Metrics for Evaluating Translation Memory Software

by
Francie Gow

School of Translation and Interpretation
University of Ottawa

Under the supervision of
Lynne Bowker, PhD
School of Translation and Interpretation

Thesis submitted to
the Faculty of Graduate and Postdoctoral Studies
of the University of Ottawa
in partial fulfillment of
the requirements for the degree of MA (Translation)

Acknowledgments

First and foremost, I would like to thank my thesis supervisor Dr. Lynne Bowker for her prompt and insightful feedback, her talent for transforming the overwhelming into the manageable and her perpetual good nature. Working with you has truly been one of the highlights of this whole adventure!

Thanks to Dr. Ingrid Meyer for steering me toward graduate work in the first place and for awakening my interest in translation technology. Thanks also to Lucie Langlois, who helped me define my project in its earliest stages and who has been providing me with interesting work experience and useful contacts ever since.

The University of Ottawa, the School of Translation and Interpretation and the Association of Translators and Interpreters of Ontario (ATIO) generously provided me with financial support. The Translation Bureau has been very supportive of my research by providing me with access to software, hardware and technical support. I am especially grateful for being granted access to the Central Archiving System, which allowed me to build the corpora I needed to test my methodology.

Another benefit of my time at the Translation Bureau was the opportunity to work with André Guyon of IT Strategies, who was very generous with his knowledge of both translation memory and evaluation. Our exchanges of ideas in the later stages of the project stimulated me to solve problems creatively and certainly resulted in an improved methodology.

Last but not least, many thanks to my friends and family who have been very supportive during the past two years. Thanks to Vanessa for walking this path ahead of me and letting me know some of what to expect. Thanks to my fellow translation students for walking this path alongside me. Thanks to my parents and sisters in Newfoundland for listening to my whoops of joy and wails of despair whenever I needed an ear. Finally, thanks to my “Ottawa family”, Inge, Julien, Anne, Alphonse and their (our) friends, for keeping me well fed and (relatively) well adjusted all this time. Additional thanks go to Alphonse for translating my abstract into French. I am privileged to have you all in my life; thanks for the “memories”!

Abstract

Translation memory (TM) tools help human translators recycle portions of their previous work by storing previously translated material. This material is aligned, i.e. segments of the source texts are linked with their equivalents in the corresponding target texts. When a translator uses a TM tool to translate a new text, the tool identifies similarities between segments of the new text and the stored source texts. The translator may then choose to insert or adapt the previous translation of that segment. Therefore, search-and-retrieval functions are an essential component of all TM tools. However, not all TM tools approach these tasks in the same way.

In conventional TM tools, the aligned texts are divided into sentence-level source and target translation units for storage in the database. Each sentence of a new source text is compared with the units stored in the database, and the tool proposes matches that are exact or similar. This is referred to as a *sentence-based approach* to search and retrieval. A different and more recently developed approach involves storing full source- and target-text pairs (known as bitexts) in the database and identifying identical character strings of any length. This is referred to as a *character-string-within-a-bitext (CSB)-based approach* to search and retrieval.

Because the second approach is more recent, traditional techniques for evaluating TM tools do not take into account this fundamental difference. Therefore, the goal of this thesis is to design and develop a new evaluation methodology that can be used to compare the two approaches to search and retrieval fairly and systematically, first by defining "usefulness" as a measurable attribute, then by measuring the usefulness of the output of each approach in an identical translation context.

The thesis is divided into three main parts. Part I defines important concepts, explains the two approaches to search and retrieval used in TM tools, and establishes a theoretical framework for the development of a comparative methodology. Part II covers the design, testing and refinement of the comparative methodology using two representative TM tools (TRADOS and MultiTrans). Part III contains an overall evaluation of the comparative methodology and suggestions for further research and development.

Résumé

Les outils de mémoire de traduction (MT) permettent aux traducteurs de recycler certaines parties de leurs travaux préalablement mis en mémoire. Les textes sont alignés, c.-à-d. que des segments des textes de départ sont mis en correspondance avec leurs équivalents dans les textes d’arrivée. Lorsqu’un traducteur utilise un outil MT pour la traduction d’un nouveau texte, l’outil retrouve les segments semblables dans les textes de départ en mémoire. Le traducteur peut alors utiliser le segment en cause tel quel, ou l’adapter au contexte. Par conséquent, les fonctions de recherche et de rappel sont une composante essentielle de tous les outils MT, chacun abordant ces questions à sa manière.

Les outils MT conventionnels divisent les textes alignés en phrases. Les phrases source sont mises en mémoire dans la base de données avec leur traduction. Chaque phrase à traduire est comparée avec les segments en mémoire, l'outil proposant des correspondances exactes ou semblables. Nous appelons cette approche à la recherche et au rappel l'approche phrastique. Une approche plus récente se distingue en ce qu'elle met en mémoire des textes entiers de départ et d'arrivée (les bitextes), ce qui permet de trouver des chaînes de caractères sans égard à leur longueur. Nous appelons cette approche à la recherche et au rappel l'approche chaîne de caractères en bitexte (CCB).

La deuxième approche étant plus récente, les techniques traditionnelles d'évaluation des outils MT ne prennent pas en compte cette différence fondamentale. L'objectif de notre thèse est donc de concevoir et d'élaborer une nouvelle méthodologie d'évaluation permettant de comparer les deux approches à la recherche et au rappel de façon équilibrée et systématique. Il s'agira d'abord de définir l'utilité de manière mesurable, pour ensuite mesurer l'utilité des segments proposés par chaque approche dans la traduction de textes donnés.

Notre thèse comporte trois parties principales. La Partie I définit les concepts importants, explique les deux approches à la recherche et au rappel utilisées dans les outils MT, et pose le cadre théorique nécessaire à l'élaboration d'une méthodologie comparative. Dans la Partie II, nous procédons à la conception de cette méthodologie comparative, pour ensuite la vérifier et la raffiner en utilisant deux outils MT représentatifs (TRADOS et MultiTrans). Dans la Partie III, nous faisons une évaluation globale de la méthodologie comparative proposée, et nous terminons en proposant des pistes d'avenir pour la recherche et le développement dans ce domaine.

Table of Contents

List of Figures

List of Tables

Chapter 0 Introduction

According to the recently published *Survey of the Canadian Translation Industry*, the sector of the translation industry that will experience the strongest growth – nearing 50% per year – is the development of technological aids for translators (1999, p.39). Translation technology covers many different types of computer aids for translators, ranging from word processors and electronic dictionaries to machine translation systems. However, one of the most popular types of tool on the market at the moment is the translation memory (TM) tool, and this will be the focus of this thesis.

0.1 Background information and motivation for research

Translation memory technology has been in use since the mid-1990s, and many translators who have not yet incorporated it into their work routines are becoming convinced that it might help them improve their productivity. As more and more TM tools become available on the market, translators, translation agencies and their clients are all asking the same question: which tool is best? It is a legitimate question, especially when a large investment of money and training is required up front. However, there is no general answer. Translation situations can vary widely, and each tool has features that could be more desirable in some contexts and less desirable in others.

Despite the significant differences between the tools on the market, they are all designed with a common purpose – to store previously translated material in an organized way and to extract from it as much useful information as possible to be recycled in future translations. Because such tools are becoming increasingly widespread, it would be useful to have a reliable way of comparing them based on their ability to perform these core functions of extracting information and presenting it to a user.

Preliminary research reveals that the approaches to these functions can be narrowed down to two fundamental categories, with all available systems falling into one or the other. One approach, referred to as the sentence-based approach, involves dividing source and target texts into corresponding sentence-length translation units, storing these paired units in a database and identifying which are identical or similar to sentences in new texts to be translated. The other approach, referred to as the character-string-in-bitext (CSB)-based approach, involves storing entire bitexts (i.e., source texts linked to their translations) in a database and searching for identical character strings of any length.

No previous research has compared the success of these two approaches in retrieving useful information. Although no study can definitively answer the question "Which tool is best?", clear data comparing the search methods would at least contribute some of the answer, and could potentially be of use to program developers. It would also address another problem: the statistics published by product manufacturers to promote their products are often inflated and are rarely comparable to each other. A systematic methodology that could be applied to several tools could provide a level field of comparison.

0.2 Objectives

The aim of the thesis is to design an evaluation methodology that could, if applied on a large enough scale, answer the question of which approach to search and retrieval, if any, extracts the most useful information given the same data. This can be a question of quantity, if one approach consistently retrieves more useful hits than the other, or a question of quality, if the two approaches tend to generate different types of hits.

The central problem will be defining usefulness in a measurable but valid way. I will begin with the assumption that a useful hit must be accurate and must save the user more time than is required to generate the hit. Producing an evaluation methodology that is both valid (produces results that accurately reflect usefulness) and reliable (repeatable by different evaluators) will be a high priority.

0.3 Methodological approach

As a first step, a literature survey will be performed to examine what researchers have already learned about evaluating translation memory tools and other related forms of translation technology. An initial evaluation methodology will be designed by adapting the most relevant information from the literature survey and combining it with personal experience.

Part of the literature survey will include research into the characteristics of test corpora suitable for TM evaluation. A source of texts meeting the criteria will be located, and a pilot corpus will be built from these texts to test and refine the initial methodology. A larger test corpus will be built for a scaled-up application of the refined methodology. The results of the scaled-up test will be discussed to illustrate the kind of information that can be obtained from applying the methodology. Finally, the general advantages and disadvantages of the methodology will be discussed and suggestions for future research will be recommended.

In summary, the following steps will be undertaken to develop the evaluation methodology:

1. literature survey;
2. design of initial evaluation methodology;
3. identification and construction of pilot and test corpora;
4. pilot test of initial evaluation methodology;
5. assessment of pilot results and refinement of evaluation methodology;
6. scaled-up application of refined evaluation methodology;
7. analysis of results;
8. assessment of refined evaluation methodology;
9. suggestions for future research.

0.4 Scope and limitations

The aim of this thesis is to develop a generic methodology for comparing sentence-based and CSB-based approaches to TM; however, owing to limitations of time and availability of software, it was not feasible to test this approach on every TM product available on the market. It was therefore necessary to choose a representative tool for each of the two categories.

Of the two approaches to TM, the sentence-based approach offers the widest selection of tools, including TRADOS, SDLX, STAR Transit and Déjà Vu, among others. TRADOS was chosen to represent sentence-based tools for two main reasons: accessibility and popularity. One purely pragmatic reason for selecting TRADOS was the fact that there were copies of the most recent version available to me for testing both at the University of Ottawa and at the Translation Bureau of the Government of Canada. Some of the other sentence-based tools were available only in demonstration form or in older versions. In addition, TRADOS is one of the best established and most widely used sentence-based tools (Benis 1999, p.16). According to the company's website, over 40,000 licenses have been sold, "representing the vast majority of the current translation technology market[1]", and major clients include the European Union, the Canadian Department of National Defence and the Canadian Forces, Microsoft, Xerox Corporation and Wal-Mart[2].

The choice of MultiCorpora's MultiTrans to represent CSB-based tools was equally straightforward. Firstly, of the few available systems that fall into the CSB category, MultiTrans features the most developed automatic search component. TransSearch by RALI[3] and LogiTerm by Terminotix also fall into this category, but both depend heavily on the user inputting search criteria. Secondly, MultiCorpora (the company that developed MultiTrans) was willing to lend me a copy of their most advanced version for research purposes and provide me with technical support.

A further limitation on the project involves the use of terminology banks. Both MultiTrans and TRADOS include highly developed terminology management components. However, term banks in TRADOS grow only when the user manually adds information. Term banks in MultiTrans grow automatically with each translation. The fair comparison of the two tools assumes that each system has access to identical information (in the form of the corpus), and this is difficult to control when one system's databases are static and the other's are dynamic. For this reason, neither terminology bank was activated during the testing.

A final limitation is related to language. Although the methodology developed for this thesis is meant to be applicable to any set of languages accommodated by the

[1] http://www.trados.com/pressshow_en.asp?lang=en&action=pressshow&cat=10&id=17&site=29&

[2] http://www.trados.com/index.asp?lang=en&cat=11&site=33&action=

[3] Le Laboratoire de Recherche Appliquée en Linguistique Informatique, Université de Montréal

systems under evaluation in any direction, I limited the tests to my own language combination and direction: French to English.

0.5 Outline

This thesis has been organized into three main parts. **Part I** is intended to provide background information about the history of translation memory technology and the benefits and drawbacks of its use (Chapter 1), a description of two distinct approaches that have emerged for searching in and retrieving information from a database (Chapter 2), and finally, a critical summary of the research that has been carried out in the area of translation memory evaluation (Chapter 3).

Part II describes the design of an initial evaluation methodology and its refinement following a small pilot test (Chapter 4). The refined methodology is then applied to a larger corpus and the results of this application are discussed (Chapter 5).

Part III provides a discussion of the strengths and weaknesses of the methodology, along with recommendations for further research (Chapter 6).

Part I

Chapter 1 Translation Memory Past and Present

1.1 From MT to TM

By now, most people have heard of machine translation (MT), a process by which a computer program carries out the task of transferring text from one language to another. The idea of machine translation has been around since the 1940s, and while the last 60 years have brought a great deal of progress in computational linguistics, users of MT are still highly dependent on human intervention in the form of pre- and post-editing to bring the output up to acceptable standards. The ultimate goal of fully automatic high-quality translation (FAHQT) still seems to be a long way off. More recently, translation researchers have begun exploring alternative uses for the computer as a support tool for human translators, which they hope will bring more immediate benefits to the translation industry.

One of these alternative avenues that has been receiving a great deal of attention in recent years is the concept of translation memory, or TM. A TM can be loosely defined as a collection of translations from which useful information can be extracted for reuse in new translation jobs. Looking up previous translations is not a revolutionary idea in itself; translators do this manually all the time. The major advantage of automating the procedure is that it removes the hit-or-miss aspect of the task, theoretically increasing consistency and productivity. It is important to note that with TM, unlike with machine translation, a human translator still does the actual work of translating.

1.2 Early incarnations of TM and related tools

In 1966, the Automatic Language Processing Advisory Committee (ALPAC) published an influential report called "Language and Machines", which concluded that future prospects for machine translation were limited. In this same report, there is a short description of a system used by the European Coal and Steel Community (CECA) that seems to qualify as an early TM system. The report describes CECA's system as

> automatic dictionary look-up with context included. […] [T]he translator indicates, by underlining, the words with which he desires help. The entire sentence is then keypunched and fed into a computer. The computer goes through a search routine and prints out the sentence or sentences that most clearly match (in lexical items) the sentences in question. The translator then retrieves the desired items printed out with their context and in the order in which they occur in the source. (ALPAC 1966, p.27)

What is interesting about the system is its bilingual output. For example, if the user underlines "aptitude au formage à froid", the output will be the nearest possible match (or matches) along with its equivalents, in this case "aptitude à la déformation au froid – cold drawing quality" (p.87). Although this particular system was intended primarily for terminological research, the process includes the elements of text alignment, automatic matching and retrieval, and keeping terms in their contexts, thus anticipating many of the essential features of modern systems.

In 1978, Peter Arthern filled in some of the blanks by expanding the idea of a "translation archive". His vision included "the storage of all source and translated texts, the ability to retrieve quickly any parts of any texts, and their immediate insertion into new documents as required" (Hutchins 1998, p.295). The quick retrieval of *any* parts of any text does not presuppose that the translator would be limited to looking up lexical units or even sentences, and the ability to insert matches into new documents adds a new dimension of usefulness for the translator.

In 1980, Martin Kay called for a complete re-evaluation of the relationship between translators and computers with his proposal of a *translator's amanuensis*. He envisioned the following:

> I want to advocate a view of the problem in which machines are gradually, almost imperceptibly, allowed to take over certain functions in the overall translation process. First they will take over functions not essentially related to translation. Then, little by little, they will approach translation itself. The keynote will be modesty. At each stage, we will only do what we know we can do reliably. Little steps for little feet! (1980, p.13)

Kay offers text editing and dictionary look-up as examples of easily mechanizable tasks that are likely to increase a translator's productivity. He then describes an outline of a TM-type system:

> [T]he translator might start by issuing a command causing the system to display anything in the store that might be relevant to [the text to be translated]. This will bring to his attention decisions he made before the actual translation started, statistically significant words and phrases, and a record of anything that had attracted attention when it occurred before. Before going on, he can examine past and future fragments of text that contain similar material. (1980, p.19)

The idea was not quite as trivial as dictionary look-up, and it seemed to fall into the area of functions "approach[ing] translation itself". However, Kay considered this task more mechanizable and more achievable in the shorter term than machine translation proper.

Alan Melby (1982, pp.217-219) picked up this theme again two years later with his "translator's workstation", functioning on three levels. The first level includes all functions that can be completed in the absence of an electronic source text, including word processing, telecommunications and terminology management. Melby's second level assumes the availability of the source text in electronic format for such functions as text analysis, dictionary look-up, and synchronized bilingual text retrieval, while the third level refers to machine translation.

Melby's description of synchronized bilingual text retrieval in his 1992 paper, "The translator workstation", begins to approach quite closely the current incarnations of TM tools:

> When a document had been translated and revised, the final version, as well as its source text, would be stored in such a way that each unit of source text was linked to a corresponding unit of target text. The units would generally be sentences, except in cases where one sentence in the source text becomes two in the target text or *vice versa*. The benefits of synchronized bilingual text retrieval are manifold with appropriate software. A translator beginning a revision of a document could automatically incorporate unmodified units taken from a previous translation into the revision with a minimum of effort (1992, p.163).

Melby's use of the words "synchronized" and "linked" is important. They refer to the concept of alignment, an important element in the design of effective TM tools. It was the appearance of TM tools like ALPS (1981) and ETOC (1988) during the 1980s (Somers 1999, pp.115-6) that pushed translation memory beyond the realm of mere academic speculation, but it was during the 1990s that TM developers were finally able to incorporate advances in corpus alignment research. Hutchins (1998, p.302) sees this as an essential step toward the viability of a TM as a useful tool for translators.

The late 1980s and early 1990s saw the development of another interpretation of "synchronized bilingual text retrieval", namely bilingual concordancing. A bilingual concordancing tool is used to search for patterns in a bitext, also called a parallel corpus, which is made up of aligned source and target texts. It retrieves the requested patterns in their immediate contexts along with their corresponding translations (Bowker 2002, pp.55-58). One such tool is RALI's TransSearch (Macklovitch et al. 2000). Users of

TransSearch must define and enter search patterns themselves, but attempts have since been made to automate the look-up process. This has led to an alternative approach to organizing and using translation memories, which will be discussed in more detail in Chapter 2, following a discussion of the more commonly used sentence-based approach.

During the early- to mid-1990s, translation memory made the next leap, from research to commercial availability. TRADOS, a German translation company, released the terminology management system MultiTerm in 1990, later following it with its TM tool Translator's WorkBench (Hutchins 1998, p.303). Atril released its in-house TM tool, Déjà Vu, in 1993 (www.atril.com). Transit, by STAR, and Translation Manager/2, by IBM, were also released around the same time (Hutchins 1998, p.303). Translation companies and freelance translators are now faced with the question of whether to embrace the new technology, which still requires a significant investment of resources, and if yes, which among the growing selection of competing brands to choose.

Angelica Zerfaß (2002a/b) and Celia Rico (2000) both provide guidelines for choosing the most appropriate TM tool for a given working environment. The available tools differ quite widely in some of their features, including editing environments, ability to handle various file formats and ability to generate statistics about databases and source texts. For good reason, TM evaluation has so far been less concerned with determining the single best tool and more concerned with helping translators and project managers evaluate the best tool for their own needs. However, this study will cover a different aspect of TM evaluation that has yet to be addressed: of the two underlying approaches to searching TM databases, which is more effective? Although the exact process varies with every tool, they all fall into one of two categories, the sentence-based approach and the CSB-based approach. These will be described in more detail in Chapter 2.

1.3 The potential of TM

The translation industry is bracing itself for a significant increase in demand over the next few years. According to the Canadian Translation Industry Sectoral Committee, the European translation industry is expected to grow by 7% annually over the next five years, and the Canadian industry by 5-10% over the next three years (1999, p.80). Canadian universities are producing only about 400 new translators each year, less than a third of the 1000 graduates required to meet the growing demand (p.19). This is exactly the kind of pressure that will favour technological innovation. Machine translation is being used in a few areas, but it is inadequate as a global solution in its current state. Instead, Computer-Assisted Translation (CAT) tools are being called upon to help human translators increase their output of high-quality translation in the short term. Translation memory tools fall easily into this category. Some companies have already begun integrating TMs into their translation process with some success (Andrés Lange and Bennett 2000). Although TMs in their present form are not a perfect solution, they are rapidly evolving in response to user needs.

In theory, when translators have organized, accessible archives of previous translations at their disposal, the result should be improvements in consistency and translation speed. This applies more to some types of documents than others; by definition, translation memory is most useful when applied to texts with repetitive content. The repetition can occur internally or across several texts in the same domain. Long texts often yield better results, since they are more likely than short texts to contain repetitive content (Austermühl 2001, p.139). Finally, certain text types, such as business/commercial, legal, scientific and technical texts, are much better candidates for TM than advertising or literary texts, which are non-repetitive in nature (Webb 1998, p.16-17).

1.3.1 Consistency

TM developers argue that users of the software will see an improvement in their consistency. Although human translators can easily come up with a variety of acceptable translations for a given passage, they cannot necessarily rely on their own memories to tell them how they have translated a particular passage before. The computer, however, is extremely reliable in this area, providing the translator with instant access to former work.

If the translation memory is on a network, several translators working on a single project can achieve greater consistency through instant access to each other's work (O'Brien 1998, p.120). To achieve similar results without TM, more resources must be invested in post-editing.

A major concern of translators is the threat of litigation in the case of mistranslation. Liability insurance was the subject of a recent article in InformATIO, the newsletter of the Association of Translators and Interpreters of Ontario (Voyer 2002, p.4). A tool that significantly improves consistency can therefore reduce the chances of litigation (Gordon 1996, p.3). All of this assumes, of course, that the database is properly maintained. Maintenance is a time consuming task; if it is neglected, errors can actually be propagated quickly throughout a document (Austermühl 2001, p.140).

1.3.2 Speed

A second reason that TM is being developed is its potential to reduce the time it takes to translate text, hence to increase a translator's productivity. Now, this can mean many things and can be measured in many ways. Sharon O'Brien (1998, p.119) states that "anything from 10% to even as high as 70% can be leveraged from translation memories". This is a rather broad range, but the kind of text being translated has an important influence on the actual increase. Lynn E. Webb (1998, p.20) quotes a 30% to 40% increase in productivity, Michael Benis (1999, p.22) quotes 30%, and Bert Esselink (2000, p.366) provides the figures 30% to 50% as the average increase in productivity in the field of software localization, a field particularly well suited to the use of TM. This clearly covers some of the distance required to meet the increased demand for translation referred to above.

The translation industry is currently in a position to make choices about how best to take advantage of this increase in productivity. One of the obvious options is to generate more income by translating more text in a given amount of time. Another increasingly popular practice is to pass on some of the savings to the client through discounts or graduated pricing systems, in which the client pays less per word for exact matches and very close matches. This can be seen as a benefit to translation vendors, since lower bids mean more contracts and improved client loyalty (Gordon 1997, p.4). However, that particular benefit will cease to exist once the industry regains some equilibrium in its practices and pricing structures. Ian Gordon (1996, p.8) also argues that freelance translators can take advantage of the rise in the promised productivity increases by "working less anti-social hours", all the while improving their earnings. This, too, may be a temporary benefit, as clients will adjust their expectations regarding price and turnaround time. In the future, we will most likely see a greater volume of texts being translated, with translators and clients sharing the financial benefits.

