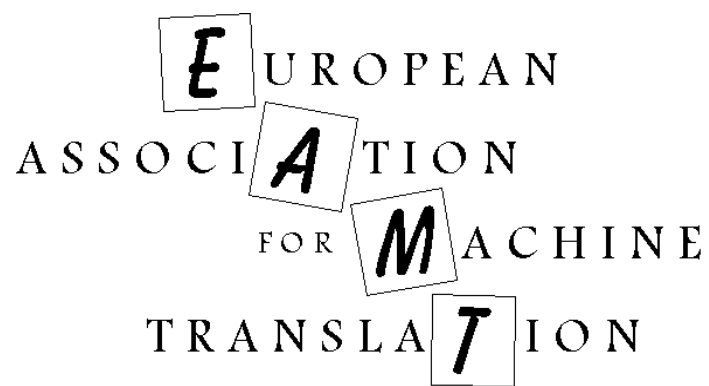


EAMT Machine Translation Workshop

TKE '96, Vienna, Austria, 29 - 30 August 1996



Proceedings

Technische Universität Wien, Elektrotechnisches Institut
Gusshausstrasse 27-29, A-1040 Wien, Österreich
Kontaktraum

For general information about the EAMT:

eamt@cst.ku.dk

EAMT Secretariat

ISSCO

54, route des Acacias

CH-1227 Carouge (Geneva)

Switzerland

Tel: +41 22 705 7115

Fax: +41 22 300 1086

The International Association for Machine Translation is a non-profit organisation registered in Switzerland.

Compiled by Dimitri Theologitis.

Thanks go to all the members of the EAMT Committee for their invaluable help with the organisation of this workshop.

D. Theologitis, November 1996

Table of Contents

What is the EAMT?	5
Introduction	
<i>John Hutchins, EAMT President</i>	7
Session 1: Controlled Languages, Localisation	
<i>Chair: John Hutchins</i>	9
Bringing Controlled Language Support to the Desktop	
<i>Drs Michiel de Koning</i>	11
Machine Translation, Translation Memories and the Phrasal Lexicon: The Localisation Perspective	
<i>Reinhard Schäler</i>	21
Session 2: Domain-Specific Lexica	
<i>Chair: Prof. Bente Maegaard</i>	35
Users' Experiences with a Set of Domain-Specific Dictionaries for the Stylus Machine Translation System	
<i>Svetlana Sokolova</i>	37
Building Term Dictionaries for Machine Translation in Practice: A User Experience	
<i>Annelise Bech</i>	45
Session 3: Experiences of a Large-scale User: The European Commission	
<i>Chair: Dimitri Theologitis</i>	51
Generalised Language Resources: EURODICAUTOM, SYSTRAN and EURAMIS - a Case Study	
<i>Jean-Marie Leick</i>	53
Machine Translation Feasibility Study at the European Commission	
<i>Dorothy Senez</i>	63
Session 4: Economics of Using Machine Translation	
<i>Chair: Viggo Hansen</i>	71
Kielikone Machine Translation Technology and Its Perspective on the Economics of Machine Translation	
<i>Dr Harri Arnola</i>	73
Use and Value of Computer-Assisted Translation in the Central Translation Service of Coop Switzerland, Basle	
<i>Martha Ebermann</i>	89
Machine Translation, Terminology and the African Languages in South Africa: An Overview	
<i>Milde Jordaan-Weiss</i>	95
Session 5: Machine Translation vs. Translation Memories: Rivals or Partners?	
<i>Chair: Dr. Jörg Schütz</i>	99

Use of Linguistic Resources like Translation Memories in Machine Translation Systems	
<i>Lee Humphreys</i>	101
Integrating Machine Translation into Translation Memory Systems	
<i>Matthias Heyn</i>	113
Translator's Workbenches: A Practical Application	
<i>Adriane Rinsche</i>	127
Session 6: Integration of Machine Translation in Information Management	
<i>Chair: Colin Brace</i>	137
"Translation and the Internet" A Sample Application Based on the Logos MT System	
<i>Joachim Meyer</i>	139
Network-based Machine Translation Services	
<i>Dr. Jörg Schütz</i>	147
Conclusion	
<i>Viggo Hansen, EAMT Secretary</i>	161

What is the EAMT?

The European Association for Machine Translation (EAMT) is an organisation that serves the growing community of people interested in MT and translation tools, including users, developers and researchers of this increasingly viable technology. The EAMT is one of three regional associations of the International Association for Machine Translation (IAMT), which counts on an increasing number of members world-wide. The EAMT is the only organisation of its kind in Europe.

