State of the Art in Translation Memory Technology

Uwe Reinke
Cologne University of Applied Sciences
uwe.reinke@fh-koeln.de

Commercial Translation Memory systems (TM) have been available on the market for over two decades now. They have become the major language technology to support the translation and localization industries. The following paper will provide an overview of the state of the art in TM technology, explaining the major concepts and looking at recent trends in both commercial systems and research. The paper will start with a short overview of the history of TM systems and a description of their main components and types. It will then discuss the relation between TM and machine translation (MT) as well as ways of integrating the two types of translation technologies. After taking a closer look at data exchange standards relevant to TM environments the focus of the paper then shifts towards approaches to enhance the retrieval performance of TM systems looking at both non-linguistic and linguistic approaches.

1 Introduction

Translation Memory (TM) systems are the most widely used software applications in the localization of digital information, i.e. the translation and cultural adaptation of electronic content for local markets. The idea behind its core element, the actual "memory" or translation archive, is to store the originals and their human translations of e-content in a computer system, broken down into manageable units, generally one sentence long. Over time, enormous collections of sentences and their corresponding translations are built up in the systems. TMs allow translators to recycle these translated segments by automatically proposing a relevant translation from the memory as a complete ("exact match") or partial solution ("fuzzy match") whenever the same or a similar sentence occurs again in their work. This increases the translator's productivity and helps ensure that the same terminology and expressions are consistently used across translations. Thus, TMs facilitate and speed-up the translation of a rapidly growing amount of specialised texts.

No other technology has changed the general conditions of translation as a professional service as radically as TM systems have done over the past 20 years. This might be due to the fact that TMs mainly support professional translators in their routine work without radically influencing cognitive translation processes in those situations that require the creativity and knowledge of the human translator.

Today most professional translators use TM technology on a regular basis (Massion 2005; Lagoudaki 2006). The most well-known commercial systems are *Across, Déjà Vu, memoQ, MultiTrans, SDL Trados, Similis, Transit* and *Wordfast*.[1]

[1] For a brief overview on TM technology see also Somers (2003) and Reinke (2006). Comprehensive investigations can be found in Reinke (2004).

Translation: Computation, Corpora, Cognition. Special Issue on Language Technologies for a Multilingual Europe, ed. by Georg Rehm, Felix Sasaki, Daniel Stein, and Andreas Witt. Volume 3, Number 1. June 2013.
ISSN 2193-6986

2 Translation Memory systems

2.1 History

The basic idea of computer-assisted reuse of human translations can be traced back to the 1960s, when the European Coal and Steel Community (ECSC) developed and used a computer system to retrieve terms and their contexts from stored human translations by identifying those sentences whose lexical items most closely matched the lexical items of a sentence to be translated:

> The translation of the sentence [i.e., the sentence stored in the data base; U.R.] is not done by the computer, but by a human translator. However, since the data produced by each query are added to the data base, the more the system is in use, the greater is the probability of finding sentences that have the desired term in the proper context. (ALPAC 1966, 27; emphasis in original)

Yet, modern TM systems differ considerably from the former ECSC application. As the quote from the ALPAC report shows, the latter was rather something like a bilingual keyword in context (KWIC) retrieval tool that mainly served the purpose of showing source language terms and their target language equivalents in their respective contexts. Retrieving previous translation units for reuse was, if at all, a secondary goal:

> The system utilized at CECA is one of 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 nearly match (in lexical items) the sentences in question. The translator then receives the desired items printed out with their context and in the order in which they occur in the source. (ALPAC 1966, 27)

A much broader reuse of existing machine-readable human translations with a clear focus on facilitating and accelerating revision processing by identifying unchanged passages was envisaged in a model developed by the translation service of the German Federal Army in the early 1970s (cf. Krollmann 1971). Apart from several lexical databases this model also envisaged subsystems for storing and analysing text corpora and translation archives stored on magnetic tape:

> […] via descriptors or keywords, large batches of text could automatically be searched for particular passages and then be displayed on video screens as an aid to the translator; […] For revised new editions of translations only the changed passages would have to be retyped. Insertion of changes and corrections into the old text would automatically be done by computer […]. (Krollmann 1971)

At the end of the 1970s EC translator Peter Arthern (1979) proposed even more far reaching computer-assisted support for the translator. His suggestions have to be seen in the context of a discussion led at that time within the European Commission about the use of terminology databases and the feasibility of introducing the MT system *Systran*. While Krollmann's model only seemed to include the reuse of identical text fragments (today known as "exact matches"), Arthern suggests a system that can also retrieve from the reference material similar source language sentences and their translations (today known as "fuzzy matches"):

> This would mean that, simply by entering the final version of a text for printing, as prepared on the screen at the keyboard terminal, and indicating in which languages to compare the new text, probably sentence by sentence, with all the previously recorded texts prepared in the organization in that language, and to print out *the nearest available equivalent for each sentence* in all the target languages, on different printers.
>
> The result would be a complete text in the original language, plus at least partial translations in as many languages as were required, all grammatically correct as far as they went and all available simultaneously. Depending on how much of the new original was already in store, the subsequent work on the target language texts would range from the insertion of names and dates in standard letters, through light welding at the seams between discrete passages, to the translation of large passages of new text with the aid of a term bank based on the organization's past usage. (Arthern 1979, 94f.; my emphasis)

While Arthern did not tackle the issue of "the nearest available equivalent" – or "similarity" - in more detail, he even envisaged the possibility of integrating TM and machine translation (MT):

> Since this form of machine-assisted translation would operate in the context of a complete text-processing system, it could very conveniently be supplemented by 'genuine' machine translation, perhaps to translate the missing areas in texts retrieved from the text memory. (Arthern 1979, 95)

Yet, it took another decade before the ideas sketched by Krollmann and Arthern became part of real applications and market-ready systems. The notion of automatically retrieving "exact matches" was first implemented in the early 1980s by ALPS Inc. (later ALPNET Corporation) in a simple component called "Repetitions Processing" as part of the company's commercial MT system called *Translation Support System* (TSS) (cf. Seal 1992). The reuse of similar sentences ("fuzzy matching") was supported by the first commercial TM systems like *IBM Translation Manager*, and *Trados Translator's Workbench II* that did not appear on the market before the early 1990s.[2]

2.2 Components

Apart from the "memory" or translation archive as its core element, a typical TM system consists of an array of tools and functionalities to assist the human translator. These usually include:

- a **multilingual editor** for reading source texts and writing translations in all relevant file formats of different word processing programs, DTP systems, etc., protecting the layout tags of these formats against accidentally being deleted or overwritten
- a **terminology management program** for maintaining termbases to store, retrieve, and update subject-, customer-, and project-specific terminology
- an **automatic term recognition feature** for automatically looking up in the termbase all terms that occur in the source text segment the translator is currently working on

[2] Hutchins (1998) and Reinke (2004, 36-41) provide further information on the history of TM systems.

- a **concordance tool** allowing users to retrieve all instances of a specific search string (single words, word groups, phrases, etc.) from a TM and view these occurrences in their immediate context
- a **statistics feature** providing a rough overview of the amount of text that can be reused from a TM for translating a new source document
- an **alignment tool** to create TM databases from previously translated documents that are only available as separate source and target text files by comparing a source text and its translation, matching the corresponding segments, and binding them together as units in a TM.

In addition, a few TM systems offer terminology extraction as an optional or an integrated feature to assist in populating termbases and setting up the terminology for an e-content localization project by extracting mono- or bilingual lists of potential terms from a selection of electronic (source and/or target) texts. Today, many TM suites also include support for machine translation, either by offering interfaces with MT systems or even by integrating their own MT component. Finally, some kind of project management (PM) support is built into most TM systems. These PM features may support:

- file handling and management (specification of all source language files, project-relevant termbases and TM databases, assistance in defining folder structures)
- management of client and translator data (addresses, contact persons, translators' skills, equipment, availability, etc.)
- workflow management (deadlines, project progress, etc.).

Figure 1 provides an overview of how the major components of a standard TM environment interact, while Figure 2 gives an example for a typical user interface of a commercial TM system.

Although professional translators often stress the need to constantly adjust to rapid technological changes in the field (some complaining about this constant pressure, others rather regarding it as a professional necessity and a challenge), it must be said that all in all the core functionalities of commercial TM systems have remained very much the same since the first – mostly still MS-DOS-based - applications became available at the beginning of the 1990s. Even the first versions contained a translation memory, a terminology management system and a (multilingual) editor, providing features like exact and fuzzy matching, pre-translation[3], concordance lookup, terminology recognition, etc. (cf. Figure 3). Of course, the matching algorithms – although still being based on simple character matching procedures – have been altered and modified to a considerable extent, and many additional features and functionalities have been added, so that a growing number of scholars, professionals and application providers now prefer to call TM systems "translation environments" or "translation environment tools (TEnT)" (cf. CERTT 2012, 8).

[3] Pre-translation refers to the batch process "of comparing a complete source text to a Translation Memory database and automatically inserting the translations of all exact matches found in the database. The result is a hybrid text containing pre-translated and untranslated segments." (eCoLoRe 2012)