1.3.3 Quality of translation experience

TM technology offers improvements in productivity, but it also has the potential to produce significant changes in the way translators see their work. Several authors have speculated about these changes.

In Kay's original vision of a *translator's amanuensis*, the computer takes over the mechanical aspects of translation, so that "the productivity of the translator would not only be magnified but the work would become more rewarding, more exciting, more human" (1980, p.3). The translator spends less time repeating old work and more time facing new, creative challenges. Some even see the mastery of the technology itself as a new and creative challenge to be embraced by translators (Heyn 1998, p.135).

The idea of TM is not as threatening to translators as the idea of machine translation, because the translator remains in control of the process (O'Brien 1998, p.120). If translators are to be convinced to adopt a new technology, it must be something that promises to help them do their work, not something that will potentially replace them.

Since translation is often one of the last steps in a production cycle, translators can be put under pressure to compensate for previous delays. This results in tight, stressful deadlines. TM can be used to alleviate this:

> A translator can even begin the translation process before the final original document is completed. If the translator is given drafts of the original document in its early stages of development, the text can be translated and stored in the TM database. Then, as updated sections of text are made available, the translator can perform fuzzy and exact matching, thus isolating the new parts from the parts that have already been translated or that are similar to the original (Webb 1998, p.15).

This scenario is especially applicable to the software localization industry, where localized products must often be released simultaneously with the original product and where delays result in lost sales (Gordon 1997, p.1).

1.3.4 Other benefits

An important aspect of translation is terminology management, and most TM software comes with fully integrated terminology databases. The software can often be used as a tool for creating glossaries and dictionaries (Webb 1998, p.13).

Project management is essential for large translation products, and "[u]tilities designed to report detailed statistics on word counts and the number of internal and external repetitions provide valuable information to project managers scheduling localization projects" (Esselink 2000, p.366).

Gordon (1997, p.4) suggests an interesting benefit that is not often discussed: the applicability of TM to minority (i.e. less widely used) languages. According to Gordon,

> MT has not served minority languages well, the main reason being the commercial reality of insufficient sales to justify the massive cost of creating the machine translation software. In contrast Translation Memory systems are highly flexible and can be customised for minority languages extremely cost-effectively. They can offer a high quality solution where none was previously available.

This is partly due to the fact that TM software is not required to have much linguistic knowledge programmed into it to work effectively (Melby 1995, p.187).

One of the greatest advantages of TM is that its usefulness increases with time. While it is possible to align old archives for use right away, it is also possible just to begin translating with the software and allow the database to build itself:

> The source-language text is entered as a whole and presented to the translator segment by segment; when the translation has been completed, the source and target are automatically saved as pairs. These texts then form the TM database. The more texts you translate, the bigger your database will become (Austermühl 2001, p.135).

The bigger your database, the more likely you are to find quality matches for new translations. While the initial investment of time, money and training may be significant, if the texts are suitable, the cost can be recovered.

1.4 Drawbacks of TM

TM products are still fairly new, so it is natural to encounter problems with their integration into the language industry. One of the most obvious is the steep learning curve translators must face when beginning to use the software (Webb 1998, p.50). The long-term benefits of TM described in the preceding paragraph are promising, but it can be discouraging to be confronted with decreased productivity in the short term while struggling to master the technology. Once the technology has been mastered, it will still require a time investment not required by traditional translation, so the time saved by using TM must continue to be greater than the time spent running and maintaining it (Esselink 2000, p.367).

A second problem is that TM systems are only useful when source texts are in electronic format. Scanning and Optical Character Recognition (OCR) software does exist to convert hard copy into electronic format, but this solution is too labour-intensive to be practical on a large scale. However, this may become less of a problem as clients become aware of the benefits of delivering source texts electronically (Bowker 2002, p.137).

Translators often work with multiple file formats, and TM systems generally require filters to preserve formatting. This step involves a certain amount of risk (Bédard 1998b, p.23), since elements of the formatting may be lost or altered when a file is converted into a TM-friendly format, and again when the file is converted back into its original format after TM processing. The programs are generally sold with a certain number of filters to cover the most commonly used formats, but these become obsolete as soon as new versions of these formats appear. Furthermore, filters for custom formats are sometimes required, and these are complicated and time consuming to program (Esselink 2000, p.367).

TM also affects the translation process itself. Exactly how it does so varies from tool to tool, but a common complaint hinges on the revision process. If it is difficult to incorporate revisions into the database, translators may find themselves doing fewer drafts than they would otherwise (Webb 1998, p.50), or they may be tempted to neglect database management. The former practice threatens the quality of the present translation, while the latter threatens the quality of future translations. This can be solved with a combination of increased awareness on the part of the translator and improved design from the TM developers.

In those programs that require one sentence to be translated at a time, translators may be discouraged, or even prevented, from making changes to the macrostructure of their translations, such as changing the order of sentences in a paragraph (Esselink 2000, p.367). There are other style issues to consider as well. For example, Heyn (1998, p. 135) notes that since fewer changes will be required if a translation unit does not contain anaphoric or cataphoric references, some translators are avoiding these structures, which may affect the overall readability of the text (see section 2.1.2). Finally, in the cases where a memory is the product of work done by many translators with different styles, users must pay extra attention to ensure that their new translations have a unified style. Translators using the software should thus be aware that they might have to pay for their improved consistency with additional effort on other fronts.

All of this assumes that the translator still has full control over the decisions made throughout the text. Claude Bédard laments the phenomenon that he calls "la sous-traitance de phrases", whereby freelance translators are sent documents with a certain number of sentences already substituted:

> Dans ce cas, le traducteur n'est pas rémunéré pour les phrases déjà traduites, bien qu'il doive en tenir compte pour les phrases, plus ou moins clairsemées, à traduire. En outre, rien ne dit que les phrases déjà traduites sont de bonne qualité ni surtout qu'elles sont cohérentes entre elles (2001, p.29).

He goes on to speculate that such conditions may be very demotivating for translators, who in this position have even less control over the final document than post-editors of machine translation.

The example mentioned above foreshadows another difficulty: should translators be paid differently for work done using TM? A practice is emerging whereby translators are paid one price for new material within a document, a lower price for material similar to segments in the memory, and a lower price still (sometimes nothing!) for identical material. While this approach is easy to calculate and looks logical to most clients, it does not necessarily provide an accurate reflection of the work involved for the translator. Even exact matches need to be checked to make sure they fit correctly into a new document. Under the above model, translators may have to give some of their work away for free. Also missing from this basic pricing structure is compensation for time spent building and maintaining databases. Although it is true that the old fixed-cost-per-word structure no longer applies in an industry based on TM technology, the new model still requires quite a bit of refinement (de Vries 2002, pp.45-47).

Finally, a memory is made of material contributed by both the client and the translator. Who owns the final product? Moral ownership is one thing, but legal ownership is important when a memory becomes a valuable commodity in the marketplace. Suzanne Topping (2000, pp.59-61) states that translators have already begun pooling their resources by exchanging translation memory databases, even though there are opponents who question the usefulness and ethics of such a practice.

Academics and practitioners alike are currently engaged in lively debate about all of these issues, and solutions are likely to emerge in a short time. Given the steady increase in demand for translation, the problems are worth overcoming to take full advantage of the benefits translation memory has to offer.

Chapter 2 Two Approaches to TM

Elliott Macklovitch and Graham Russell (2000, p.1) provide two definitions of translation memory, a narrow definition limited to sentence-level processing, and a broader definition that allows for alternative levels of processing. This distinction has a significant effect on the strategies that can be employed for searching the database of a given TM system, making it a central issue for this thesis. Section 2.1 will explore the narrower, sentence-based definition in detail, and the broader definition will be examined more closely in section 2.2.

2.1 Sentence-based approach

Under the narrow definition, which is also the most common, a translation memory system is "a particular type of support tool that maintains a database of source and target-language pairs, and automatically retrieves the translation of those sentences in a new text which occur in the database" (Macklovitch and Russell 2000, p.1). In other words, when a given bitext is aligned, each source sentence is linked to its equivalent in the target language, and the pair of sentences is stored as a separate translation unit in a database. If the translator translates a new text using this database, the translation memory system will compare each sentence of the text with the contents of the database to find an exact or close match with a previously translated source sentence. If one is found, the linked target sentence can be inserted into the new text in the appropriate place, and the translator can make any necessary modifications. If no match is found, the translator simply translates from scratch. Once the new sentence has been translated, the new source/target sentence pair is added to the database. In this way, the database grows bigger (and in theory, more useful) every time it is used. Claude Bédard (1998a, p.25), Sharon O'Brien (1998, p.116) and Marie-Claude L'Homme (1999, p.213) all use close variations of this definition. O'Brien even cites "sentence memory" as a synonym for translation memory (p.116). Michael Benis (1999, 2000) does not explicitly define translation memory in this way in his comparative reviews of TM software, but all of the tools he reviews fall into this category. See Appendix A for a list of sentence-based TM tools currently available on the market.

2.1.1 Advantages of the sentence-based approach

Why is the sentence the privileged unit of translation in these tools? Quite simply, it is easier for a computer program to identify sentences than it is to identify other types of translation units. Sentences generally begin with capital letters and end with strong punctuations marks. Problems arise when the texts include abbreviations with periods in the middle of sentences, but this can be addressed to some extent with the use of stop lists to help the program identify and ignore such abbreviations (Bowker 2002, p.95). Hard returns and tabulations are also used to delimit sentences, with the result that textual elements such as titles and list items, not technically sentences at all, are included in the database (L'Homme 1999, p.215). However, this hardly constitutes a drawback.

Fixing the boundaries of the translation unit in this way permits the search algorithm to retrieve a variety of results. The most desirable result is the **exact** or **perfect match**, which is "identical to the sentence the translator is currently translating, both linguistically and from a formatting point of view" (O'Brien 1998, p.117). Another possible result is the **full match**, which is identical to the input sentence except for certain "variable elements" like numbers, dates or currencies (Bowker 2002, p.98). Of course, a database often contains sentences that provide useful information even if they do not fall under either of the previous definitions. Therefore, sentence-based TM tools are also designed to retrieve **partial** or **fuzzy matches**, which are similar to the input sentence. Similarity is relative, and each tool has its own way of measuring it, but overall this is a useful feature of search algorithms. For example, the following pair of sentences has a fuzzy match relation:

A. Capacité d'organiser son travail et d'établir des priorités.
B. Capacité d'organiser efficacement son travail et de fixer des priorités.

In my test database, when sentence A was the input sentence, sentence B was retrieved by TRADOS Translator's Workbench and labelled a 72% fuzzy match.

Any translation memory database that is in the form of a collection of linked sentences also has the advantage of being exchangeable between several different TM tools. This is possible thanks to the efforts of the Localization Industry Standards Association (LISA), which led the development of TMX (Translation Memory Exchange), a standard format that allows translation units to be exported from one TM program and reopened in another without the loss of information attached to those units, such as creation date or formatting attributes. This is important in an industry where it is impractical for a translator to own every possible TM tool that potential clients might require.

Having a database of sentences that grows as you translate allows for the possibility of networking. If a team of translators is working on a very large project, each translator on the network can have instant access to the work of all his or her colleagues, which can save time and improve consistency.

Although a tool that could access segments smaller than the sentence is likely to generate more matches, a significant benefit of using the sentence as the basic unit is that the matches that do come up are much more likely to be relevant. This is considered “an extreme form of high-precision, low-recall search” (Simard and Langlais 2000, p.1). Using this logic, some users even opt to adjust the system to retrieve only paragraph matches (an option in some sentence-based tools), so that they do not have to spend as much time verifying that the matches are appropriate in their new contexts (Esselink 2000, p.363). Obviously, this approach gives the best results when applied to updates of previously translated texts, such as the second version of a previously translated user manual.

2.1.2 Disadvantages of the sentence-based approach

Although it is usually possible to translate one sentence at a time, in the same order as in the original, this is not always the way translators work. Sometimes it is desirable or necessary to collapse two source-language sentences into one target-language sentence, or to expand one source-language sentence into two target-language sentences (Bédard 1998a, p.25). In addition, the source text may contain an alphabetical list of terms, which has to be reordered in the target text.

Situations such as these complicate the automatic alignment process considerably and sometimes result in badly matched units in the database. Database management is therefore essential, and for translators getting paid by the word, it represents time for which they are not compensated in the short term[4].

[4] This is one reason that translators are beginning to charge by the hour (Cohen 2002, pp.16-17).

The assumption built into sentence-based tools that there exists a one-to-one relationship between source and target sentences can also cause problems while the translator is working on new texts. Although it is often possible to get around the constraints of the tool and alter the text structure in the desired way, the translator must be aware that this seriously limits the reusability of the units created (Bédard 2001, p.29). What works well in a given context may be unusable in any other context. Heyn provides another example of the same problem:

> The existence of translation memory technology may also influence the way translators formulate texts. For example, since retrieved translation units normally require fewer changes if they do not contain anaphoric and cataphoric references, translators are tending to avoid the use of such devices. The effect is a more technical style, and sometimes a less readable text. In the end it is up to the translator to decide whether text cohesion should be compromised in order to facilitate the translation memory (1998, p.135).

This example was briefly mentioned in section 1.4, but bears repeating here because it is strictly a consequence of the sentence-based approach. In both examples, the translator has to choose between producing the best possible translation in the short term and building the most profitable (i.e. reusable) database in the long term.

Bédard (2001, p.29) warns that under these circumstances, translation professionals are reduced to mere “traducteurs de phrases”. He reminds us that “tout traducteur qui se respecte ne traduit pas des phrases, mais un message, et il gagne pour ce faire à s’affranchir des frontiers artificielles que constituent les points de fin de phrase.” This implies that good work is penalized and sloppy work is rewarded within a sentence-based TM tool.

Although the possibility of networking is listed above as an advantage, it is not without its drawbacks. François Lanctôt (2001, p.30) evokes a scenario in which translators with different working styles are required to work together from a central translation memory. Translator A likes to get a first draft done as quickly as possible, then clean up the text on the second pass. Translator B likes to move through slowly and carefully, getting everything right the first time. If Translator A is sending the lower-quality draft sentences into the memory, Translator B has no way of knowing whether or not they have been revised, and may feel obliged to use them in the name of consistency.

When it comes to searching the database, the central focus of this thesis, the translator must be aware of the limitations of fuzzy matching. A human can look at a sentence and determine at a glance whether it is similar to another sentence in a useful way. A programmer designing a matching algorithm must first establish a statistical model of similarity. For example, the program might consider one sentence similar to another if it can be modified to match it exactly within a maximum number of keystrokes, or if both sentences have a minimum number of words in common. These models will always be loose approximations, and the resulting algorithms run the risk of generating useless matches, otherwise known as "noise", or missing sentences that do not meet the criteria but would nevertheless be useful to a translator. The latter phenomenon is called "silence" (Bowker 2002, pp.99-100).

For an example of noise, compare the following segments from the test corpus:

a. Prendre des mesures de dotation et de classification.
b. Connaissance des techniques de rédaction et de révision.

Obviously, having access to the translation of the first segment is of no use whatsoever to somebody trying to translate the second sentence. However, TRADOS Translator's Workbench rates the match at 56%, presumably since half of the words are identical and in exactly the same positions. If the program is set to propose all matches greater than 50%, this pair meets the criteria.

On the other hand, the database may contain a helpful sub-clause that is not retrieved by the system because it does not account for a high enough percentage of the sentence in which it occurs. Suppose that the segment to be translated is “Connaissance de l’organisation du Ministère et des ministères clients", a heading pulled from one of my test texts. My database of previous translations happens to include the list item "**Connaissance de l'organisation du ministère**, de ses politiques, objectifs et priorités ;" [emphasis added] along with its translation. The same database also contains the list item "Connaissance des lois administrées par les **ministères clients** et la Loi sur l'administration financière" [emphasis added] and its translation, which means that there is enough information stored in the database to help a translator translate the new segment in its entirety. However, neither of the stored segments would be retrieved by a sentence-based system retrieving matches of 70% and higher. This is an example of silence.

One might presume that exact matches are always safe, but even here the translator must remain alert. If the database is poorly aligned, or if it contains poor-quality translations, exact matches are unreliable. This can be solved in part by careful database maintenance. However, the solution is not always that simple:

> Indeed, there are times when the proposed translation may not be appropriate, such as when the client has expressed a preference for using a particular style or term. Even though a segment may be identical, translators are concerned with translating complete texts rather than isolated segments, so it is important to read the proposed translation in its new context to be certain that it is both stylistically appropriate and semantically correct (Bowker 2002, p.97).

Homonymy causes some difficulty in this respect (Bowker 2002, p.97). Take the sentence, “The key is stuck.” This might appear in the database with a perfectly good translation referring to a “clé”. However, it is possible that the new text is referring to a “touche” on a keyboard, requiring a completely different translation. With the sentence-based approach, the translator has no access to contextual information to judge the validity of a given segment.

All of this can be overcome by diligence on the translator’s part. However, what if the translator is not given full control of the text to be translated? In cases where translators are paid less (or even nothing) for exact matches, clients with access to TM software may be tempted to batch translate the text, which means automatically replacing all exact matches with the translations proposed by the TM. The translator then receives a sort of hybrid or partially translated text and is forced either to translate around potentially inappropriate sentences and deliver lower-quality work or to take the time to clean up the text without fair compensation. Claude Bédard (2001, p.29) sees this kind of situation as demoralizing:

> On imagine la démotivation professionnelle que de telles conditions de travail peuvent susciter chez le traducteur. Il est ironique de constater, un *postéditeur* (réviseur de texts traduits par un système de TA) conserve la maîtrise d’œuvre et la responsabilité de la traduction finale, car il prend en charge l’intégralité de texte, ou du moins de sa tranche de texte… Cela reste, malgré tout, la traduction telle qu’on la connaît. Tandis qu’avec la sous-traitance des phrases, la traduction change de visage.

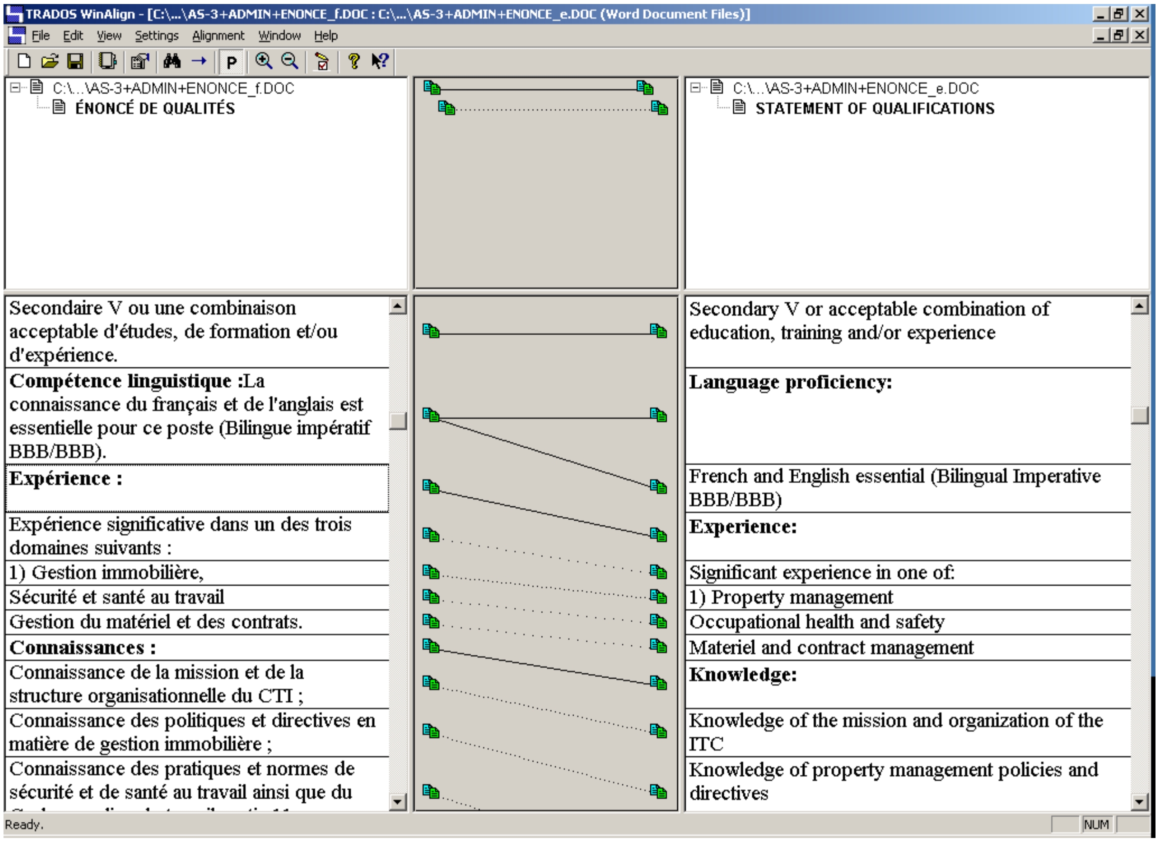
Common sense dictates that any system that lowers the quality of the translation experience will be met with resistance.

While it may appear at first glance that the drawbacks outweigh the benefits when it comes to the sentence-based approach, it is important to note that, while there are relatively few advantages, these are quite significant. The disadvantages are greater in number, but they are comparatively minor and can eventually be overcome with some effort.

2.1.3 An example of the sentence-based approach: TRADOS

TRADOS is a TM tool made up of separate but integrated components, including WinAlign, an alignment tool, MultiTerm, a terminology management tool, and Translator's Workbench (TWB), a tool that allows the user to search the translation memory database and retrieve segments for insertion into a new text. Other components are available, but these three are sufficient to describe the system's basic functioning. The following descriptions of these components are based on personal experience with TRADOS, with reference to the user manuals and on-line help that accompany the tool. As explained in section 0.4, TRADOS was chosen to represent the sentence-based approach for this study because of its popularity in relation to the other tools in this category and because I had access to a copy for testing purposes.

WinAlign is used to create translation memories out of previously translated texts. Both source and target texts must be in electronic format. Once the user defines which texts belong together, WinAlign automatically divides each text pair into segments, which, as noted above, roughly correspond to sentences, but may also include units such as titles, subtitles, items on a list or cells in a table, and creates a link between the source segment and its corresponding target segment. Each pair of linked source and target segments makes up a translation unit (TU). As shown in Figure 2-1, WinAlign displays the link as a dotted line stretched between the segments, with the source segments running down the left side of the screen and the target segments down the right.


TRADOS WinAlign - [C:\...\AS-3+ADMIN+ENONCE_f.DOC : C:\...\AS-3+ADMIN+ENONCE_e.DOC (Word Document Files)]
File Edit View Settings Alignment Window Help
C:\...\AS-3+ADMIN+ENONCE_f.DOC
ÉNONCÉ DE QUALITÉS
C:\...\AS-3+ADMIN+ENONCE_e.DOC
STATEMENT OF QUALIFICATIONS
Ready.
NUM

Secondaire V ou une combinaison acceptable d'études, de formation et/ou d'expérience.

Compétence linguistique : La connaissance du français et de l'anglais est essentielle pour ce poste (Bilingue impératif BBB/BBB).

Expérience :

Expérience significative dans un des trois domaines suivants :

1) Gestion immobilière,

Sécurité et santé au travail

Gestion du matériel et des contrats.

