Automation of Translation: Past, Presence, and Future

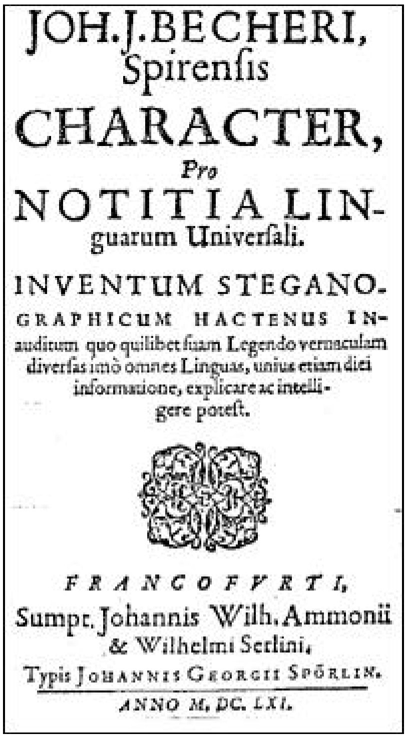
Karl Heinz Freigang, Universität des Saarlandes, Saarbrücken

Introduction

First attempts in "automating" the process of translation between natural languages can be traced back until the middle of the seventeenth century, when a German monk in the town of Speyer, Johannes Becher, wrote a booklet on his invention of a mathematical meta-language designed to describe the meaning of sentences written in any language.

This meta-language, consisting of strings of numbers for the meaning of words and some other numbers expressing the semantics of inflectional endings, was accompanied by lists of equations assigning e.g. German and Latin words to mathematical expressions of meanings, so that sentences in one of these languages could be translated into the other language in a "mechanical" way via these equations.

The first real "machines" for mechanical translation were invented between 1930 and 1940 by men like Georges Artsruni a French engineer ("inventor of the "Mechanical Brain") and the Russian engineer P. Trojanskij. Independently of each other they invented mechanical devices for scanning a punch tape containing expressions in one natural language and mapping these "words" with words of another language punched on a second tape.



JOH.J.BECHERI,
Spirenſis
CHARACTER,
Pro
NOTITIA LIN-
guarum Univerſali.
INVENTUM STEGANO-
GRAPHICUM HACTENUS IN-
auditum quo quilibet ſuam Legendo vernaculam
diverſas imò omnes Linguas, unius etiam diei
informatione, explicare ac intelli-
gere poteſt.

FRANCOFVRTI,
Sumpt. Johannis Wilh. Ammonii
& Wilhelmi Serlini,
Typis JOHANNIS GEORGII SPÖRLIN.
ANNO M. DC. LXI.

Fig. 1: Facsimile of booklet cover

Only after World War II at the end of the forties when the first large electronic calculating machines began to be used for mathematical purposes, scientists began to think about using these machines also for non numerical purposes, e.g. for decoding encrypted messages and translating them into natural language. In his well-known Memorandum to the Rockefeller Foundation in 1949 Warren Weaver made his famous analogy between translation and decoding:

> "I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded in some strange symbols. All I need to do is strip off the code in order to retrieve the information contained in the text."

In the following years, however, machine translation turned out to be much more complicated, because it is not only a matter of one-to-one correspondence between codes or symbols but rather of analyzing the grammatical and semantic meaning of language in order to be able to translate from one language into another (more details about the history of machine translation can be found e.g. in Hutchins 1986).

From Georgetown to ALPAC

The first "successful" demonstration of a machine translation system took place on January 7, 1954 at Georgetown University in Washington, D.C. A system translating from Russian into English was implemented on an IBM main frame computer including a bi-lingual dictionary with about 250 entries. A text corpus consisting of 60 simple sentences in Russian was successfully translated by the system, based on a word-by-word substitution process with some additional rules for the generation of correct word order in the English sentences.

The results of the experiment was regarded as evidence for the feasibility of machine translation, and at that time some researchers were convinced that the final success of machine translation was only a matter of building huge dictionaries. "Fully Automatic High-Quality Machine Translation" (FAHQT) was considered as a goal which could be reached within a few years.

About ten years later, the American government installed a commission with the task of analyzing the translation market and the state of the art reached in machine translation research and development. This Automatic Language Processing Advisory Committee published its final report in 1966, the famous ALPAC-Report, which also became known as the "Black Book on machine translation" ? not only because of the black cover of the booklet. The Commission came to the result that there was no need for further support of research and development in machine translation and that it would make much more sense to spend money for improving the quality of traditional translation by human translators:

> "The committee sees, however, little justification at present for massive support of machine translation per se, finding it ? overall ? slower, less accurate and more costly than that provided by the human translator..." (ALPAC 1966)

As a consequence of this report, government funding for machine translation was almost totally stopped in the USA, in Europe there were some research projects like those at the University of Grenoble in France and at the University of Saarbrücken in Germany. Only in the mid-seventies research and development was again restarted also in the USA, but above all also in Japan, a development which was later on intensified again when since the eighties more and more powerful micro-computers (PCs) were developed.

The situation today

Researchers and developers all over the world are in the meantime convinced that the ambitious goal of Fully Automatic High-Quality Translation will not be reached within the near future and have therefore concentrated in the last few years not so much on machine translation systems in the proper sense of the word but rather on the development of tools assisting the translator in his everyday work.