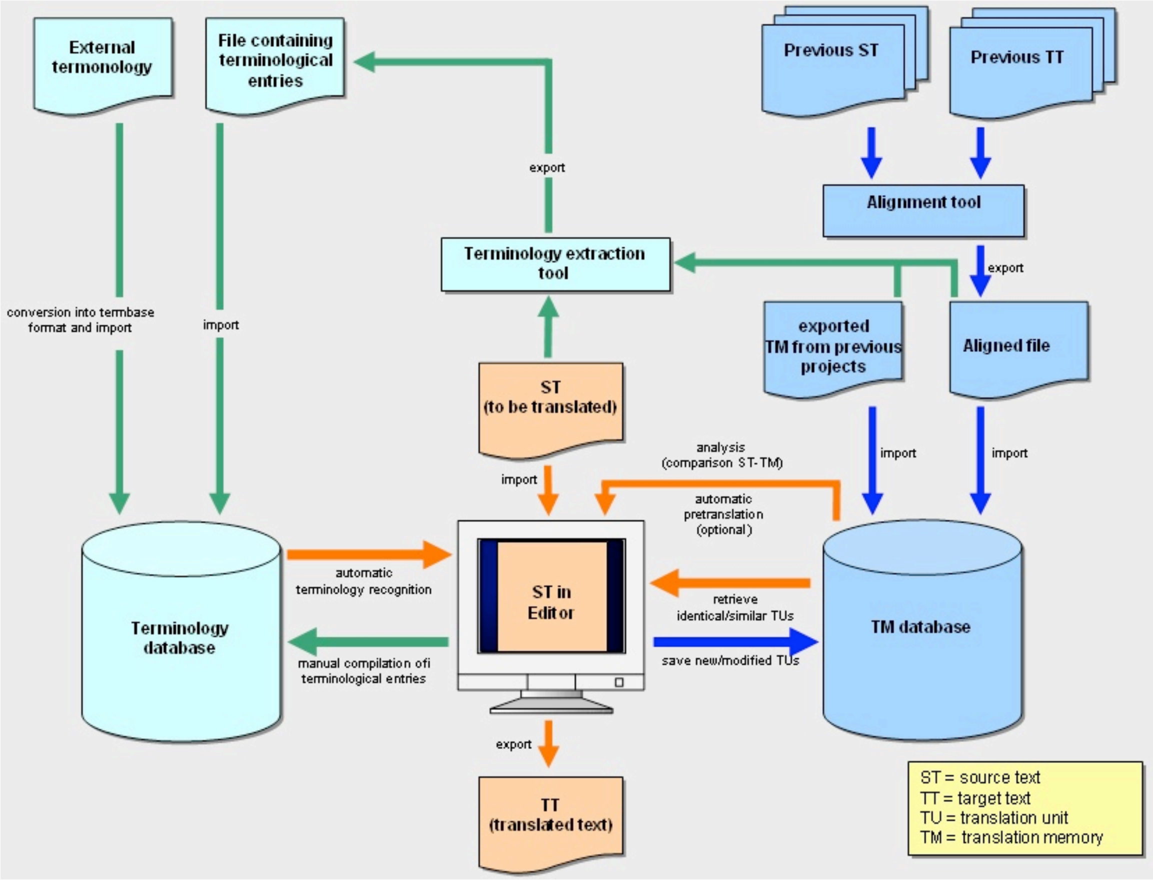
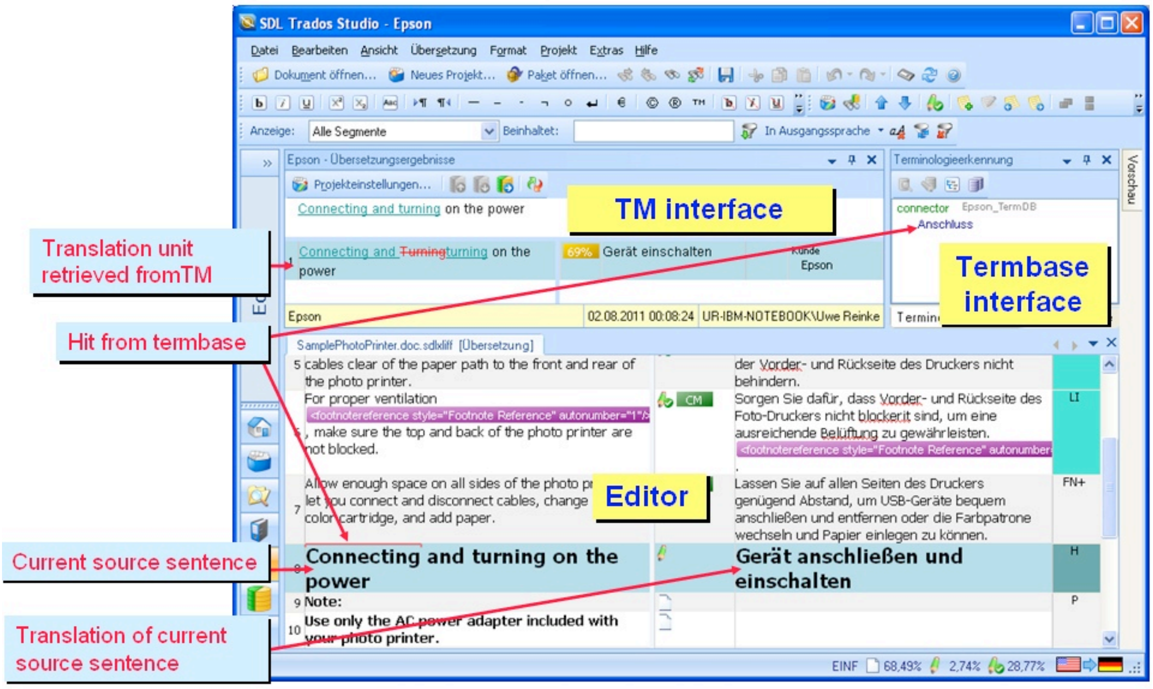

External termonology
File containing terminological entries
Previous ST
Previous TT
export
Alignment tool
Terminology extraction tool
export
conversion into termbase format and import
import
exported TM from previous projects
Aligned file
ST (to be translated)
analysis (comparison ST-TM)
import
import
import
automatic pretranslation (optional)
Terminology database
automatic terminology recognition
ST in Editor
retrieve identical/similar TUs
TM database
manual compilation ofi terminological entries
save new/modified TUs
export
TT (translated text)
ST = source text
TT = target text
TU = translation unit
TM = translation memory

Figure 1: Components and processes in a translation memory (TM) system (excluding project management and machine translation functionalities


SDL Trados Studio - Epson
Datei Bearbeiten Ansicht Übersetzung Format Projekt Extras Hilfe
Dokument öffnen... Neues Projekt... Paket öffnen...
Anzeige: Alle Segmente Beinhaltet: In Ausgangssprache
Epson - Übersetzungsergebnisse
Projekteinstellungen...
Terminologieerkennung
Vorschau
TM interface
Termbase interface
Editor
Translation unit retrieved fromTM
Hit from termbase
Current source sentence
Translation of current source sentence
connector Epson_TermDB
Anschluss
Kunde Epson
Epson 02.08.2011 00:08:24 UR-IBM-NOTEBOOK\Uwe Reinke
SampleePhotoPrinter.doc.sdlxliff [Übersetzung]
EINF 68,49% 2,74% 28,77%

Connecting and turning on the power

1 Connecting and ~~Turning~~turning on the power 69% Gerät einschalten

5 cables clear of the paper path to the front and rear of the photo printer. der Vorder- und Rückseite des Druckers nicht behindern.

6 For proper ventilation <footnotereference style="Footnote Reference" autonumber="1"/, make sure the top and back of the photo printer are not blocked. CM Sorgen Sie dafür, dass Vorder- und Rückseite des Foto-Druckers nicht blockert sind, um eine ausreichende Belüftung zu gewährleisten. <footnotereference style="Footnote Reference" autonumber LI

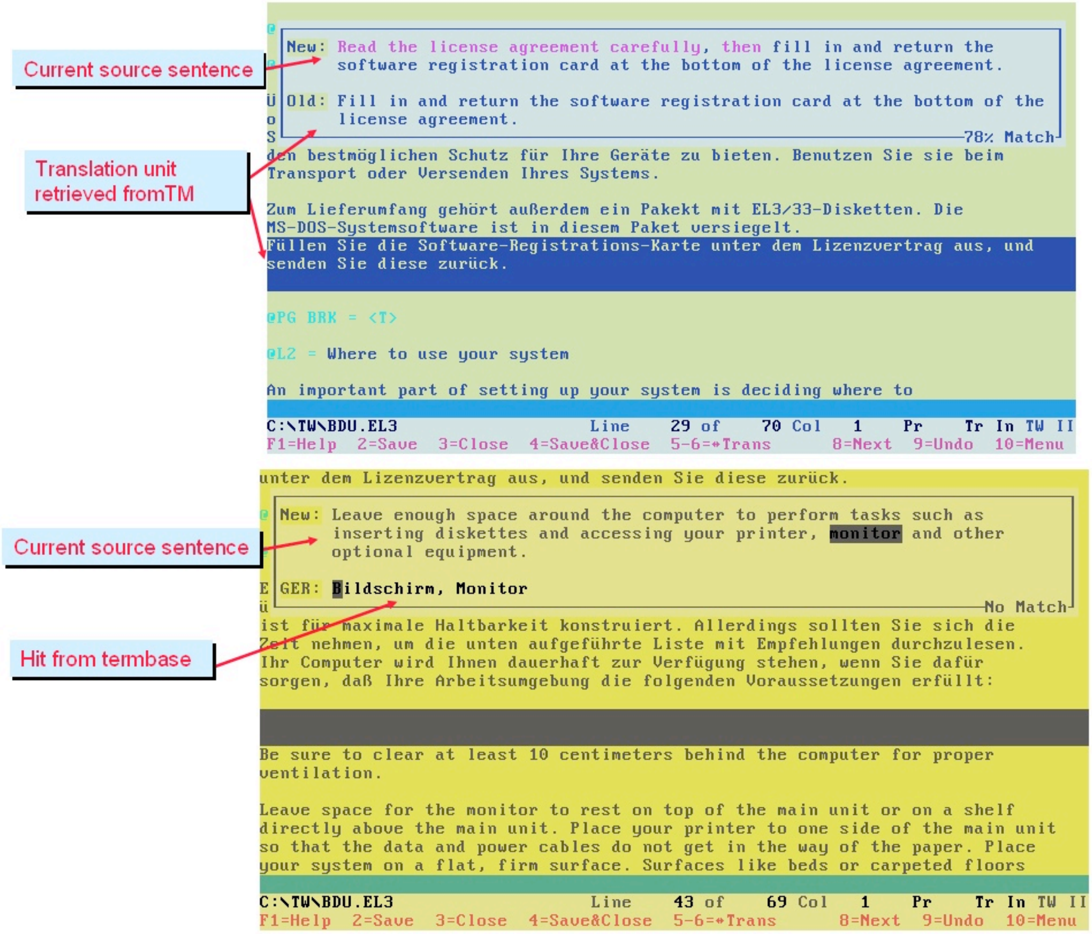
7 Allow enough space on all sides of the photo pr let you connect and disconnect cables, change color cartridge, and add paper. Lassen Sie auf allen Seiten des Druckers genügend Abstand, um USB-Geräte bequem anschließen und entfernen oder die Farbpatrone wechseln und Papier einlegen zu können. FN+

8 Connecting and turning on the power Gerät anschließen und einschalten H

9 Note: P

10 Use only the AC power adapter included with your photo printer.

Figure 2: User interface of SDL Trados Studio



```
@
 New: Read the license agreement carefully, then fill in and return the
@      software registration card at the bottom of the license agreement.

Ü Old: Fill in and return the software registration card at the bottom of the
o      license agreement.
S                                                                  78% Match
den bestmöglichen Schutz für Ihre Geräte zu bieten. Benutzen Sie sie beim
Transport oder Versenden Ihres Systems.

Zum Lieferumfang gehört außerdem ein Pakekt mit EL3/33-Disketten. Die
MS-DOS-Systemsoftware ist in diesem Paket versiegelt.
Füllen Sie die Software-Registrations-Karte unter dem Lizenzvertrag aus, und
senden Sie diese zurück.

@PG BRK = <T>

@L2 = Where to use your system

An important part of setting up your system is deciding where to

C:\TW\BDU.EL3                  Line    29 of    70 Col    1     Pr     Tr In TW II
F1=Help  2=Save  3=Close  4=Save&Close  5-6=+Trans       8=Next  9=Undo  10=Menu
```

```
unter dem Lizenzvertrag aus, und senden Sie diese zurück.

@ New: Leave enough space around the computer to perform tasks such as
        inserting diskettes and accessing your printer, monitor and other
        optional equipment.

E GER: Bildschirm, Monitor
ü                                                                  No Match
ist für maximale Haltbarkeit konstruiert. Allerdings sollten Sie sich die
Zeit nehmen, um die unten aufgeführte Liste mit Empfehlungen durchzulesen.
Ihr Computer wird Ihnen dauerhaft zur Verfügung stehen, wenn Sie dafür
sorgen, daß Ihre Arbeitsumgebung die folgenden Voraussetzungen erfüllt:


Be sure to clear at least 10 centimeters behind the computer for proper
ventilation.

Leave space for the monitor to rest on top of the main unit or on a shelf
directly above the main unit. Place your printer to one side of the main unit
so that the data and power cables do not get in the way of the paper. Place
your system on a flat, firm surface. Surfaces like beds or carpeted floors

C:\TW\BDU.EL3                  Line    43 of    69 Col    1     Pr     Tr In TW II
F1=Help  2=Save  3=Close  4=Save&Close  5-6=+Trans       8=Next  9=Undo  10=Menu
```

Figure 3: Fuzzy matching and terminology recognition in TRADOS Translator's Workbench II

What has changed dramatically indeed during the last two decades is the translation workflow, i.e., the way the translation processes are organized and the way the parties involved in these processes interact and collaborate. The introduction of client/server solutions after the turn of the millennium enabled new ways of real-time collaboration among distributed teams but led to even more controversial discussions about property rights on TM data collections and liability issues. The near future will reveal to which extent new buzzword technologies and forms of collaboration like "cloud computing" and "crowd sourcing" will actually affect translation workflows and work situations.