Connaissances :

Connaissance de la mission et de la structure organisationnelle du CTI ;

Connaissance des politiques et directives en matière de gestion immobilière ;

Connaissance des pratiques et normes de sécurité et de santé au travail ainsi que du

Secondary V or acceptable combination of education, training and/or experience

Language proficiency:

French and English essential (Bilingual Imperative BBB/BBB)

Experience:

Significant experience in one of:

1) Property management

Occupational health and safety

Materiel and contract management

Knowledge:

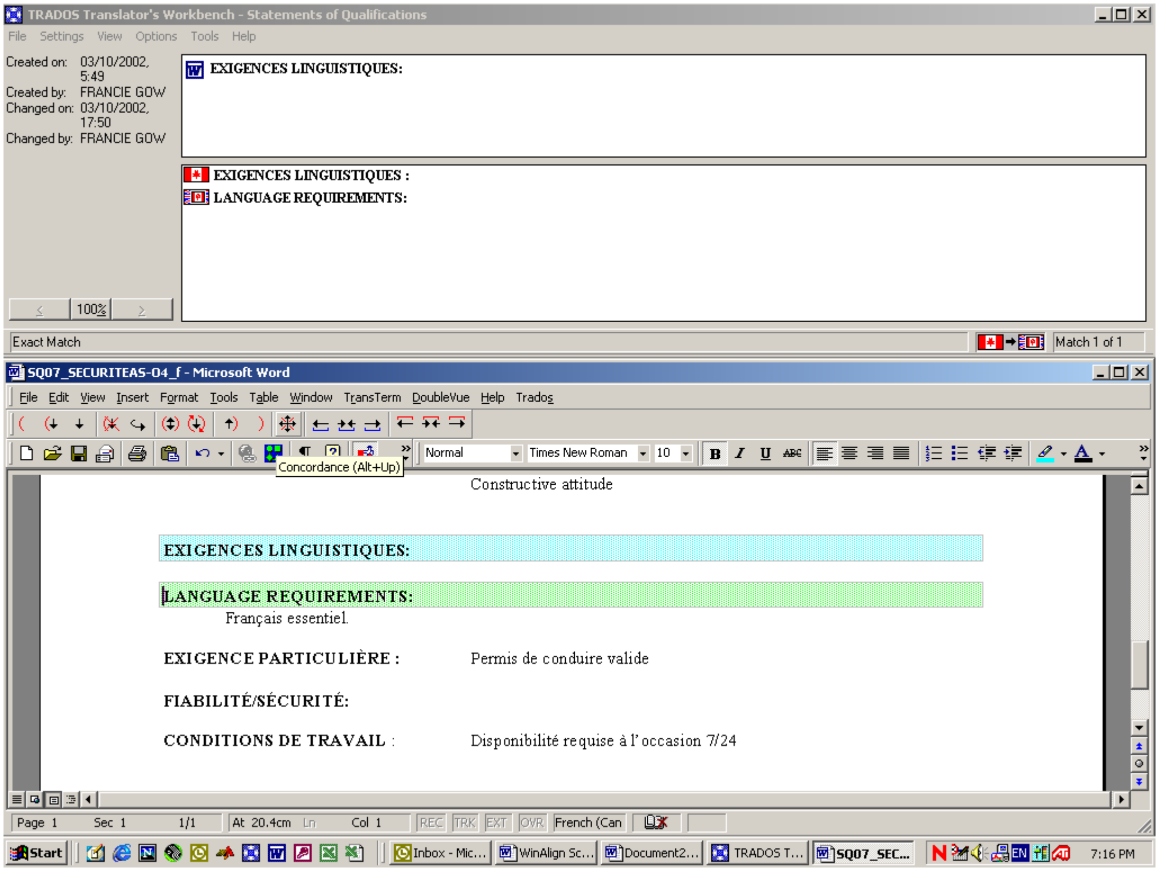
Knowledge of the mission and organization of the ITC

Knowledge of property management policies and directives

Figure 2-1 TRADOS WinAlign

The user can scroll down the screen, checking the alignments visually, correcting the mistakes, and confirming, or “committing”, correct alignments. Any alignment that has been committed by the user is linked with a solid line (see Figure 2-1). An alignment that has not been committed by a human is given a slight penalty by the system, so that an automatically generated translation unit will never be labelled an exact match and inadvertently inserted into the new translation without verification.

Once a memory has been created in WinAlign, it can be imported into TWB, the TRADOS component that interacts with Microsoft Word. As shown in Figure 2-2, in a typical set-up, the user will have the Word window taking up the bottom two thirds of the screen, and the TWB window taking up the top third. Clicking the “Open” command from a special TRADOS menu in Word, the user "opens" the first sentence in the text to be translated. This prompts the tool to search through the TM database and display in the TWB window any exact, full or fuzzy matches it finds. It assigns a percentage to each, representing the degree of similarity with the source sentence, and proposes all matches to the user from the highest percentage to the lowest, along with their translations. If there is a 100% match, nothing else is presented (see Figure 2-2).



Created on: 03/10/2002, 5:49
Created by: FRANCIE GOW
Changed on: 03/10/2002, 17:50
Changed by: FRANCIE GOW

EXIGENCES LINGUISTIQUES:

EXIGENCES LINGUISTIQUES :
LANGUAGE REQUIREMENTS:

Exact Match

Match 1 of 1

Constructive attitude

EXIGENCES LINGUISTIQUES:

LANGUAGE REQUIREMENTS:

Français essentiel.

EXIGENCE PARTICULIÈRE : Permis de conduire valide

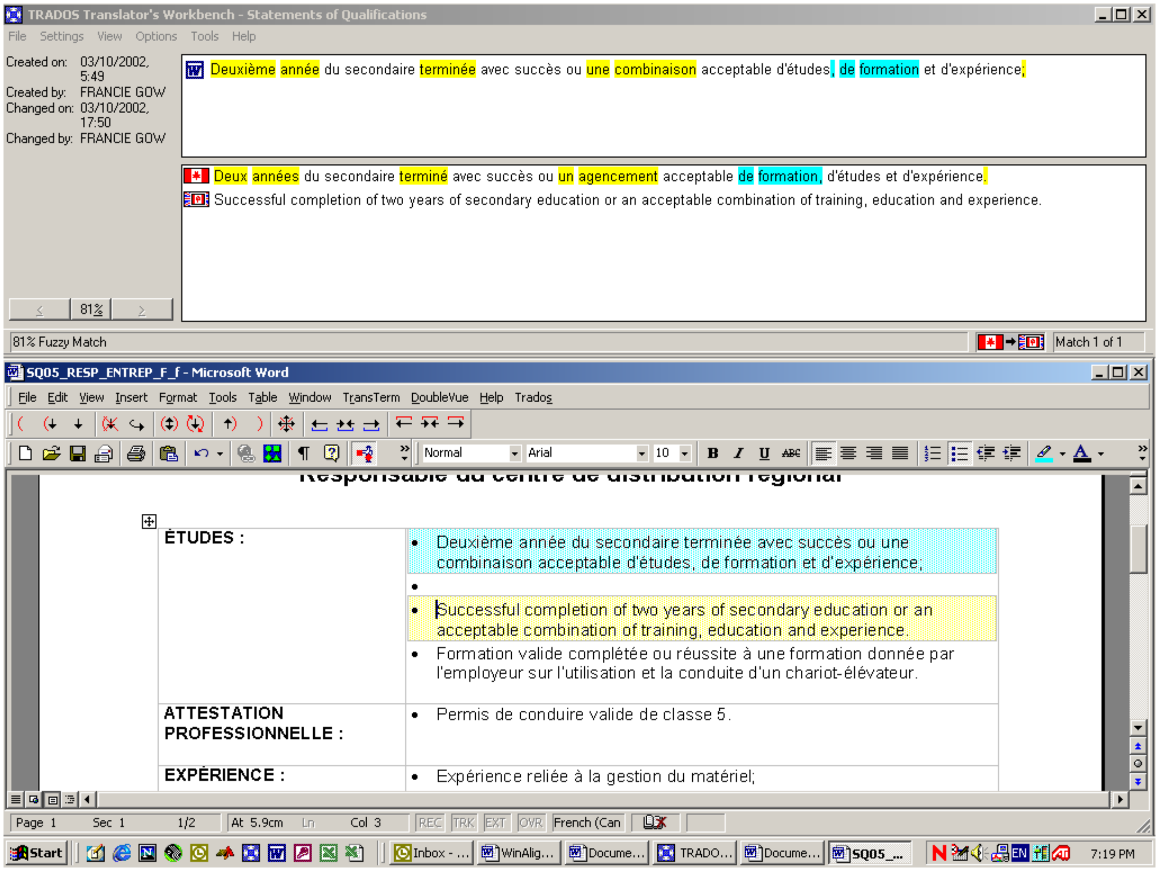
FIABILITÉ/SÉCURITÉ:

CONDITIONS DE TRAVAIL : Disponibilité requise à l'occasion 7/24

Page 1 Sec 1 1/1 At 20.4cm Ln Col 1 REC TRK EXT OVR French (Can

Figure 2-2 A 100% match in TRADOS Translator's Workbench (TWB)

The user selects the best option and clicks the "Get" command, which inserts the translation directly into the text in Word. Now both the original sentence to be translated and the translation proposed by TWB are visible in Word, in two coloured windows. The original is always open in a blue window. The target window directly below is green if the proposed translation comes from a 100% match and yellow for a fuzzy match (see Figure 2-3). The user can make any necessary changes in this window (or insert a translation from scratch if no match is found), and then select “Set/Close” from the menu to add the new translation unit to the TM database. The user now sees only the target sentence on the screen and can proceed in the same fashion with the next sentence. In this way, the database grows with every new sentence translated, and internal repetitions can be exploited. Internal repetition refers to linguistic material that appears more than once in a single text. For example, a sentence like “Press the ENTER key to continue” might appear several times in the same computer manual. External repetition, on the other hand, refers to linguistic material that is repeated in different documents.


TRADOS Translator's Workbench - Statements of Qualifications
File Settings View Options Tools Help

Created on: 03/10/2002, 5:49
Created by: FRANCIE GOW
Changed on: 03/10/2002, 17:50
Changed by: FRANCIE GOW

Deuxième année du secondaire terminée avec succès ou une combinaison acceptable d'études, de formation et d'expérience;

Deux années du secondaire terminé avec succès ou un agencement acceptable de formation, d'études et d'expérience.

Successful completion of two years of secondary education or an acceptable combination of training, education and experience.

81% Fuzzy Match

Match 1 of 1

81%
SQ05_RESP_ENTREP_F_f - Microsoft Word
File Edit View Insert Format Tools Table Window TransTerm DoubleVue Help Trados
Normal
Arial
10
Responsable du centre de distribution régional

ÉTUDES :	• Deuxième année du secondaire terminée avec succès ou une combinaison acceptable d'études, de formation et d'expérience; • • Successful completion of two years of secondary education or an acceptable combination of training, education and experience. • Formation valide complétée ou réussite à une formation donnée par l'employeur sur l'utilisation et la conduite d'un chariot-élévateur.
ATTESTATION PROFESSIONNELLE :	• Permis de conduire valide de classe 5.
EXPÉRIENCE :	• Expérience reliée à la gestion du matériel;

Figure 2-3 A fuzzy match in TRADOS Translator's Workbench (TWB)

Users can create term records in MultiTerm and access these records through TWB. If any term or phrase of a new text matches an entry of the same language in MultiTerm, the record will be displayed in the TWB window and can be inserted automatically into the new text. The MultiTerm window does not have to be visible to be accessible, simply open and running in the background. Further discussion of the MultiTerm feature is beyond the scope of this thesis since MultiTerm was not used as part of the experiment. The reasoning behind this decision is explained in more detail in section 0.4.

2.2 Character-string-within-a-bitext (CSB)-based approach

According to Macklovitch and Russell (2000, p.1), the broader definition of translation memory

> regards TM simply as an archive of past translations, structured in such a way as to promote translation reuse. This definition, notice, makes no assumptions about the manner in which the archive is queried, nor about the linguistic units that are to be searched for in the archive.

This definition is actually a little bit *too* broad for the purposes of this study, as it would include bilingual concordancers, which “can be used to investigate the contents of a parallel corpus” (Bowker 2002, p.55) in response to queries input by the user. We will assume for the present study that a TM tool is distinguished from a simple bilingual concordancer by the presence of an *automated* search component, which can analyze an input text against the stored previous translations and somehow signal to the user the presence of potentially useful matches. However, one of the underlying principles of the

Macklovitch/Russell definition is that the sentence is not the only possible storage unit for TM databases, and this remains valid.

It was Brian Harris (1988a, pp.8-11) who first coined the term “bitext”, referring to “the juxtaposition of a translation’s source (ST) and target (TT) texts on the same page or screen” (1988b, p.41). In the previous paragraph, “parallel corpus” is used in this sense. When the bitext is in electronic form, it can be aligned to facilitate bilingual searches. The principal difference between a database of aligned segments and a bitext is that in the latter case, the entire text remains intact. This is another way to store texts in a translation memory.

If the memory is stored in this way, it is possible to search at levels other than the sentence or paragraph. For example, the tool can search for any string of characters. Simard and Langlais (2000, p.1) point out that “just because a sentence has not been translated before does not necessarily mean that the TM does not contain smaller segments that could be useful to the translator.”

One of the original incarnations of this approach was TransSearch, an initiative of Université de Montréal’s Laboratoire de recherche appliquée en linguistique informatique (RALI). The translation databases used in TransSearch are called TransBases; the TransBase used in the version of TransSearch that was freely available on the Internet for several years was made up of Canadian parliamentary debates from 1986 to 1993, containing about 50 million words (Macklovitch et al. 2000).

TransSearch’s query system is highly flexible, able to find exact words, flexional variants of words, expressions, and groups of expressions separated by other words. It also recognizes negations and disjunctions. It has both unilingual and bilingual search options. Once a query has been entered, TransSearch presents a table of matches along with their translations. By making the system available on the Web and tracking its use, RALI has been able to demonstrate the demand for sub-sentential searching in a bitext: by March 2001, five years after it first appeared, TransSearch was processing more than 50 000 queries every month (Simard and Langlais 2000, p.1).

The main drawbacks of the TransSearch approach are that the user must enter search strings manually and cannot insert the results directly into the text to be translated (beyond traditional cut-and-paste methods). While TransSearch is an excellent reference tool, it cannot claim the degree of integration with the text that other tools can. Companies such as Terminotix and MultiCorpora have since designed tools that are capable of sub-sentential searching in bitexts and that are fully integrated with the word processor, which means that searches are automated and matches can be entered directly into the text. LogiTerm is the tool sold by Terminotix. MultiCorpora's contribution, MultiTrans, will be described in more detail in section 2.2.3.

2.2.1 Advantages of the CSB-based approach

The primary advantage of searching for character strings within a bitext instead of looking for matches in isolated sentence pairs is the preservation of context beyond the sentence level. Users maintain access to the global properties of a text, such as "who originally translated a document, when, for which client, as part of which project, who revised and approved the translation, etc." (Macklovitch and Russell 2000, p.9). Texts also have a global style. A user may need to verify that a translation proposed by the translation memory is drawn from a text whose style is similar to that of the new text to be translated. For a short segment, even a little bit of context around the proposed text may be helpful for validating the appropriateness of the proposed translation in a new context (Macklovitch and Russell 2000, p.9).

The bitext approach is particularly interesting for novice translators or translators working in new domains. Bitext can be used for preparatory background reading (Macklovitch and Russell 2000, p.9), and inexperienced translators benefit from access to sub-sentential information, which can provide clues for handling tricky expressions or structures.

If the text to be translated has a large proportion of repeated material concentrated in chunks, as may be the case for an update or revision, a bitext-based system has the advantage of being able to identify and process several consecutive identical sentences (or paragraphs, or even pages) at once (Macklovitch and Russell 2000, p.9). A sentence-based system is limited to evaluating one sentence at a time, which would waste time in this circumstance[5].

In a sentence-based system, accurate alignment is crucial. This means that significantly more time must be spent creating and maintaining databases. In contrast, when context is preserved—as it is in the CSB-based approach—a faulty alignment can be corrected with little extra effort during the translation process, and there is no danger of automatically inserting “false” 100% matches into a text (Arrouart and Bédard 2001, p.30).

[5] TRADOS does have a Translate-to-Fuzzy feature, which, when activated, replaces all 100% matches with their translations until it hits the next fuzzy match or non-match. However, TRADOS is unable to alert the user ahead of time that a string of consecutive 100% matches exists in a text. Furthermore, there is a possibility that TRADOS will draw 100% matches from multiple texts with different styles.

2.2.2 Disadvantages of the CSB-based approach

The CSB-based approach does have a few disadvantages when it comes to searching. A system designed to identify identical character strings is likely to miss some useful passages that would be picked up by a system that allows for fuzzy matching. Also, while it is an advantage to be able to pick up short strings below the sentence level, this is counterbalanced by the increased noise from unreliably small units. The system is likely to identify many two- or three-word strings, which the user must take time to evaluate, whether or not they turn out to be useful. Finally, the CSB-based approach makes it more difficult to recycle internal repetitions in a new text. With the sentence-based approach, new sentence pairs are added to and are accessible in the database as soon as they are translated. In the CSB-based approach, content from the new text cannot be searched in bitext format until the entire translation is complete and added to the database. Even if terms and expressions are added to term banks during the translation process, these will not be identified as internal repetitions unless the user repeats the search process after each addition.

This also presents a limitation on networking. In large projects where it is necessary to have several translators working at once, the CSB-based approach does not provide the same level of accessibility to one another's work as the sentence-based approach. Additionally, the TMX standard format was designed with the sentence-based approach in mind, so users of CSB-based tools may not be able to take advantage of it in the same way to share translation memories.

2.2.3 An example of the CSB-based approach: MultiTrans

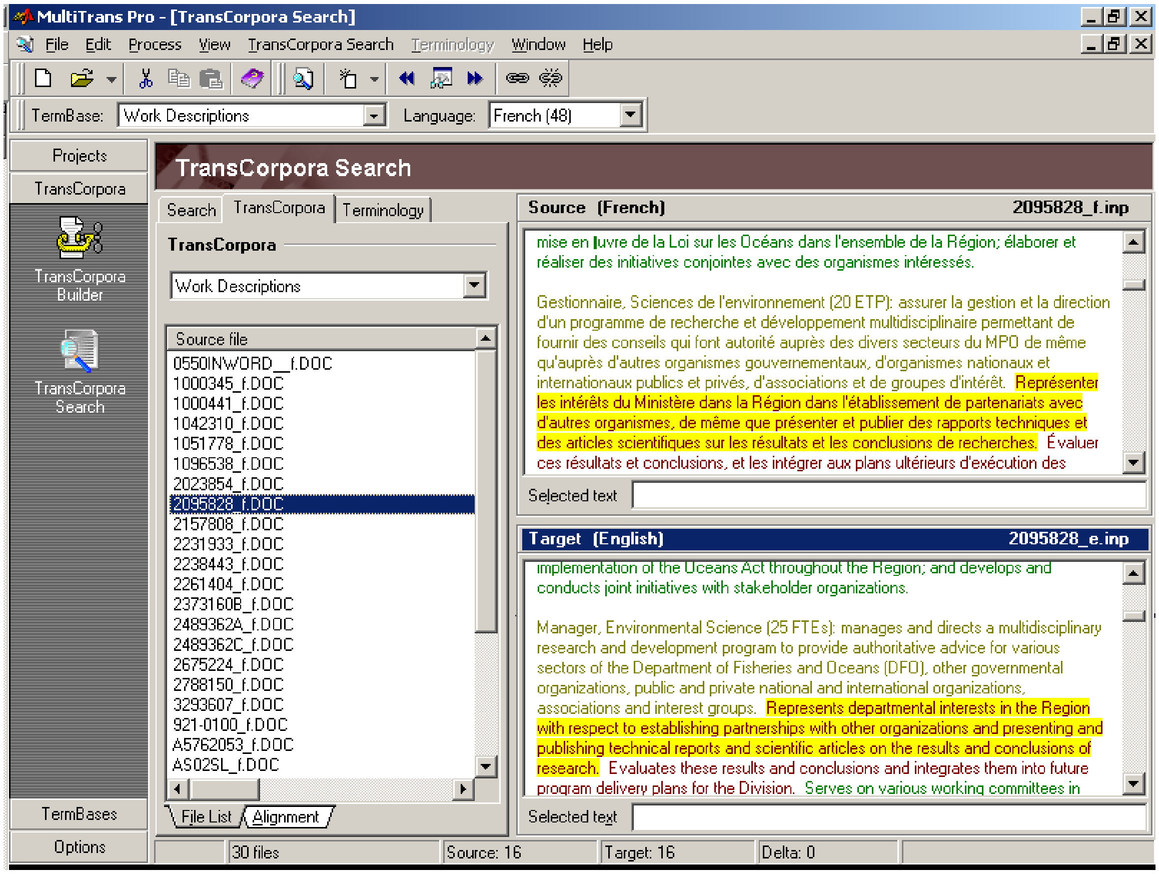
While there are many examples of sentence-based TM tools, MultiCorpora's MultiTrans is one of very few CSB-based tools currently available. It is important to note that MultiCorpora does not actually advertise its product as a TM tool. Its Web site states that MultiTrans uses "full-text corpus technology" that provides "translation support and language management solutions" (http://www.multicorpora.ca). Many potential clients associate TM with the narrow definition described in section 2.1, so this is a way to differentiate MultiTrans from the sentence-based TM tools on the market. I refer to it as a type of TM tool in this thesis on the basis of the broader definition of TM given in section 2.2, that is, "an archive of past translations, structured in such a way as to promote translation reuse". MultiTrans is integrated with the word processor and features a strong automatic search component, which makes it an excellent illustration of the CSB-based approach to searching a database of previously translated texts.

As with TRADOS, the following description of MultiTrans is based on my personal experience with the tool, with reference to the MultiTrans user manuals. MultiTrans comprises three modules: the TransCorpora module, the TermBase module and the TransTerm module. The TransCorpora module is used to index and align reference documents and to search for expressions in and extract terminology from the resulting bitexts[6]. The TermBase module is used to create terminology banks. The TransTerm module is a menu in Microsoft Word that connects the document to be translated to all of the reference material stored in MultiTrans (MultiCorpora, p.13).

[6] MultiCorpora calls these source and target text pairs "TransCorpora", but I will continue to use the more generic term "bitext".

The indexing function of the TransCorpora module is called TransCorpora Builder. The user designates the source language, the target language and the pairs of texts that will form bitexts. At this point, the user can also choose to create a terminology extraction file, which is a list of all the expressions “containing two words (or more) with a frequency of more than 1 in the reference documents” (MultiCorpora, p.33).

Also part of the TransCorpora module is the MultiTrans search environment, which is called TransCorpora Search (see Figure 2-4). On the right side of the screen are two windows displaying the source and target portions of the selected bitext. These are aligned sentence by sentence. As the user scrolls down through the source text, each sentence is highlighted in turn, along with its proposed alignment in the target window. Sometimes the alignment is off by one or more sentences, but this is easy to correct manually and does not cause problems during the translation process. On the left side of the screen, the user can choose between three tabbed windows: a search window for performing manual searches in the database, a TransCorpora window pointing to all of the bitexts in the open TransCorpora file, and a Terminology window containing the terminology extraction file.



TermBase: Work Descriptions

Language: French (48)

TransCorpora

Work Descriptions

Source file

0550INWORD__f.DOC
1000345_f.DOC
1000441_f.DOC
1042310_f.DOC
1051778_f.DOC
1096538_f.DOC
2023854_f.DOC
2095828_f.DOC
2157808_f.DOC
2231933_f.DOC
2238443_f.DOC
2261404_f.DOC
2373160B_f.DOC
2489362A_f.DOC
2489362C_f.DOC
2675224_f.DOC
2788150_f.DOC
3293607_f.DOC
921-0100_f.DOC
A5762053_f.DOC
AS02SL_f.DOC

Source (French) 2095828_f.inp

mise en ▌uvre de la Loi sur les Océans dans l'ensemble de la Région; élaborer et réaliser des initiatives conjointes avec des organismes intéressés.

