Written by Błażej Palacz

A Comparative Study of CAT tools (MAHT workbenches) with Translation Memory Components

Master Thesis
Written under guidance of
prof. Włodzimierz Sobkowiak
The School of English
Adam Mickiewicz University

Poznań, Poland 2003

Acknowledgements

I would like to express my gratitude to Marcin Feder Ph.D. from School of English at Adam Mickiewicz University in Poznań who played a crucial role in arousing my interest in CAT and was always ready to promptly answer all my queries as well as to provide insightful comments.

Introduction

We live in the global village, an interrelated world, in which the need to communicate effectively is growing fast and acquires a new significance.

In the era of globalization the demand for an unimpeded flow of information among various languages has grown considerably. Many businesses operate on the worldwide market and need to localize their products and services.

As a result, the market for translation services is booming and its extensive needs are still largely unsatisfied. In turn, more and more is demanded from translators in terms of both productivity and quality.

Fortunately, the advances in technology offer support to the translator. Ever more sophisticated and powerful Computer Assisted Translation software comes into play and offers the opportunity to meet these demands. Powerful computer technology can enhance the uniquely human abilities of translators by coupling them with raw computing power. Consequently, a translator may truly enjoy the best of both worlds.

However, even though translation technology enjoys a growing popularity and its study has flourished, in Poland its investigation and application still remains very limited and many, even truly seasoned translators, approach this technology with distrust.

The present thesis, as limited as it may be, is an attempt at presenting the basics of CAT and providing a sample comparison of CAT tools available on the market.

In the first chapter the reader will find a general introduction to concepts fundamental to translation technology in question, including information on the history of their development.

The second chapter is devoted to an overview of functionalities provided by the latest in CAT software and is meant as a general presentation of modules contained in a typical CAT software package.

In the third chapter the author discusses CAT evaluation and presents a feature checklist used to perform a sample comparative evaluation of four CAT software packages.

The final chapter presents the results of the sample evaluation.

Chapter 1 Introduction to key concepts.

1.1 Computer Assisted Translation (CAT)

The present thesis may be seen as a part of research in the broad domain known as Computer Assisted Translation.

Marcin Feder states that "Computer Assisted/ Aided Translation is a very broad and general term used to describe various machine, mainly computer, techniques employed to (fully or partially) automate or assist human translation" (Feder 2001: 49). Definitely the stress should be put on the word *assisted*, as there is some confusion concerning the inclusion of Machine Translation into CAT (cf. Feder 2001: 49, Schüller 1995: 8). The author would argue that even though CAT is a term inclusive of Machine Assisted Human Translation defined below, Machine Translation should not be seen as a part of CAT as it is understood in this thesis.

1.2 Machine Assisted Human Translation (MAHT)

Another, much more specific term central to the present work is that of Machine Assisted Human Translation (MAHT). Machine Assisted Human Translation may be defined as a "type of translational activity where a human remains the pivotal part of the translation process (the translation proper is still performed by a human being and the computer technology only provides assistance)" (Feder 2001: 49). The relation between the human and machine involvement is a crucial characteristic feature of MAHT, which sets it apart from Machine Translation (also referred to as Automatic Translation) or its sub domain of Human Assisted Machine Translation (cf. Arnold et al. 1994, Feder 2001). One could venture to say that MAHT tools, in a broader context, fully conform to the philosophy of ideal collaboration

of man and machine put forward by Donald A. Norman[1], in his words: "Technologies have traditionally forced people to conform to the needs of machines. But in this era of advanced information processing, the power of the machine can readily be tailored to the needs of the user" (Norman in Schüller 1995: 1).

MAHT incorporates a broad spectrum of technologies. According to Feder "MAHT embraces a variety of tools offered to support translation, e.g. terminology management systems; translation memories; multilingual corpora; text alignment techniques, often (...) within an integrated working environment" (Feder 2001: 49-50).

Even though the development of MAHT systems may be considered as an offshoot of MT research, and the distinction between Machine Translation (MT) and MAHT is anything but clear, the author would argue that one could try and base such a distinction on who or what is the dominating party in the human-machine relationship. Hence, to distinguish MAHT from MT, it could be argued that the central role which the ***human translator*** performs in the MAHT process is the crucial distinctive feature of MAHT, and as such, cannot be stressed enough. In the context of MAHT, the translation proper is performed by the ***human***, while the machine plays only an auxiliary role.