2.3 Types of TM systems

In most systems available on the market the TM is a database. Each record in a TM database contains a translation unit (TU) consisting of a pair of source and target text segments.[4] In addition to the TU there may be further information on the creation

[4] In most TMs translation units consist of source language sentences and their target language equivalents. Apart from 1:1 equivalences, where a sentence from the source text is transferred into one sentence in the target text, this can also include 1:n and n:1 relations, depending on the decisions taken by the individual translator. Moreover, smaller TUs having the size of clauses or phrases, larger units based on paragraphs, or nested units starting at paragraph level and then assigning further relations at sentence level may also occur.

and modification dates, the person who created or modified the entry, the project(s) or customer(s) the TU is used for, etc.

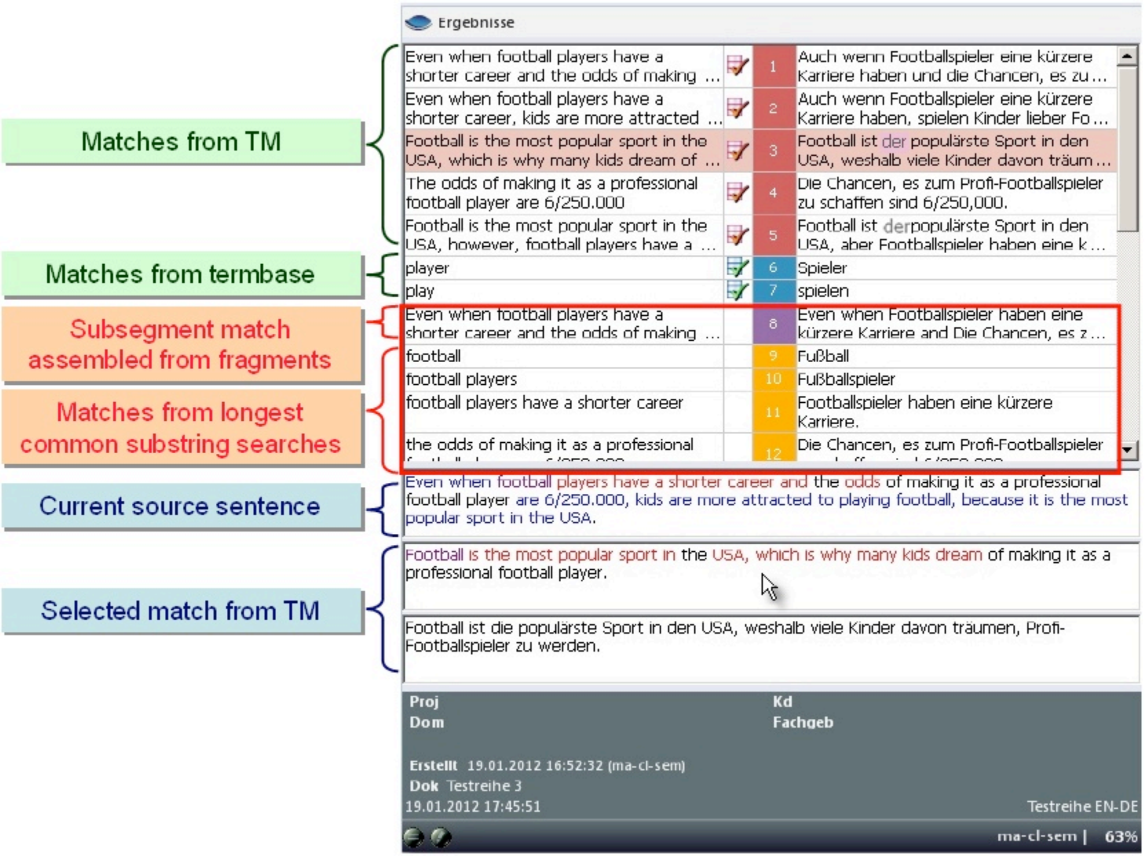
A major feature of a typical TM database is the fact that it grows incrementally. The database is 'dynamic' because new TUs – no matter whether they are created from scratch or by adjusting the translation of a similar TU retrieved from the TM - are added during the translation process.

Basically, there are three ways of feeding a TM:

- While translating: When translating a text using a TM database each segment from the source text will be automatically stored in the database along with its translation.
- By importing another TM database: This can either be a TM created with the same TM system or a TM available in the Translation Memory eXchange format (TMX), which is supported by all commercial systems.
- By aligning existing translations and their original texts: With the help of an alignment tool it is possible to create TM databases from the source and target text files of previous translation projects.

Some TM systems do not make use of the database approach but store entire source and target text pairs in their proprietary formats as reference material for future reuse in related translation projects. While TM databases constitute an amalgamation of translation units that isolates each segment from its context, the reference text approach makes it easier to take context into account during the matching process. On the other hand, this approach is rather static, i.e., it is not possible immediately to reuse TUs that have just been created. Therefore, systems based on the reference text approach also create a so called temporary "fuzzy index", which is a kind of temporary database providing access to recently created TUs as well as fuzzy-match functionality. In turn, TM systems following the database approach have tried to overcome the complete decontextualisation of their TUs by adding so-called "context matches" or "perfect matches", where an exact match is preceeded and/or followed by another exact match, i.e., the segment to be translated and the match retrieved from the TM have the same textual environment. This is achieved by simply storing in the TM database the relevant context segments together with the actual TUs and sometimes by additionaly taking into account information obtained from style sheets, document templates or structural document markup (cf. Chama 2010). Some database-oriented TM systems have also included the reference text approach as an additional option to retrieve translaslation units for reuse by allowing to specifiy bilingual files from previous translation projects and combining them with TM databases. In general, it seems that the developers of commercial TM systems more and more try to combine the advantages of both the database-oriented and the reference text-oriented approaches.

Another major issue in TM technology is the retrieval of fragments below sentence level. Most commercial TM systems now offer some kind of subsegment matching. The simplest form of subsegment matching is to look for complete TM database and termbase units that are part of the current souce language segment and automatically insert their target language sections, thus usually creating suggestions that form a mix of source and target language fragments and require further adaptation ("fragment assembly"). A more complex way of finding subsegments is to retrieve longest common substrings (LSCs) from TM database units (Figure 4). Finally, a third – and probably the most productive – way of subsegment matching that can be found in commercial TM systems is to automatically suggest target language fragments while typing a translation (auto-completion; Figure 5). These fragments are retrieved from bilingual lexicons that were statistically generated from TM databases (cf. Chama 2010).


Matches from TM
Matches from termbase
Subsegment match assembled from fragments
Matches from longest common substring searches
Current source sentence
Selected match from TM

Ergebnisse

Even when football players have a shorter career and the odds of making …		1	Auch wenn Footballspieler eine kürzere Karriere haben und die Chancen, es zu …
Even when football players have a shorter career, kids are more attracted …		2	Auch wenn Footballspieler eine kürzere Karriere haben, spielen Kinder lieber Fo …
Football is the most popular sport in the USA, which is why many kids dream of …		3	Football ist der populärste Sport in den USA, weshalb viele Kinder davon träum …
The odds of making it as a professional football player are 6/250.000		4	Die Chancen, es zum Profi-Footballspieler zu schaffen sind 6/250,000.
Football is the most popular sport in the USA, however, football players have a …		5	Football ist derpopulärste Sport in den USA, aber Footballspieler haben eine k …
player		6	Spieler
play		7	spielen
Even when football players have a shorter career and the odds of making …		8	Even when Footballspieler haben eine kürzere Karriere and Die Chancen, es z …
football		9	Fußball
football players		10	Fußballspieler
football players have a shorter career		11	Footballspieler haben eine kürzere Karriere.
the odds of making it as a professional		12	Die Chancen, es zum Profi-Footballspieler

Even when football players have a shorter career and the odds of making it as a professional football player are 6/250.000, kids are more attracted to playing football, because it is the most popular sport in the USA.

Football is the most popular sport in the USA, which is why many kids dream of making it as a professional football player.

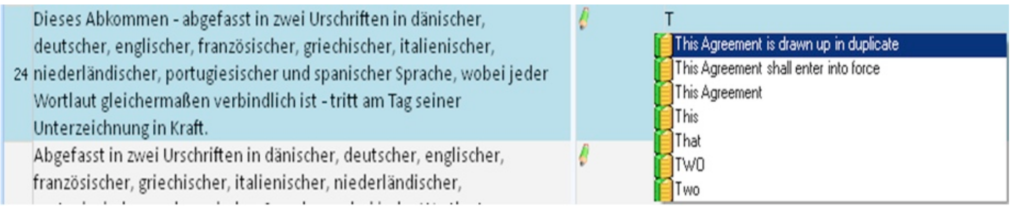
Football ist die populärste Sport in den USA, weshalb viele Kinder davon träumen, Profi-Footballspieler zu werden.

Proj Kd
Dom Fachgeb

Erstellt 19.01.2012 16:52:32 (ma-cl-sem)
Dok Testreihe 3
19.01.2012 17:45:51 Testreihe EN-DE

ma-cl-sem | 63%

Figure 4: Subsegment matching in Kilgray MemoQ…



24 Dieses Abkommen - abgefasst in zwei Urschriften in dänischer, deutscher, englischer, französischer, griechischer, italienischer, niederländischer, portugiesischer und spanischer Sprache, wobei jeder Wortlaut gleichermaßen verbindlich ist - tritt am Tag seiner Unterzeichnung in Kraft.

Abgefasst in zwei Urschriften in dänischer, deutscher, englischer, französischer, griechischer, italienischer, niederländischer,

T

- This Agreement is drawn up in duplicate
- This Agreement shall enter into force
- This Agreement
- This
- That
- TWO
- Two

Figure 5: … and SDL Trados Studio

2.4 Translation Memory and Machine Translation

2.4.1 Distinction between TM and MT

TM technology is not to be confused with machine translation. Whereas MT translates without human intervention, TM systems provide features and tools to store and retrieve segments translated by a human translator. Despite this essential distinction between TM and MT, TM technology shares certain commonalities with both "example-based machine translation" (EBMT), an approach first suggested in Japan in the early 1980s (Nagao 1984), and "statistical machine translation" (SMT), an approach developped at IBM in the late 1980s (cf. Brown et. al. 1988) that did not have its breakthrough before the turn of the millenium and today has to be considered the state-of-the art paradigm in MT (cf. Koehn 2010, 17f.). Both TM and EBMT/SMT try to retrieve "best matches" for the sentences of the text to be translated from a bilingual text archive or database containing sentence-level alignments of existing translations and their original texts.[5] Yet, there are fundamental differences between the purposes of EBMT/SMT and TM systems. A TM is mainly an information retrieval system that leaves decisions about whether and how to reuse and adjust the retrieved results – and thus the actual translation task – to the human translator. EBMT and SMT aim at producing translations by automatically selecting suitable fragments from the source language side of the retrieved TUs and building the translation from the corresponding elements of the target language side. Due to the complexity of this recombination task, not every TU contained in a translation archive is equally suited for reuse in TM systems and EBMT or SMT environments.

2.4.2 Integration of TM and MT

For good reason MT has so far been used very little in high quality e-content localization. MT is only suited for a very limited range of text types, and source texts have to be carefully tailored to the capabilities and restrictions of an MT system to minimize the amount of time and effort needed for post-editing.

Nevertheless, TM suites increasingly offer support for MT. Basically, there are two possible ways of combining MT and TM:

1. Batch processing (usually during project preparation):

In a batch scenario, all segments of the source text that do not produce an exact or high percentage "fuzzy match" when being compared with the TM database may be exported for processing by MT. After the unknown segments have been translated by the MT application, the new translation units can be merged into the TM database. When the translator works on the text, the units generated by the MT system will be presented as candidate translations, possibly with a pre-defined matching penalty.