Gestionnaire, Sciences de l'environnement (20 ETP): assurer la gestion et la direction d'un programme de recherche et développement multidisciplinaire permettant de fournir des conseils qui font autorité auprès des divers secteurs du MPO de même qu'auprès d'autres organismes gouvernementaux, d'organismes nationaux et internationaux publics et privés, d'associations et de groupes d'intérêt. Représenter les intérêts du Ministère dans la Région dans l'établissement de partenariats avec d'autres organismes, de même que présenter et publier des rapports techniques et des articles scientifiques sur les résultats et les conclusions de recherches. Évaluer ces résultats et conclusions, et les intégrer aux plans ultérieurs d'exécution des

Selected text

Target (English) 2095828_e.inp

implementation of the Oceans Act throughout the Region; and develops and conducts joint initiatives with stakeholder organizations.

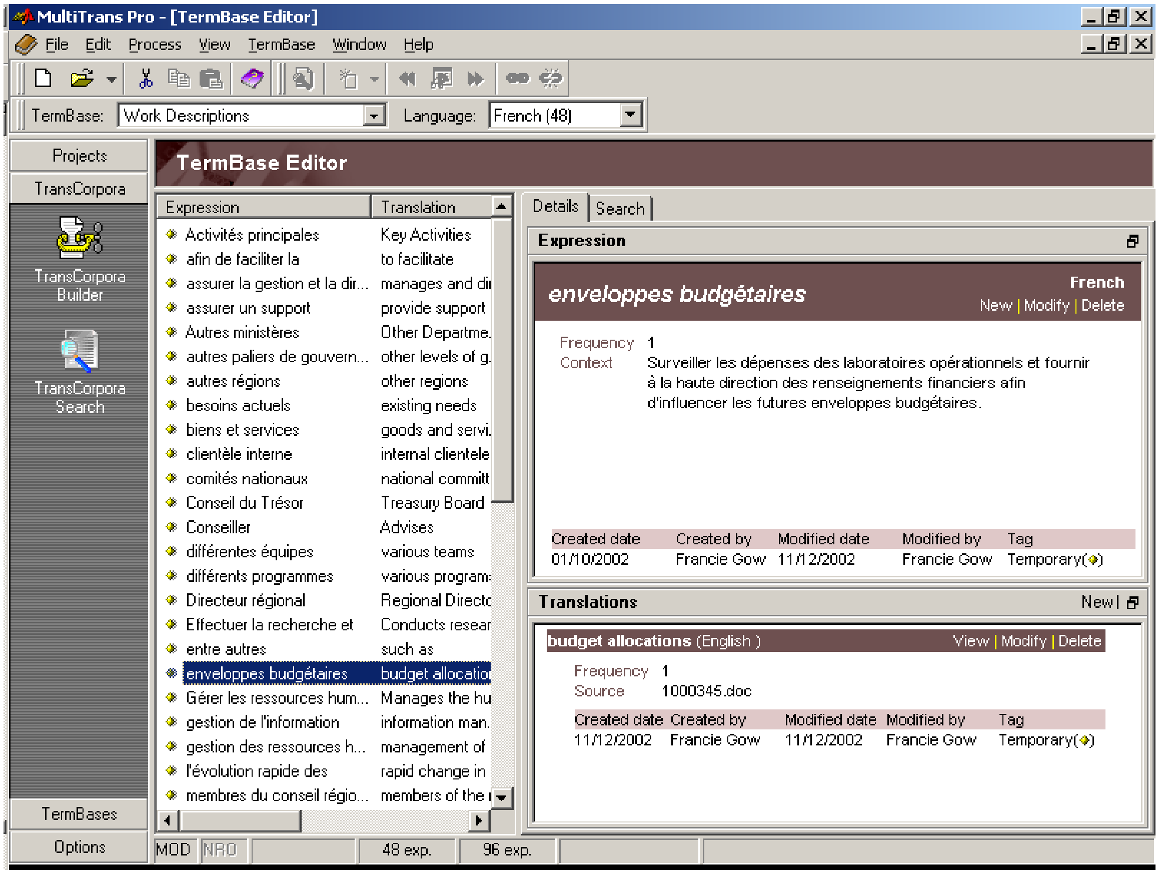
Manager, Environmental Science (25 FTEs): manages and directs a multidisciplinary research and development program to provide authoritative advice for various sectors of the Department of Fisheries and Oceans (DFO), other governmental organizations, public and private national and international organizations, associations and interest groups. Represents departmental interests in the Region with respect to establishing partnerships with other organizations and presenting and publishing technical reports and scientific articles on the results and conclusions of research. Evaluates these results and conclusions and integrates them into future program delivery plans for the Division. Serves on various working committees in

Selected text

30 files | Source: 16 | Target: 16 | Delta: 0

Figure 2-4 TransCorpora Search module of MultiTrans

The TermBase module allows the user to create and search in terminology banks (see Figure 2-5). On the left side of the screen is an alphabetical list of all the expressions stored in the TermBase. On the right are two tabbed windows, a Search window for entering manual queries, and a Details window displaying the individual term records, which can include user-defined fields. Terminology can be imported from external files, or it can be inserted directly into the TermBase from the reference material in the TransCorpora module.



TermBase: Work Descriptions Language: French (48)

Expression	Translation
Activités principales	Key Activities
afin de faciliter la	to facilitate
assurer la gestion et la dir...	manages and di
assurer un support	provide support
Autres ministères	Other Departme.
autres paliers de gouvern...	other levels of g.
autres régions	other regions
besoins actuels	existing needs
biens et services	goods and servi.
clientèle interne	internal clientele
comités nationaux	national committ
Conseil du Trésor	Treasury Board
Conseiller	Advises
différentes équipes	various teams
différents programmes	various program:
Directeur régional	Regional Directo
Effectuer la recherche et	Conducts resear
entre autres	such as
enveloppes budgétaires	budget allocatio
Gérer les ressources hum...	Manages the hu
gestion de l'information	information man.
gestion des ressources h...	management of
l'évolution rapide des	rapid change in
membres du conseil régio...	members of the i

Expression

enveloppes budgétaires — French — New | Modify | Delete

Frequency 1
Context Surveiller les dépenses des laboratoires opérationnels et fournir à la haute direction des renseignements financiers afin d'influencer les futures enveloppes budgétaires.

Created date	Created by	Modified date	Modified by	Tag
01/10/2002	Francie Gow	11/12/2002	Francie Gow	Temporary(◆)

Translations New |

budget allocations (English) View | Modify | Delete

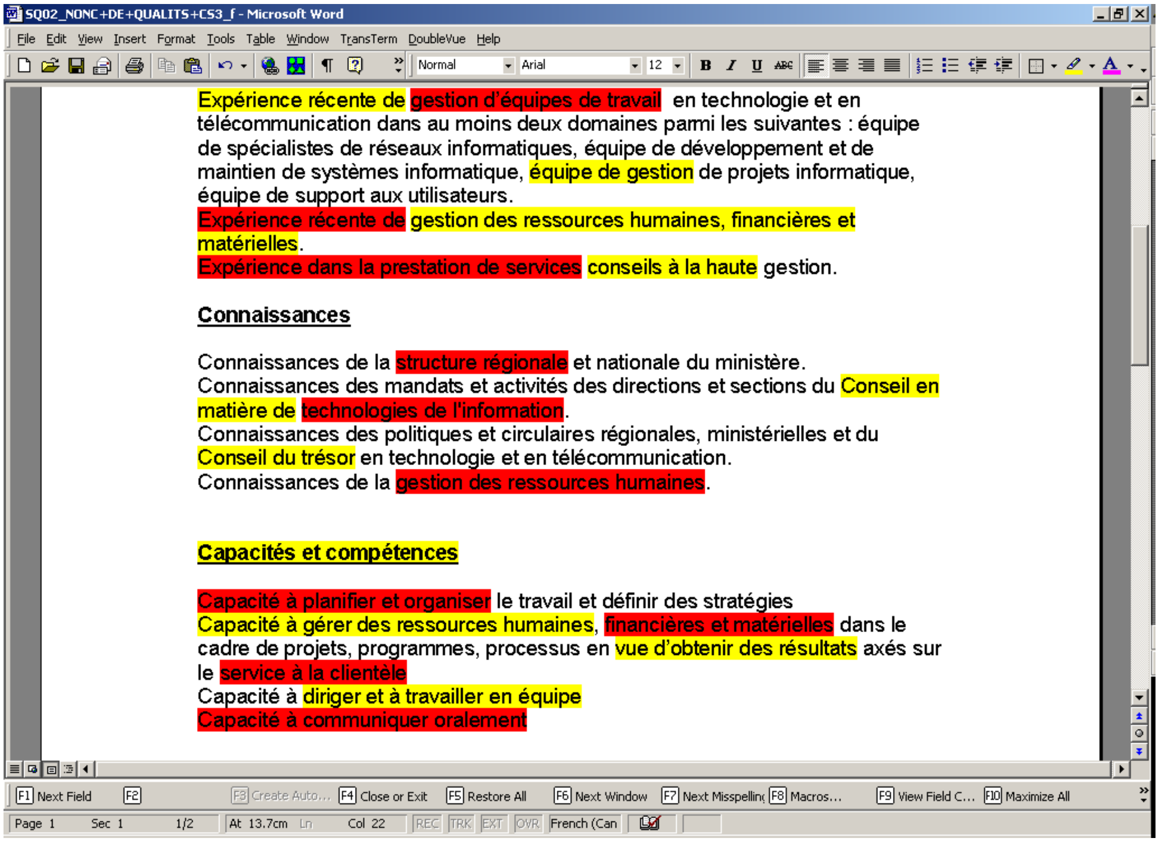
Frequency 1
Source 1000345.doc

Created date	Created by	Modified date	Modified by	Tag
11/12/2002	Francie Gow	11/12/2002	Francie Gow	Temporary(◆)

48 exp. 96 exp.

Figure 2-5 TermBase module of MultiTrans

The TransTerm module serves as the link between the translation environment (Microsoft Word) and the other two modules. Upon opening a new source document in Word, the user connects to the appropriate database in MultiTrans. The two principle functions that can be carried out on the source document are the TermBase Process and the TransCorpora Process. The first is essentially a pretranslation process; it identifies all of the expressions in the document that occur in the TermBase and automatically replaces them with their equivalents (in blue text for easy identification). The second compares the source text with the indexed reference material and highlights every matching character string containing at least two words (see Figure 2-6).



Expérience récente de gestion d'équipes de travail en technologie et en télécommunication dans au moins deux domaines parmi les suivantes : équipe de spécialistes de réseaux informatiques, équipe de développement et de maintien de systèmes informatique, équipe de gestion de projets informatique, équipe de support aux utilisateurs.
Expérience récente de gestion des ressources humaines, financières et matérielles.
Expérience dans la prestation de services conseils à la haute gestion.

Connaissances

Connaissances de la structure régionale et nationale du ministère.
Connaissances des mandats et activités des directions et sections du Conseil en matière de technologies de l'information.
Connaissances des politiques et circulaires régionales, ministérielles et du Conseil du trésor en technologie et en télécommunication.
Connaissances de la gestion des ressources humaines.

Capacités et compétences

Capacité à planifier et organiser le travail et définir des stratégies
Capacité à gérer des ressources humaines, financières et matérielles dans le cadre de projets, programmes, processus en vue d'obtenir des résultats axés sur le service à la clientèle
Capacité à diriger et à travailler en équipe
Capacité à communiquer oralement

Figure 2-6 The result of a TransCorpora Process carried out on a new source text

To see how a particular string has been translated before, the user can select it and click "Fetch from TransCorpora" from the TransTerm menu. This will bring the user directly to the TransCorpora Search screen, where the sentence containing the character string will be displayed in the source window, with the aligned sentence displayed in the target window. The user then scans the target text for the appropriate equivalent, which may be at or below the sentence level, and inserts the equivalent directly into the Word document. The source and target expressions can be inserted into the TermBase at the same time.

Any insertions in the new document will be coloured blue, whether they come from a TermBase or a TransCorpora Process. The translator translates the remaining text, then resets the original colours with a command from the TransTerm menu. The new text and its translation can then be indexed and added to the TransCorpora file to be used as reference material for the next translation. In the meantime, the TermBase has grown larger (and in theory, more useful) with every insertion.

The differences between the two approaches described in the chapter make the task of comparative evaluation difficult. The next chapter will examine whether traditional evaluation methods can adequately deal with these differences and, if not, how they can be modified to do so.

Chapter 3 Existing Research in TM Evaluation

Evaluation of translation tools is of interest to everyone involved in the translation process, from the tool developers and the translators to the clients who pay for the translated documents. It is relevant at numerous stages. During development, researchers use ongoing evaluation methods to find ways to improve emerging tools. Before purchasing a tool, buyers must evaluate whether a single tool will meet their requirements or evaluate which of many similar tools meets those requirements best. Even after a tool is integrated into a translation work cycle, the users must evaluate whether it is meeting those requirements adequately, and if not, what to change.

Unfortunately, comprehensive evaluation of a computer-assisted translation tool is costly and time consuming, with no guarantee that the ultimate findings will be satisfactory. Furthermore, each potential user has such different needs that there is no detailed general methodology that can be universally applied (Cormier 1992, p.384).

Although many potential buyers of translation memory tools will ask which of the available tools is "best", the question they should really be asking is which tool is best suited to their particular circumstances (Zerfaß 2002a, p.49). The answers will vary from buyer to buyer. In fact, even a single buyer's requirements may change from one job to another, depending on the text types in question and the various clients' preferences. A prioritized list of context-specific requirements is the essential foundation for any evaluation attempt.

Evaluation strategies are divided into two major categories: black box evaluation and glass box evaluation. Trujillo (1999) differentiates these in the following manner:

> In black box evaluation, the [machine-assisted translation] system is seen as a black box whose operation is treated purely in terms of its input-output behaviour, without regard for its internal operation. Black box evaluations are particularly suited to evaluation by users and translators. […] By contrast, in glass box evaluation the various components that make up the system are inspected and their effect on the overall workings of the system is assessed. Glass box evaluation is particularly relevant to researchers and developers, who need to identify modules whose operation is deficient in some way (p. 256).

It is important to note that all of the evaluation methods discussed in this thesis are of the black box variety, since I am approaching this evaluation as a user/translator, and users/translators make up the target audience of this study.

3.1 General evaluation of TM tools

A few authors discuss general methodologies for TM evaluation. In 1993, King outlined four essential steps:

> Il faut d'abord identifier les besoins qu'on espère satisfaire et ce qu'on attend du système. Cette réflexion amènera à la définition d'une liste de critères qu'on ordonnera en fonction de leur importance relative. Ensuite, vient la recherche d'une technique qui permet de réunir des données relatives à chacun de ces critères […]. La troisième étape rassemble ces données et leurs analyses. La dernière formule un jugement sur la base des informations fournies par l'analyse des données recueillies (p.266).

This is a good start, but clearly requires further development. In 1995, the Expert Advisory Group on Language Engineering Standards (EAGLES) published a report that describes a more rigorous evaluation method for TM tools[7].

Rico (2000, p.36) summarizes the EAGLES methodology in her article entitled "Evaluation Metrics for Translation Memories". The first step involves identifying the context in which a tool will be used. Potential contextual features include translation volume, text types, languages and quality control management. Each feature must be

[7] http://issco-www.unige.ch/projects/ewg96/node157.html#SECTION00104300000000000000

assigned a value or weight relative to all the other features. Then the measurable attributes associated with each feature should be listed. Rico (p.37) offers a list of potential attributes including accuracy, security, efficiency, pricing policy, customization and updates. The final step is simply the execution of the evaluation. This particular approach is well suited to verifying the adequacy of a single tool, but it could also be adapted for comparison testing.

All of the above articles are strictly theoretical; the authors do not apply their methods to particular tools. Other authors attempt to identify particular measurable features of TM tools. Webb (1998) examines various scenarios of TM use from a cost/benefit perspective, while Lynch and Heuberger (2001) propose a systematic method for measuring the return on investment provided by TM tools.

Benis (1999 and 2000), Zerfaß (2002b) and Höge (2002) provide more specific information in the form of comparative analyses of a specific group of commercially available TM tools. Zerfaß's brief overview compares the tools based on TM model[8], translation environment, supported file formats, TMX compliance, fuzzy-match quality and handling of special elements such as abbreviations and acronyms. Benis's reviews are more detailed but less systematic, adopting a conversational style to discuss more or less measurable features of each tool, such as user-friendliness, cost, filters and potential for productivity increases. In her PhD thesis, Höge (2002) provides a more detailed and systematic comparative study of four different TMs, in which she focuses principally on user-oriented testing. For example, she examines criteria such as usability (e.g. by

[8] Zerfaß distinguishes between two models of TM tools: the database model and the reference model. These resemble the sentence-based and CSB-based approaches to a certain degree, but Zerfaß does not go beyond a simple description of them in her review.

investigating how easy it is for users to learn the tool and how many mistakes they make). However, all three of these authors deal exclusively with sentence-based tools, and none provide any suggestions for designing an approach that can be used to compare sentence-based and CSB-based approaches to TM.

3.2 Evaluation of automatic search and retrieval: edit distance

When Zerfaß (2002b) discusses fuzzy-match quality, she is referring to the validity of fuzzy matches, or whether they contain any material that can contribute to the translation of an input sentence. Another element in determining match quality is time. If two matches are equally valid, the better of the two is the one that saves the user the most time. Usefulness, then, is a function of both validity and time. In an attempt to measure fuzzy-match quality (one possible definition of usefulness) objectively, researchers have borrowed the concept of "edit distance", originally developed for spell-checking technology. Researchers have tried to find objective ways of measuring fuzzy-match quality and often use the concept of edit distance as a measurable approximation.

The US-based National Institute of Standards and Technology defines edit distance as the "smallest number of insertions, deletions, and substitutions required to change one string […] into another"[9]. This metric, or variants of it, is used within some existing TM tools to measure the distance between a sentence stored in the database and a new input sentence (Simard and Langlais 2000, p.2). In theory, a smaller edit distance implies a higher degree of similarity between sentences.

Edit distance could also potentially be used to approximate the usefulness of proposed translations by measuring them against a model translation. In this case, a small

[9] http://www.nist.gov/dads/HTML/editdistance.html

edit distance between a proposal and a model implies that the proposal is probably useful. There are, however, limitations to this approach in its simplest form, the first being the assumption that the model translation is the only valid translation (Simard and Langlais 2000, p.5). Akiba et al. (2001, p.2) address this issue in their study of automatic evaluation of machine translation (MT) output by incorporating several acceptable model translations into their metrics.

Even if there were only one correct model sentence, there still exists the problem of whether to calculate the edit distance between two sentences on a strictly character-by-character basis. This works well for spell-checkers but gives less accurate results in TM technology where formatting and word order must be taken into account. In an attempt to address this problem, Planas and Furuse (1999, p.332) have proposed an array structure, instead of a linear structure, for storing translation units. The top layer of the array is made up of individual text characters, and lower levels include whole words, lemmas, parts of speech, and formatting tags, among other possibilities. Each layer of a stored unit is compared to its equivalent layer in a potential match, thus producing more accurate results.

Finally, even if the relatively simple character-by-character comparison were valid, the exact formula for determining edit distance is always open to debate. Does insertion refer to the insertion of a single character or a whole word? What if two words are simply reversed? Does that count as two substitutions? In their study of MT evaluation, Akiba et al. (2001, p.3) use *sixteen* variations of an edit distance formula and take an average of the resulting scores.

There are two major advantages to using edit distance as a measure of TM output. The first is objectivity, since the measure is a statistical one. The second is reduced cost to the evaluator, since significant amounts of time and resources are required to implement subjective human evaluation of every TM tool to be tested. However, while a small edit distance is intuitively a good indication of a translation's usefulness, it remains an approximation and has the potential to generate both silence and noise. The evaluator must also be prepared to pay the high initial cost of programming an algorithm that can produce acceptable results.

3.3 Evaluation of related types of translation technology

Because TM technology is quite new, there has been relatively little work carried out to date with regard to the evaluation of TM tools. For this reason, it is worth investigating evaluation-related work that has been carried out on other types of translation technology. For example, machine translation (MT) is much more established than TM, so researchers working in this field may have grappled with some of these problems before. One kind of MT in particular, Example-Based Machine Translation (EBMT), is very similar to TM in its approach. Likewise, TransType is another Computer-Assisted Translation tool that resembles TM, and its creators have published an in-depth reflection on the evaluation processes they used while developing the tool.

3.3.1 Example-Based Machine Translation

The main difference between EBMT and TM is that in the former case, the machine selects the appropriate segment from the examples found and inserts it into the new text. With TM, it is still the human who makes the final decision. The three components of EBMT are “matching fragments against a database of real examples, identifying the corresponding translation fragments, and then recombining these to give the target text” (Somers 1999, p.116). The first component is the most important here because it is the characteristic that EBMT shares with TM.

In a section entitled “Evaluating the Matcher”, Somers (1999, pp.148-149) illustrates two basic approaches to the evaluation problem: a subjective approach based on human ranking vs. an objective approach based on the calculation of edit distance. He also identifies the ideal measure of a matching algorithm: “In each case, an attempt is made to quantify not only the number of examples retrieved, ***but also their usefulness for the translator in the case of a TM***, or the effort needed by the next part of the translation process in the case of EBMT” (p.147) [emphasis added].

The subjective measures generally involve rating tools, where human evaluators assign a category to each example provided by the matching algorithm. The following examples illustrate this approach:

> Both Sato (1990) and Cranias et al. (1994) use 4-point scales. Sato’s “grades” are glossed as follows: (A) exact match, (B) “the example provides enough information about the translation of the whole input”, (C) “the example provides information about the translation of the whole input”, (F) “the example provides almost no information about the translation of the whole input”. Sato apparently made the judgments himself, and so was presumably able to distinguish between the grades. More rigorously, Cranias et al. (1994) asked a panel of five translators to rate matches proposed by their system on a scale ranging from “a correct (or almost) translation”, “very helpful”, “[it] can help” and “of no use”. Of course these evaluations could be subject to criticism regarding subjectivity and small numbers of judges (Somers 1999, p.147).

The advantage of this approach is that human translators are more likely than programs to be able to identify useful matches. One of the reasons that computers cannot measure usefulness accurately is that usefulness is partly relative to the individual translator.

Although a group of evaluators will likely agree on the majority of cases, there would probably be slight variations between the results. The larger problem is the time that it takes humans to perform a comprehensive evaluation. The objective, repeatable measures involve edit distance, but Somers confirms what was discussed in section 3.2, that there is no agreed upon formula for measuring edit distance and that it can never be more than an approximation for usefulness. However, if an approximation is all that is necessary, running an algorithm is certainly less time consuming than having humans perform the evaluation.

3.3.2 TransType

TransType is a kind of interactive machine translation software created by researchers at RALI[10] that calls on MT technology to predict what a translator will type next and display a proposal (Langlais et al. 2000). The translator can either ignore the proposal altogether and continue typing or accept the proposal to reduce the number of keystrokes necessary to complete the translation. The evaluation of TransType is not the same as the evaluation of a TM matching system, but it is relevant in that the researchers attempt to measure the usefulness of a translation-aid tool.