Undoubtedly, MT research had a considerable influence on the emergence and development of MAHT systems (described in detail below), especially in the work of Alan Melby who made an attempt to transfer the conclusions arrived at during his work on MT systems to the realm of human translation (cf. Schüller 1995: 8). Undeniably, the failed attempts at creating an MT system capable of Fully Automatic High Quality Machine Translation proved to be a driving force behind MAHT research and development. In 1964 the US National Academy of Sciences established the Automatic Language Processing Committee (ALPAC) in order to evaluate the results of MT research that consumed around 20 million

[1] Professor of Computer Science at Northwestern University, prolific publicist, member of, among others, Human Factors & Ergonomics Society, advocate of 'human-centered design'. For a full biography access http://www.jnd.org/bio-sketch.html

dollars worth of funds since 1954. In 1966 ALPAC produced its (in)famous report that literally crushed MT research and halted funding of MT research for the next 20 years. The conclusions presented by the committee meant the end of the then-MT research hopes: "(...) we do not have useful machine translation. Further, there is no immediate or predictable prospect of useful machine translation." (ALPAC 1966: 32) Some would argue (cf. Hutchins 1996) that the report did not do justice to MT research, and was flawed in its nature, but regardless of its weaknesses its impact on NLP (Natural Language Processing) research is undeniable. What is most important in the context of MAHT are the final recommendations listed in this document, suggesting that research should be supported on:

"1. practical methods for evaluation of translations;

2. *means for speeding up the human translation process*;

3. evaluation of quality and cost of various sources of translations;

4. investigation of the utilization of translations, to guard against production of translations that are never read;

5. study of *delays* in the over-all translation process, and *means for eliminating them*, both in journals and in individual items;

6. *evaluation of the relative speed and cost of various sorts of machine-aided translation*;

7. *adaptation of existing mechanized editing and production processes in translation*;

8. the over-all translation process; and

9. production of adequate reference works for the translator, *including the adaptation of glossaries that now exist primarily for automatic dictionary look-up in machine translation.*" (ALPAC 1966: 34)

From the above it clearly follows that even though the report proved disastrous for MT research of the time, it led to fruitful reformulation of NLP research targets, with more stress put on supporting the human translator (ALPAC 1966: 32): "machine aids may be an important adjunct to human or machine-aided translation, (…) Machine-aided translation may be an important avenue toward better, quicker, and cheaper translation." In the wake of the ALPAC

report one of the approaches to MT system development focused on "general purpose systems (…) as an *aid to translation.*" (Warwick in Schüller 1995: 7).

It clearly transpires from the history and the present state of the art of MT that a fully automated high quality translation thought to be at an arms reach in the 50s, still remains, at best, a song of the future (cf. Arnold et al. 1993), therefore the shift in NLP research in the direction of MAHT is easily understandable.

While discussing MAHT, it is also important to mention the distinction put forward by Feder (2001: 50) between "*proper MAHT tools* which (…) are the most sophisticated translation support tools, namely workbench packages equipped with translation memory systems and offering terminology management, dictionary as well as automatic translation facilities" and the various so-called *authoring tools* "word processors; spell, grammar, style checkers; electronic dictionaries, encyclopedias and other reference works; terminological databases; text retrieval tools, email and Internet services; context-sensitive searching utilities; automated dictionary update interfaces and morphological analysis tools". The latter often form a part of the former, but proper MAHT tools possess the distinctive feature of integrating these particular facilities bringing about advantages resulting in a certain synergy effect. The tools evaluated in the present thesis all conform to the aforementioned definition of proper MAHT tools.

1.3 MAHT Workbench, also referred to as an Integrated Translation System

A translator's workbench, also referred to as a workbench package or an integrated translation system, may be defined as a group of software systems integrated in a single working environment that provide a translator with various functionalities that support his/her work (cf. Spies 1995: 2).

Despite the fact that MAHT technologies in general are deeply rooted in MT research, the development of integrated translation support systems may be characterized as a bottom up

process, grown out of the needs of the translation community (cf. Feder 2001, Webb 1999), which had to face the growing volume of translation, shorter deadlines, as well as the ever higher demands concerning consistency of terminology in, and quality of, translated documents.