The rapid development in computer hardware and software in the last few years has led to a situation, where professional translation is no longer possible without support of various translation tools. All steps in the process of translation can nowadays be made more comfortable by using a variety of software solutions for research of terminology and background material at the very beginning of a translation project, for assistance in writing and editing a translation or for management of project and customer data.

Research for terminology and background information

Traditional places to search for terminology and background information material concerning the subject area of a translation project were public and private libraries. Today, the Internet provides access to information via search engines, online terminology databases or via online research in library catalogues, often also making available information in full text databases. A number of web sites are concentrating on providing information about and links to terminology resources and other sites relevant for translation, e.g.:

- the Terminology Forum of the University of Vaasa (http://www.uwasa.fi/comm/termino/) with links to a variety of terminological online-databases and dictionaries
- the terminology portal of the Institute of Translating and Interpreting at the University of Innsbruck (http://info.uibk.ac.at/c/c6/c613/termlogy/termsrch.html) with useful information on terminology in the Internet
- URODICAUTOM, the terminological database of the Commission of the European Union, which is freely available in the Web (http://eurodic.ip.lu/cgi-bin/edicbin/EuroDicWWW.pl)
- the web site of the Department for Applied Linguistics, Translating and Interpreting of the University of the Saarland with a comprehensive collection of translation-oriented links (http://www.uni-saarland.de/fak4/fr46/deutsch/www.htm) to other training institutes, professional organizations, language industry etc.
- the web sites of the International Federation of Translators (FIT) (http://www.fit-ift.org/) and of national organizations like the German Bundesverband der Dolmetscher und Übersetzer (BDÜ) (http://www.bdue.de/)

Other sources of information and at the same time opportunities to exchange experience and knowledge between translators are discussion lists in the Internet which can easily be subscribed and which offer discussions on a variety of translation-relevant subjects, e.g.:

- Lantra-L, an international forum on all aspects of translation and interpretation (for more information refer to http://www.geocities.com/Athens/7110/lantra.htm)
- u-forum, a German speaking mailing list on general issues concerning professional translation and interpretation (for more information refer to http://www.techwriter.de/thema/u-forum.htm) and u-cat, a German discussion forum especially oriented towards all questions in connection with translation tools (cf. http://www.techwriter.de/thema/u-cat.htm)
- TW_Users, an international discussion forum on all aspects connected with the use of Trados translation tools (Translator's Workbench); this discussion group belongs to the Yahoo-groups and can be accessed via http://groups.yahoo.com/.

Multilingual editing and word processing

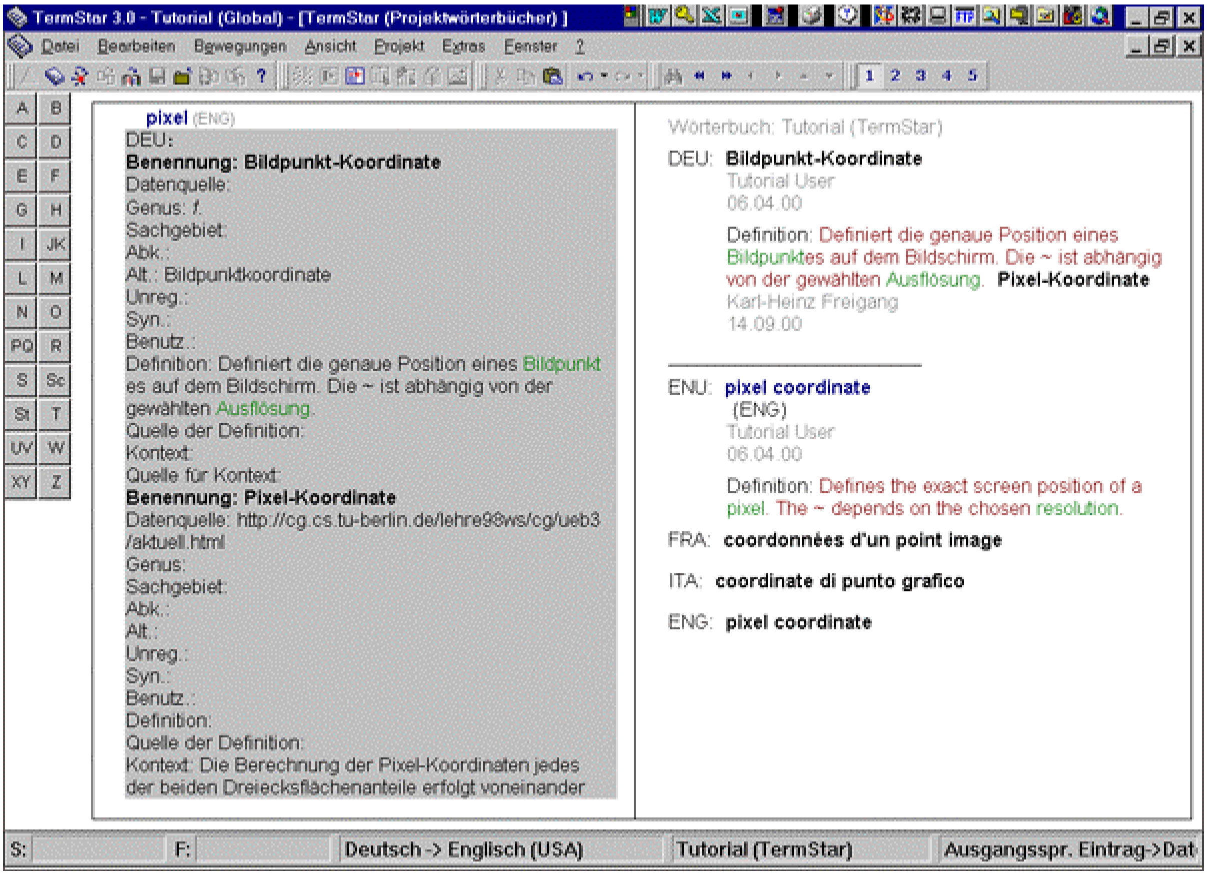
Editors or word processing systems which have to be used in the translation process must offer the possibility to edit and enter texts in all languages, which are relevant for the respective translator. Some years ago, using the traditional 8-bit character sets (ASCII or ANSI), only 256 different characters were available, and if the translator had to work in languages with non-Latin character sets like Russian or Greek switching between different code pages was necessary. Nowadays, modern Windows-systems offer support for Unicode, which in its 16-bit version offers more than 65 thousand different characters within one character set. So the translator has only to install the Windows language support and can switch between different keyboard layouts even within one document. Current multilingual word processing systems, of course, also support country-specific formats for dates, currency, measurement units etc.