And What Does it Do?

1. Conferences and Workshops

Every two years, the IAMT organises the MT Summit, a unique conference dedicated to the world of translation technology that alternates between Europe, North America and Asia. In 1995, the MT Summit was held in Luxembourg; in 1997, it will take place in San Diego, California. In years when the Summit is not being held in Europe, the EAMT organises an annual workshop on some facet of MT, usually in conjunction with another event. EAMT members can attend EAMT- and IAMT-sponsored events at reduced rates.

2. Expertise and Networking

The EAMT is an excellent way to learn more about the practical aspects of MT. MT is still an imperfect technology and its members have a lot of collective wisdom to share about the use of MT in working environments. Through the EAMT, you may be able to reach people who have “been there before”.

3. MT News International

Together with its sister associations, the EAMT publishes a lively and informative newsletter, *MT News International*, three times a year. It includes news of upcoming events such as workshops and conferences, reports on previous events, company and product news and updates of research developments. In short, *MTNI* is an excellent source of information on the world of MT.

4. Other Publications

The IAMT regularly produces publications which are available to members at reduced rates, such as conference proceedings, a directory of MT systems and a “yellow-pages” of the members of the MT community.

5. Bibliographic Service

The EAMT offers a bibliographic service to its members. For a nominal amount, members can order photocopies of articles on MT from a wide range of publications in the EAMT archives.

Introduction

John Hutchins, EAMT President

The European Association for Machine Translation (EAMT) is one of the three regional associations which make up the International Association for Machine Translation. Each regional association organises activities, conferences and workshops in its own geographical area with the aim of bringing together users, developers, researchers and others interested in this increasingly important field of computer aids and systems for translation.

In the last few years, there have been a number of far reaching and important changes in what is traditionally called Machine Translation (MT). Developments have been made possible by the increased power and storage capacity of computers, the increased availability of machine-readable documentation, and the expansion and wider use of global telecommunications (most recently the internet). Recent years have seen a trend away from mainframe systems (installed mainly in multinational, organisations and large translation services) towards computer-based translation aids for professional translators, PC-based software for individual translation needs, and networked services for translators and for occasional translation.

The large mainframe MT systems were originally intended to produce translations fully automatically, but in practice they are rarely used without human intervention. Only if MT systems are restricted to narrow subject domains is it possible to take unedited source documents and produce good quality output without assistance from translators. Invariably, MT output must be revised if it is to be of publishable quality. However, rough unedited versions can have some value for information assimilation, as drafts for correspondence, or for assessing the need for later human or human-aided translation. In large multinational organisations, where a single source text (e.g. a technical manual) must be translated in many languages, it is increasingly common practice for input texts to be 'pre-edited': i.e. prepared for machine translation by reducing ambiguities of vocabulary or grammatical structures, and/or already written in a 'controlled language' (e.g. a simplified form of English) to ensure that the output does not need major revision.

The appearance of cheaper and more accessible PC-based translation software is satisfying primarily occasional translation needs, mainly by non-professional translators or by people with little or no knowledge of source or target languages. These systems have often relatively small dictionaries and sometimes lack of specialist vocabulary — although the deficits are becoming less significant as more powerful PC-based systems appear (many of originally mainframe systems are now available in PC versions.)

In general, professional translators do not like to work with systems (whether mainframe or PC) which attempt fully automatic translation. Output is rarely good enough for use in professional contexts: too much has to be changed, and translators do not want to be human assistants to machines. For their needs the development of the translator's workbench is far more attractive. Workbenches offer a wide range of facilities, from multilingual (and multi-character) text processing and editing, telecommunication facilities, management of terminology, access to previous translations, and the option of using MT when desired. The translator remains fully in control: the workbench is an extension and elaboration of traditional translation

practice. In particular, they offer 'translation memories', corpora of aligned bilingual texts (source documents and their translations), allowing whole segments of previously translated material to be easily incorporated. Such translation memories are made possible only with the increasing availability of machine-readable documents, whether created internally or externally.

The need for translation is itself growing ever more urgent: with shorter production times demanding more quickly produced translations (a feature particularly strong in the area of software localisation), with wider global markets requiring translations into more languages, and with the growth of telecommunications bringing to more people an awareness of resources and databases in other languages. To meet the demands there is a widening range of types of systems (and of potential types of uses); and translation facilities are increasingly available on networks (as services for both professional and occasional translators.) The question, as always, is what are the most appropriate systems for each particular need or situation. There are no perfect solutions (the problems of linguistic processing and of translation are alone too great); the best use must be made of those systems that are available, and the producers and developers must be encouraged to improve and introduce new facilities to meet user needs. Users and potential users need to be informed about what is possible and what is not.