Schüller traces the development of MAHT workbenches back to the already mentioned report of ALPAC from 1966, which takes note of two systems that were already in existence in 1965, and were an early attempt at automating the process of translation: one was the system used in a Translation Agency of the German Federal Air Force in Mannheim. The translator could use a text specific glossary containing terms he marked and searched for. Another system was that of the European Coal and Steel Community consisting of a terminology database storing terms with the context (text-based) in which they appeared (Schüller 1995: 12).

But the true theoretical foundations of translation workbenches were laid down by Melby in 1982 when he presented a concept of an ideal Individual Translator Workstation (cf. Schüller 1995: 9-11). Such an ideal system should contain three levels of translation support tools:

a) level 1 should consist of a simple text editing system with a customizable list of often reoccurring words, terminology management system with its own database stored on a local PC and access to a centrally managed databank, means of accessing pairs of Source Language and Target Language documents.
b) level 2 should provide access to machine readable SL documents and include tools enabling the study of these texts (for instance the frequency at which a certain term occurs in the SL document). The system should allow for a screen area “suggestion box option” in which the system would automatically suggest TL equivalents of SL terms contained in the term bank. These suggestions could be easily entered into the TL document by the translator. Optionally, a morphological mechanism should suggest the proper inflexion of a term.

c) level 3 tools should provide the integration of a complete MT component into the Translator Workstation. Melby sees the MT component as a self-evaluating system, which would be capable of evaluating the quality of the suggested translation equivalents, and would divide them into five categories: A, B, C, D and E. The translation manager could choose what category, or grade, could be either suggested to the translator, or automatically entered into the TL document.

The main difference between b) and c) is such that in c) whole sentence equivalents are presented, while in b) the suggestions consist in specific term equivalents. The optional character of support provided by the system stressed by Melby is important in that it does not force the translator to accept the suggestions provided by the system. They can be utilized when they are useful, and if they are not, they do not unnecessarily hinder the translator's performance. Melby sees a number of advantages of such a level-based approach: thanks to a prompt introduction of these support systems the communication between the people involved in the translation process is much faster. Furthermore, the translators play an active role in the translation process and are not frustrated by being "garbage collectors" improving on the output provided by machines in post-editing.

These ideas were further developed in 1990 when Melby presented an upgraded model of an Individual Translator Workstation with two levels of translation support tools (cf. Melby 1990):

Level 1

The selection of an operating system. The OS should be versatile. Melby suggested the use of MS-DOS + MS Windows combination, as GUI (Graphical User Interface[2]) based operating

[2] An interface that enables you to choose commands, start programs, and see lists of files and other options by selecting from windows, icons, and menus on the screen. Choices can be made either with keyboard commands or by using the mouse to move an on-screen pointer. Graphical user interfaces take full advantage of the bit-mapped graphics displays of personal computers. GUIs are easier for most people to learn to use than command-driven interfaces. Also, a GUI uses a standard format for text and graphics, so that different applications running under a common GUI can share data. A graphical user interface is used on all Macintosh computers. Many applications for

systems could utilize the greatest selection of available applications, and Windows was capable of using non-Latin characters

a) Improved text correction support. Text editing systems should include text correction facilities for a number of languages and offer the possibility of adding support for more languages.

b) Support of SGML (Standard Generalized Markup Language)[3] as an exchange format. Translators should not be forced to work in a number of different text editing systems. SGML should be the exchange format, which all other text formats could be reformatted into, and all text editing systems should include tools enabling such a reformulation.

c) Utilization of telecommunication technology. Telecommunication technology should be used to a greater extent to facilitate the transfer of documents and glossaries, as well as to improve access to databases containing useful research data.

d) Terminology management. Melby underlined the importance of both a freely customizable data storage structure and the possibility of accessing various databases. He also stressed the significance of functionalities such as: filtering of information contained in an entry to be presented, improved search and analysis.

Level 2

a) Text analysis tools. A dynamic concordancing system should, for example, index all the words in a text to provide information about the frequency of a particular term, or possibly context in which it was used. Text analysis tools could also help compare the text database with the text to be translated, in order to find and study new terms.