Terminology Management

Terminology plays an important part in translation of LSP texts (Language for Special Purposes) from various subject fields. Besides terminology which can be found in dictionaries or online-databases, translators have to compile and manage terminological data from various other sources, e.g. terminology collections from customers or colleagues. This terminological data have to be imported into user-defined terminological databases which are managed with special Terminology Management systems. In order to be able to use this terminology also while editing a translation in a word processing system, interfaces between the terminology system and the word processor are installed which allow the translator to look up terms in the database from within the word processor and also to paste translations from the database into the text.

There are different types of terminology management systems available on the market, which differ in the complexity of the entries, in the number of languages or language pairs which can be stored in one database and in the flexibility of the entry structure, i.e. the possibility of creating one's own entry structure.

An example for a terminology management system with a complex, predefined entry structure, allowing concept-oriented, multilingual terminology management is the system TermStar from STAR AG (http://www.star-group.net/). Figure 2 shows the structure of a TermStar entry in edit mode.


TermStar 3.0 - Tutorial (Global) - [TermStar (Projektwörterbücher)]
Datei
Bearbeiten
Bewegungen
Ansicht
Projekt
Extras
Fenster
?
1
2
3
4
5
A
B
C
D
E
F
G
H
I
JK
L
M
N
O
PQ
R
S
Sc
St
T
UV
W
XY
Z
S:
F:
Deutsch -> Englisch (USA)
Tutorial (TermStar)
Ausgangsspr. Eintrag->Dat

pixel (ENG)
DEU:
Benennung: Bildpunkt-Koordinate
Datenquelle:
Genus: *f.*
Sachgebiet:
Abk.:
Alt.: Bildpunktkoordinate
Unreg.:
Syn.:
Benutz.:
Definition: Definiert die genaue Position eines Bildpunktes auf dem Bildschirm. Die ~ ist abhängig von der gewählten Ausflösung.
Quelle der Definition:
Kontext:
Quelle für Kontext:
Benennung: Pixel-Koordinate
Datenquelle: http://cg.cs.tu-berlin.de/lehre98ws/cg/ueb3/aktuell.html
Genus:
Sachgebiet:
Abk.:
Alt.:
Unreg.:
Syn.:
Benutz.:
Definition:
Quelle der Definition:
Kontext: Die Berechnung der Pixel-Koordinaten jedes der beiden Dreiecksflächenanteile erfolgt voneinander

Wörterbuch: Tutorial (TermStar)

DEU: **Bildpunkt-Koordinate**
Tutorial User
06.04.00

Definition: Definiert die genaue Position eines Bildpunktes auf dem Bildschirm. Die ~ ist abhängig von der gewählten Ausflösung. **Pixel-Koordinate**
Karl-Heinz Freigang
14.09.00

ENU: **pixel coordinate**
(ENG)
Tutorial User
06.04.00

Definition: Defines the exact screen position of a pixel. The ~ depends on the chosen resolution.