For the theoretical evaluation, the testers rely on an automatic measure of the number of keystrokes saved, assuming that the hypothetical translator accepts each correct proposal as soon as it appears. However, they also perform a more involved and realistic evaluation with human translators in an attempt to measure the tool's usefulness more accurately. A qualitative survey is described, but it is the quantitative analysis that

[10] Le Laboratoire de Recherche Appliquée en Linguistique Informatique, Université de Montréal

is most relevant here. The testers calculate a final score of efficiency, which is defined as the ratio of productivity over effort. The latter values are calculated as follows:

> The productivity is computed as the typing speed of a subject, that is, the ratio of the characters produced in the translation over the time spent to accomplish it. […] The effort is the ratio of any action (keystrokes or mouse click) produced over the time spent to translate (Langlais et al. 2000, p.645).

This idea of taking into account the number of primitive actions required by the user to produce a certain result is important and is easily applicable to the evaluation of TM software.

3.4 Relevance of existing research to this thesis

Although none of the research described above accomplished exactly what I am seeking to do (i.e. develop a methodology for comparing sentence-based and CSB-based approaches to TM), it did provide useful guidance for designing my own methodology. The following sections will describe some of the ways in which I considered adapting existing research methods to my own work and, where relevant, my reasons for rejecting existing research methods in favour of a newly designed approach.

3.4.1 General framework

Of the general evaluation research, the EAGLES report will be the most useful for establishing a framework for the evaluation methodology. The first requirement is to define the context in which the tools will be used. I have adapted this interpretation of context to mean the context in which the evaluation methodology will be developed. The evaluation that will be performed for this thesis will focus on one particular feature of TM tools, namely the automatic search-and-retrieval function, will be comparative in nature, and must be flexible enough to account for differences in the presentation of data by each tool. It will be applied to TRADOS and MultiTrans in this case but should not be limited to these tools. The next step is defining the measurable attributes associated with the features to be tested. There is only one feature in this case; therefore relative weights or values need not be determined. The measurable attributes associated with the automatic search-and-retrieval function might include the time it takes to generate output, the validity of the output, and the time gained or lost by the user as a result of evaluating and/or incorporating the output. This essentially comes down to an attempt to measure the usefulness of the output of a given TM tool.

For measuring the usefulness of the output, the research described above suggests two possible avenues, as outlined by Somers: a more objective approach involving edit distances or a more subjective approach involving human rating systems.

3.4.2 Exploring the objective approach

A strong argument in favour of applying edit distance is that it makes it possible to automate the evaluation process. This allows users to measure large quantities of data with lower cost, and human testers are spared from a potentially tedious job. Unfortunately, there is a high cost involved in creating the program in the first place, and it was not within the scope of this thesis to design an edit distance program from scratch.

However, I did attempt to exploit the tools I already had at my disposal. TRADOS, like many TM tools, has a built-in edit distance program for measuring fuzzy matches. A 96% match is supposed to be more useful than a 62% match because the edit distance is much smaller. TRADOS normally compares an input source sentence to the source units stored in the translation memory database. I considered feeding the initial output of each tool back into TRADOS, but this time artificially comparing it to the model translations of the source texts. That would generate a percentage match for each sentence of output, compared with a previously approved translation of the input sentence.

This idea was ultimately rejected for two reasons. Firstly, it assumed that the model translation was the only possible translation, and this was unrealistic. Secondly, it was biased towards the sentence-based approach, since the fuzzy matches produced by that approach are complete sentences in the target language. The CSB-based approach may generate sentences with sub-segments replaced here and there, leaving some parts in the source language. The TRADOS edit distance algorithm was not designed to accommodate mixed-language input.

3.4.3 Exploring the subjective approach

Sato's grading system, as described by Somers (1999, p.147) in section 3.3.1, seemed like another useful starting point for comparing the output of the two tools. However, this method was also biased towards a sentence-based approach. The output of one approach is characterized by full sentences and the output of the other by terms and short phrases; therefore the grades cannot be applied equally to both.

This brings up another potential problem: how is a "hit" to be defined? A fuzzy or exact match produced by TRADOS in the form of a sentence can logically be considered a hit. However, if the same sentence is processed in MultiTrans, and three separate "chunks" of that sentence are replaced, does that constitute a single hit or three? The current evaluation methods are all based on the sentence-based approach and do not account for this distinction.

A final problem with this approach for measuring usefulness is that it can only be applied after the human evaluator has seen the proposal made by either tool. One of the factors evaluators must consider in deciding the usefulness of a proposal is whether they already knew the information or whether they would have had to spend time looking for it. Once a proposal is in their heads, it is impossible to mentally "erase" it, making it difficult in many cases to judge whether they would have considered that suggestion before it was proposed.

Despite these problems, many elements of this type of grading system can still be incorporated into a more valid, reliable methodology that accounts for the fundamental differences between sentence-based TM and CSB-based TM. An adapted methodology will be discussed in detail in Chapters 4 and 5.

3.4.4 Adequacy testing vs. comparative testing

Much of the research that has been done on the evaluation of TM tools involves establishing criteria for adequacy in a given context and determining whether a given tool meets these criteria. Of the comparative studies that exist, all are limited to the comparison of various sentence-based tools. What is missing is a method for comparing the output of sentence-based tools with that of their CSB-based counterparts. This thesis represents an attempt to go some way towards filling this gap, and the initial evaluation methodology that I designed will be described in the following chapter.

Part II

Chapter 4 Designing a Methodology for Comparing Sentence-based to CSB-based TM

When a TM tool receives a new input text, it searches its database of previously translated material and proposes to the user chunks of text that meet certain criteria. These criteria are designed to more or less approximate the concept of usefulness. Because each approach uses different criteria to perform this function – the sentence-based approach uses fuzzy matching and attempts to match chunks of text at the sentence level, while the CSB-based approach looks for repeated character strings of any length – each may offer different results, even given the same data. One logical basis of comparison for the two approaches would be the proposed translations that are output by each, given identical content in each TM database and identical input texts for translation.

As outlined in section 3.4.4, there currently exists no recognized methodology for comparing the usefulness of the output of these two underlying approaches to TM. In an effort to bridge this gap, this thesis aims to propose a valid and reliable evaluation methodology that is general enough to be applied fairly to both a sentence-based tool and a CSB-based tool. As described in section 0.3, the overall approach used to develop this methodology consists of two main steps: 1) the design of an initial methodology based on a literature survey of existing evaluation techniques and a pilot study; and 2) the refinement of this initial methodology based on the results of lessons learned during the pilot study. The literature survey has already been discussed in Chapter 3. The remainder of this chapter will focus on the pilot study, while the refinement of the methodology based on the outcome of the pilot study will be described in Chapter 5.

4.1 Choice of representative tools

To design and test the methodology, it was necessary to work with representative tools from each of the two categories. TRADOS 5.5 was chosen to represent the sentence-based category of TM tools, and MultiTrans 3.0 was chosen to represent the CSB-based category. The reasoning behind the selection of these two particular tools is described in detail in section 0.4.

4.2 Pre-conditions for designing a methodology to compare sentence-based and CSB-based approaches to TM

To make the comparison as fair as possible, exactly the same input data must be used in both MultiTrans and TRADOS. Input data takes two forms: the corpus of bitexts that makes up the TM database and the new source texts to be translated.

With regard to the corpus, it must first be noted that MT and CAT systems can be evaluated with a test suite, a test corpus or a combination of the two. A test suite is a "carefully constructed set of examples, each testing a particular linguistic or translational problem" (Trujillo 1999, p.257). A test suite is particularly suited to the type of diagnostic study performed to identify problems in a tool that need addressing, perhaps by the program developer. A test corpus, in contrast, is constructed of real texts, possibly specialized, which are used as input to the tool being evaluated (Trujillo 1999, p.257).

Although a test corpus approach is not always as systematic as a test suite approach, a sufficiently large corpus can offer a good reflection of how a tool might perform in a real situation. It also has the advantage of being easier to construct. To evaluate the adequacy of a tool, a case can be made for using either approach, or both.

However, for a purely comparative study, the test corpus is entirely suitable on its own. Furthermore, unlike MT systems, TM tools need not actively solve translation problems. They simply match patterns in a corpus, making the test corpus an obvious choice. Zerfaß (2002b) takes this approach in her comparison of sentence-based TM tools. She recommends that potential users take their time to

> evaluate the tools and use some real-life examples, not the sample files that are provided with the tools. They are great for getting to know how the tools work, but they do not give you the real-life picture (p.14).

For these reasons, the methodology developed for this thesis involves the use of a test corpus rather than a test suite. The following sections will describe the design and construction of the pilot corpus and the selection of the new source texts to be used in the pilot study.

4.3 Corpora

Two corpora were required to carry out this project. Initially, a small pilot corpus was designed and constructed to act as a resource and test bed for designing the initial methodology. This corpus will be described in the rest of section 4.3. A larger test corpus, made up of entirely different texts, was then constructed in order to conduct more rigorous testing of the refined methodology. This larger test corpus will be described in section 5.1. In the case of both the pilot corpus and the test corpus, the bitexts were taken from the Central Archiving System at the Translation Bureau of the Government of Canada.

4.3.1 Pilot corpus: design and construction

The first step in creating a pilot corpus was to establish a set of criteria with which to filter the approximately three million bitexts stored in the Central Archiving System. Two groups of texts were required to carry out the pilot test: a set of bitexts with which to construct a corpus for the TM databases, and a smaller set of source texts of the same type to be used as input. In a real translation environment, these source texts would be the texts that the user wants to translate. It is not necessary for these texts to have been previously translated, but for testing purposes it is useful to have model translations for reference, so they were drawn from the Archive of bitexts along with the corpus texts.

4.3.2 Text selection

The criteria used to select corpus texts are listed and discussed in Table 4.1 below.

Table 4.1 Text selection criteria

Criterion	Comments
Quality	The quality of the matches proposed by any TM tool depends on the quality of the input. Although it is reasonable to assume that users will not always have flawless material in their databases, maintaining a high standard of quality in the test corpus allows more focus on the performance of each tool's search method. Using texts translated by human translators from the Translation Bureau ensures a minimum standard.
Language Direction	The language direction of the bitexts should be the language direction of the evaluator, in this case French to English.
Format	Both source and target texts must be available in electronic format to be used in any TM tool. This is true of many of the texts in the Central Archiving System. All texts must be in Microsoft Word format. Both MultiTrans and TRADOS are integrated with Word, and using documents created in this format avoids potential problems related to file conversion.
Subject Field	All texts in the pilot corpus, plus the associated source input texts, must belong to the same subject field. TM tools are designed to capitalize on repetition, and restricting the corpus to a particular field increases the chances of particular terms and phrases being repeated. In a regular evaluation context, it is not a problem if the tool generates few proposals. It simply means that the tool is performing poorly. However, when an evaluation methodology is being developed, it is essential to guarantee enough proposals with which to test ideas. Limiting the test to a single subject field is one strategy that can be used to address this issue. From the user's point of view, it also reduces the risk of homonymy (see section 2.1.2 for discussion).
Text Type	TM technology is helpful when the text type being used contains either internal or external repetition (see section 1.3). The text type chosen for *evaluation* purposes does not need to be highly repetitive (the majority of texts are not), but must be repetitive enough that the tools being tested are likely to generate proposals. The corpus should contain a single text type for the same reason that it should contain a single subject field – so that more proposals are likely to be generated.
Confidentiality	The documents contained in the Central Archiving System are not necessarily in the public domain. One of the conditions I had to respect in exchange for being given access to the system was that I use no confidential material in my examples, and that any names of individuals be censored. This is reasonable but restricts the choice of texts.
Number	30 bitexts + 3 source texts in the same field

4.3.3 Finding the texts

At the beginning of the project, I spent several weeks exploring the Central Archiving System in search of any group of texts that met all of the above criteria. There was no need to filter for quality, since all translations entered into the system had been translated and revised by humans, and met the standards of the Translation Bureau. It was possible to filter by file type, so the search was restricted to MS Word files from the start. There were no explicit filters available for text type or confidentiality; those I had to determine on a text-by-text basis. However, I was able to address this problem somewhat through keyword searches. Some, though not all, text types explicitly identify themselves. For example, the pilot corpus is composed of work descriptions, which usually contain the words "work description" somewhere in their titles or bodies.

There was no explicit filter for subject field, but there was a Service filter, which served as a close equivalent. For example, I could choose to browse only translations from Life Sciences or Criminology. This was my initial plan, but the lack of a filter for language direction posed a serious problem. The vast majority of translations done at the Translation Bureau are from English to French, which meant that I had to sift through several pages of hits before finding a single translation in my required language direction (French to English). It was obvious from the start that I would never find a large number of suitable texts of comparable text type in a reasonable amount of time, if at all, using this approach. The solution I settled on was to restrict my search to the Montreal Regional Service, one of the few services in which the majority of texts are translated toward English. The one drawback of this decision was that the texts listed in this service covered a vast range of subject fields and were not sorted. As was the case for text type and confidentiality, I relied on keyword searches to address the problem.

I eventually identified a suitable group of bitexts to create my pilot corpus. The small corpus was in the subject field of employment and consisted of 30 work descriptions, which were 3234 words long (10 pages) on average. The texts were divided into sections with similar headings, and complete sentences and paragraphs were used. There was a reasonable amount of repetition in the texts.

Once the general text type was determined, the texts had to be downloaded from the Central Archiving System and aligned in MultiTrans and TRADOS. This was an extremely tedious process. One of the Central Archiving System's limitations is that texts must be downloaded one at a time, each language separately. This does not usually pose problems because the texts are normally consulted on screen. It took a total of two hours to download the 33 bitexts required for the design phase (30 for the corpus and 3 to be used as input texts).

4.3.4 Building the corpora

Because the two TM tools store and process information differently, it was necessary to create two versions of the pilot corpus: one for use with MultiTrans and one for use with TRADOS. Note that the textual content of the two versions remained identical; the only difference was in the way the two tools stored this information[11].

[11] This is much like saving a text as both a Word and a WordPerfect file; the content remains identical, but the underlying file format differs.

4.3.4.1 MultiTrans

I chose to begin by building the corpora in MultiTrans for two reasons. Firstly, based on my experimentation with the two TM tools, I determined that the process is faster and easier in MultiTrans than in TRADOS. Secondly, the way in which MultiTrans displays its aligned texts allowed me to inspect all of my bitexts at a glance. This was important, since a small number of the files that I had downloaded turned out to be unusable. The most common reason was that both sides of the bitext appeared in the same language, an error generated when the texts were originally uploaded into the Central Archiving System. It was difficult to catch all of these instances while I was downloading files, so it was good to have a second opportunity to weed out and replace useless texts. By the time I began building the corpora in TRADOS, all of the corpus content problems were solved, which saved a lot of time.

The main reason that the corpus building process is faster in MultiTrans is that it is not necessary to verify the alignment. All one has to do is create a list of bitext pairs. In both TRADOS and MultiTrans, the automatic alignment is rarely perfect but generally very good. With TRADOS, a considerable amount of time must be spent checking the alignments before a database is used, since each misalignment makes the tool less useful at the time of translation. Misalignments in MultiTrans, however, do not diminish the tool's usefulness during the translation process, and they can be identified and repaired with minimal effort at any time.

Once all of the faulty bitexts were discovered and replaced (a process that took several hours), the process of generating the MultiTrans version of the pilot corpus was very short, requiring only ten minutes.

4.3.4.2 TRADOS

Using the same set of bitexts, I employed TRADOS WinAlign to generate a second version of the pilot corpus. Identifying bitexts is very simple in WinAlign; any number of texts can be paired up with a single mouse click. Verifying the alignments within the texts, on the other hand, is time consuming, and longer texts require more time than shorter texts. One unpleasant feature of WinAlign is that it is relatively easy to click the wrong button and undo several minutes' worth of effort, especially during the learning process. Once I became accustomed to the tool and stopped having to repeat my work, I was able to align the TRADOS version of the pilot corpus in about an hour.

It is important to note that I did not take the time to align the bitexts perfectly. This is a very time-consuming process, which can involve correcting misalignments, merging segments that have been inappropriately split and editing the content or formatting of source and target units. The amount of time spent maintaining the database affects the retrieval results, but it is unlikely that all TRADOS users maintain their databases to perfection. In an attempt to reflect realistic usage of the tool, I corrected all misalignments, merged most of the inappropriately split source segments (merging target segments is not necessary for the tool to function well), and occasionally edited source and target units.

4.4 Pilot test: designing the metrics

The 1995 EAGLES report[12] identifies, among others, the following properties of a good evaluation method:

[12] http://issco-www.unige.ch/projects/ewg96/node155.html#SECTION00104100000000000000

- *reliable*: an evaluation should produce similar or identical results when repeated in the same context by different evaluators
- *valid:* end users should be able to infer from the measurement values obtained what effect the tool will have on their productivity
- *efficiently applicable*: the evaluation should be performable with the least effort possible, especially by the end users

Objective approaches to evaluation as described in Chapter 3 are generally strong in the first and third categories (reliability and efficient applicability), but weak in the area of validity. Subjective approaches are generally stronger in their validity, are perhaps slightly weaker in their reliability (although this can be prevented through good design) and are generally weakest in their efficient applicability. In developing a new methodology that can be used to compare sentence-based and CSB-based TM tools, priority will be given to validity and reliability, although an effort will be made to respect all three properties as much as possible.

4.4.1 A possible solution to the "subjectivity vs. objectivity" problem

It became obvious from early considerations of the objective and subjective approaches that the subjective approach would make the best starting point, despite the hurdles associated with it (see section 3.4.3) The biggest problem that needed to be solved was the difficulty in determining how useful a proposal might be to the translation process after being influenced by that proposal. A solution to this apparent paradox was to shift the determination of usefulness to an earlier stage of the evaluation process, preferably before the application of the TM tools.

To accomplish this, an analysis and mark-up procedure was developed in which the evaluator reviews the input text independently of the TM tool and decides which items he or she would take the time to research in a real translation situation. The items in question may be terms, titles, phrases, sentences or paragraphs. In the context of this thesis, I myself acted as the evaluator, and I identified three reasons that would prompt me to mark an item in the input text:

a) I do not know how to translate it.
b) I could come up with a reasonable translation myself, but I suspect that there might be a similar or identical item in the translation memory. If there is, I want to make sure that I am consistent with it.
c) I am relatively sure that I know what the translation is, but I would feel more confident if this could be verified by another source.

Terms and short phrases could fall under all three scenarios. When entire sentences are marked, they usually fall under the second scenario. See Appendix B for an example of a marked up input text. For the work descriptions, the analysis and mark-up procedure required 5-10 minutes per page.

This step provides an unbiased snapshot of what proposals one translator would consider useful, against which the output of the tools can be measured. Of course, this snapshot would vary depending on the translator creating it, but for comparative purposes, it is enough that the same snapshot be applied to all TM tools being tested. It essentially provides a checklist of questions to be answered. Scores can be generated depending on whether Tool X answers each question completely, partially, or not at all, and these scores can be compared formally against a similar set of scores generated for Tool Y.

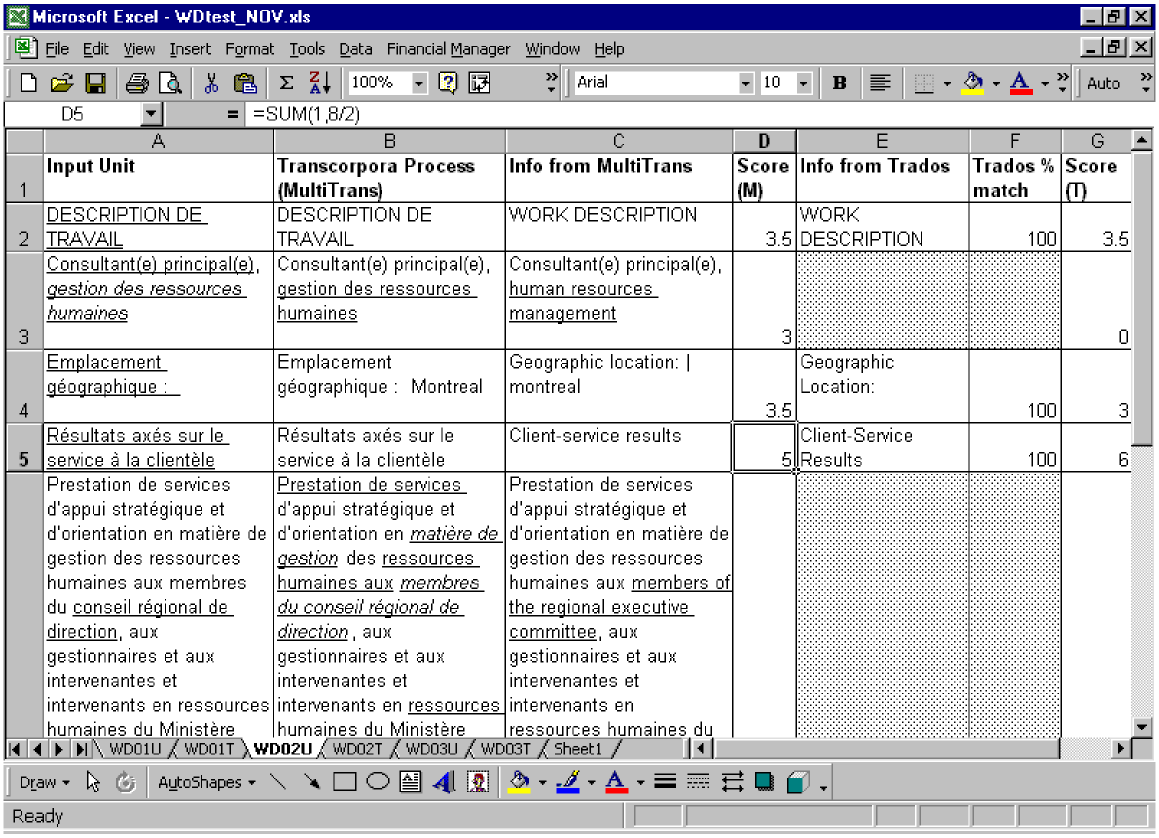
Retaining an element of human judgment in the process provides the groundwork of validity for any usefulness scores that will later be applied. However, shifting the most subjective phase of the testing to the earliest point, before the tools are used, limits any bias and allows the rest of the evaluation process to be more evenly applied to the tools in question. Once the input texts have been analyzed and marked up, the scoring system applied to the output of each tool can be designed to be repeatable by any evaluator, hence more reliable.

4.4.2 Principal difference between output of each tool

As explained in sections 3.4.2 and 3.4.3, the sentence-based and CSB-based approaches to automatic search and retrieval result in fundamentally different presentations of output. When an input text is pre-translated in a sentence-based tool (without any “improvements” on the part of the user), the result is a text in which all the sentences for which no match is found remain in the source language, and the rest are entirely in the target language. If, on the other hand, one were to pre-translate a text using a CSB-based tool, accepting proposals for every character string that matches something in the database regardless of length, the resulting text would contain hybrid sentences, with some chunks in the source language and other chunks in the target language. It is essential to account for this difference in applying a fair comparative test.

4.4.3 Recording the output of each tool

At this point, I had to decide on a basic format for recording evaluation results. I made an early attempt to record data in a Microsoft Access database. However, I required enough flexibility to make frequent changes to the layout of information during the pilot phase, and the database structure proved to be too rigid to make changes easily. I decided to store my data in a Microsoft Excel spreadsheet instead. The entire pilot test constituted a single project, and each of the three input texts was assigned its own worksheet. The input texts were labelled WD01, WD02 and WD03, with WD denoting the text type (work description). Figure 4-1 illustrates part of a sample worksheet. Column A is reserved for units of the source text, Columns B through D are reserved for scoring MultiTrans output and Columns E through G are reserved for scoring TRADOS output.