IBM PC and compatible computers use GUIs, most of them based either on Microsoft Windows or Presentation Manager (for OS/2 systems). UNIX systems such as X-Windows also use GUIs. *Compact American Dictionary of Computer Words © 1995, 1998 by Houghton Mifflin Company*

[3] A coding system for marking up a text so that different types of information are identified. For instance, in an encyclopedia the headwords, main text, and cross references can be tagged in a standard way, irrespective of how they may be typeset in printed material or displayed in electronic form. Hypertext Markup Language (HTML) is a particular application of SGML used in electronic publishing. Xtensible Markup Language (XML) is a simplified version designed for use on the Internet. *The Macmillan Encyclopedia 2001*

b) Automatic terminology search. It should make the TL equivalent terms contained in the text database available for simple insertion in the TL text.

c) Generating of, and access to, bilingual databases. Calling it *synchronized bilingual retrieval* Melby describes the storage of already translated texts mentioned in 1982 in more detail. After a translation is completed, corresponding SL and TL segments should be stored in such a way that the revision of the original document and, respectively, the creation of an updated version would be significantly simplified. With the help of appropriate routines the differences between the new document version and its old, already translated, version could be quickly found, leaving the translator with the task of translating only those new and different parts of the text.

The systems of synchronized bilingual retrieval envisaged by Melby, came into existence later, and became the fundamental technology behind all MAHT workbenches known as Translation Memory (defined in 1.4)

In 1980, before Melby presented his ideas, Martin Kay had already foreseen a shift in MT research towards a "translator's workstation equipped with an intelligent text editor (split screen), terminology lookup facility with morphological analysis, global editing functions and additional modules. He also postulated that some form of a history archive of the translation process be kept" (Kay 1997[4]: 13 ff. in Feder 2001: 16).

Feder stresses the role of translators and general developments of electronic technology in the rapid development of MAHT systems in the early eighties, and traces it to "the actual advent of MAHT technologies, quite independently from MT research and ALPAC recommendations (…) when translators, alongside the general public, were first introduced to personal computers on a large scale" (Feder 2001: 13).

[4] The article was originally written in 1980 as a Xerox working report and was reproduced in 1997 in *Machine Translation.*

Translators themselves, regardless of ideas stemming from MT research, at least conceptually contributed to the development of MAHT workbenches, as they always wanted to translate faster and better. All these factors led to the emergence of the present state of the art in MAHT workbenches: "With the later addition of terminology management tools, resident or external terminological databases, concordancing, tag protection and alignment features, the translator's word processor coupled with translation memory and all (…) add-on modules became an integrated working environment known popularly as *workbench*" (Feder 2001: 15).

The period between 1992-1994 saw important developments in MAHT tools history, when the major software suppliers (such as: the IBM with its Translation Manager, used internally in the company at first; GlobalWare with XL8; Eurolang with Optimizer; SDL International with SDLX, Atril with Déjà Vu) released their first fully-fledged proper MAHT tools, that slowly but surely started gaining popularity among translators.

It is difficult to determine which company was at the forefront of these changes, but it may be safely assumed that these developments owe a lot to a German company TRADOS, one of the major MAHT software suppliers today. The company founded in 1984, worked on the "grandfather of all translation memory systems" (Kingscott 1999: 9-11), ALPS Translation Support System, in cooperation with the IBM.

The system could not handle large amounts of text, had excessive hardware requirements for that time, and was not economically viable, even though it had been further developed (addition of a Terminology Management System) and marketed by INK Netherlands, it did not sell (one has to be aware that computers were extremely rare and expensive at that time and most translators still used typewriters).

TRADOS acquired rights to this product and decided to develop its own package on its basis, and in 1987 it developed the first split-screen translation editor (with one half reserved for SL text, and the other for TL text). In 1992 TRADOS introduced its first proper MAHT package – Translator's Workbench 2 for DOS.

1.4 Translation Memory (TM)

Translation Memory is the key component of MAHT workbenches, and as a fundamental technology behind all MAHT systems it is a well-defined and understood concept (cf. Feder 2001: 57).

According to the EAGLES (Expert Advisory Group on Language Engineering Standards) Evaluation Working Group reports, "translation memory is a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions.(…) Unlike other language engineering systems, a TM does not come provided with linguistic data (…) rather it is a shell to be populated with translation equivalents" (EAGLES 1995: 140, EAGLES 1999: 106).

