexts of native origin exclusively (Göpferich 1995:23) frequently validated following consultations with experts in given subject areas. Thus, the reliability of linguistic data included is much higher.

One more argument in favor of using terminology management tools in translation is that usually there are no specialist dictionaries in the new and quickly developing fields of knowledge or 'the production of up-to-date reference works is lagging behind (Špela: 2001). The reason for this situation is that the process of compilation and publishing of printed dictionaries takes much longer and is more costly than in the case of electronic terminology collections. Therefore, machine-readable sources can reach the users much faster. Moreover, it is much easier to update and modify an electronic termbase than a printed dictionary.

Another obvious advantage, perhaps the most significant one, is that the results of terminology research once carried out are saved and kept for reuse in later projects. The electronic form allows also for easier exchange and sharing of resources among translators. Consequently, teams of translators working on large translation projects are equipped with tools ensuring greater terminological consistency, and therefore higher quality of translation.

Also, it is instructive to point out that using a terminology management tool is beneficial, even if other CAT tools are of little assistance, e.g. when the source text does not contain many repetitions, and there are no parallel texts which can help build a translation memory. In such cases, the translation memory module may turn out useless, while terminology management system may be of utmost assistance, offering a significant enhancement of terminological consistency (cf. Benis 1998).

As it is pointed out by the specialist of man-machine interaction in translation process:

> '[…] an individual translator cannot carry out the task of managing an entire project alone in a reasonable amount of time unless he or she works in a team; second, that this team needs to automate as many parts as possible of the

process if it is to provide a quick response to the client; and; finally, that translators need to adapt themselves to this new environment and learn new skills.' (Rico Pérez 2001)

In conclusion, termbanks, and terminology management tools are indispensable in enhancing the translator's ability to transmit a correct message in the target language to the recipient. They ensure better quality and consistency and boost the speed of translation, reducing the time spent on performing such pre-translation tasks as terminology extraction or collection and validation.

5. CONCLUSION:

Efficient terminology management is a prerequisite of a good translation service. The increasing volume of translation resulting from the processes of globalization and internationalization sets new challenges to translators. In order to meet them, all translators, either freelance or corporate, should take advantage of the new solutions increasing the speed of translation work while maintaining or improving the quality. Bearing in mind the above-mentioned arguments it seems obvious that efficient terminology management in translation can be implemented only through modern terminology management tools, tailored to the needs of particular working environments.

However, on many occasions translation memory tools are perceived as more productive and worth investment than terminology management tools. The cost-benefit ratios in the case of terminology managers must therefore be calculated very carefully. It should be born in mind that the benefits drawn from using the specialist terminology management software become transparent in a long-term perspective (Wright:10).