FRA: **coordonnées d'un point image**

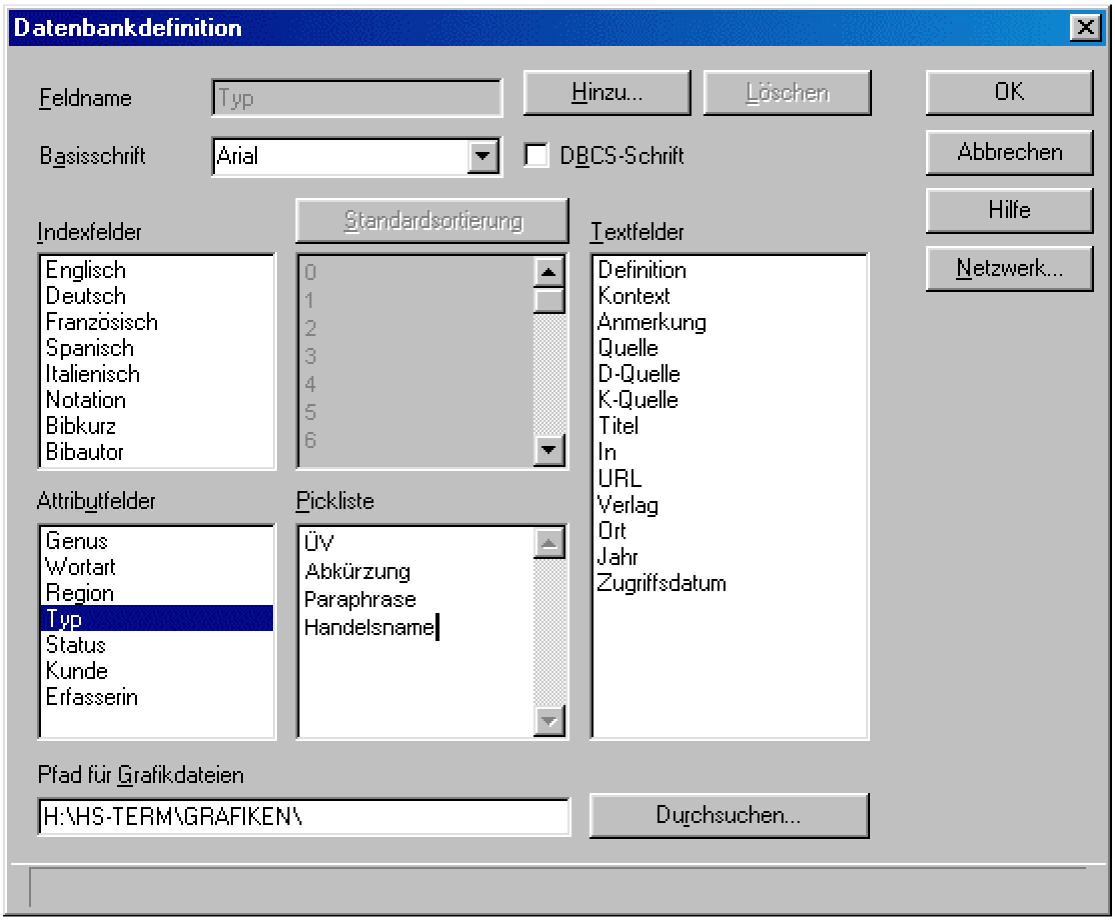
ITA: **coordinate di punto grafico**

ENG: **pixel coordinate**

Fig. 2: TermStar entry

An example for a terminology management system allowing users to define their own entry structure is Multiterm from TRADOS GmbH (http://www.trados.com/).

Here users are free to define data categories they want to use, distinguishing between "index fields" (languages and other categories which will be used for sorting the entries), "text fields" (categories like Definition, Context etc. containing text with variable length) and "attribute fields" (categories with a fixed, predefined set of values, e.g. Grammar, Part of Speech etc.). Figure 3 shows an example of user-defined data categories in Multiterm.



Feldname: Typ

Basisschrift: Arial

☐ DBCS-Schrift

Indexfelder:
Englisch
Deutsch
Französisch
Spanisch
Italienisch
Notation
Bibkurz
Bibautor

0
1
2
3
4
5
6

Textfelder:
Definition
Kontext
Anmerkung
Quelle
D-Quelle
K-Quelle
Titel
In
URL
Verlag
Ort
Jahr
Zugriffsdatum