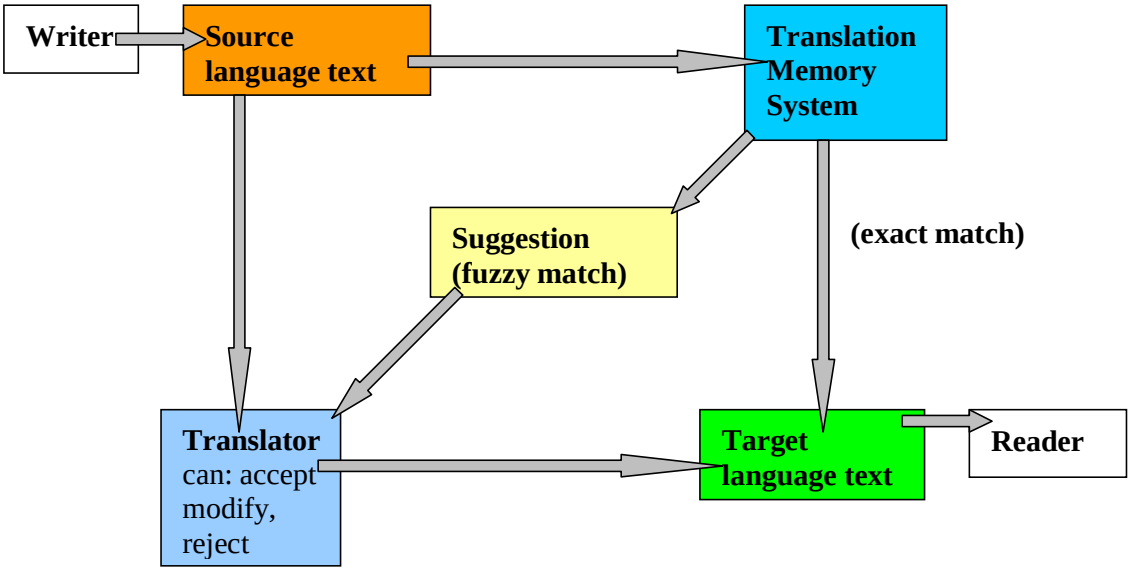

Figure 1: Task model for document translation with support of a TM. (cf. EAGLES 99: 106)

In other words, and according to many sources, a Translation Memory, also referred to as Sentence Memory or Translation Archive, is basically a database containing pairs of translation units in the target and source languages, that is, e.g. a sentence in the original language alongside with its translation. Translation workbenches provide the user with the so-

called *alignment* tool, which lets him create a translation memory from previously translated documents, by coupling SL and TL translation units (TUs) and entering them into the translation memory database. The process of creating TU pairs is referred to as ***alignment.*** Even though definitions vary, the basic translation unit is usually defined as a sentence, but it can also be a bigger (a paragraph) or a smaller portion of a text (a phrase, a word, or even a number). Furthermore, what constitutes a translation unit may be set by the translator himself, even though a sentence is definitely the most common type of a TU.

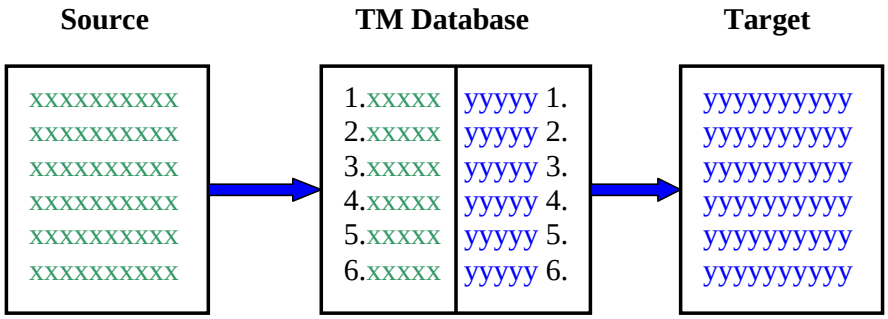

Figure 2: Basic Translation Memory Process Model (Webb 1999: 7)

To put it in a nutshell, the main advantage of translation memory is the fact that it significantly boosts the translator's productivity by providing leverage from already translated material in the process that is often fashionably called recycling. Once a particular sentence, or a phrase, has been translated and entered into the translation memory, every time the same or a similar phrase or sentence in the SL occurs, it will be (partially or fully) automatically translated (retrieved from the TM and entered into the TL document), relieving the translator of the tedious task of jogging his own, human, often faulty, memory. As a result, the translation with the help of TM is highly consistent and cost effective (cf. for example Ray, Eric 1999).