2. Interactive processing (during the translation stage proper):

In an interactive scenario translators can invoke the MT system each time there is no match with the TM database. If the result from the MT system proves helpful, it can be edited as necessary. The resulting translation unit will then be stored in the TM database for future reuse.

Commercial TM systems like *Across* or *SDL Trados Studio* offer interfaces to both RBMT and SMT systems. Large IT companies like *Sybase* report productivity gains by combining SMT and TM, provided that the MT system has been trained with a large-enough company-specific bilingual corpus (cf. Bier 2012). Like other large companies *Sybase* has carried out experiments using the freely-available SMT system *Moses* (cf. Koehn et al. 2007) interactively together with a TM system. Bier (2012) mentions faster turnaround (delivery time decreased by an average of 50%), 20-30% cost reductions for updates, stable translation quality (no visible impact on style with full post-editing, less content errors, slight increase in minor linguistic errors) and a rise

[5] Both EBMT and SMT are corpus-based approaches, so that the term corpus-based MT (CBMT) is used as an umbrella for both as opposed to rule-based MT (RBMT) (Carl and Way 2003, xviii). The major difference between EBMT and SMT is that SMT considers translation as a "statistical optimization problem" (Koehn 2010, 17) and is based on probability calculations over large bilingual corpora, while EBMT tries to find analogies between an input sentence and examples from a bilingual corpus applying more "traditional" linguistic means like (morpho-)syntactic analysis and thesauri. For an extensive overview on EBMT see Carl and Way (2003) and Somers (2001). A comprehensive introduction to SMT can be found in Koehn (2010).

in productivity between 5 and 70% (depending on the kind of source texts, the terminology used and the performance of individual translators).[6]

2.5 Data exchange standards for TM systems

2.5.1 Overview

A versatile TM system must be able to handle the full range of proprietary and standard file formats in which e-content can be produced and exchanged. One of the major meta-standards that play a central role in technical documentation is the eXtensible Markup Language (XML) (W3C 2008). XML provides a framework to create markup languages for all kinds of individual document types, and there is a growing number of XML-based standards and formats to support various aspects of the documentation and localization process. While standards like DocBook (OASIS 2006), DITA (OASIS 2007), and XLIFF (OASIS 2008) are related to the creation and exchange of localizable content, TMX (LISA 2005), SRX (LISA 2008) and TBX (ISO 2008) serve the purpose of facilitating the exchange of reference material (TM databases and termbases).

Current efforts like Linport (Language Interoperability Portfolio, (Linport 2012)) and TIPP (Translation Interoperability Protocol Package, (Interoperability Now! 2012)) focus on the development of a standard for the exchange of complete translation projects between different translation environments.

2.5.2 Supporting standards for the exchange of localizable e-content

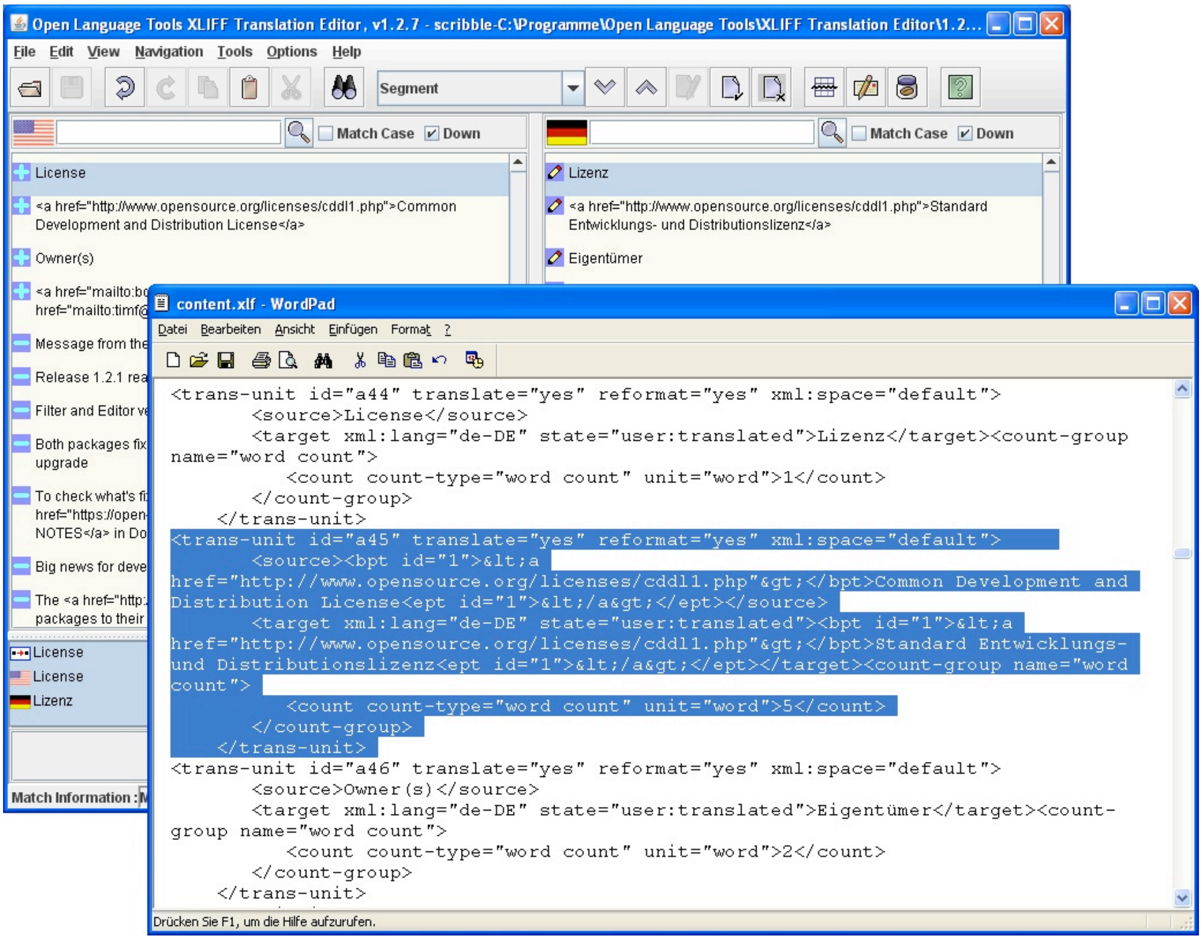
For public XML-based standards like DocBook, DITA und XLIFF TM systems should include import routines that provide an automatic distinction between so-called “external” XML markup elements that need not be modified during the translation process and “internal” elements, which the translator may need to move, add or delete. Translatable and non-translatable attribute values should be distinguished automatically as well.

For proprietary XML-based formats, TM systems should provide a feature to create import routines from a combination of various sources, i.e., XML document type definitions (DTDs), XML schema definition files (XSDs) and localizable XML content files, keeping the effort for manually correcting translation-related settings for the indiviudal XML elements and attributes as small as possible.

Content in formats like XLIFF, which mainly serve the purpose of exchanging bilingual files during the localization process, must be diplayed correctly in the TM system's multilingual editor, i.e., for editors using separate windows or table columns for source and target languages the <source> and <target> elements of an XLIFF file must be placed into the correct windows or columns (Figure 6). Moreover, metadata like translation comments and information on the processing status of translation units should be adequately imported, displayed and exported without any loss of information (Figure 7).

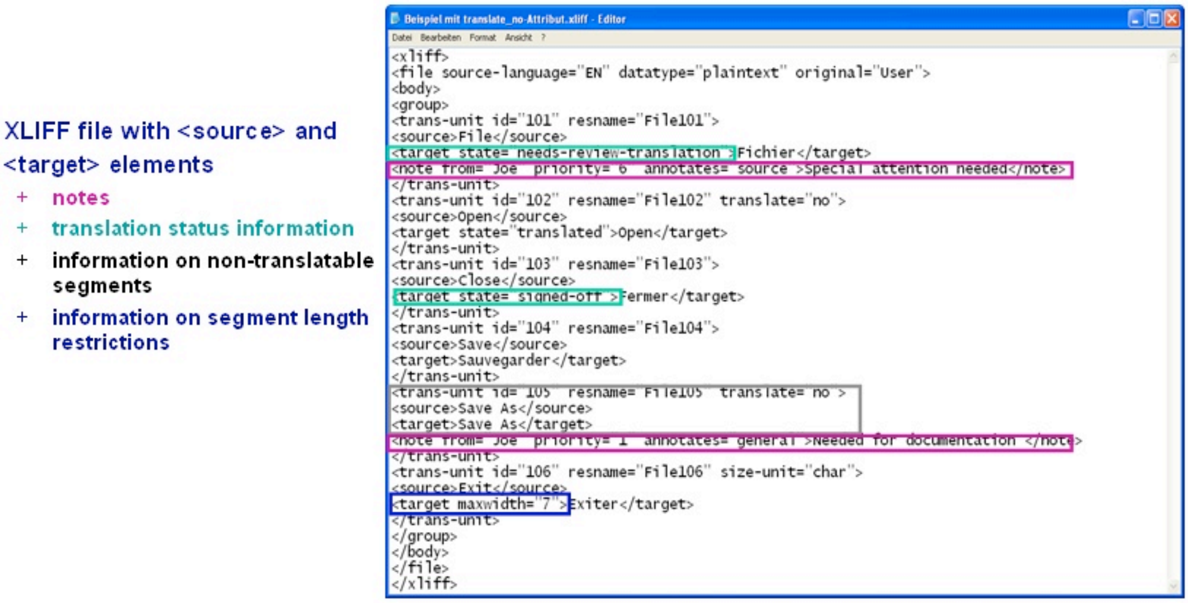
Finally, it must be taken into account that XLIFF is a kind of hybrid format, because apart from localizable content XLIFF files can also contain bilingual reference material from previous versions or related documents. TM systems must be able to recognize this reference material in an XLIFF file and store it in a TM database together with relevant metadata also contained in the XLIFF file, like information on match values, authors, systems used to create the material, etc. (Figure 8).

[6] For a comparison of TM and SMT output see also Offersgaard et al. (2008). Offersgaard et al. report high productivity gains of more than 65% for certain domains and for situations in which the TM database does not produce matches for two thirds or more of the sentences to be translated. Guerberof (2009) also reports higher processing speed for post-editing SMT output compared with TM matches, but also points to the fact that deviation between individual subjects is very high.