This EAMT workshop was devoted to the exploration of machine translation and computer-based translation tools ('translation technologies') in actual use and under development. In publishing the following collection of papers given at the workshop, EAMT hopes to encourage further collaboration and discussion between those with practical experience of translation aids and systems, the developers of new systems, and those wanting to find out what is available now and what may be coming in the near future.

Session 1: Controlled Languages, Localisation

Chair: John Hutchins

Introduction

The presentations in this session will be introducing a number of the practical and economic issues involved in setting up translation services based on MT and MAT systems.

Interest in controlled languages has witnessed a sharp increase in the last few years - earlier this year, for example, there was the first international conference devoted exclusively to this topic - held at the University of Leuven, Belgium. The relationship between MT and controlled language has been close for many years. The earliest examples date from the late 1970s, with the SYSTRAN implementation at Xerox Corporation, and the specific customised systems by Smart Communications Inc. And the approach itself derives in part from the even earlier observation that MT is always more successful when input texts are stereotyped or limited in vocabulary and grammar: the best known illustration being the Meteo system in Canada which since 1976 has been translating daily weather reports.

The second theme in this first session is MT and localisation. Here the developments have been more recent, but equally dramatic. Prompted by pressures to adapt computer software to local conditions and languages, many companies have set up services based on MT and computer-aided translation systems - many, in fact, in Ireland with financial incentives from the European Union. But localisation is not restricted to software products - the lessons from the use of MT in this area should be valid in many other parts of the global market.

John Hutchins

John Hutchins is president of the European Association for Machine Translation. He is the author of a number of articles on machine translation, he has given presentations to many MT conferences since the late 1970s, and has written two books on the subject: a history published in 1986, and a textbook (jointly with Harold Somers) published in 1992. He is also chief editor of **MT News International**, the thrice-yearly newsmagazine of the International Association for Machine Translation.

John Hutchins
The Library, University of East Anglia
Norwich NR4 7TJ, UK

E-mail: J.Hutchins@uea.ac.uk; or: 100113.1257@compuserve.com

Bringing Controlled Language Support to the Desktop

Drs Michiel de Koning

Abstract

This paper discusses the work being carried out at Cap Volmac on tools and services involving controlled languages and MT pre-editing. It focuses on the planning involved in implementing projects for authoring environments, based on experiences of the last 5 years.

It describes the steps to be taken for a successful implementation. These steps involve typical activities such as analysis, design, product selection, development, and integration, both on the IT and linguistic side.

The main focus of the paper will be on the following aspects: language specification (company specific, industry specific, application specific, generic), active intervention/ author support during document writing/editing and organisational requirements.

This paper will conclude a brief discussion of expected development, in terms of links to other automated (sub)systems, such as mark-up languages (structure of the document), document Management and Workflow Management (status of document), PDM (external data)

Drs Michiel de Koning

Mr. M. C. de Koning is project manager at Cap Volmac in the Netherlands. For the last three years, he has been involved in the development of tools and services based on the application of controlled languages for MT projects. Prior to that, he was active as a IT consultant on various projects.

Mr. de Koning has a Masters degree in Linguistics at the Utrecht University.

Cap Volmac

Active Documentation is a collective term for the consultancy, services and tools provided by Cap Volmac which support the production process of documentation. Active Documentation is part of Cap Volmac Team Support Technology B.V., which incorporates document production, document usage (retrieval, publication), document storage (DMS), groupware (Lotus Notes), and workflow automation (DIS/WFM).

Active Documentation focuses on content management and, more specifically, language- and text technology. Automated correction and machine translation systems, text structuring (SGML), terminology management and document conversion are a few examples.

Cap Volmac in the Netherlands is a member of the Cap Gemini Group, a world-wide IT consultancy and services firm.