As the translation progresses and a SL unit similar or identical to the one stored in translation memory occurs, the system retrieves the corresponding TL unit, (with the threshold

level of correspondence, or similarity, usually set by the user - rarely less than 70%), and enters it into the SL document, leaving the decision to accept, reject, or modify (edit) the suggested translation to the translator.

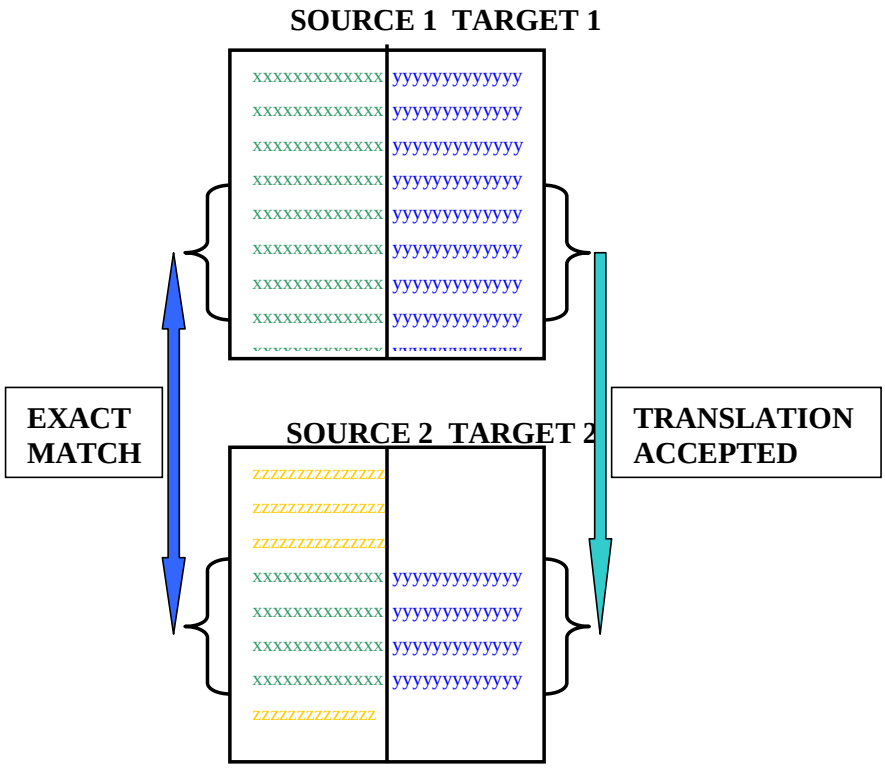

Figure 3: Exact matching (cf. Webb 1999: 10)