Open Language Tools XLIFF Translation Editor, v1.2.7 - scribble-C:\Programme\Open Language Tools\XLIFF Translation Editor\1.2...
File Edit View Navigation Tools Options Help
Segment
Match Case
Down
Match Case
Down
License
<a href="http://www.opensource.org/licenses/cddl1.php">Common Development and Distribution License</a>
Owner(s)
Lizenz
<a href="http://www.opensource.org/licenses/cddl1.php">Standard Entwicklungs- und Distributionslizenz</a>
Eigentümer
License
License
Lizenz
Match Information :
content.xlf - WordPad
Datei Bearbeiten Ansicht Einfügen Format ?
Drücken Sie F1, um die Hilfe aufzurufen.

```
<trans-unit id="a44" translate="yes" reformat="yes" xml:space="default">
        <source>License</source>
        <target xml:lang="de-DE" state="user:translated">Lizenz</target><count-group
name="word count">
            <count count-type="word count" unit="word">1</count>
        </count-group>
    </trans-unit>
<trans-unit id="a45" translate="yes" reformat="yes" xml:space="default">
        <source><bpt id="1">&lt;a
href="http://www.opensource.org/licenses/cddl1.php"&gt;</bpt>Common Development and
Distribution License<ept id="1">&lt;/a&gt;</ept></source>
        <target xml:lang="de-DE" state="user:translated"><bpt id="1">&lt;a
href="http://www.opensource.org/licenses/cddl1.php"&gt;</bpt>Standard Entwicklungs-
und Distributionslizenz<ept id="1">&lt;/a&gt;</ept></target><count-group name="word
count">
            <count count-type="word count" unit="word">5</count>
        </count-group>
    </trans-unit>
<trans-unit id="a46" translate="yes" reformat="yes" xml:space="default">
        <source>Owner(s)</source>
        <target xml:lang="de-DE" state="user:translated">Eigentümer</target><count-
group name="word count">
            <count count-type="word count" unit="word">2</count>
        </count-group>
    </trans-unit>
```

Figure 6: Fragment from an XLIFF file in a text editor and in an XLIFF translation editor

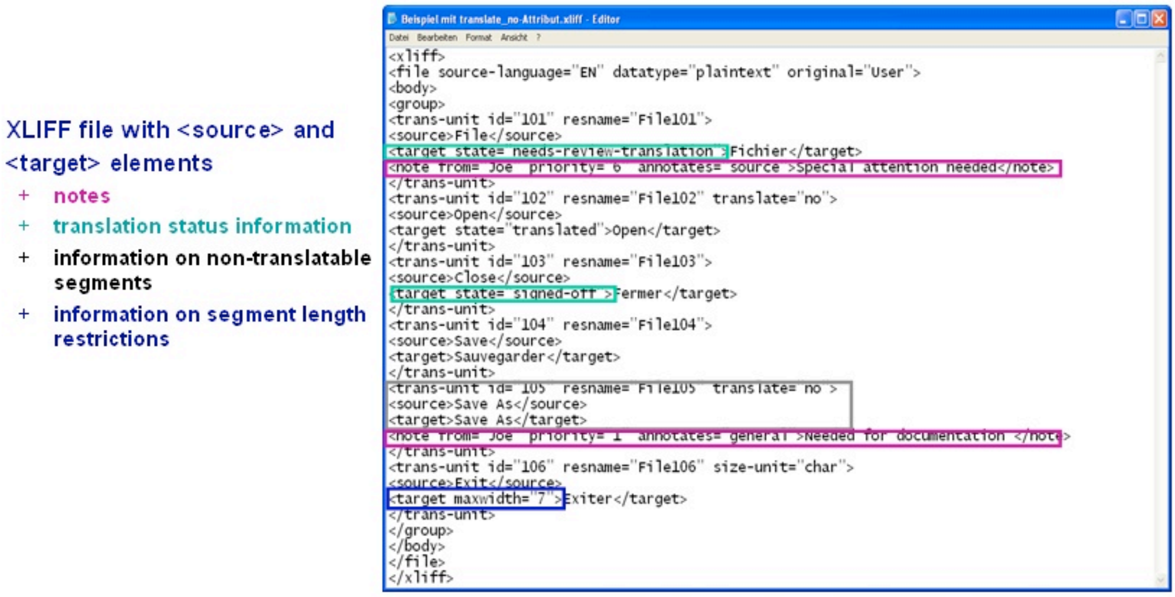

Beispiel mit translate_no-Attribut.xliff - Editor
Datei Bearbeiten Format Ansicht ?

XLIFF file with <source> and <target> elements

- notes
- translation status information
- information on non-translatable segments
- information on segment length restrictions

```
<xliff>
<file source-language="EN" datatype="plaintext" original="User">
<body>
<group>
<trans-unit id="101" resname="File101">
<source>File</source>
<target state="needs-review-translation">Fichier</target>
<note from="Joe" priority="6" annotates="source">Special attention needed</note>
</trans-unit>
<trans-unit id="102" resname="File102" translate="no">
<source>Open</source>
<target state="translated">Open</target>
</trans-unit>
<trans-unit id="103" resname="File103">
<source>Close</source>
<target state="signed-off">Fermer</target>
</trans-unit>
<trans-unit id="104" resname="File104">
<source>Save</source>
<target>Sauvegarder</target>
</trans-unit>
<trans-unit id="105" resname="File105" translate="no">
<source>Save As</source>
<target>Save As</target>
<note from="Joe" priority="1" annotates="general">Needed for documentation </note>
</trans-unit>
<trans-unit id="106" resname="File106" size-unit="char">
<source>Exit</source>
<target maxwidth="7">Exiter</target>
</trans-unit>
</group>
</body>
</file>
</xliff>
```

Figure 7: Complex XLIFF file containing various metadata


XLIFF file with <source> and <target> elements
+ notes
+ translation status information
+ information on non-translatable segments
+ information on segment length restrictions
Beispiel mit translate_no-Attribut.xliff - Editor
Datei Bearbeiten Format Ansicht ?

```
<xliff>
<file source-language="EN" datatype="plaintext" original="User">
<body>
<group>
<trans-unit id="101" resname="File101">
<source>File</source>
<target state="needs-review-translation">Fichier</target>
<note from="Joe" priority="6" annotates="source">Special attention needed</note>
</trans-unit>
<trans-unit id="102" resname="File102" translate="no">
<source>Open</source>
<target state="translated">Open</target>
</trans-unit>
<trans-unit id="103" resname="File103">
<source>Close</source>
<target state="signed-off">Fermer</target>
</trans-unit>
<trans-unit id="104" resname="File104">
<source>Save</source>
<target>Sauvegarder</target>
</trans-unit>
<trans-unit id="105" resname="File105" translate="no">
<source>Save As</source>
<target>Save As</target>
<note from="Joe" priority="1" annotates="general">Needed for documentation </note>
</trans-unit>
<trans-unit id="106" resname="File106" size-unit="char">
<source>Exit</source>
<target maxwidth="7">Exiter</target>
</trans-unit>
</group>
</body>
</file>
</xliff>
```

Figure 8: XLIFF file containing various reference material

2.5.3 Supporting standards for the exchange of reference material

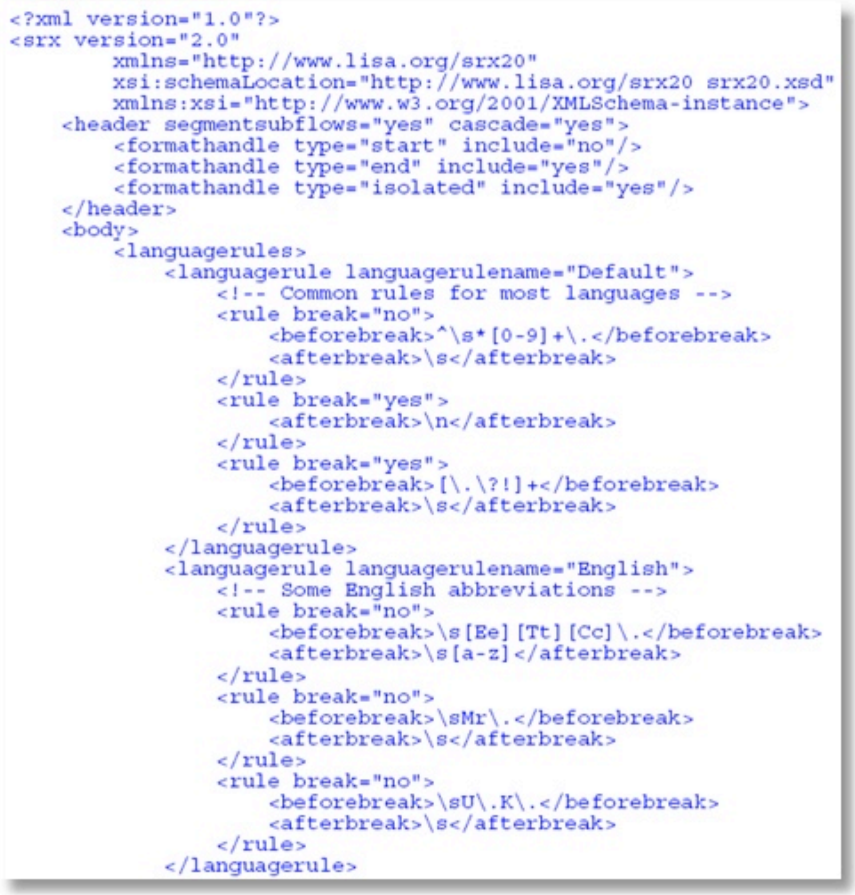
The exchange of TM databases mainly causes problems with respect to the maintenance of layout information and dynamic fields, i.e., placeholders for embedded objects and automatically adjustable content like cross-references and other variables, contained in TUs and the exchange of information on rules used for the segmentation of text into TUs.

To keep the loss of layout-related information and placeholders for embedded objects and dynamic fields contained in TUs as small as possible when exchanging TMs between different applications most TM systems support TMX Level 2. The TMX standard has been available since 1998. It has been developed by the Localization Industry Standards Association (LISA), which was an interest group of major information technology companies and localization service providers. After LISA became insolvent in 2011 TMX is now being maintained by the Localization Industry Standards (LIS) Industry Specification Group (ISG) of the European Telecommunications Standards Institute (ETSI) (cf. GALA 2012) and the standard is freely available from the website of the Globalization and Localization Association (GALA).[7]

Breaking up text into smaller TUs requires segmentation rules that may differ between languages as well as text types and file formats. Examples include individual punctuation characters like the quotation mark in Spanish or the different treatment of colons, semi-colons and other characters depending on language and text type. In order to overcome a loss in reusability of TUs due to different segmentation rules applied in different TMs the Segmentation Rules eXchange (SRX) standard was introduced in 2004. The segmentation rules contained in an SRX file (Figure 9) must be applied when exporting and importing TMs as well as during the actual translation process when the current source text has to be split up into TUs.

Like TMX SRX was developed by LISA and is now being maintained by ETSI. It can also be downloaded from the GALA website.

[7] GALA is a non-profit organization of localization and translation service providers, language technology developers and other companies involved in language services or technology. The former LISA standards can be found at http://www.gala-global.org/lisa-oscar-standards.



```
<?xml version="1.0"?>
<srx version="2.0"
        xmlns="http://www.lisa.org/srx20"
        xsi:schemaLocation="http://www.lisa.org/srx20 srx20.xsd"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <header segmentsubflows="yes" cascade="yes">
        <formathandle type="start" include="no"/>
        <formathandle type="end" include="yes"/>
        <formathandle type="isolated" include="yes"/>
    </header>
    <body>
        <languagerules>
            <languagerule languagerulename="Default">
                <!-- Common rules for most languages -->
                <rule break="no">
                    <beforebreak>^\s*[0-9]+\.</beforebreak>
                    <afterbreak>\s</afterbreak>
                </rule>
                <rule break="yes">
                    <afterbreak>\n</afterbreak>
                </rule>
                <rule break="yes">
                    <beforebreak>[\.\?!]+</beforebreak>
                    <afterbreak>\s</afterbreak>
                </rule>
            </languagerule>
            <languagerule languagerulename="English">
                <!-- Some English abbreviations -->
                <rule break="no">
                    <beforebreak>\s[Ee][Tt][Cc]\.</beforebreak>
                    <afterbreak>\s[a-z]</afterbreak>
                </rule>
                <rule break="no">
                    <beforebreak>\sMr\.</beforebreak>
                    <afterbreak>\s</afterbreak>
                </rule>
                <rule break="no">
                    <beforebreak>\sU\.K\.</beforebreak>
                    <afterbreak>\s</afterbreak>
                </rule>
            </languagerule>
```

Figure 9: Section from an SRX file

Exchanging data between the terminology management components of various TM systems can be much more difficult than sharing TMs among various applications. This is due to the fact that the structure and complexity of termbases may differ severely from system to system and – in the case of user-definable entry structures – even among termbases created with the same application. It has taken a long time since the efforts to define a universal exchange format for terminological data have lead to the Termbase eXchange Standard (TBX). Although TBX has become an ISO standard in 2008 (cf. ISO 2008) the format is still not properly supported by all TM systems.