Drs Michiel de Koning,
Cap Volmac Active Documentation
Daltonlaan 400
Pobox 2575, 3500GN Utrecht, The Netherlands

Tel: +31 30 2527127, Fax: +31 30 2527045
E-mail: MCdeKoning@CapVolmac.nl

Introduction

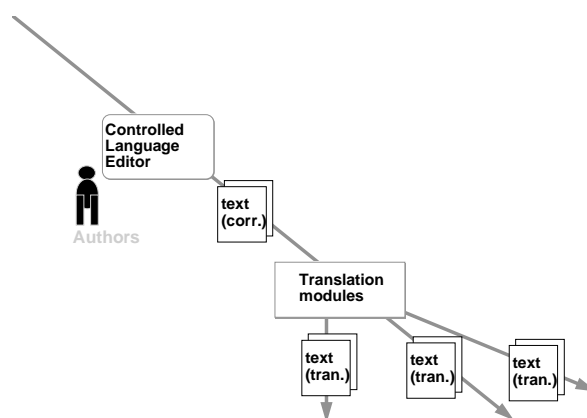
For the purpose of this paper, we will use a rather practical definition of controlled language. In terms of what we want to achieve with language control, a controlled language is the specification of a language that will improve processing of information (be it a human or machine) later on. Thus, a controlled language will most likely be implemented for the purpose of eliminating interpretation problems. As such, it can be seen as a kind of sieve, that allows only a subset of a natural language.

Interpretation problems may occur in different situations, real time (speech) or written (documentation). The typical effect of interpretation problems is that the recipient will check back with the originator to make sure he understood properly. In speech this normally is not a problem, but in delayed communication, this will take much more effort.

Controlled languages have been around now for a number of decades. Probably the best known example is Simplified English, used in the aerospace industry as a means to improve the ‘understandability’ of maintenance manuals, particularly for non-native audiences. This means that certain types of sentence structures, phrases or words are not allowed in maintenance manuals, because they may be understood incorrectly.

If we take this approach a step further, we could also use language control to improve performance of machine translation systems. Similar interpretation problems that, in this case, determine the performance of MT systems, may be reduced by providing input written in controlled language. This has been done for existing MT systems (Xerox Multinational Customized English), and is done for new systems that are optimized for the combination of controlled language input and machine translation.

When we look at using controlled language as a preparatory phase for using MT systems, we can draw the following figure:



Traditionally, MT systems produced output that had to be post-edited. When used in a specific domain and tuned to this domain, the amount of post-editing becomes less. We can further improve the performance of the MT system by pre-editing the input

(according to specification of a controlled language). Considering that pre-editing only occurs once (in the source language), and post-editing must be done for all target languages, the extra effort of pre-editing may very well pay off.

Controlled language specification

The principle factor that must be taken into account for the specification of any controlled language is the purpose. Why do we want to control the expressive power and variation of language? If we look at Simplified English (or similar writing rules found in text books), some constructions are not approved so that people better or more quickly understand a written message. However, for MT systems, these same constructions may not cause any problems. Take, for example, the passive construction. In Simplified English, the passive may not be used in text segments that deal with procedures. An MT system does generally not have difficulty translating a passive construction.

Clearly, we can see that for different purposes, we can specify different types and levels of control. For this paper, we will take a closer look at MT systems, and the way in which language control can affect the performance of such systems.

Language requirements

Until now, MT systems are used in two ways: human aided machine translation, and machine aided human translation. The first is represented by systems that provide rough translations that must be post-edited, the second by systems that provide functions such as dictionary look-up or translation memory to a translator. Combinations of the two types of systems in one system have also been designed and implemented, such as a translation memory with a translation engine for unknown phrases.

Language control can be used to improve such types of translation systems. Let us take a look at one example. An MT system that would have to translate the sentence "Remove panels and open doors." into French will most likely produce at least two translations: (1) "Enlever panneaux et ouvrir portes" or (2) "Enlever panneaux et portes ouvertes". If we make the input sentence a bit more specific, we can solve this problem. Assume we have a rule in our controlled language that we must use determiners whenever possible. We would then write our English sentence as (1) "Remove the panels and open the doors" or (2) "Remove the panels and the open doors." Now, when we feed one of these alternatives to an MT system, we will get a proper translation.

When we want to specify a controlled language, we need to find out which grammatical and lexical phenomena must be controlled. This is not an easy task. We must also make sure that the communication style of the documents for which the language is targeted, is suitable for language control. For example, legislative texts often contain ambiguities to allow for interpretation by a judge. Clearly, this type of language usage should not be controlled. For that reason, most applications of controlled language are found in industrial documentation, where the primary objective is to pass information unambiguously. Examples are manuals, help texts, course material, et cetera.