Microsoft Excel - WDtest_NOV.xls
File Edit View Insert Format Tools Data Financial Manager Window Help
100%
Arial
10
Auto
WD01U
WD01T
WD02U
WD02T
WD03U
WD03T
Sheet1
Draw
AutoShapes
Ready

D5 = =SUM(1,8/2)

	A	B	C	D	E	F	G
1	**Input Unit**	**Transcorpora Process (MultiTrans)**	**Info from MultiTrans**	**Score (M)**	**Info from Trados**	**Trados % match**	**Score (T)**
2	DESCRIPTION DE TRAVAIL	DESCRIPTION DE TRAVAIL	WORK DESCRIPTION	3.5	WORK DESCRIPTION	100	3.5
3	Consultant(e) principal(e), *gestion des ressources humaines*	Consultant(e) principal(e), gestion des ressources humaines	Consultant(e) principal(e), human resources management	3			0
4	Emplacement géographique :	Emplacement géographique : Montreal	Geographic location: | montreal	3.5	Geographic Location:	100	3
5	Résultats axés sur le service à la clientèle	Résultats axés sur le service à la clientèle	Client-service results	5	Client-Service Results	100	6
	Prestation de services d'appui stratégique et d'orientation en matière de gestion des ressources humaines aux membres du conseil régional de direction, aux gestionnaires et aux intervenantes et intervenants en ressources humaines du Ministère	Prestation de services d'appui stratégique et d'orientation en *matière de gestion* des ressources humaines aux *membres du conseil régional de direction* , aux gestionnaires et aux intervenantes et intervenants en ressources humaines du Ministère	Prestation de services d'appui stratégique et d'orientation en matière de gestion des ressources humaines aux members of the regional executive committee, aux gestionnaires et aux intervenantes et intervenants en ressources humaines du				

Figure 4-1 Sample worksheet

4.4.4 Input units

The first column of each worksheet was reserved for recording the results of the input text analysis. The text had to be divided into manageable translation units (TUs) so that it could be distributed among the cells, and the sentence was the obvious choice. TRADOS is inflexible about its TU boundaries; output must be evaluated one sentence at a time. MultiTrans is more flexible about its TU boundaries, which can occur at, above or below the sentence level. The sentence was therefore fixed as the unit of evaluation for

MultiTrans as well, bearing in mind that compensatory factors would have to be built into the scoring system to make sure that this requirement never put MultiTrans at a disadvantage. (This did not turn out to be difficult.)

If a full unit of the input text was highlighted (indicating that a proposed translation of the entire segment would be considered useful), it was copied and pasted into a cell in the first column, labelled "Input Unit", and underlined in full. If only a subsection of a sentence was marked in the input text, the entire sentence was still pasted into the cell, but only the relevant subsection was underlined. This means that any unit that does not appear at all in the "Input Unit" column is completely ignored during the pre-translation process. This may appear to put MultiTrans at a disadvantage. However, it is a realistic reflection of the way in which a translator would use MultiTrans, since it is generally inefficient to look up and insert short segments that do not pose translation problems. Table 4.2 illustrates how input units are recorded.

Table 4.2 Samples of marked-up text recorded in input column

Marked-up Text	Equivalent in Input Text Column
J'ai eu l'occasion de formuler des commentaires sur cette description de travail.	J'ai eu l'occasion de formuler des commentaires sur cette description de travail.
Fournir des conseils et des orientations stratégiques en fonction des priorités et des objectifs régionaux et influencer les décideuses et décideurs sur l'évolution et le développement de la gestion des ressources humaines afin de supporter la prestation de services aux clients et l'atteinte des résultats.	Fournir des conseils et des orientations stratégiques *en fonction des* priorités et des objectifs régionaux et influencer les décideuses et décideurs sur l'évolution et le développement de la gestion des ressources humaines afin de supporter la prestation de services aux clients et l'atteinte des résultats.
Gestion des ressources humaines	<no entry>

4.4.5 TRADOS output

Other columns were reserved for recording the output generated by each of the two tools. This was relatively straightforward for TRADOS. First I set the minimum match rate to 70%, which is the system default and the rate recommended by the Translation Bureau[13]. I pre-translated the input text in TRADOS, accepting what I considered to be the best proposal (usually, although not necessarily, the one with the highest match percentage assigned to it), but not making any additional changes. If there was no match for a particular input unit, I left the corresponding "Info from TRADOS" cell blank. If multiple matches were found, I pasted the best one into the cell (see Column E in Figure 4-1).

4.4.6 MultiTrans output

The process was slightly more complex for MultiTrans, but still achievable when done in two columns (see Columns B and C in Figure 4-1). MultiTrans pre-translates an entire text at once, without making replacements right away, but uses highlights to indicate to the user which segments have exact matches in its translation memory database. In the first MultiTrans column, the source text was simply copied into the cell, and the highlighting was indicated[14]. Once the appropriate replacements were made, the mixed-language sentences were pasted into an adjacent MultiTrans column, with the replacements emphasized by underlining.

[13] Personal communication with André Guyon, IT Strategies, Translation Bureau, Government of Canada (October 2002)

[14] For ease of reading, the alternating red and yellow highlighting seen on the MultiTrans screen was replaced by alternating underlining and underlining/italics when transcribed into the Excel worksheet.

Choosing which highlighted segments to look up and which to ignore can be a subjective process, so I developed a selection process closely tied to the marked up input text. Under this system, a highlighted segment is looked up in the following cases:

a) An entire unit is highlighted (e.g. "Énoncé de qualités" is replaced with "Statement of qualifications");
b) The highlighted section corresponds to a section marked in the input text.

Consider the following example: "Représenter la région pour négocier des ententes ou régler des litiges avec les *organismes centraux*". If "négocier des ententes" is not underlined in the first column (Input Text), but "organismes centraux" is underlined (which means that it has been previously designated as a unit that would have been looked up by the translator), then only the latter is substituted. The first is ignored and is not eligible to receive points. It is assumed that the translator already knows the translation and can type it faster than he or she can look it up and paste it in. The resulting replacement will be "Représenter la région pour négocier des ententes ou régler des litiges avec les central agencies".

There are exceptional cases when the entire term is underlined, but no substitution is made because the text is divided inappropriately. Consider the following input unit: "Expérience dans la prestation de services conseils à la haute gestion."

This indicates that the desired item is "prestation de services conseils". A TransCorpora Process generates the following:

"Expérience dans la prestation de services *conseils à la haute* gestion."

The entire term is underlined but divided between two separate character strings. In this case, any segments fetched from the database will contribute nothing to the validation of the desired item.

Two more columns are reserved for recording the scores for each tool, but first an appropriate scoring system must be developed.

4.4.7 Measurable attributes

In Rico's list of measurable attributes (see section 3.1), two are directly related to the match identification and retrieval process: accuracy and efficiency. The former is a measure of "system performance in terms of precision (percentage of valid segments from all those retrieved) and recall (percentage of segments retrieved from all those valid in the TM database)" (2000, p.37). The latter is a measure of "time behaviour in terms of retrieval time" (2000, p.37). Precision and recall are related to the concepts of noise and silence described in section 2.1.2: a tool that generates noise has low precision, while a tool that generates silence has low recall.

Accuracy and efficiency are really about measuring usefulness and time saved, which are highly relevant to users. While usefulness and time saved are not exactly the same concepts, they are undeniably linked (see section 3.2). A match may be valid or invalid in a given context. If it is invalid, then it goes without saying that it is useless to the user, and it automatically wastes time. This is a serious charge, since one of the main purposes of TM technology is to speed up the translation process (see section 1.3.2). However, the inverse is not necessarily true. A valid match is useful if and only if it saves time: research time if it provides information the user does not already know, typing time if it is faster to insert a match than to translate it from scratch, or revision time if the reviser spends less time making the translation consistent with previous work or with the work of other team members. Even a valid match is useless if it takes more time to find and retrieve it than to translate the segment from scratch.

In designing a methodology to compare the usefulness of output from sentence-based searches vs. CSB-based searches, it is important to keep in mind this relationship between validity and time.

For any given chunk of text to be translated, each tool will produce one of five results:

1. No proposal
2. An invalid proposal (that necessarily wastes time)
3. A valid proposal that wastes time
4. A valid proposal that has no effect on time
5. A valid proposal that saves time

Category 5 is the only desirable result. Category 1 cannot be measured, Category 4 can be ignored, and Categories 2 and 3 must be penalized. The tool will require a certain amount of effort on the part of the user to generate results even in Category 5, and this effort must also be considered (in terms of time).

4.4.8 Scoring system

To develop a scoring system, the output units from TRADOS and MultiTrans were analyzed and divided into the following categories:

- Multiple words inserted with no changes required
- Multiple words inserted with minor changes required (punctuation or capitalization)
- Multiple words inserted with major changes required (change in word order, additions or deletions, etc.)
- Single word inserted with no changes required
- Single word inserted with minor changes required (punctuation or capitalization)
- Provides answers to questions indicated by underlining in Column A (Input Unit), even if major changes are required

The task at this point was to assign relative values to each category that reflected their usefulness to the user. As explained in the section above, usefulness is a function of validity, time gain and time loss.

A first attempt to account for all three elements in a single value for each of the six categories proved too complicated. It was almost impossible to apply consistently and did not clearly illustrate the relationship between the category and the score. For example, two output units might have the same low score, despite the fact that one was very useful and required a lot of time to implement, while the other did not require much time to implement but was only marginally useful to the user.

Separating validity, time gain and time loss into three separate scores was not ideal either, since validity had already been filtered earlier in the process. Of all the proposals offered by either tool, only those that were valid were inserted into the pre-translation in the first place. There was also a problem associated with all of the segments for which no proposal was accepted. They required time, but they were not accounted for on the scoring spreadsheet.

A solution to these problems was inspired by RALI's testing of TransType (see section 3.3.2), specifically their definition of effort as "the ratio of any action (keystrokes or mouse click) produced over the time spent to translate" (Langlais et al. 2000, p.645). Since time gain is only applicable to valid matches that are accepted by the user, a positive time-gain score could be applied to the output listed on the spreadsheets, while a separate negative time-loss penalty could be applied to every output unit. Time loss could be measured in two ways: time spent performing actions such as mouse clicks and time spent evaluating proposals by each tool[15]. The latter could be measured by assigning negative values based on the number and type of proposals presented by a tool for each input unit.

4.4.9 Time-gain scores to be tested with pilot corpus

The relative scores in Table 4.3 can be understood intuitively. Output that fits into any of these categories will save the user time, and for the first five categories, the descending scores clearly illustrate which types of input are the most useful. There is no need to consider how much time it takes to process each result, since that will be accounted for in the separate time-loss penalty.

Table 4.3 Time-gain scores to be applied to both tools

Output from TM tool	Score
Multiple words inserted with no changes required	3
Multiple words inserted with minor changes required (punctuation or capitalization)	2
Multiple words inserted with major changes required (change in word order, additions or deletions, etc.)	1
Single word inserted with no changes required	1
Single word inserted with minor changes required (punctuation or capitalization)	0.5
Provides answers to questions indicated by underlining in Column A (Input Unit), even if major changes are required	2

The sixth category is slightly different. Sometimes a match proposed by the tool is very awkward to insert into the input text, requiring a great deal of manipulation (and time) on the part of the user. If the user already knows how to translate the segment in question, it is usually faster in this case to type in the translation from scratch. However,

[15] For example, the more proposals a tool generates, regardless of whether these proposals are valid or constitute noise, the more time a user must spend evaluating which is the most useful match.

if the data in the Input Unit column indicates that the user would have spent time researching the segment in question, then any proposal that provides a satisfactory answer to the user's "question" must be saving some of the time required to do research.

4.4.10 Bonus points

Because of the comparative nature of this study, there was no need to set a benchmark score to measure a tool's adequacy. In fact, the possible range of scores will vary depending on the length of each input text being processed. The scores generated by TRADOS and MultiTrans (or any other TM tool that might be tested) are only significant in relation to each other.

With this comparative framework in mind, a bonus feature was added to the scoring system. If neither tool generates a proposal for a given input unit, there is no penalty, since it is possible that the information is simply not present in the TM database. If both tools generate proposals, these proposals will be scored normally based on the amount of time they save. However, in cases where one tool generates a valid proposal and the other does not, the latter tool has clearly missed information that exists in the database (an example of silence). In such cases, the tool that generates a proposal is awarded one bonus point over and above the score it receives based on the category of the output.

4.4.11 Time-loss penalties to be tested with pilot corpus

Two separate scoring tables must be used to measure time loss in TRADOS and MultiTrans because of the fundamental difference between the output units being measured. In both cases, the user deals with one chunk of text at a time. In TRADOS, this chunk is the unit that is opened by the tool for processing, usually a sentence. In MultiTrans, each character string identified as a match during the TransCorpora Process (see section 2.2.3) is a discrete chunk, which may be at, above or below the level of the sentence, and is marked by highlighting. Each character string highlighted in yellow on the screen constitutes a single chunk, and each character string highlighted in red on the screen also constitutes a single chunk.

Table 4.4 MultiTrans time-loss penalties

Amount of input unit highlighted	Proposals generated	Penalty	Justification
Full input unit highlighted as a single chunk. No evaluation time required to perform Fetch; it is the obvious choice under the circumstances	Fetch yields one match	-1	Evaluation time required to determine whether match is worth inserting
	Fetch yields multiple matches	-2	More evaluation time required to determine which, if any, match is worth inserting
Portion of input unit highlighted as a single chunk	No fetch performed	-1	Evaluation time required to determine whether it is useful to fetch
	Fetch yields one match	-2	Evaluation time required to determine whether match is worth inserting
	Fetch yields multiple matches	-3	More evaluation time required to determine which, if any, match is worth inserting
More than one consecutive input unit highlighted as a single chunk[16]	n consecutive units provided by MultiTrans	+n	Having two or more units inserted at once saves time, and generally requires less context verification

[16] The following observations are based on my own experience with MultiTrans. Occasionally, two heading-type units on a single line will be highlighted as one chunk and can be fetched together. The bonus applies to this scenario. Identical paragraphs (or even multiple sentences) are not highlighted as a single chunk in MultiTrans. Because they must be fetched one sentence at a time, the bonus does not apply to this

Table 4.5 TRADOS time-loss penalties

Proposals generated	Penalty	Justification
Unit opened, no match	-0.5	User does not know in advance which units have matches; must click to open each unit
100% match	-1	Takes a short time to insert, but no evaluation time required
One fuzzy match	-2	Evaluation time required to determine whether it is worth inserting
Multiple fuzzy matches	-3	Evaluation time required to determine which, if any, is worth inserting

4.5 Results of pilot study

Various shortcomings in the initial scoring system were revealed during its application to the pilot corpus, and improvements were made throughout the process to address these problems. At 14 pages, WD03 was the longest source text, so it was processed first to generate the maximum number of proposals. All the major shortcomings were identified during the processing of this one text, so the shorter WD01 and WD02 were simply used as backups on which to practise before applying the more refined methodology to the main corpus.

4.5.1 Problems with time-gain score

The first problem that came to light regarding the time-gain score was the treatment of single-word insertions. Because of the lack of context, it always takes extra time to validate a single-word substitution. This is especially true of polysemous words, which are the very words for which the translator requires the most assistance. The time

scenario. Of course, if an entire paragraph or section is highlighted sentence by sentence, all of the matches will almost certainly be found together in the same bitext, so the user always has the option of selecting the entire paragraph's translation and inserting it all at once into the input text, then deleting the "leftover" source sentences that follow. Also, if the user knows the paragraph will come up frequently, a TermBase entry can be created for it.

spent validating the single-word substitution generally negates any potential time gains, so the score was adjusted to zero.

The second problem was the reverse of the first, applying instead to the first two categories involving multiple-word substitutions. The more words there are in a single substitution, the more likely it is to belong to the same context, and the less time is required to evaluate its validity. Thus the scores for the first two categories should increase in direct proportion to the length of the unit. It was decided that the variable n would be incorporated into the score, where n would be equal to the number of words in the source input unit[17]. It would have been legitimate simply to add n to the basic time-gain score, but it was decided that n/2 would be added instead to prevent the numbers from growing too high.

A final problem related to the sixth category, proposals that provide answers to implied questions marked in the input text. In the original scoring table, there was no way to adjust for cases in which more than one question was answered. The original value of 2 was therefore multiplied by the variable q, with q representing the number of questions answered within a given unit.

4.5.2 Problems with time-loss penalty

The biggest problem with the time-loss penalty was the lack of acknowledgment that large numbers of proposals take more time to evaluate than small numbers of proposals. In the original score, the only options were no proposals, one proposal and

[17] It would also be possible to make n equal to the number of words in the target substitution, but considering that two substitutions of differing length could both be considered equally valid, the scoring would be biased against the tool that produced the most concise substitution.

more than one proposal. However, it takes more time to sort through and evaluate five proposals than two proposals.

To compensate for this, the scoring was refined so that the penalty would increase by 0.25 points for each number of proposals between two and five. With the penalty for time spent evaluating a single proposal set at -1, two proposals generated a penalty of -1.25, three a penalty of -1.5, four a penalty of -1.75, and five and above a penalty of -2. Based on experience, it was assumed that when more than five proposals were generated, a suitable one would be found within the first five, so the penalty was capped at that point.

A second problem involved the bonus points awarded to MultiTrans for identifying consecutive units. It became apparent that the "Translate-to-Fuzzy" feature in TRADOS could perform a similar function. However, it was not justifiable to award the same bonus to TRADOS, since MultiTrans is forced to take its consecutive units from a single context (which helps ensure validity), while TRADOS could potentially draw 100% matches from various unrelated locations in the database to make up its consecutive units. This increases the need for validation.

Along the same lines, the highlighting system associated with MultiTrans makes it immediately clear which units are stored consecutively in the translation memory database. With TRADOS, the user has to click the Translate-to-Fuzzy option every time, with no guarantee that consecutive 100% matches exist. With these two drawbacks in mind, the bonus for consecutive units in TRADOS was given half the value of the bonus for consecutive units in MultiTrans.

4.6 Refinements to methodology based on results of pilot study

The following sections describe the refined methodology that resulted from the pilot study. These will be applied to a scaled-up corpus in Chapter 5. The time-gain score is outlined in Table 4.6, and the time-loss penalties are outlined in Tables 4.7 and 4.8.

4.6.1 Score 1 (Time gain)

Table 4.6 applies to both MultiTrans and TRADOS.

Table 4.6 Time Gain

Information provided by TRADOS or MultiTrans	Score
A multiple-word unit inserted with no changes, where n=number of words in the source-language unit	2+n/2
A multiple-word unit inserted with only minor changes required (punctuation or capitalization)	1+n/2
Multiple words inserted with major changes required (changes to word order, additions or deletions)	1
Single word (the translator will always have to verify whether the suggestion is suitable in the new context, even in the case of 100% matches, so usefulness is negligible)	0
Provides answers to implied questions marked in Column A (Input Unit), regardless of changes required (q=number of questions answered within a given unit)	2q
Bonus added to score if one tool pulls information from the memory that the other tool misses entirely	1

4.6.2 Score 2 (Time loss)

For TRADOS, a penalty is applied to every unit opened. For MultiTrans, a penalty is applied to every highlighted chunk. Tables 4.7 and 4.8, the time-loss penalties, are based on the categories listed below. The penalty for each scenario is the sum of the penalties for each category involved. A scenario that involves one action and the evaluation of a single match would receive a penalty of -1.5, or B + C. The variable n represents the number of matches proposed.

A. Evaluation of whether to fetch= -0.5
B. Action (open a unit or perform a fetch)= -0.5
C. Evaluation of a single match= -1
D. Evaluation of 2 to 4 matches= -1 – (n – 1)/4
E. Evaluation of 5 or more matches= -2

Table 4.7 Time loss in MultiTrans (Score 2)

Amount of input unit highlighted	Proposals generated	Penalty	Justification
No highlight	N/A	0	N/A
Full unit highlighted (unit defined as in Column A)	Fetch yields one match	-1.5	B+C
	Fetch yields 2 to 4 matches	-1.5 – (n – 1)/4	B+D
	Fetch yields 5 or more matches	-2.5	B+E
Portion of unit highlighted	No fetch performed	-0.5	A
	Fetch yields one match	-2	A+B+C
	Fetch yields 2 to 4 matches	-2 – (n – 1)/4	A+B+D
	Fetch yields 5 or more matches	-3	A+B+E
Bonus awarded for consecutive units (CUs), where x equals the number of CUs highlighted	N/A	+x	See Table 4.4

Table 4.8 Time loss in TRADOS (Score 2)

Proposals generated	Penalty	Justification
Unit opened, no match	-0.5	B
100% match	-0.5	B
One fuzzy match	-1.5	B+C
2 to 4 fuzzy matches	-1.5 – (n – 1)/4	B+D
5 or more fuzzy matches	-2.5	B+E
Bonus awarded for consecutive units (CUs), where x equals the number of CUs generated by the Translate-to-Fuzzy option	+x/2	See section 4.5.2

4.6.3 Score 3 (Composite)

The composite score is a function of the time-gain score and the time-loss penalty and is meant to provide an overall measure of usefulness. The time-gain score is given more weight than the time-loss penalty, since one useful match saves more time in research and revision than one useless lookup wastes. In theory, a text considered suitable for translation memory should generate a positive composite score, unless the memory is underdeveloped. Unsuitable texts should generate negative scores, requiring too much time to process and retrieving too little useful information from the memory.

For the purposes of this test, the time-gain score (Score 1) will be given twice the value of the time-loss penalty (Score 2):

$$\boxed{\text{Score 3} = 2(\text{Score 1}) + \text{Score 2}}$$

More testing would be required to determine whether this is the optimal relative value. However, even if it is not, it should still be fair if equally applied to both tools. The tool that generates the highest composite score is the tool whose automatic search-and-retrieval function presents the most useful information to the translator in the least amount of time.

Chapter 5 Testing the Methodology

In this chapter, the refined version of the methodology developed in Chapter 4 will be applied to a larger corpus, and the results will be discussed. It is important to note that this application is designed to test the applicability of the evaluation methodology. Many more tests, using a wider range of corpora, would be necessary to provide convincing evidence regarding the superiority of one approach over another. In addition, this methodology evaluates the efficiency of search-and-retrieval techniques only. A TM tool that performs less well than another in this area may still be a good choice for a job on the strength of its other features (e.g. terminology or project management components). Rather than give a definitive answer to the question of which approach is better, this limited application will demonstrate various kinds of information that can be obtained by using this evaluation methodology.

5.1 Design and compilation of main corpus

Most of the criteria used for compiling the main corpus were identical to those used for the pilot corpus, as described in Table 4.1. The requirements for quality, language direction, format, subject field and confidentiality remained the same. As with the pilot corpus, it was necessary to base the main corpus on a single text type. The text type selected for the main corpus had to be different from that selected for the pilot corpus in order to ensure that the methodology was applicable to more than one text type. For this scaled-up test, the required number of bitexts was increased to 110: 100 to build the corpus, and 10 additional texts to be used as input for obtaining search-and-retrieval data.

The broad subject field (employment) was the same in both cases. However, the main corpus was entirely made up of statements of qualifications, a text type with different characteristics than those of the work descriptions used for the pilot corpus. At an average length of 228 words (1.5 pages), they were ideal for testing purposes. They were written mostly in sentence fragments (e.g. "Driver's licence required") and although they contained little internal repetition, they had a high degree of external repetition, which made them good candidates for use with TM tools.