2.6 Advantages and limitations of TM systems

The advantages of using TM systems are fairly obvious: they increase the translator's productivity and enhance translation quality by ensuring that terminology and expressions are used consistently within and across translations. Users in industry and international organizations usually claim a 25 to 60 per cent rise in productivity (cf. Reinke 2004, 113f.). However, at least in some industries productivity gains seem to come to an end after a certain time. Thus, at Sybase "[t]raditional TM technology [is] almost fully exploited" with "ca. 80% of costs spent on 'new' words" and "only 20% spent on recycling" (Bier 2012). Bier also states that there are "[n]o more improvements in turnaround times" as the average productivity of translators has remained at a maximum level of 2.400 words per day for years.

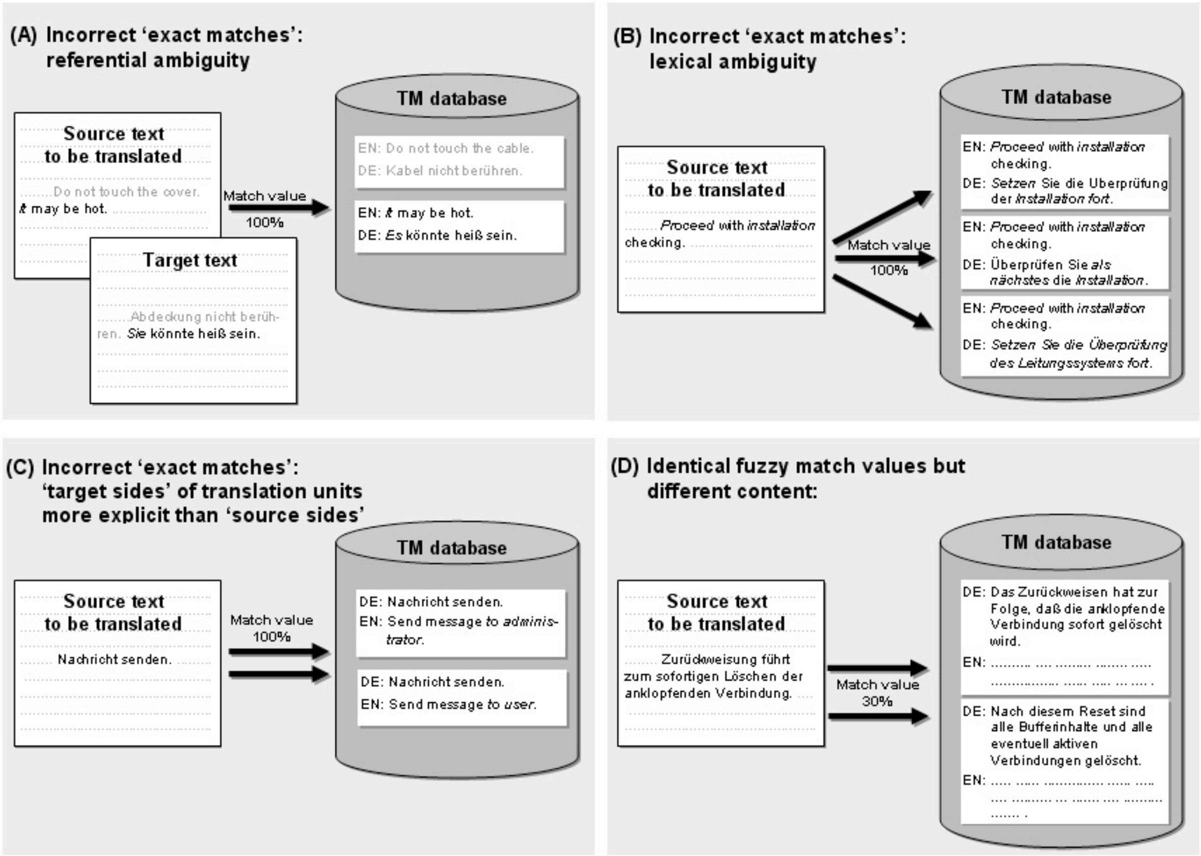

(A) Incorrect 'exact matches': referential ambiguity
Source text to be translated
Do not touch the cover. It may be hot.
Match value 100%
TM database
EN: Do not touch the cable.
DE: Kabel nicht berühren.
EN: It may be hot.
DE: Es könnte heiß sein.
Target text
Abdeckung nicht berühren. Sie könnte heiß sein.
(B) Incorrect 'exact matches': lexical ambiguity
Source text to be translated
Proceed with installation checking.
Match value 100%
TM database
EN: Proceed with installation checking.
DE: Setzen Sie die Uberprüfung der Installation fort.
EN: Proceed with installation checking.
DE: Überprüfen Sie als nächstes die Installation.
EN: Proceed with installation checking.
DE: Setzen Sie die Überprüfung des Leitungssystems fort.
(C) Incorrect 'exact matches': 'target sides' of translation units more explicit than 'source sides'
Source text to be translated
Nachricht senden.
Match value 100%
TM database
DE: Nachricht senden.
EN: Send message to administrator.
DE: Nachricht senden.
EN: Send message to user.
(D) Identical fuzzy match values but different content:
Source text to be translated
Zurückweisung führt zum sofortigen Löschen der anklopfenden Verbindung.
Match value 30%
TM database
DE: Das Zurückweisen hat zur Folge, daß die anklopfende Verbindung sofort gelöscht wird.
EN: ....
DE: Nach diesem Reset sind alle Bufferinhalte und alle eventuell aktiven Verbindungen gelöscht.
EN: ....

Figure 10: Examples in English (EN) and German (DE), demonstrating shortcomings of fuzzy match algorithms (Reinke 2006, 64)

Furthermore, it must be stated that the use of TM systems may also have negative effects on translation quality. One of the major disadvantages of TM systems is that they usually operate at sentence level. Thus, there is a serious danger that the translator will focus too much on isolated sentences, possibly disregarding the contexts they are embedded in (cf. Reinke 2004, 136ff.).

Examples (A) and (B) in Figure 10 examplify this problem with respect to referential and lexical ambiguity. In example (A) the pronoun *it* is an anaphoric reference to the noun phrase *the cover* in the previous sentence. As the German translation *die Abdeckung* is female, the pronoun should be female as well (i.e., *sie*). In the same English sentence in the TM the pronoun *it* refers to a different noun phrase with a German translation using a neuter noun like *das Kabel,* so that *it* has to become *es*. Thus, an exact match for *It can be hot* yields a translation that does not fit the current context. In example (B) terms like *installation* or general language words like *prcoeed* are lexically ambiguous. *Installation* could, for instance, refer the installation of a piece of software or to a piping system, while *to proceed with s.th.* might mean *to continue a process that has been interrupted* or *to go on with the next step of a process*. These different meanings require different translations. Therfore, an exact match from the TM might produce an incorrect translation.

The matching algorithms of TM systems are based on very simple formal criteria like the similarity of character strings. Thus, the human translator's notion of the degree of similarity between a segment to be translated and a segment retrieved from the database may differ considerably from the degree of similarity calculated by the TM system. This may lead to situations where "exact matches" yield wrong translations (examples (A) to (C) in Figure 10) or one translation of a "fuzzy match" requires little or no adjustment, while another "fuzzy match" with the same similarity value is not useful at all, e.g., because the content belongs to a different (sub-)domain (example (D) in Figure 10).

Despite these drawbacks, it should be noted that TM systems generally integrate into the translation workflow comparatively smoothly. As opposed to MT, they leave human translators in control of the actual translation process, while relieving them from routine work and maintaining translation as a creative act whenever the linguistic resourcefulness of a human being is required.

3 Approaches to enhance the information retrieval performance of TM systems

3.1 Approaches not applying "linguistic knowledge"

Although commercial TM systems have been available for over two decades, their retrieval performance has not improved considerably in terms of quality and quantity. Of course, the matching algorithms have been altered and modified over the time, but they still rely on simple character- or token-based matching procedures without taking into account linguistic aspects like morphosyntactic, syntactic or semantic features that may determine the "similarity" of translation units.[8] Even rather straightforward approaches that do not rely on "linguistic knowledge" but could, for instance, easily improve the retrieval performance for TUs containing so-called placeable and localizable elements[9] are not yet a matter of course in commercial TM systems.

Azzano (2011) presents a detailed analysis of the question in how far the occurrence of placeable and localizable elements influence the retrieval performance of commercial TM systems. He found that placeable elements sometimes lead to comparatively low fuzzy match values because some systems treat them like standard text when comparing the lengths of source language segments (SegSL) to be translated and source language segments from a TM (SegSL_{TM}). Instead, it would be more reasonable to use a fixed penalty when SegSL and SegSL_{TM} only differ with respect to the placeable elements they contain while the remaining standard text is identical.

Azzano (2011) also reports that some systems yield exact matches when SegSL and SegSL_{TM} contain both identical text and identical placeable elements and just differ in the order or position of the placeable elements. This is a serious mistake because in most cases these modifications will also be relevant to the new translation if the target language segment from the TM (SegTL_{TM}) will be reused.

Comparatively simple methods could also be applied to improve the retrieval of TM segments containing localizable elements. Instead of treating them like plain text they should be seen as special elements that follow certain patterns. These patterns can be recognized with the help of regular expressions. For the calculation of match values the same principles already suggested for placeable elements could be applied (i.e., using a fixed penalty if SegSL and SegSL_{TM} differ in terms of localizable elements). Azzano (2011) found that to a certain extent commercial TM systems do apply regular expressions to identify localizable elements, but for some elements like

[8] For a brief overview on similarity measures relevant to TM systems see Trujillo (1999, 61-68), Reinke (2004, 193-198), Sikes 2007.

[9] Placeable elements like tags, inline graphics and dynamic fields usually do not contain translatable text. They can often be copied ("placed") into the target text without any need for further modifications. Tags are markup elements in HTML and XML files; inline graphics and dynamic fields typically occur in DTP formats and Microsoft Word files. Localizable elements like numbers, dates, URLs or e-mail addresses, in turn, consist of plain text following a certain pattern, so that they can be identified without any "linguistic knowledge". The localization of these elements follows given rules and often does not influence the remaining parts of a TU.

complex numerical patterns they still show severe weaknesses, while other elements are not recognized at all. Although there are useful and well-known regular expressions, e.g. for identifying URLs in plain text (cf. Goyvaerts and Levithan 2009), these are hardly implemented in commercial TM systems. Azzano (2011) suggests a number of regular expressions to improve the recognition of various localizable elements.