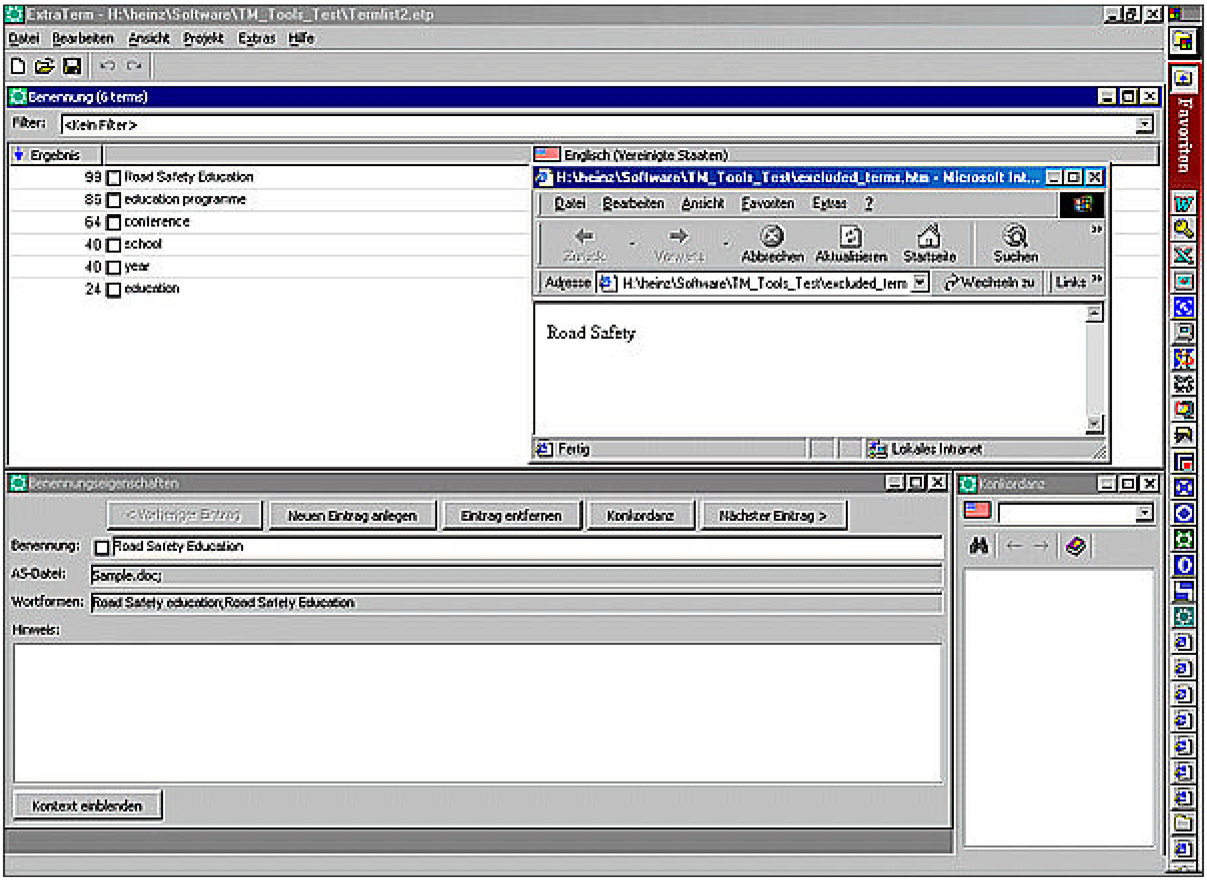
Attributfelder:
Genus
Wortart
Region
Typ
Status
Kunde
Erfasserin

Pickliste:
ÜV
Abkürzung
Paraphrase
Handelsname

Pfad für Grafikdateien: H:\HS-TERM\GRAFIKEN\

Fig. 3: User-defined data categories in Multiterm

As a means to compile terminology collections, software solutions have been developed which try to extract mono- or bi-lingual terminology candidates from existing texts. In the case of extraction of bi-lingual terminology the respective source and target language texts have to be aligned before they can be processed by terminology extraction tools, so that there is a translation relation between the source and target language segments (sentences) containing the term candidates. When extracting term candidates from monolingual texts, an existing terminology database can be used during extraction, in order to exclude terms which already exist in the data base. Figure 4 shows a list of terms extracted from a monolingual document together with "excluded terms" which are already present in the respective Multiterm database; the extraction tool used is ExtraTerm from TRADOS GmbH.


ExtraTerm - H:\heinz\Software\TM_Tools_Test\Termlist2.etp
Datei Bearbeiten Ansicht Projekt Extras Hilfe
Benennung (6 terms)
Filter: <Kein Filter>
Ergebnis
Englisch (Vereinigte Staaten)
H:\heinz\Software\TM_Tools_Test\excluded_terms.htm - Microsoft Int...
Datei Bearbeiten Ansicht Favoriten Extras ?
Zurück Vorwärts Abbrechen Aktualisieren Startseite Suchen
Adresse H:\heinz\Software\TM_Tools_Test\excluded_term
Wechseln zu Links
Fertig
Lokales Intranet
Benennungseigenschaften
< Vorherige Eintrag
Neuen Eintrag anlegen
Eintrag entfernen
Konkordanz
Nächster Eintrag >
Kontext einblenden
Konkordanz
Favoriten

99 □ Road Safety Education
85 □ education programme
64 □ conference
40 □ school
40 □ year
24 □ education