The specification of a controlled language can be done in several ways. We can consider to create a controlled version of each existing natural language. But again,

when we look at the areas where we want to apply controlled languages, a more focused approach can be more beneficial. Document types, as mentioned above, often comply to a sort of sublanguage, because of the nature of the document. On-line help text has certain characteristics, so have maintenance manuals. Furthermore, each type of industry (automotive, aerospace, IT) has its own language characteristics. At Cap Volmac, it is our belief that we should try to specify controlled language for specific applications, be it to improve communication or improve automated processing (or both).

The use of a controlled language has a large impact on the way documents are created (see also below). Because of this, we must make sure that the extra effort will be worth while. Improved performance of MT systems will reduce the amount of work done in the post-editing phase. This ‘balancing’ of pre-editing versus post-editing is important in specifying the rigorousness of the controlled language. For example, MT systems typically have problems with PP-attachment¹. A control rule might be to not use any PPs at all. This is however hardly attainable in most types of communication. We have to consider whether to reduce the input (pre-editing) or leave it for the post-editing phase.

Specification in steps

If we choose to specify specific controlled languages for specific applications, we have to go through a number of steps to arrive at such a specification. Below the activities are discussed briefly as they are implemented in projects at Cap Volmac. As before, we do not include any aspects relating to readability, we focus mainly on the translatability. Readability is of course one of the implementation issues of controlled language that must be considered since the documents are to be published in the controlled source language, but is outside the scope of this paper.

The objective of these steps is to come to a controlled language specification that can be implemented in an automated system.

Purpose

We must first determine the purpose for which we want to use controlled language. Which target languages are involved (that affect the type of control in the source language), and which type of MT system—if any, language control can also be used to facilitate manual translation, particularly Western to Asian— will be used.

Document analysis

The next step is to get a good idea of the sublanguage as it is used in existing documentation (on paper or electronic). Three aspects are particularly important: lexicon (terminology), grammar, and style. In a representative set of documents, we first look at specific terminology used. This could be just nouns, but also expressions (phrases or verb clusters) that are only found in these documents.

Next, we look at grammatical constructions. All sentence structures are documented in global terms, e.g. S-V-O-(A), A-S-V-O, NP-V-O. Finally, in this phase, the type of constructions that are preferred (active over passive, adverb placement, etc.) are marked separately.

¹ e.g. He saw the girl with binoculars. Who has the binoculars?

These findings are discussed with the organization to come to an understanding on the language usage as it occurs (the sublanguage).

Analysis (descriptive) grammar

From the complete set as specified in the previous phase, we must arrive at what is presumably a smaller set that is tuned to the input requirements of processes to follow. Here, we much check the result of the previous phase with the requirements of the MT system. At Cap Volmac, we also develop translation modules which makes it possible to very precisely select the kind of constructions or terms that cause problems. When developing controlled languages for other MT systems, finding out the shortcomings is not a trivial matter. Some manufacturers even claim their system cannot be analyzed to that effect. In that case, general phenomena should be tested in the translation module to check whether they are processed correctly or should be avoided.

Correction (prescriptive) rules

Once the core specification of the controlled language has been determined, it must be considered whether specific correction rules for non-approved structures or words are included in the specification. This can never be a complete set of rules. Correction rules can be two-fold. Either they regulate the conversion of a non-approved form to an approved alternative, or they give hints on how to come to an approved alternative. Correction rules can occur for all three levels: style, grammar and lexicon (terminology).

Lexicon encoding

Based on the grammatical specification of the analysis grammar (the approved grammar) and correction rules, we can now start to build our lexicon. Lexical entries will require specific syntactic and semantic attributes to be able to check its validity. For example, if the word *display* may only be used as a noun, yet not as a verb, the lexicon must contain that information. Likewise, if agreement between subject and verb is required (which is very likely), we need number information for both nouns and verbs. The specification of the grammar will dictate the required attributes for the lexical entries.

To create a proper lexicon, all existing documentation must be analyzed. Apart from closed class items (prepositions, pronouns, etc.) at least all approved words must be encoded. Preferably, all non-approved words are included and marked as such, with a reference to an approved alternative. Compared to the work done for terminology management systems, all words are encoded, as opposed to primarily nouns and adjectives (typically 90% of terminology databases). Yet, from the point of view of a controlled language, less pragmatic information is stored, such as number of occurrences, context, parent-child relations, etc.)²

To make this work easier, at Cap Volmac we use concordance programs for this purpose (to determine the category of a word, a list of occurrences with a limited context is used). The number of entries required is roughly determined by two things:

² Unless of course you combine these two in one application

the type of application (approved words) and the proficiency of authors (non-approved words)

Testing

The specification must be tested. At Cap Volmac, we used a methodology similar to software development. Grammar rules must all be tested, to see whether their correction produces approved alternatives. A good practice here is to include examples for all rules, and use these examples to test. In the lexicon, the non-approved words must be checked. This can be done by generating short phrases from all entries in the lexicon and have these tested against the specification using the grammar. For example, assuming a noun phrase as a sentence is accepted in controlled language, you take all nouns from the lexicon and produce NPs by adding a determiner and closing with a period ('door' → 'The door.').