3.2 Approaches applying "linguistic knowledge"

3.2.1 Current approaches in commercial and research systems

Linguistics-driven efforts on enhancing retrieval in TM systems are basically motivated by two different goals:

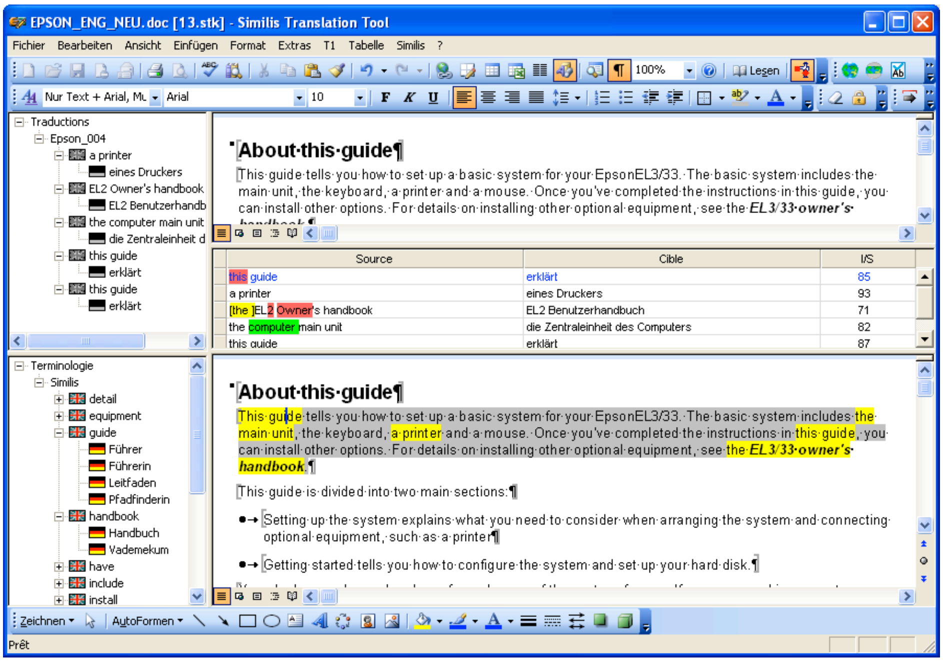
(1) improving recall and precision of (monolingual) retrieval, i.e. enhancing quantity, quality and ranking of matches, at segment level and at subsegment level (retrieval of "chunks", (complex) phrases, clauses) by enriching the retrieval algorithms of TM systems with "linguistic knowledge"

(2) automized adjustment of fuzzy matches to enhance reusability and reduce post-editing efforts by integrating SMT technology into TM systems.

With *Similis* the French company *Lingua et Machina* produces one of the very few commercial TM systems that do not only rely on character-based matching algorithms but try to integrate linguistic methods by using morphosyntactic analysis and shallow parsing to identify fragments below segment level (cf. Planas, 2005). Planas (2005) describes his system as "second generation translation memory software". Of course, this kind of linguistically enhanced application is only available for a restricted number of language pairs.[10] Investigations indicate that at least for certain language combinations like English-German the system only identifies rather short phrases like simple NPs but cannot retrieve larger syntactical units, which would be desirable for the support of professional computer-assisted human translation (cf. Kriele 2006; Macken 2009). Figures 11 and 12 illustrate these findings for an English-German example shown the *Similis* translation and alignment editors.

Linguistically enhanced TM systems have mainly been developed and tested as research systems (cf. Gotti, Fabrizio et al., 2005; Hodász and Pohl, 2005; Mitkov and Corpas 2008). Like *Similis* they mostly integrate morphosyntactic analysis and shallow syntactic parsing. However, there are even efforts to include semantic information to improve the retrieval of sentence-level praraphrases that differ lexically and syntactically (cf. Mitkov and Corpas 2008). Due to the rather restricted availability of semantic data in relevant subject areas, the relevance of these approaches for commercial implementations is still rather small.

[10] Currently *Similis* supports combinations between English, German, French, Italian, Spanish, Portugese and Dutch (http://similis.org/linguaetmachina.www/index.php?afficher=10&sel=40&info=Spezifikationen).



About·this·guide¶

This·guide·tells·you·how·to·set·up·a·basic·system·for·your·EpsonEL3/33.·The·basic·system·includes·the·main·unit,·the·keyboard,·a·printer·and·a·mouse.·Once·you've·completed·the·instructions·in·this·guide,·you·can·install·other·options.·For·details·on·installing·other·optional·equipment,·see·the·***EL3/33·owner's·handbook***.¶

Source	Cible	I/S
this guide	erklärt	85
a printer	eines Druckers	93
[the]EL2 Owner's handbook	EL2 Benutzerhandbuch	71
the computer main unit	die Zentraleinheit des Computers	82
this guide	erklärt	87

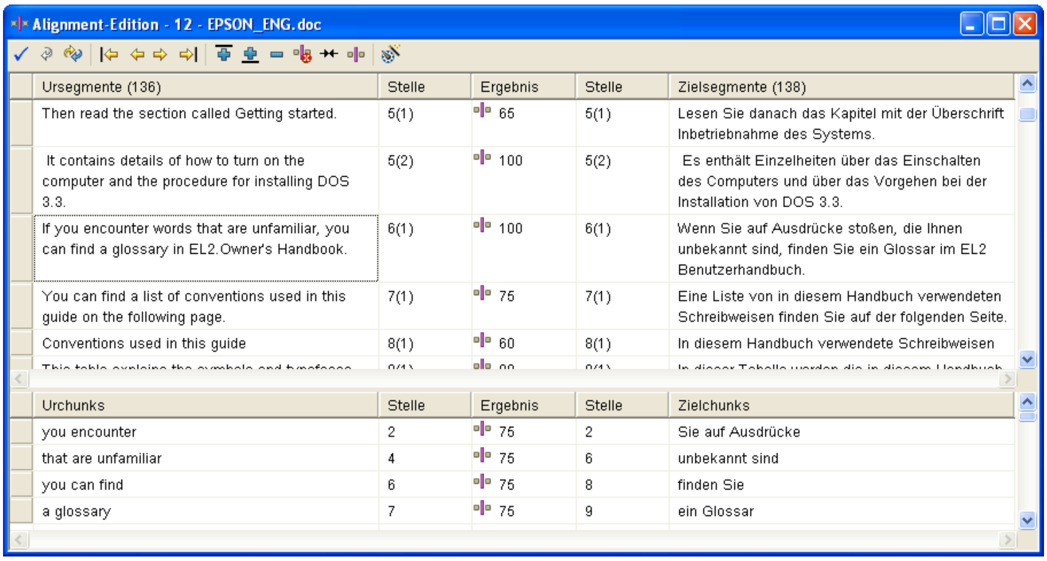
About·this·guide¶

This·guide·tells·you·how·to·set·up·a·basic·system·for·your·EpsonEL3/33.·The·basic·system·includes·the·main·unit,·the·keyboard,·a·printer·and·a·mouse.·Once·you've·completed·the·instructions·in·this·guide,·you·can·install·other·options.·For·details·on·installing·other·optional·equipment,·see·the·***EL3/33·owner's·handbook***.¶

This·guide·is·divided·into·two·main·sections:¶

- Setting·up·the·system·explains·what·you·need·to·consider·when·arranging·the·system·and·connecting·optional·equipment,·such·as·a·printer¶
- Getting·started·tells·you·how·to·configure·the·system·and·set·up·your·hard·disk.¶

Figure 11: English-German example for subsegment retrieval in Similis



Ursegmente (136)	Stelle	Ergebnis	Stelle	Zielsegmente (138)
Then read the section called Getting started.	5(1)	65	5(1)	Lesen Sie danach das Kapitel mit der Überschrift Inbetriebnahme des Systems.
It contains details of how to turn on the computer and the procedure for installing DOS 3.3.	5(2)	100	5(2)	Es enthält Einzelheiten über das Einschalten des Computers und über das Vorgehen bei der Installation von DOS 3.3.
If you encounter words that are unfamiliar, you can find a glossary in EL2.Owner's Handbook.	6(1)	100	6(1)	Wenn Sie auf Ausdrücke stoßen, die Ihnen unbekannt sind, finden Sie ein Glossar im EL2 Benutzerhandbuch.
You can find a list of conventions used in this guide on the following page.	7(1)	75	7(1)	Eine Liste von in diesem Handbuch verwendeten Schreibweisen finden Sie auf der folgenden Seite.
Conventions used in this guide	8(1)	60	8(1)	In diesem Handbuch verwendete Schreibweisen
This table explains the symbols and typefaces	9(1)	90	9(1)	In dieser Tabelle werden die in diesem Handbuch

Urchunks	Stelle	Ergebnis	Stelle	Zielchunks
you encounter	2	75	2	Sie auf Ausdrücke
that are unfamiliar	4	75	6	unbekannt sind
you can find	6	75	8	finden Sie
a glossary	7	75	9	ein Glossar

Figure 12: Subsegment alignment in Similis

More recent research on enhancing retrieval in TM systems mainly seems to focus on improving the reusability of fuzzy matches by applying methods from SMT (cf. Biçici and Dymetman 2008; Zhechev and van Genabith 2010; Koehn and Senellart 2010). The aim is to identify those fragments that make the difference between a segment to be translated and a fuzzy match retrieved from a TM database and adjust their translations automatically using SMT procedures. Ideally, for the human translator there would be no additional post-editing effort for these matches. However, one should have a careful "empirical look" at the question how this "fusion" of human translation and machine translation at segment level actually affects the post-editing of fuzzy matches and in how far it really enhances the productivity of human translators as well as text quality.

3.2.2 Integrating robust linguistic procedures into existing commercial systems

Ways of integrating standard methods and procedures known from computational linguistics into commercial TM systems are currently analyzed at Cologne University of Applied Sciences in a research project supported by the German Federal Ministry of Education and Research (BMBF) (cf. Azzano, Reinke and Sauer 2011). The focus of the project lies on enhancing the performance of commercial TM systems with respect to the retrieval of paraphrase patterns and subsegment fragments as well as on improving term recognition and validation with the help of robust procedures for morphosyntactic and sentence syntactic analysis. The goal is to develop interface models and prototypical interfaces between commercial TM systems and "lingware" using SDL Trados Studio 2009 and the morphosyntactic analysis tool MPRO (Maas, Rösener and Theofilidis 2009) as a prototypical environment and German and English as prototypical languages to gain experiences for the development of further language modules and for applying the results to other TM systems.

In the first phase relevant similarity patterns were identified and classified using authentic multilingual technical documentation (user manuals and operating instructions from various areas). For this purpose, TM databases were created and compared with "related" texts (updates, texts on closely related items of communication, texts belonging to related text types and dealing with the same item of communication). Currently the master TM database contains 51.000 segments. Both the segments from the TM databases and the texts "related" to the TM material were morphosyntactically annotated with MPRO. To identify relevant similarity patterns the "related" texts were automatically matched with the TM databases using the pre-translate function. In many cases the resulting match values and the similarity judgments of human translators differed considerably. In a further step, the linguistic differences between the segments of the new, "related" texts and the matches from the TM were described and categorized in order to identify linguistic features that may help to enhance the retrieval performance of commercial TM systems.

To integrate morphosyntactical information into the commercial TM a stand-alone SQL database was developed. This "linguistic TM" is built from the morphosyntactically annotated segments of the commercial TM and – apart from the tokens of the text surface – mainly contains information obtained from lemmatization, compound analysis and word class recognition. The segments of the "linguistic TM" are linked to the "originals" in the commercial TM via unique IDs. To accelerate the retrieval of relevant TUs from the SQL database the data is stored in the form of suffix arrays (cf. Aluru 2004).

When looking up TUs in the "linguistic TM" during the translation process each SLSeg first need to be morphosyntactically analyzed and annotated. The actual retrieval process then consists of two steps. First, the tokens found in the SLSeg to be translated are compared with the tokens in the $\text{SLSeg}_{\text{TMling}}$ to determine whether one or more $\text{SLSeg}_{\text{TMling}}$ completely or partially contain SLSeg. A second query searches the "linguistic TM" for all $\text{SLSeg}_{\text{TMling}}$ with morphosyntactic patterns similar to those of the SL_{Seg} to be translated. For all results of both queries the Longest Common Substrings (LCS) between SL_{Seg} and $\text{SLSeg}_{\text{TMling}}$ are calculated using Generalized Suffix Arrays (GSA) (cf. Rieck, Laskov and Sonnenburg 2007). In order to rank the results a formula will be developed that combines the matches obtained from the two queries taking into consideration the number and the length of LCS as well as their position in SL_{Seg} and $\text{SLSeg}_{\text{TMling}}$ (cf. Hawkins and Giraud-Carrier 2009).

4 Conclusions and outlook

This paper has tried to give an overview of the state of the art in TM technology, explaining the major concepts and looking at recent trends in both commercial systems and research. As TM and MT "have been developed very much in isolation" because "different communities played a role in each technology's development" (Koehn and Senellart 2010) and computational linguistics has long ignored the relevance of TM as the major language technology used in professional translation, there is still ample scope for further research as well as for closer collaboration between academia and the language industry.