Road Safety

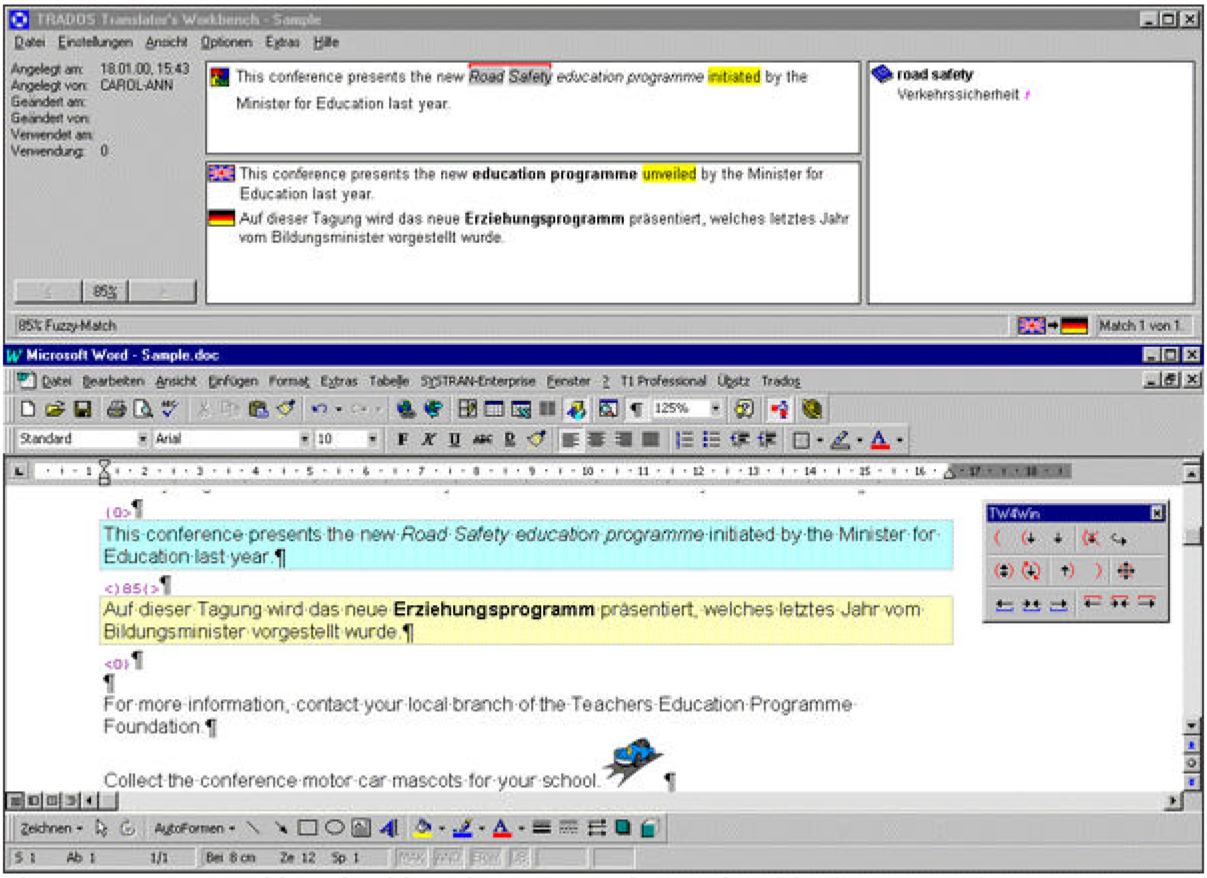
Benennung: □ Road Safety Education
AS-Datei: Sample.doc;
Wortformen: Road Safety education;Road Safety Education
Hinweis:

Fig. 4: Monolingual term list extracted by ExtraTerm

Translation Memory Tools

In most cases, these terminology management systems are a component of an integrated translation environment containing besides the terminology tool an editor for entering and editing the translation and a software tool for recycling former translations of identical or similar source language texts. These "Translation Memory" tools are used during the translation process to search for translation units ("segments", mostly sentences) which are identical or similar to the segments which are currently being translated. The material contained in this translation memory, i.e. source language texts and their target language translations, is either organized as pairs of segments (sentences) in a bi-lingual sentence database or as pairs of "reference texts" where source and target language sentences are associated with each other. The search algorithms are not only able to find exactly identical segments ("exact matches") but also similar segments ("fuzzy matches"). When translating a new text using the translation memory tool, all translation units can be stored in the memory and thus be re-used again when the same or similar units occur again in the same document or in another document.

An example for such an integrated translation environment is TRADOS Translator's Workbench using Word as an editor and Multiterm as terminology component (Fig. 5). In this system complete segments are looked up in the Translation Memory and displayed in the upper part of the screen, whereas single or multi-word terms are found in Multiterm and displayed in the upper right corner of the screen.


TRADOS Translator's Workbench - Sample
Datei Einstellungen Ansicht Optionen Extras Hilfe
85%
Microsoft Word - Sample.doc
Datei Bearbeiten Ansicht Einfügen Format Extras Tabelle SYSTRAN-Enterprise Fenster ? T1 Professional Übstz Trados
125%
Standard
Arial
10
TW4Win
Zeichnen
AutoFormen

Angelegt am: 18.01.00, 15:43
Angelegt von: CAROL-ANN
Geändert am:
Geändert von:
Verwendet am:
Verwendung: 0

This conference presents the new *Road Safety education programme* initiated by the Minister for Education last year.

This conference presents the new **education programme** unveiled by the Minister for Education last year.

Auf dieser Tagung wird das neue **Erziehungsprogramm** präsentiert, welches letztes Jahr vom Bildungsminister vorgestellt wurde.