It took three hours to download the 110 bitexts required for the application phase. It took 30 minutes to generate the corpus in MultiTrans, three times as long as it took to generate the pilot corpus. Once the corpus was generated in MultiTrans, the same verification was performed to identify and replace corrupt or unusable texts, which required an additional hour. Then a second version of the main corpus was generated for use with TRADOS. Interestingly, this process took just over an hour, the same amount of time required to build the pilot corpus, despite the fact that there were more than three times as many bitexts. This can be explained by the fact that the statements of qualifications were on average much shorter than the work descriptions and were structured in such a way that there were fewer errors with the automatic alignment. In fact, more than a third of the statements of qualifications were perfectly aligned from the start, which was never the case with the work descriptions.

5.2 Application of methodology to main corpus

The 10 input texts were labelled SQ01 through SQ10, with SQ standing for Statement of Qualifications. All were analyzed and marked up to generate the input units for Column 1 (i.e. those sections of the text for which the translator was seeking help).

This process required between one and five minutes for each page. Less time was required for these SQ texts than for the work descriptions used in the pilot study simply because there was less text on each page.

The SQ texts were then pre-translated in MultiTrans and TRADOS, and the output was graded according to the evaluation methodology. Through trial and error, it was determined that the most efficient way to gather data was to alternate between recording time-gain and time-loss data for each unit of a text. The first unit was evaluated; if no replacement was made, the time-loss category (see Table 5.2 for description of categories) was noted in the appropriate column on the time-loss spreadsheet. If a replacement was made, the time-loss category was still noted, then the replacement was pasted into the appropriate column of the time-gain spreadsheet, and its category was noted. This process was repeated for each unit until the end of the input text was reached. After all the replacements and categories were recorded, the categories were replaced with their corresponding scores[18].

It was most efficient to process all 10 texts in one tool before moving on to the next one. Due in part to frequent interruptions and also to the learning curve involved, it was difficult to record the length of time required to process each text. However, I estimate that with practice it took approximately twenty minutes to process each text in TRADOS and thirty minutes in MultiTrans. The SQ texts were 1.5 pages long on average, but the time required to perform the evaluation would probably increase in direct

[18] The scores could have been recorded directly, but mentally identifying the category then calculating the score on the spot slowed down the process considerably. By simply noting the category as an intermediate step, the scores could be later be inserted through a quick search-and-replace process.

proportion to the length of the texts being processed. The data gathered for SQ01 through SQ10 are listed in Appendices C and D.

5.3 Summary of results and discussion

This section presents both the overall performance of each of the two tools tested and a detailed breakdown of the data collected during the evaluation process.

5.3.1 General results

Table 5.1 provides a summary of all three scores for each of the 10 input texts in both TRADOS and MultiTrans, with the last row containing the sums of all the scores generated in each category. This last row gives an overall picture of the comparative performance of the automatic search-and-retrieval functions. In this case, there is very little difference between the two tools' success in retrieving useful information, with TRADOS scoring only 0.02% higher than MultiTrans in the time-gain score. However, there is a significant difference between the time-loss penalties, with TRADOS requiring much less time to accomplish the same task. TRADOS's higher composite score implies that it is the more suitable tool for this particular text type.

Table 5.1 Overall performance of TRADOS and MultiTrans

Input Text	Time-gain Score		Time-loss Penalty		Composite Score	
	MultiTrans	TRADOS	MultiTrans	TRADOS	MultiTrans	TRADOS
SQ01	13	16	-20.5	-18	5.5	14
SQ02	41.5	44.5	-33	-13.75	50	75.25
SQ03	53	55	-39.25	-11	66.75	99
SQ04	152	156	-55.75	-30.25	248.25	281.75
SQ05	69.5	75	-48.25	-15.25	90.75	134.75
SQ06	17.5	26	-15.75	-18	19.25	34
SQ07	51	49	-48	-15.25	54	82.75
SQ08	33.5	24	-47.25	-22	19.75	26
SQ09	25	28.5	-31.75	-15.75	18.25	41.25
SQ10	34.5	28	-20.75	-16	48.25	40
All texts	**490.5**	**502**	**-360.25**	**-175.25**	**620.75**	**828.75**

5.3.2 Breakdown of scores and penalties

The general scores answer the broad question of which TM tool retrieves the most useful information in a given situation, but it can also be informative to look a little more closely at trends occurring at lower levels of the evaluation results. Tables 5.2 to 5.8 provide a detailed look at exactly what kind of information is being extracted by each tool from its TM database.

5.3.2.1 Detailed time-gain scores

In many cases, the two tools generated identical scores for a given unit. These occurrences are illustrated in Table 5.3. However, the cases where the two tools generated different scores (see Table 5.4) offer more insight into the essential differences between the sentence-based and the CSB-based approaches.

Table 5.2 Separate categories applied for time-gain score

Category	Description
1	A multiple-word unit inserted with no changes
2	A multiple-word unit inserted with only minor changes required (punctuation or capitalization)
3	Multiple words inserted with major changes required (changes to word order, additions or deletions)
4	Provides answers to "questions" underlined in Column A (Input Unit), regardless of changes required
5	Single Word
6	No proposal
7	Bonus points awarded

Table 5.3 Occurrences per category where scores were identical in both tools

	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Total
SQ01	3		1			21		25
SQ02	7					13		20
SQ03	4	1				14		19
SQ04	10					5		15
SQ05	1	4	1			8		14
SQ06	1					8		9
SQ07	2	1				7		10
SQ08	3	1				16		20
SQ09	3		1			9		13
SQ10	1	1				11		13
Total	35	8	3			112		158

With regard to similarity of the time-gain scores generated by the two tools, the highest concentrations of identical scores occur in the first and sixth categories, which correspond to exact matches and no matches. This is not very surprising, since true exact matches will generally be replaced in the same way by any TM tool, while more variation in performance is to be expected in the middle categories, such as Categories 2 and 3, where only a limited number of identical scores were observed. There are no entries for Categories 4, 5 and 7 because the only time that the scores were equal in these cases was when both were zero. These occurrences are already accounted for in Category 6.

Table 5.4 Occurrences per category where scores differed between tools

	Tool	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7
SQ01	TRADOS	1						
	MultiTrans		1					
SQ02	TRADOS	3		1				1
	MultiTrans	1	2				1	
SQ03	TRADOS	7	1	2			1	1
	MultiTrans	2	6	1	1		1	1
SQ04	TRADOS	12		1				1
	MultiTrans		12				1	
SQ05	TRADOS	5	7	1			1	2
	MultiTrans	8	4				2	1
SQ06	TRADOS	2	2			3	1	3
	MultiTrans		1	3	1		3	4
SQ07	TRADOS	10	1			2	2	
	MultiTrans	1	11	2	1			4
SQ08	TRADOS		2			5	2	
	MultiTrans	2	2	4	1			7
SQ09	TRADOS	3		2		1		1
	MultiTrans	1	3	1			1	1
SQ10	TRADOS		2				1	
	MultiTrans	3						1
Totals	TRADOS	43	15	7	0	11	8	9
	MultiTrans	18	42	11	4	0	9	19

When observing the differences that occur with regard to the time-gain scores generated by the two tools (Table 5.4), a number of initial conclusions can be drawn from the totals recorded in the last two rows. The first observation is the inverse scoring pattern in the first two categories: for the TRADOS units, the ratio of Category 1 hits to Category 2 hits is 43:15 (approximately 3:1). The corresponding MultiTrans ratio is 18:42 (approximately 1:3). This implies that when a full unit is replaced, it is much more likely to require minor editing in MultiTrans than in TRADOS. MultiTrans, on the other hand, is slightly stronger in Category 3, replacements requiring major changes, and is the only tool that offers any proposals whatsoever for Category 4, replacements that answer implied questions in the marked up input text. In the first four categories, all of which generate positive scores (as opposed to the zero scores generated by Categories 5 and 6), MultiTrans generates more hits overall, with a total of 75, compared with only 65 for TRADOS. This is also reflected in the higher number of bonus points awarded to MultiTrans, for retrieving useful units missed by TRADOS. However, TRADOS gets a slightly higher score overall (see Table 5.1) because of its concentration of high-value Category 1 hits.

The 11 occurrences for TRADOS in Category 5, single words, reveal an interesting phenomenon. MultiTrans is designed to ignore single-word units because of their lack of reliability. In most cases where the input unit consisted of a single word, both tools received a score of zero. Even when TRADOS generated a proposal, this was not awarded any points because the time needed to evaluate the unit negated any gain achieved by retrieving the unit (see section 4.5.1). However, Table 5.4 only records the cases when the scores are different. This means that there were 11 cases where TRADOS retrieved a single word (receiving a score of zero), while MultiTrans generated a positive score. This seemingly impossible result is explained by the fact that MultiTrans does replace single-word headings when they appear within consecutive units. These were always correct, since they came bundled with their contexts, so were awarded one point each. However, the "consecutive unit" bonus was not added. The formatting generally required a lot of adjustment, which could not easily be reflected in the time-loss penalty, so it was decided that these concerns cancelled each other out. Examples occur in texts SQ06 through SQ09. It is likely that this phenomenon is a peculiarity of this text type.

5.3.2.2 Detailed time-loss penalties

The time-loss penalties of the two tools are less directly comparable than the time-gain scores, since their units are defined differently. The input units defined for the time-gain scores are irrelevant to the calculation of the time-loss penalties because a user must evaluate and process the discrete chunks presented by each tool. Based on the design of their search-and-retrieval functions, MultiTrans tends to identify shorter (sub-sentence) units in greater numbers, while TRADOS tends to identify longer (sentence-length) units, but not all of them have proposals associated with them. This pattern reflects a central difference between the CSB-based and sentence-based approaches, as explained in section 4.4.2.

The time-loss penalties were calculated using two measurable attributes: the time it takes to perform the primitive actions, such as mouse clicks, associated with search-and-retrieval functions, and the time it takes to evaluate the proposals generated by each tool. Because of the comparative nature of this study, it was not necessary to calculate the actual time required, but only to assign relative weights to each category. Clicking the mouse requires the same amount of time in any tool, and it was also assumed that it requires on average the same amount of time to decide between two proposals whether TRADOS or MultiTrans is being used.

Table 5.5 shows that TRADOS generally identifies more units than MultiTrans for the SQ input texts. This is because in TRADOS, the user must “click open” each unit to launch a search of the database. This means that although TRADOS may not find a match for every unit opened, each unit will generate at least the minimum time penalty associated with the clicking action. In MultiTrans, a chunk of text is only identified as a unit if a match does appear in the database. Therefore, there are usually fewer units, but these units will always generate an evaluation penalty in addition to a primitive action penalty.

Table 5.5 Number of units per text identified by each tool

	SQ01	SQ02	SQ03	SQ04	SQ05	SQ06	SQ07	SQ08	SQ09	SQ10	Total
MultiTrans	24	31	25	27	32	13	24	40	34	15	265
TRADOS	33	28	31	28	34	23	32	36	33	23	301

5.3.2.2.1 Presentation of results in MultiTrans

A key to the categories used for evaluating time-loss penalties in MultiTrans is presented in Table 5.6. In Table 5.7, the number of hits in each category is presented for each input text.

Table 5.6 Separate categories applied for time-loss penalty in MultiTrans

Amount of input unit highlighted	Proposals generated	Category
No highlight	N/A	N/A
Full unit highlighted (unit defined as in Column A)	Fetch yields one option	A
	Fetch yields 2 to 4 matches	B
	Fetch yields 5 or more matches	C
Portion of unit highlighted	No fetch performed	D
	Fetch yields one match	E
	Fetch yields 2 to 4 matches	F
	Fetch yields 5 or more matches	G
x consecutive units highlighted	N/A	H

Table 5.7 Occurrences per category in MultiTrans

MultiTrans	Cat. A	Cat. B	Cat. C	Cat. D	Cat. E	Cat. F	Cat. G	Cat. H	Total
SQ01	1		3	19	1				24
SQ02	2		6	21		2			31
SQ03	3	2	8	9	2	1			25
SQ04	2	4	12	4	1	2	2		27
SQ05	3	1	10	14	2		2		32
SQ06			1	8	2	1	1		13
SQ07	2	1	11	5	1	1	3		24
SQ08	2	4	6	24	3		1		40
SQ09	1	1	4	25	2		1	(1*)	34
SQ10	2		2	8		1	2		15
Total	18	13	63	137	14	8	12		265

*Consecutive unit bonuses are included for information but not counted in total

5.3.2.2.2 Presentation of results in TRADOS

Table 5.8 shows how many exact or fuzzy matches were proposed by TRADOS for each unit opened in a given input text. For example, in the input source text SQ09, which was divided by TRADOS into 33 units, TRADOS proposed one exact match for twelve units, one fuzzy match for three units, four fuzzy matches for one unit and no matches for the remaining seventeen units. There were also three runs of consecutive 100% matches.

Table 5.8 Type and number of matches proposed by TRADOS

TRADOS	Exact	1F*	2F	3F	4F	5F	6F	7F	8F	None	Total	CUs**	Bonus Points
SQ01	5	3								25	33	1	1
SQ02	11	1	3			2				11	28	1	2
SQ03	17	3	2			1				8	31	3	6.5
SQ04	14	1	1	2	2	7				1	28	2	3.5
SQ05	16	5		1						12	34	3	5
SQ06	6		4		1	1				11	23	1	1
SQ07	14	3	1				1			13	32	2	4
SQ08	11		3	1	1			1		19	36	2	2
SQ09	12	3			1					17	33	3	2.5
SQ10	1	1		1					1	19	23	0	0
Total	107	20	14	5	5	11	1	1	1	136	301	18	27.5

*F = Fuzzy Match(es)
**CUs = the number of runs of consecutive 100% matches identified in a text

5.3.2.2.3 Comparison of time-loss penalties

Despite the differences between the presentations of the results, a certain number of comparisons can be made. TRADOS generated a total of 136 hits in the column labelled “None”, indicating units that were opened by the user but for which no matches were found (Table 5.8). The closest equivalent to this in MultiTrans is Category D, designating the highlighted units that the evaluator examined but for which no fetch was performed. For example, if “Capacité à utiliser le module de gestion du matériel du système” occurred in a pre-translation, many translators would find it easier to type the translation from scratch than to take the time to fetch the match from the database and insert it, regardless of the validity of the match[19]. In MultiTrans, 137 units fell into this category (Table 5.7). Each of these categories represents the minimum penalty for each tool, but the large numbers of hits in each case add up to a significant loss of time for the user, with no associated benefit. It is more serious in the case of MultiTrans, since it takes

[19] As explained in section 4.4.6, the decision of whether or not to fetch depends entirely on the information marked for look-up in the input text, removing any subjectivity from this stage of the evaluation.

more time to evaluate the value of fetching a unit than it does simply to click one open in TRADOS. Hence, as illustrated previously in Table 5.1, this contributed to the fact that MultiTrans generated a more significant overall time-loss penalty (-360.25) than did TRADOS (-175.25).

Tables 5.7 and 5.8 also reveal a second difference between the tools. TRADOS generated proposals for just over half the units identified. Of these, 64.9% constituted one exact match, 12.1% constituted one fuzzy match, 8.5% constituted two fuzzy matches, and the percentages continue to decrease for larger numbers of fuzzy matches. The only exception is the category representing five fuzzy matches, which at 6.7% is higher than the categories for three or four fuzzy matches. TRADOS presented just one match (exact or fuzzy) 77% of the time, and rarely produced large numbers of proposals.

MultiTrans follows a completely different pattern in this respect. A fetch was performed in approximately half the cases (128 out of 265). Categories A and E represent the cases where only one match was proposed and account for 25% of all cases where a fetch was performed. Categories B and F together represent cases where two, three or four matches were proposed and account for 16.4% of all cases. Finally, Categories C and G represent cases where five or more matches were proposed and account for 58.6% of all cases. This indicates that when users do decide to perform a fetch on a unit, there is often a lot of information to sort through[20]. This can be positive, allowing translators to choose the best of many ideas in many contexts, but it also has a higher time cost.

It is possible for the user to adjust the settings in TRADOS to set an upper limit on the number of proposals. The maximum is fifty and the default is five; I set the limit to

[20] The multiple proposals are identical in many cases; they simply appear in different contexts.

eight for the purposes of this evaluation. In the one case where eight matches were proposed, I raised the limit and reprocessed the unit, confirming that there were indeed exactly eight matches identified in the database. MultiTrans, on the other hand, often generated much greater numbers of proposals. The highest number was 74, but this was for the unit "Statement of Qualifications", which is the title of almost every text in the corpus[21]. Of the 75 occurrences in Categories C and G, exactly two thirds fell between 5 and 25 proposals, and a full third represented proposals numbering from 26 to 75. This shows that whenever at least one match exists in the database, TRADOS is much more likely to offer a single proposal, while MultiTrans is much more likely to offer multiple proposals. This makes sense, as shorter segments are more likely to reoccur in a variety of contexts than full sentences. In addition, the single proposal offered by TRADOS is likely to be longer and more useful because of the nature of the sentence-based segments used in this tool.

5.3.3 Additional observations

Once a few texts had been processed, the performance of each tool for similar texts became relatively predictable. However, over the course of the evaluation, a few individual proposals emerged that revealed interesting or unexpected aspects of the sentence-based and CSB-based approaches. These examples will be discussed below, along with my observations about their significance.

[21] This did not occur in TRADOS, in which users have the option of "reorganizing" the database, or eliminating redundant translation units.

5.3.3.1 Left-to-right processing

When MultiTrans searches through an input text, it processes from left to right. This means that it searches until it finds the longest character-by-character match possible, and then it begins searching again from immediately after the end of that match. It can cut across sentence boundaries at a sub-sentence level, as illustrated by the second unit in the following example, extracted from the test corpus, but modified for brevity:

> Répondre à des demandes urgentes et multiples formulées par les gestionnaires ou les *clients. Il faut* fournir un effort psychologique et émotionnel pour *faire face aux* innombrables problèmes qui surgissent tous les jours.

However, this is not necessarily a good thing. The phrase "clients. Il faut" and others following this pattern are unlikely to be of any use to a translator, but worse, they may obscure another potential match. In this case, "fournir un effort psychologique" is almost as good as "Il faut fournir un effort psychologique", but consider the following example:

> Direction générale : Direction des *ressources humaines*

The first unit is only useful up to the second word, but the fact that the third word is attached to it prevents the tool from finding the entire unit "Direction des ressources humaines", which does appear in the test corpus. Of course, an astute translator who understands what is happening and guesses that the information is indeed in the database can simply select the desired text and do a fetch anyway. Admittedly, such examples are very rare. However, this remains a small flaw in the system, and it may be worth investigating search methods other than left-to-right processing. For example, it may be helpful to apply a method that includes some form of backtracking. André Guyon of the

Translation Bureau is currently investigating a bi-directional approach that involves processing from the centre of a sentence to its extremities[22].

5.3.3.2 To fuzzy or not to fuzzy?

The fact that MultiTrans performs as well as it does without fuzzy matching makes one question the value of fuzzy matching. It is certainly a very difficult concept to capture in a logical algorithm and has the potential to generate a lot of noise. In theory, if a sentence has only small changes in it, MultiTrans should be able to pick up the identical character strings around the changed parts, and so provide a translation of much of the sentence anyway. TRADOS, by limiting itself to full sentences[23], risks missing a number of potentially helpful phrases and terms.

The biggest argument in favour of fuzzy matching seems to be related to time. Consider the following sentence, which contains a typographical error:

> Croire en sa capacité de réussir, même en situation difficile, et assumer la responsabilité es résultats obtenus grâce à ses efforts.

Clearly, “es” should read “des”. The same sentence, without the error, appears in the TM database. TRADOS identifies this sentence as a 99% fuzzy match, then inserts its perfectly acceptable translation:

> Believe in your ability to succeed, even in difficult situations, and take responsibility for the results obtained through your efforts.

MultiTrans, which does not perform any fuzzy matching, identifies two character strings (all the words preceding “es” and all the words following “es”) as having perfect matches

[22] Personal communication with André Guyon, IT Strategies, Translation Bureau, Government of Canada (February 2003)

[23] TRADOS can also identify and replace terms, which are below the sentence level, but only when those terms have been manually entered into the MultiTerm component of the tool. As explained in section 0.4, an investigation of MultiTerm is beyond the scope of this study.

in the database. Of course, both of these matches come from exactly the same sentence in the database, so one can fetch the entire sentence the first time, resulting in the following:

> Believes in his or her ability to succeed, even in difficult situations, and assumes responsibility for results achieved through his or her efforts. es résultats obtenus grâce à ses efforts.

The remaining portion of the source text can then be deleted. The minor variations in the translation simply show that there is more than one valid translation in the database; they are unimportant for this example. What is important is that MultiTrans finds the same information, but the additional deletions required mean slightly more time must be spent editing the final result. The main reason that TRADOS performed better than MultiTrans in this test is that TRADOS proposed significantly more translations that required no editing whatsoever[24]. The difference in scores is very small in each case but it adds up over a long text.

However, the fact that MultiTrans is not limited to the sentence level gave it a significant advantage in acquiring bonus points: it required slightly more editing time, but it was able to find sub-sentential information that TRADOS missed entirely. It was the only tool that made proposals in the fourth category, answering terminology or phraseology questions, without requiring the user to spend any time manually feeding a terminology bank[25]. With this in mind, the MultiTrans approach might have advantages for novice translators who depend on their TM databases more for the translation

[24] When a segment is inserted into a new text with MultiTrans, proper nouns often have their capital letters suppressed and so may need to be corrected by the translator. This does not occur with TRADOS.

[25] Both TRADOS and MultiTrans would probably have performed equally well in this regard if their term banks had been pre-filled and contained the same information. That, however, would have defeated the purpose of testing the automatic search-and-retrieval functions.

solutions they contain than for the time they save. An experienced translator may be more interested in the approach that requires the least amount of editing time.

Based on the design of the two tools being tested, one gets the impression that fuzzy matching and "character-string-within-a-bitext" searching are mutually exclusive functions. However, given the advantages of each, it seems worthwhile to explore ways of combining the two in a single tool. Other companies, such as Atril and Terminotix, have recently begun to advertise this very combination of capabilities in their tools. For example, in an advertisement for the most recent version of its Déjà Vu translation memory tool featured on the back cover of the October 2002 edition of *Language International*, Atril asks:

> How often do you find yourself translating exactly the same sentence? While remembering previous translations of complete sentences is useful, repetition is much more common at the sub-sentence level. If your tool's idea of translation reuse stops at retrieving entire sentences, then you aren't fully exploiting the wealth of information contained in your translation memory database.

At the time of writing up this thesis I have not been able to confirm Atril's claim for their tool Déjà Vu, but my testing of LogiTerm by Terminotix leads me to conclude that they have successfully integrated both approaches. However, as explained in section 0.4, the automatic search function of LogiTerm is much more limited than those of TRADOS and MultiTrans, and the retrieval system seems only slightly more advanced than traditional cut-and-paste methods.

TM technology is out of its infancy, but it is still in an early stage of development. Many interesting solutions have emerged to a variety of the challenges associated with text recycling, and translators can only benefit as these solutions are blended into functional hybrids.

Part III

Chapter 6 Conclusion

6.1 General comments about the proposed methodology

The original purpose of this thesis was to determine whether it was possible to compare two fundamentally different approaches to automatic search and retrieval—the sentence-based approach used by TRADOS and similar tools and the CSB-based approach used by MultiTrans—in a systematic and unbiased way.

In section 4.4, I listed three properties of a good evaluation methodology as defined by EAGLES: reliability, validity and efficient applicability. How does the methodology developed for this thesis measure up against these standards?