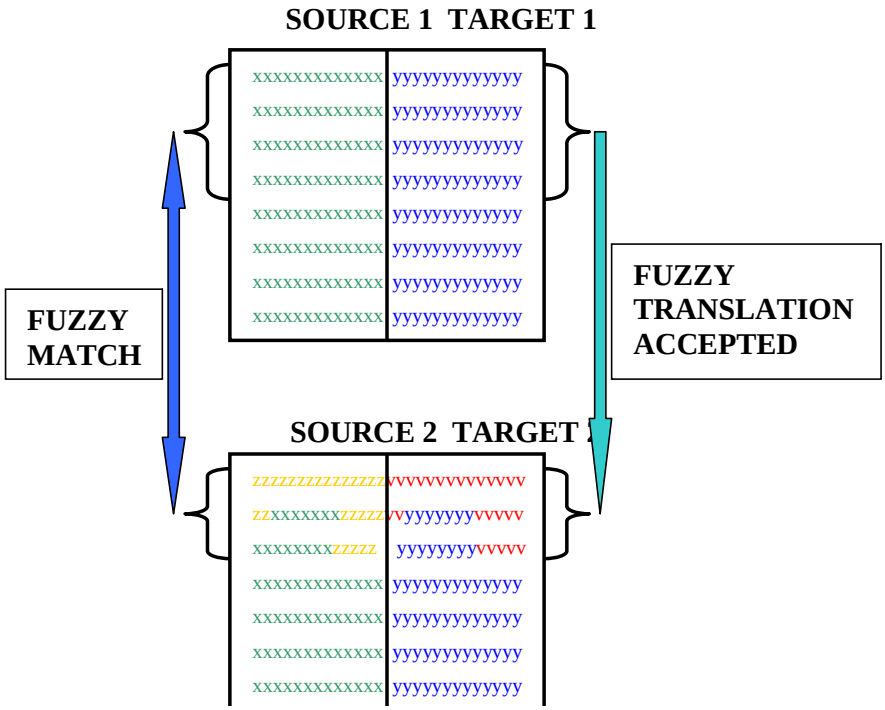

Figure 4: Fuzzy matching (cf. Webb 1999: 10)

At this juncture it is important to define the notions of *exact match* (vide Figure 3) and *fuzzy match* (vide Figure 4). Those two kinds of matches may be distinguished by the level of correspondence, or similarity, between the SL unit in a new translation project and the SL unit stored in the translation memory. 100% (identical) matches are called exact (perfect) matches, while all other (i.e. non-100%) matches are labeled as fuzzy matches. Translation support systems can enter exact matches automatically, which greatly enhances the translator's output (cf. Webb 1999, Benis 2002), but, along other drawbacks, increases the probability of accepting and propagating erroneous translation (cf. Webb 1999).

Obviously the increase in the translator's output offered by TM (coupled with a number of advantages offered by MAHT workbenches, terminology consistency among other cf. Webb 1999) depends on a number of factors, such as (without going into details) the internal (reoccurrence of words, phrases and sentences in a single document) and external (external reoccurrence of those repetitions among two or more documents) repeatability of a source language text (cf.1.5) or texts, type of the text (it is argued that technical documentation lends itself best to TM supported translation, even though the question of what constitutes an "ideally TM translatable text" is still open to a certain extent (cf. Feder 2002b)), or the size and contents of the TM at the translator's disposal.

The most optimistic estimates give a staggering number of 20,000 to 40,000 words increase in daily output (cf. Benis 1998: 5) (compared with around 2000 to 3000 words of estimated daily translation output an unaided translator is capable of). Lynn A. Webb discusses

the issue of TM advantages in detail and her case studies and survey confirm a significant increase in a translator's daily output (cf. Webb 1999: 40-46).

Another advantage of a more general nature consists in the fact that the translator is relieved of the tedious, repetitive tasks that computers excel at and can safely focus on the more creative part of his work (for an in-depth discussion of advantages and disadvantages cf. Webb 1999; Ray, Eric 1999).

1.5 Internal and external repeatability (reoccurrence)

The phenomenon of repeatability could be called the second most important motivating force behind the development of MAHT systems, along the conclusions of MT research and the needs of translators.

As mentioned before, the key advantage of TM systems is the fact that they let the user recycle his old translations, by finding similar (fuzzy match) or identical (exact match) TUs in the new text that has to be translated and replacing them with equivalents contained in the TM. But for the TM system to serve its purpose, it needs to be able to find such an identical or similar item. Hence, if the new text has nothing to do with the previously translated material, (as far as its type is concerned) , the use of a TM is basically meaningless (cf. Hüberger 2002, Webb 1999). Therefore for a text to be a good material for TM-supported translation it needs to possess the crucial characteristic of ***repeatability*** (apart form other characteristics, cf. Feder 2002b).

Feder aptly defines repeatability as "(...) the degree of repetition of textual material within a given text or across a body of texts.(...) repeatability may be measured within one text or across a corpus of texts (of course, not all texts, but original documents and their subsequent versions or updates, documents related to the same subject domain or documents translated for the same client)." (Feder 2002).

Schüller, mentioning the results of a study of text repeatability in software handbooks performed by Magnus Merkel from the Department of Computer and Information of the University of Linköping, Sweden, distinguishes two types of repeatability: internal and external. Internal repeatability is the repeated occurrence of words, phrases and whole sentences in a single document, while external repeatability may be defined as the reoccurrence of these items across two or more documents (cf. Schüller 1995: 13).

Repeatability is definitely the most important feature of texts to be MAHT- translated that can make or brake the computer assisted translation process.