road safety
Verkehrssicherheit

85% Fuzzy-Match

Match 1 von 1

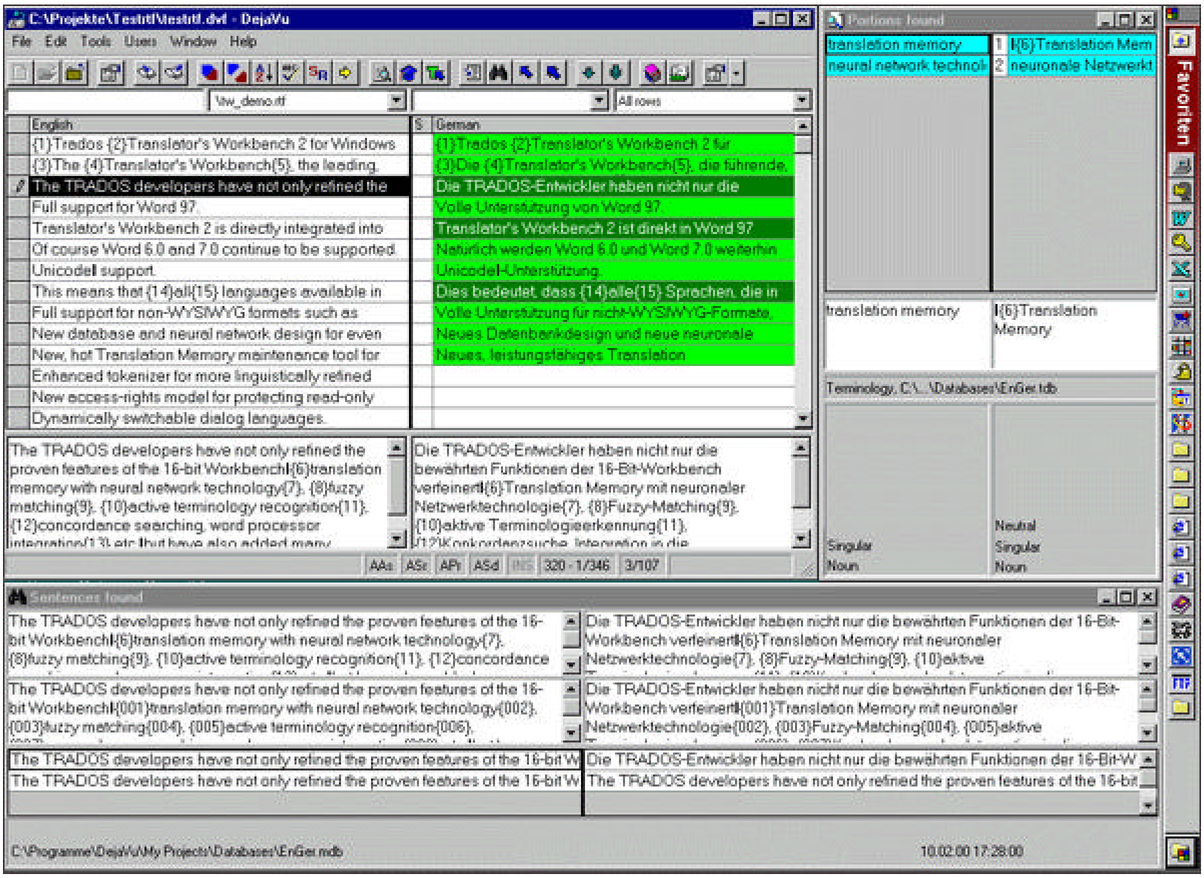
{0>¶
This conference presents the new *Road Safety education programme* initiated by the Minister for Education last year.¶
<}85{>¶
Auf dieser Tagung wird das neue **Erziehungsprogramm** präsentiert, welches letztes Jahr vom Bildungsminister vorgestellt wurde.¶
<0}¶
¶
For more information, contact your local branch of the Teachers Education Programme Foundation.¶

Collect the conference motor car mascots for your school. ¶

S 1 Ab 1 1/1 Bei 8 cm Ze 12 Sp 1

Fig. 5: TRADOS Workbench with active segment in Word and in the Translation Memory

In the system Déjà Vu from Atril, Spain (http://www.atril.com/) the search algorithms not only look up and find complete segments (sentences) but also so-called "Portions", which normally are word groups or sub-clauses. Thus, even if there is no correspondence found for the whole segment (sentence) in the translation memory database, Déjà Vu tries to "assemble" a translation from "portions" found inthe memory and terms found in the term base. The editor in Déjà Vu is organized in form of a table with the source language segments in the left and the translations in the right column; the results of look-up in the translation memory are shown on the bottom part of the screen, portions and terminology found are displayed on the right hand side of the screen (Fig. 6).



	English	S	German
	{1}Trados {2}Translator's Workbench 2 for Windows		{1}Trados {2}Translator's Workbench 2 für
	{3}The {4}Translator's Workbench{5}, the leading		{3}Die {4}Translator's Workbench{5}, die führende
	The TRADOS developers have not only refined the		Die TRADOS-Entwickler haben nicht nur die
	Full support for Word 97.		Volle Unterstützung von Word 97.
	Translator's Workbench 2 is directly integrated into		Translator's Workbench 2 ist direkt in Word 97
	Of course Word 6.0 and 7.0 continue to be supported		Natürlich werden Word 6.0 und Word 7.0 weiterhin
	Unicode support.		Unicode-Unterstützung.
	This means that {14}all{15} languages available in		Dies bedeutet, dass {14}alle{15} Sprachen, die in
	Full support for non-WYSIWYG formats such as		Volle Unterstützung für nicht-WYSIWYG-Formate,
	New database and neural network design for even		Neues Datenbankdesign und neue neuronale
	New, hot Translation Memory maintenance tool for		Neues, leistungsfähiges Translation
	Enhanced tokenizer for more linguistically refined		
	New access-rights model for protecting read-only		
	Dynamically switchable dialog languages.		