An important field that could not be touched upon in this paper for reasons of space and time is empirical research on how TM and MT and the combination of both actually integrate into the translation workflow and how they influence the work of the translator. Christensen and Schjoldager (2010, 99) identify three different areas of empirical TM research, namely "technology-oriented", "workflow-oriented" and "translation-theoretical", and conclude that

> Empirically documented knowledge about the nature and applications of TM systems and translators' interaction with them is both scarce and fragmented. In particular, more research is needed on how translators interact with TM technology and on how it influences translators' cognitive processes. The translation profession itself will also welcome more knowledge about the translators' perspective on TM technology. (Christensen and Schjoldager 2010, 99)

Research into these areas has only begun and it is to be hoped that in the near future more funding will be made available in this direction, because language technology for a multilingual society must, like any technology, serve the needs of its users.

5 References

ALPAC Automatic Language Processing Advisory Committee (eds). 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, D.C.: National Research Council (Publication 1416).

Aluru, Srinivas. 2004. Suffix Trees and Suffix Arrays. In *Handbook of Data Structures and Applications*, ed. by Mehta, Dinesh P. and Sartaj Sahni, 29-1–29-21. Boca Rayton: Chapman & Hall/CRC.

Arthern, Peter J. 1979. Machine Translation and computerized terminology systems: A translator's viewpoint. In *Translating and the computer, proceedings of a seminar. London, 14th November, 1978*, ed. Snell, Barbara M., 77-108. Amsterdam: North-Holland.

Azzano, Dino. 2011. Placeable and localizable elements in translation memory systems. Dissertation. Ludwig-Maximilians-Universität München.

Azzano, Dino, Uwe Reinke, and Melanie Sauer. 2011. Ansätze zur Verbesserung der Retrieval-Leistung kommerzieller Translation-Memory-Systeme. In *Multilingual Resources and Multilingual Applications. Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology (GSCL) 2011*, ed. by Hedeland, Hanna, Thomas Schmidt, and Kai Wörner, 123-128. Hamburg: Universität Hamburg, Sonderforschungsbereich Mehrsprachigkeit.

Biçici, Ergun and Marc Dymetman. 2008. Dynamic Translation Memory: Using Statistical Machine Translation to improve Translation Memory Fuzzy Matches. In *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science 4919*. ed. Gelbukh, Alexander F., 454-465. Berlin, Heidelberg: Springer.

Bier, Kerstin. 2012. *An MT journey: MT in use at Sybase, a SAP company*. TAUS open source machine translation showcase. Paris, June 4, 2012. http://www.slideshare.net/TAUS/4-june-2012-taus-moses-open-source-mt-showcase-paris-kerstin-bier-sybase

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jellinek, John D. Lafferty, Robert L. Mercer, Paul S. Roossin. (1988): A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88), Budapest, August 1988*, 71-76.

Carl, Michael and Andy Way (ed.). (2003). *Recent Advances in Example-Based Machine Translation.* Dodrecht, Boston, London: Kluwer.

CERTT. 2012. Glossary of translation tool types. In *Collection of Electronic Resources in Translation Technologies*. University of Ottawa, School of Translation and Interpretation. http://aix1.uottawa.ca/~certt/Glossary of translation tool types_F_FINAL.pdf

Chama, Ziad. 2010. Vom Segment zum Kontext. *technische kommunikation*, 32(2):21-25.

eCoLoRe. 2012. Glossary of Terms Related to eContent Localisation. University of Leeds, Centre for Translation Studies. http://ecolore.leeds.ac.uk/xml/materials/overview/glossary.xml?lang=en

GALA. 2012. LISA OSCAR Standards. http://www.gala-global.org/lisa-oscar-standards

Gotti, Fabrizio, Philippe Langlais, Elliott Macklovitch, Didier Bourigault, Benoit Robichaud and Claude Coulombe. 2005. 3GTM: A Third-Generation Translation Memory. In *Proceedings of the 3rd Computational Linguistics in the North-East (CLiNE) Workshop, Gatineau, Québec, August 2005*. http://www.iro.umontreal.ca/~felipe/Papers/paper-cline-3gtm-2005.pdf

Goyvaerts, Jan and Stephen Levithan. 2009. *Regular expressions cookbook*. Sebastopol, O'Reilly.

Guerberof, Ana. 2009. Productivity and quality in MT post-editing. In *MT Summit XII – Workshop: Beyond Translation Memories: New Tools for Translators MT*, August 29, 2009, Ottawa, Ontario, Canada.

Hawkins, Brian E. und Giraud-Carrier, Christophe G. 2009. Ranking search results for translated content. In *IRI '09 - Proceedings of the 10th IEEE international conference on Information Reuse & Integration*, ed. Zhang, Kang and Reda Alhajj, 242-245. Piscataway NJ: IEEE Press.

Hodász, Gábor and Gábor Pohl. 2005. MetaMorpho TM: a linguistically enriched translation memory. In *International Workshop: Modern Approaches in Translation Technologies*, 26-30. Borovets, Bulgaria. http://www.mt-archive.info/RANLP-2005-Hodasz.pdf

Hutchins, John W. 1998. "The origins of the translator's workstation". *Machine Translation* 13:287-307.

Interoperability Now! (2012). The TMS Interoperability Protocol Package (TIPP). Version 1.4.1. http://code.google.com/p/interoperability-now/downloads/detail?name=The_TMS_Interoperability_Protocol_Package-1.4.1.pdf&can=2&q=

ISO. 2008. ISO 30042:2008: Systems to manage terminology, knowledge and content – TermBase eXchange (TBX). Genf: International Organization for Standardization.

Koehn, Phillip. (2010). *Statistical Machine Translation*. Cambridge et. al: Cambridge University Press.

Koehn, Phillip and Jean Senellart. 2010. Convergence of Translation Memory and Statistical Machine Translation. In *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"* [Workshop at AMTA 2010: the Ninth conference of the Association for Machine Translation in the Americas, Denver, CO, 4 November 2010], ed. Zhechev, Ventsislav, 21-31. http://www.mt-archive.info/JEC-2010-Koehn.pdf

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. (2007). Moses: Open Source Toolkit for Statistical Machine Translation, in *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007*. http://homepages.inf.ed.ac.uk/pkoehn/publications/acl2007-moses.pdf

Kriele, Christian. 2006. *Vergleich der beiden Translation-Memory-Systeme TRADOS und SIMILIS*. Diploma thesis. Saarbrücken: Saarland University [unpublished].

Krollmann, Friedrich. 1971. "Linguistic data banks and the technical translator". *Meta* 16(1-2):117-124.

Lagoudaki, Elina. 2006. Translation Memory systems: Enlightening users' perspective. London: Imperial College.

Linport. (2012). Linport: The Language Interoperability Portfolio Project. http://www.linport.org/

LISA. 2005. Translation Memory eXchange (TMX), version 1.4b. http://www.gala-global.org/oscarStandards/tmx/tmx14b.html

LISA. 2008. Segmentation Rules eXchange (SRX), version 2.0. http://www.gala-global.org/oscarStandards/srx/srx20.html

Maas, Heinz-Dieter, Christoph Rösener, and Axel Theofilidis. 2009. Morphosyntactic and semantic analysis of text: The MPRO tagging procedure" In *State of the Art in Computational Morphology: Workshop on Systems and Frameworks for Computational Morphology. SFCM 2009, Zurich, Switzerland, September 4, 2009, proceedings,* ed. Cerstin Mahlow and Michael Piotrowski, 76-87. New York: Springer.

Macken, Lieve. 2009. In search of the recurrent units of translation. In *Evaluation of Translation Technology*, ed. by Daelemans, Walter and Véronique Hoste, 195-212. Brussels: Academic and Scientific Publishers.

Massion, François. 2005. *Translation-Memory-Systeme im Vergleich*. Reutlingen: Doculine.

Mitkov, Ruslan and Gloria Corpas. 2008. Improving Third Generation Translation Memory Systems Through Identification of Rhetorical Predicates. In: *LangTech 2008, Rom, 28-29 Februar 2008. Proceedings*. http://langtech.fub.it/en/poster/07_MITKOV.pdf

Nagao, Makoto. 1984. "A framework of a mechanical translation between Japanese and English by analogy principle". In *Artificial and human intelligence. Edited review papers presented at the International NATO Symposion on artificial and human intelligence, Lyon, 1981,* ed. Alick Elithorn and Ranan Banerji, 173-180. Amsterdam, New York, Oxford: North Holland.

OASIS. 2006. *The DocBook Document Type.* Billerica, MA (USA): Organization for the Advancement of Structured Information Standards (OASIS). http://www.oasis-open.org/docbook/specs/docbook-4.5-spec.html

OASIS. 2007. *DITA Version 1.1. Architectural Specification.* Billerica, MA (USA): Organization for the Advancement of Structured Information Standards (OASIS). http://www.oasis-open.org/committees/download.php/24944/dita1.1.zip

OASIS. 2008. *XLIFF Version 1.2.* Billerica, MA (USA): Organization for the Advancement of Structured Information Standards (OASIS). http://docs.oasis-open.org/xliff/xliff-core/xliff-core.html

Offersgaard, Lene, Claus Povlsen, Lisbeth Almsten, and Bente Maegaard. 2008. "Domain Specific MT in Use". In *Proceedings of the Twelfth EAMT conference, 22-23 September 2008*, ed. Hutchins, John and Walter von Hahn, 148-157. Hamburg: HITEC e.V.

Paulsen Christensen, Tina and Anne Schjoldager. 2010. Translation-Memory (TM) Research: What Do We Know and How Do We Know It?. In *Hermes – Journal of Language and Communication Studies* 44:89-101.

Planas, Emmanuel. 2005. SIMILIS: Second-generation translation memory software. In *Translating and the Computer 27: Proceedings of the Twenty-seventh International Conference on Translating and the Computer*. London: Aslib.

Reinke, Uwe. 2004. *Translation Memories: Systeme – Konzepte – linguistische Optimierung*. Frankfurt am Main: Lang.

Reinke, Uwe. 2006. Translation Memories. In *Encyclopedia of Language and Linguistics*, ed. Brown, Keith, 61-65. Oxford: Elsevier.

Rieck, Konrad, Pavel Laskov, and Sören Sonnenburg. 2007. Computation of Similarity Measures for Sequential Data using Generalized Suffix Trees. In *Advances in Neural Information Processing Systems 19*, ed. Schölkopf, Bernhard, John Platt, and Thomas Hoffman, 1177-1184. Cambridge, MA: MIT Press.

Seal, Thomas. 1992. ALPNET and TSS: The commercial realities of using a computer-aided translation system. In *Translating and the computer 13. Proceedings from the Aslib conference 1991*, 120-125. London: Aslib.

Sikes, Richard. 2007. Fuzzy matching in theory and practice. *MultiLingual*, 18(6): 39-43.

Somers, Harold L. 2001. Review article: Example-based Machine Translation. *Machine Translation* 14:113-157.

Somers, Harold L. 2003. Translation memory systems. In *Computers and translation: A translator's guide*, ed. Somers, Harold L., 31-47 Amsterdam/Philadelphia: John Benjamins.

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

W3C. 2008. Extensible Markup Language (XML) 1.0 (Fifth Edition) W3C Recommendation 26 November 2008. World Wide Web Consortium. http://www.w3.org/TR/REC-xml/

Zhechev, Ventsislav and Josef van Genabith. 2010. Maximising TM Performance through Sub-Tree Alignment and SMT. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas*. http://amta2010.amtaweb.org/AMTA/papers/2-19-ZhechevVanGenabith.pdf