It is always necessary to prioritize the goals of a project, and in this case, validity was given the highest priority. A TM tool is only successful if it provides useful information to its users, so in a comparative study, the "best" tool is by definition the most useful tool. The challenge was finding a way to redefine usefulness as a set of measurable attributes. It was concluded early in the project that traditional edit distance measures are inadequate reflections of usefulness, although their value as approximation tools are high in situations where evaluation must be automated (e.g. when large volumes of data must be processed in a short time). Defining usefulness as a function of validity, time gain and time loss went a long way towards solving this problem. The other factor that supports the soundness of the methodology proposed in this thesis is its non-reliance on the model translation, reflecting the reality that several valid translations may exist for any given segment of text.

Reliability was another property that was given high priority during the development stage. The drawback with most subjective evaluations is that they can produce variable results when performed by different human evaluators. A degree of subjectivity is necessary for validity, so it could not be removed entirely; however, by limiting the subjective measure to the analysis and mark-up phase, in which the evaluator marks the sections of the input texts that he or she would be likely to research in a translation situation, the potential for bias towards one tool or another is limited. It also solves the problem discussed in section 3.4.3 of determining the usefulness of a proposal after one has seen it. If the evaluator's translation knowledge (or lack of knowledge) about the input text has not been captured in a "snapshot" before the evaluation takes place, it becomes very easy for the evaluator to confuse true usefulness with simple validity. The evaluator may no longer know whether a particular proposal would have come to mind without prompting, and so may be tempted to say that the match is useful just because it is valid. The methodology proposed here accounts for the fact that a valid match that saves time is more useful than a valid match that does not save time, which in turn is more useful than a valid match that wastes time. Once the analysis phase is complete and marked-up input texts exist as benchmarks, any evaluator can apply the scoring system and generate identical results, which is a good indication of reliability.

One potential argument against the reliability of the scoring system is that no two evaluators are likely to mark up the input texts in exactly the same way. Moreover, an experienced translator is likely to mark fewer sections than a novice translator. The same input text marked in two different ways will result in two different sets of scores for the tools being tested. However, this is an advantage to an individual translator applying the methodology, since the results would truly reflect that individual's needs. When the evaluation is being performed on behalf of a larger group of translators, the evaluator selected to perform the tests should be representative of the translators who will be using the tool. The group may determine the definition of "representative"; it could refer to years of experience, familiarity with translation technology, use of reference materials such as term banks and parallel texts, or any other variable that is deemed important. Rather than showing the methodology to be unreliable, the analysis and mark-up phase provides a desirable element of customizability.

The final property outlined in the EAGLES report is efficient applicability, and this is the greatest weakness of this approach. The very features that ensure its validity make it impossible to automate, and it is very time consuming to perform the evaluation by hand. The report does acknowledge this paradox, stating that despite the desirability of a methodology that requires little effort from the user, "validation will probably always remain somewhat problematic, so evaluations will probably always involve some degree of user activity."[26] It also takes time to build suitable test corpora. Then again, this is a problem common to all TM evaluation methods, and it may be addressed somewhat as more and more people take an interest in evaluation and begin to pool their resources, including test corpora.

Overall, the methodology presented in this thesis shows that it is indeed possible to compare sentence-based TM tools to CSB-based TM tools. It is probable that one approach would generate higher scores for certain text types, while the other approach would do better for other text types. Researchers and developers could use the

[26] http://issco-www.unige.ch/projects/ewg96/node155.html#SECTION00104100000000000000

methodology to gain information about general approaches or specific TM tools by testing a wide variety of text types, while translators and translation companies hoping to choose a tool can limit their evaluations to the text types they most frequently encounter. The composite scores would provide a general comparison at a glance, while a breakdown of the results would provide useful insights into the relative strengths and weaknesses of each approach being tested.

Above all, the evaluator must keep in mind that the automatic search and retrieval function is only one component of a TM tool, and that there are many more factors to consider before deciding which tool is most appropriate for a given context. For example, it was implied in Chapter 5 that the sentence-based approach provides information in significantly less time than the CSB-based approach. However, a potential user must consider the time costs related to other aspects of the tool. MultiTrans, for example, requires much less time for database maintenance than does TRADOS. Only the user can choose whether he or she prefers to invest the extra time in retrieving proposals or in maintaining databases. Margaret King (1993) summarizes the situation well:

> Bien sûr, la performance en termes d'exactitude de la traduction, ou de temps nécessaire pour produire une traduction utilisable, sera toujours d'une importance capitale, mais il y aura toute une série d'autre considérations liées au confort de l'utilisation : **aucun système ne peut être efficace si personne n'accepte de l'utiliser** (p.265) [emphasis added].

6.2 Recommendations for further research

This methodology could be applied and augmented in a variety of ways to answer several different kinds of questions. The most obvious application would be to different text types. TM researchers already have a general idea of what types of texts are suitable for TM processing (see section 1.3), but applying this evaluation technique could further refine this knowledge by determining which approaches are best suited to processing specific text types.

Evaluations could be performed using different language combinations to find out whether a given approach is more suitable for particular languages or language directions.

Evaluations could be performed using tools other than TRADOS and MultiTrans. These would be useful to verify that the methodology is not tool-dependent and eventually to evaluate new or blended approaches to search and retrieval that have yet to be developed.

Finally, the analysis and mark-up phase, which was discussed as a potential reliability problem in the previous section, could be exploited in a study involving translators with different levels of experience. A tentative conclusion based on the test described in section 5.3 was that the CSB-based approach might be better for novice translators, while the sentence-based approach could potentially be of more use to experienced translators. This is because the former extracts more translation information from the database and the latter extracts more matches that require no editing. This suggestion could be investigated by performing the evaluation with input texts submitted by translators falling into predefined categories of experience.

A further study involving the design of the methodology itself might be desirable, in which different equations for the composite score are compared. In this case, the time-gain score was assigned twice the weight of the time-loss penalty, which was adequate for the comparative test described in Chapter 5, but further tests might determine whether the optimal multiple is indeed 2. As stated in section 4.6.3, the optimal composite score would produce a positive number when text types suitable for TM are tested and a negative number when unsuitable types are tested. If this were the case, the evaluation could be adapted for use as a benchmark test for evaluating the adequacy of a single tool.

Since technology is changing rapidly, it would be necessary for future researchers to update and adapt this methodology to take into account any new features added to TM tools. It may also be possible in the future to incorporate it into a larger, more comprehensive evaluation methodology that tests not only the automated search-and-retrieval function, which was the focus of this study, but also other features of the tools such as terminology management, database maintenance, etc.

6.3 Concluding remarks

Technology is becoming more and more important in translation, but translators have limited resources. It is not feasible to buy and spend time mastering every tool that is currently available or that will become available in the future. If not used properly, technology can actually slow a translator down, so it is important to be able to choose the best tool for a job. To do this, users and developers need reliable methods for comparing and evaluating different types of tools. I hope that this thesis has gone some way towards providing a framework within which they can carry out such evaluations.

Glossary

alignment: the process of building a translation memory from previously translated material by linking corresponding source and target segments

bitext: the juxtaposition of a translation's source (ST) and target (TT) texts on the same page or screen; in the case of an electronic bitext, corresponding source and target segments are linked to each other (aligned)
(synonym: parallel corpus)

character-string-within-a-bitext (CSB)-based approach to automatic search and retrieval: an approach to translation memory in which full bitexts are aligned and stored in a database; new texts are compared to the texts in the database and identical character strings of any length are retrieved

exact match: a perfect character-by-character match between a segment stored in a translation memory and a given input segment
(synonyms: 100% match, perfect match)

external repetition: linguistic material that is occurs more than once in several documents

fuzzy match: a match identified by a translation memory tool as being similar but not identical to a given input segment; the required degree of similarity can be set by the user and can fall anywhere between 1% and 99%
(synonym: partial match)

input unit: a segment of source text designated by the evaluator for measurement during the implementation of the evaluation methodology

internal repetition: linguistic material that occurs more than once within a single document

noise: the retrieval by a translation memory tool of inaccurate or unhelpful matches from its database

output unit: a target-language proposal generated as a result of a match in one of the tools being evaluated
(synonym: proposal)

sentence-based approach to automatic search and retrieval: an approach to translation memory in which sentence-length translation units are aligned and stored as discrete records in a database; each sentence of a new text is compared to all the sentences in the database, and identical or similar sentences are retrieved

silence: the failure of a translation memory tool to retrieve useful data stored in its database

translation memory: a database of previously translated text in which source-language segments are linked with their corresponding translations and from which information can be retrieved during the translation of new texts
(synonyms and abbreviations: TM, translation memory database, TM database, memory database, memory)

translation memory tool: translation support software that allows users to recycle repetitive translation material through the creation of translation memory databases, from which it retrieves segments that are similar or identical to new segments being translated, and inserts them into the new translation
(synonyms and abbreviations: TM tool, translation memory system, TM system)

translation unit: a pair of segments stored in a translation memory, made up of a source segment and its translation
(abbreviation: TU)

useful proposal: a valid proposal that saves the user more time than is required to generate it[*]

valid proposal: a proposal that provides the user with accurate information about how to translate a given source segment*

[*] This definition applies only in the context of the evaluation methodology described in this thesis

References

Akiba, Yasuhiro, Imamura, Kenji and Sumita, Eiichiro. (2001), ‘Using Multiple Edit Distances to Automatically Rank Machine Translation Output’, Proceedings of MT Summit VIII: Machine Translation in the Information Age, Santiago de Compostela, Spain (Proceedings on CD-ROM).

ALPAC. (1966), *Language and Machines: Computers in Translation and Linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, DC.

Andrés Lange, Carmen and Bennett, Winfield Scott. (2000), ‘Combining Machine Translation with Translation Memory at Baan’, in: Sprung, Robert C. (ed) *Translating into Success: Cutting-edge strategies for going multilingual in a global age*, John Benjamins Publishing Co., Amsterdam/Philadelphia, pp.203-218.

Arrouart, Catherine et Bédard, Claude. (2001), ‘Éloge du bitexte’, *Circuit*, vol. 73, p.30.

Atril. (2002, October 6 – last update), ‘About Us’, (Déjà Vu - Translation memory and productivity system), Available: http://www.atril.com (Accessed: 2002, October 6).

Austermühl, Frank. (2001), *Electronic Tools for Translators*, St. Jerome Publishing, Manchester.

Bédard, Claude. (1998a), ‘Ce qu’il faut savoir sur les mémoires de traduction’, *Circuit*, vol. 60, pp.25-26.

Bédard, Claude. (1998b), ‘« Jamais plus vous ne traduirez… » ou les mémoires des traduction, deuxième partie’, *Circuit*, vol. 61, p.23.

Bédard, Claude. (2001), ‘Une nouvelle profession : traducteur de phrases’, *Circuit*, vol. 70, p.29.

Benis, Michael. (1999), ‘Translation memory from O to R’, (TransRef – The Translation Reference Center), Available: http://transref.org/default.asp?docsrc=/u-articles/Benis3.asp (Accessed: 2003, March 2).

Benis, Michael. (2000), ‘How the memory measured up’, (TransRef – The Translation Reference Center), Available: http://transref.org/default.asp?docsrc=/u-articles/Benis4.asp (Accessed: 2003, March 2).

Bowker, Lynne. (2002), *Computer-Aided Translation Technology*, University of Ottawa Press, Ottawa.

Canadian Translation Industry Sectoral Committee. (1999), *Survey of the Canadian Translation Industry: Human Resources and Export Development Strategy*, Final Report of the Canadian Translation Industry Sectoral Committee, September 30th 1999, Ottawa/Montreal.

Cohen, Betty. (2002), 'Mémoires et tarification, un débat à finir', *Circuit*, vol. 76, pp.16-17.

Cormier, Monique C. (1992), 'L'évaluation des systèmes de traduction automatique', *META*, vol. 37, no. 2, juin 1992, pp.383-384.

EAGLES. (2001, October 2 – last update), 'Benchmarking Translation Memories', (Evaluation of Natural Language Processing Systems: Final Report), Available: http://issco-www.unige.ch/projects/ewg96/node157.html #SECTION00104300000000000000 (Accessed: 2002, October 6).

Esselink, Bert. (2000), *A Practical Guide to Localization*, John Benjamins Publishing Co., Amsterdam/Philadelphia.

Gordon, Ian. (1996), 'Letting the CAT out of the bag – or was it MT?', in: *Translating and the Computer 18: Papers from the Aslib conference held on 14 & 15 November 1996*, Aslib, London.

Gordon, Ian. (1997), 'The TM Revolution – What does it really mean?', in: *Translating and the Computer 19: Papers from the Aslib conference held on 13 & 14 November 1997*, Aslib, London.

Harris, Brian. (1988a), 'Bi-text, a new concept in translation theory', *Language Monthly*, no. 54, March 1988, pp.8-11.

Harris, Brian. (1988b), 'Are You Bitextual?', *Language Technology*, no. 7, May-June 1988, p.41.

Heyn, Matthias. (1998), 'Translation Memories: Insights and Prospects', in: Bowker, Lynne, Cronin, Michael, Kenny, Dorothy and Pearson, Jennifer (eds) *Unity in Diversity? Current Trends in Translation Studies*, St. Jerome Publishing, Manchester, pp.123-136.

Höge, Monika. (2002), 'Towards a Framework for the Evaluation of Translators' Aids Systems', Ph.D. thesis, Helsinki University, Helsinki.

Hutchins, John. (1998), 'The Origins of the Translator's Workstation', *Machine Translation*, vol. 13, no. 4, pp. 287-307.

Hutchins, W. John and Somers, Harold L. (1992), *An Introduction to Machine Translation*, Academic Press Limited, London.

Kay, Martin. (1980/1997), 'The Proper Place of Men and Machines in Language Translation', Research report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, CA. Reprinted in *Machine Translation* (1997), vol.12, nos. 1-2, pp.3-23.

King, Margaret. (1993), 'Sur l'évaluation des systèmes de traduction assistée par ordinateur', in: Bouillon, Pierrette and Clas, André (eds) *La Traductique*, Les Presses Université de Montréal, Montréal, pp.261-269.

Lanctôt, François. (2001), 'Splendeurs et petites misères… des mémoires de traduction', *Circuit*, vol. 72, p.30.

Langé, J.M., Gaussier, É. and Daille, B. (1997), 'Bricks and Skeletons: Some Ideas for the Near Future of MAHT', *Machine Translation*, vol. 12, nos. 1-2, pp.39-51.

Langlais, Philippe, Sauvé, Sébastien, Foster, George, Macklovitch, Elliott, and Lapalme, Guy. (2000), 'Evaluation of TransType, a Computer-aided Translation Typing System: A Comparison of a Theoretical- and a User-Oriented Evaluation Procedures', *LREC 2000 Second International Conference on Language Resources and Evaluation*, Athens, pp.641-648.

Lavallé, François. (2002), 'MultiTrans, un outil nouveau genre', *Circuit*, vol. 74, p.32.

L'Homme, Marie-Claude. (1999), *Initiation à la traductique*, Linguatech éditeur inc, Brossard, Québec.

Lynch, Clove and Heuberger, Andres. (2001), 'Calculating Return on Investment in Tools', *MultiLingual Computing and Technology*, vol. 12, Issue 3, pp.41-44.

Macklovitch, Elliott and Russell, Graham. (2000), 'What's Been Forgotten in Translation Memory', in: White, John S. (ed) *Envisioning Machine Translation in the Information Future: Proceedings of the 4th Conference of the Association for Machine Translation in the Americas (AMTA 2000)*, October 10-14 2000, Cuernavaca, Mexico, pp.137-146.

Macklovitch, Elliott, Simard, Michel and Langlais, Philippe. (2000), '*TransSearch*: A Free Translation Memory on the World Wide Web', in: *LREC 2000 Second International Conference on Language Resources and Evaluation*, Athens, pp.1201-1208.

Melby, Alan. (1982), 'Multi-level Translation Aids in a Distributed System' in: *COLING 82, Proceedings of the Ninth International Conference on Computational Linguistics*, Prague, July 5-10, 1982, NorthHolland Publishing Company, Amsterdam, pp.215-220.

Melby, Alan. (1992), ‘The translator workstation’, in: Newton, John (ed) *Computers in Translation: A Practical Approach*, Routledge, London/New York, pp.147-165.

Melby, Alan with Warner, C. Terry. (1995), *The Possibility of Language*, John Benjamins Publishing Co., Amsterdam/Philadelphia.

MultiCorpora R&D Inc. (No date) *User Manual: MultiTrans Pro*, MultiCorpora R&D Inc., Gatineau, Canada.

MultiCorpora R&D Inc. (2003, January 7 – last updated), ‘ MultiTrans™: Translation Support and Language Management Solutions for Everyone’ (MultiCorpora Translation Tools) Available: http://www.multicorpora.ca (Accessed: 2003, March 2).

National Institute of Standards and Technology. (2002, January 29 – last updated), ‘Edit Distance’ (Dictionary of Algorithms and Data Structures) Available: http://www.nist.gov/dads/HTML/editdistance.html (Accessed: 2003, January 11).

O’Brien, Sharon. (1998), ‘Practical Experience of Computer-Aided Translation Tools in the Software Localization Industry’, in: Bowker, Lynne, Cronin, Michael, Kenny, Dorothy and Pearson, Jennifer (eds) *Unity in Diversity? Current Trends in Translation Studies*, St. Jerome Publishing, Manchester, pp.115-122.

Planas, Emmanuel and Furuse, Osamu. (1999), ‘Formalizing Translation Memories’, in: *Proceedings of MT Summit VII*, Singapore, pp.331-339.

Rico, Celia. (2000), ‘Evaluation Metrics for Translation Memories’, *Language International*, vol. 12, no. 6, December 2000, pp.36-37.

Rode, Tony. (2000), ‘Translation Memory: Friend or Foe?’, *International Journal for Language and Documentation*, April 2000, pp.12-13.

Schwab, Wallace. (2001), ‘Translation Memories in Transition’, *Circuit*, vol. 71, p.30.

SDL International. (2002, October 6 – last updated), ‘SDLX Translation Suite’, (SDL International: Translation - Localization - Globalization - Computer Aided Translation) Available: http://www.sdlintl.com/sdlx (Accessed: 2002, October 6).

Shadbolt, David. (2001), ‘The Translation Industry in Canada’, *MultiLingual Computing & Technology*, # 16, vol. 13, Issue 2, pp.30-34.

Simard, Michel and Langlais, Philippe. (2000), ‘Sub-sentential Exploitation of Translation Memories’, *LREC 2000 Second International Conference on Language Resources and Evaluation*, Athens (Proceedings on CD-ROM).

Somers, Harold. (1999), ‘Review Article: Example-based Machine Translation’, *Machine Translation*, vol. 14, no. 2, pp.113-157.

Sprung, Robert C. (ed). (2000), *Translating Into Success: Cutting-edge strategies for going multilingual in a global age*, John Benjamins Publishing Co., Amsterdam/Philadelphia.

STAR. (2002 – last update), ‘Transit – Product Overview’, (STAR Language Technology for Professional Translators), Available: http://www.star-transit.com/products/transit/en/ (Accessed, 2002, October 6).

Topping, Suzanne. (2000), ‘Sharing Translation Database Information’, *MultiLingual Computing & Technology*, #33, vol. 11, Issue 5, pp.59-61.

TRADOS. (2002 – last update), ‘TRADOS – Language Technology For Your Business’, (TRADOS – Language Technology For Your Business), Available: http://www.trados.com/ (Accessed: 2002, October 6).

TRADOS. (2001), *TRADOS Specialist Guide*. TRADOS Ireland, Ltd., Dublin.

TRADOS. (2001), *Translation Guide*. TRADOS Ireland, Ltd., Dublin.

Trujillo, Arturo. (1999), *Translation Engines: Techniques for Machine Translation*, Springer, London.

Voyer, Louise. (2002), ‘Prudence’, *InformATIO*, vol. 31, no. 2, p.4.

de Vries, Arjen-Sjeord. (2002), ‘Getting Full or Fuzzy? The payment issue with full matches generated by translation memory systems’, *Language International*, June 2002, pp.44-47.

Webb, Lynn E. (1998), ‘Advantages and Disadvantages of Translation Memory: A Cost/Benefit Analysis’, M.A. thesis, Monterey Institute of International Studies, Monterey, CA.

Zerfaß, Angelika. (2002a), ‘Evaluating Translation Memory Systems’, in: *LREC 2002: Language Resources in Translation Work and Research*, May 28 2002, Las Palmas de Gran Canaria, pp.49-52.

Zerfaß, Angelika. (2002b), ‘Comparing Basic Features of TM Tools’, *Language Technology Supplement, MultiLingual Computing and Technology*, #51, vol.13, Issue 7, pp. 11-14.

Appendix A – List of Available Tools

Traditional sentence-based TM tools

Tool	Developer	URL
Déjà vu	Atril	www.atril.com
SDLX	SDL International	www.sdlintl.com
Transit	STAR	www.star-transit.com
TRADOS	TRADOS	www.trados.com
Translation Manager	IBM	www-4.ibm.com/ software/ad/translat/tm/

CSB-based TM tools

Tool	Developer	URL
LogiTerm	Terminotix	www.terminotix.com
MultiTrans	MultiCorpora	www.multicorpora.com

Software localization tools that include a TM component

Tool	Developer	URL
Passolo	Passolo	www.passolo.com
Catalyst	Alchemy	www.alchemysoftware.ie

Bilingual Concordancers

Tool	Developer	URL
TransSearch	RALI	www-rali.iro.umontreal.ca

Appendix B – Sample Document Marked Up by Evaluator

Descriptions de tâches

Directeur régional, services administratifs et installations (AS-07)

Résultats axés sur le service à la clientèle :

Assurer la gestion et la direction des programmes et services suivants au sein du Ministère : immobilier, télécommunications, acquisitions des biens et services (installations), services administratifs (bureautique, courrier, gestion de l'information, formulaires, politiques environnementales).

Activités principales :

Assurer un support opérationnel et une orientation fonctionnelle dans les domaines en question pour l'ensemble des points de services de DRHC dans la région du Québec.
Gérer les ressources humaines et les différents budgets attribués à sa division, nécessaires à la réalisation des différents programmes ; négocier avec les administrations régionales, nationales et le Conseil du Trésor les enveloppes budgétaires nécessaires. Mobiliser les différentes équipes spécialisées dans la prestation d'une variété de programmes et de services destinés à la clientèle interne du Ministère ; tenir compte des différentes orientations régionales et nationales quant à la gestion des ressources humaines et des ressources financières.
Développer et orienter la région dans l'exécution et la prise de décision touchant les installations et services administratifs.
Représenter la région pour négocier des ententes ou régler des litiges avec les organismes centraux, les autres ministères, les autres régions, les autres paliers de gouvernement et les organismes extérieurs ; siéger à titre d'expert, sur des comités nationaux, régionaux et locaux.
Développer et émettre des procédures, normes, directives et stratégies, concevoir des programmes relatifs aux installations et aux services administratifs, comme entre autres le plan d'aménagement des locaux, en tenant compte des orientations et de la vision du ministère.
Établir la planification à moyen et long terme dans les champs de compétence en tenant compte de l'évolution des différents marchés externes et l'évolution rapide des différentes technologies et les besoins actuels et futurs de l'organisation (immobilier, télécommunications, gestion de l'information, etc.).
Établir des partenariats internes et externes afin de faciliter la prestation des programmes et services.

Conseiller les membres du Conseil régional de Direction en regard de l'optimisation de l'utilisation des locaux, des aménagements, les systèmes de télécommunications, des biens, l'équipement, etc.; fournir à ce titre l'expertise et le support technique.

Effectuer la recherche et le développement dans les champs d'expertise propres afin de mettre en œuvre les meilleures solutions dans le but d'améliorer l'efficacité et réduire les coûts.

Appendix C – Time-Gain Scores

Appendix D – Time-Loss Penalties