The TRADOS developers have not only refined the proven features of the 16-bit Workbench{6}translation memory with neural network technology{7}, {8}fuzzy matching{9}, {10}active terminology recognition{11}, {12}concordance searching, word processor

Die TRADOS-Entwickler haben nicht nur die bewährten Funktionen der 16-Bit-Workbench verfeinert{6}Translation Memory mit neuronaler Netzwerktechnologie{7}, {8}Fuzzy-Matching{9}, {10}aktive Terminologieerkennung{11},

translation memory	1	{6}Translation Mem
neural network technol	2	neuronale Netzwerkt

translation memory
{6}Translation Memory

Terminology, C:\...\Databases\EnGer.tdb

	Neutral
Singular	Singular
Noun	Noun

The TRADOS developers have not only refined the proven features of the 16-bit Workbench{6}translation memory with neural network technology{7}, {8}fuzzy matching{9}, {10}active terminology recognition{11}, {12}concordance

Die TRADOS-Entwickler haben nicht nur die bewährten Funktionen der 16-Bit-Workbench verfeinert{6}Translation Memory mit neuronaler Netzwerktechnologie{7}, {8}Fuzzy-Matching{9}, {10}aktive

The TRADOS developers have not only refined the proven features of the 16-bit Workbench{001}translation memory with neural network technology{002}, {003}fuzzy matching{004}, {005}active terminology recognition{006},

Die TRADOS-Entwickler haben nicht nur die bewährten Funktionen der 16-Bit-Workbench verfeinert{001}Translation Memory mit neuronaler Netzwerktechnologie{002}, {003}Fuzzy-Matching{004}, {005}aktive

The TRADOS developers have not only refined the proven features of the 16-bit W	Die TRADOS-Entwickler haben nicht nur die bewährten Funktionen der 16-Bit-W
The TRADOS developers have not only refined the proven features of the 16-bit W	The TRADOS developers have not only refined the proven features of the 16-bit

C:\Programme\DejaVu\My Projects\Databases\EnGer.mdb
10.02.00 17:28:00

Fig. 6: Déjà Vu with editor, translation memory results, portions and terminology

The use of "portions" and the assemble feature in Déjà Vu make this system similar to computer-aided translation systems which are also called "example-based" machine translation systems. Instead of trying to implement a complete and comprehensive linguistic analysis within a machine translation system, this approach tries to base the machine translation process on a huge collection of "translation examples", structured e.g. according to types of surface structure phrases (noun phrases etc.).

Automatic Translation Today

"Example-based" machine translation is, however, still in development and no machine translation system available on the market uses this approach, although some of the systems try to make use of translation memory technology. These systems like Langenscheidt T1 (http://www.langenscheidt.de/deutsch/index.html) or Personal Translator from Linguatec (http://www.linguatec.de/topics/mt2001.shtml) combine traditional machine translation approaches (T1 formerly METAL developed by University of Austin and Siemens and Personal Translator formerly LMT developed at IBM) with the concept of translation memory. Before processing a sentence by their linguistic analysis routines, these systems can look up the whole sentence in a translation memory ("translation archive"); only if no matching sentence is found in the memory linguistic analysis is started.

Similar to these two systems which have their roots in veterans of machine translation history there are other systems available today which are PC-based versions of old mainframe translation systems. One of the dinosaurs in machine translation, SYSTRAN, which is still in use at the Commission of the European Union and at some other places, is in the meantime available as SYSTRAN Personal, Professional or Enterprise running under Windows NT/2000 (http://www.systransoft.com/). The oldest PC-based machine translation system with its origins in the Georgetown system, Globalink, is now available as a Windows-based software, PowerTranslator (http://www.lhsl.com/powertranslator/).

Besides these PC-based systems, there are still machine translation systems in use running on mainframe computers or under UNIX operation system. SYSTRAN, as mentioned above, is still used by the Commission of the European Union, producing rough translation mostly for information purposes only. TAUM METEO, developed at the University of Montreal, has been in permanent use over the last decades for translating weather forecasts from English into French at the Canadian Meteorological Center. These weather forecasts contain a restricted vocabulary and only few types of syntactic structures; METEO has been optimized for dealing with such a restricted language.

Conclusions

Computer-aided translation tools are indispensable at the translator's workplace of today. Whereas the linguistic capabilities of automatic translation systems have not been very much improved within the last two decades, translation memory and terminology management tools are widely used by technical translators having to deal with specialized texts which are often repetitive and subject to frequent updating and which require coherent usage of terminology.

But not only in the translation process in the proper sense of the word automation is constantly increasing, also the whole process of researching and managing information retrieval is governed by computer technology, the most important area being the use of the Internet. This also includes communication between translators and customers as well as project management.

Efficient use of all tools for the translator's working environment also may require new workflow organization by the translator. Since mistakes made in the initial phase of introducing tools in one's own environment may have disastrous consequences later on, it is recommended to spend some hours or days getting acquainted with the tools before using them in a real translation project. Efficient training is offered by the distributors of the software, by professional organizations and by training institutes.

Bibliography

Hutchins, W.J. (1986): Machine Translation: Past, Presence, Future. Ellis Horwwod/Wiley, Chichester/New York.

ALPAC (1966): Languages 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.

Octubre 2001