All errors must be evaluated and corrected in an iterative process.

Implementation issues

Two major issues must be tackled when implementing a solution based on controlled language.

The first is that documentation in controlled language must be accepted by the audience, the users of the information. Again, we believe communication in controlled language can only be implemented for industrial documentation. In the case of Simplified English, it was a regulation for the entire industry to comply to this controlled language. In a more consumer oriented environment (e.g. consumer electronics) it may not always be possible to start using controlled language if this will affect the consumer's perception of the product in a negative sense.

The second is that the people involved in the document creation process must be willing and capable of accepting such a change in the way they work. In our experience, larger organizations with a number of professionally trained authors will make this transition much easier than smaller companies with only one or two authors. This is mostly due to the fact that larger organizations have already begun to implement Quality Assurance programs and are ready to accept the consequences (less 'freedom', procedures) of the introduction of controlled language.

When a controlled language is introduced in the authoring environment, the author must be ready to adopt the new way of working. This will involve training. Not only in working with new tools (see below), but also learning to write in controlled language. Part of this can be done in a training course, but the author will have to go through a learning curve. Basically, he/she has to learn a new language. It turns out only full time authors are able to come up to speed in writing in controlled language; it is not feasible for people who occasionally write documentation to learn to write in this way.

The implementation of controlled language as pre-editors for MT systems involves another issue. Apart from the fact that a company will have to think of which documents will be written in controlled language (to implement a certain style of communicating with its customers), it must also be decided which existing documentation must be rewritten. Typically, this is done on an on-demand basis. If further processing of the document (translation to a new language, changes or

updates, etc.) requires conformance to the new controlled language standard, the document must be rewritten.

When controlled language is introduced in an organization, this will also introduce new tasks regarding the maintenance of resources dealing with the controlled language specification. Most importantly, the (company specific) terminology or more generally the lexicon (in case of automated controlled language support) must be kept up-to-date. This is a very important job, since it will manage the style of communication (consistency, preference, etc.) of an organization.

Other resources, such as grammars, when built right, tend not to change. Building them requires specialist linguistic knowledge not found in industrial organizations. So this remains the responsibility of the supplier of the tools (when relevant) or the regulatory body (in case of industry wide specifications).

Organizations quite easily adopt new tasks, tools, and responsibilities, when they are directly related to the existing processes. In this perspective, changes to the document creation process can be implemented properly. However, when it comes to translation, they prefer to keep this as it was, in the hands of subcontractors. This means that the introduction of controlled language to improve translation efficiency, should not imply organizations must also start to use and manage MT systems themselves. A good way to deal with this is to offer a package deal. When an organization delivers its documentation in controlled language, specialized translation agencies can provide high quality translation with an improved efficiency (through the use of MT tools). The translation agency manages the MT system and has the capability to verify the quality of the output.

User support and interaction in authoring tools

When a company decides to implement a controlled language, a number of methods and/or tools can be used to help authors in their activities. We already mentioned training of authors, because they must learn the 'new' language.

The most elementary support is to have all related documentation (the controlled language specification, word lists, guide lines, do's and don't's) on paper available on the desktop.

The next stage would be to make this information available electronically.³ Depending on the platform, this can be done in an on-line help, hyper-linked style. A major advantage here is that it is easier to manage updates to the specification. This type of support can exist independently of the actual authoring environment (word processor, DTP-package).

More advanced solutions can be provided, but these will require integration with the authoring environment. The most common application is a spell checker, that is tuned to the CL specifications. Obviously, this will only provide support on part of the specifications, namely the lexical requirements (most likely it will not be able to do this task properly when lexical specifications are context sensitive, c.f. use 'display' only as a noun, otherwise use 'show')

³ One issue to be taken into account when starting to make things electronic, is distribution of resources. If an organization does not yet have a means of sharing resources (no network, people working away from the office), it generally is more difficult to make sure everyone is using the same (version of the) resources.

At Cap Volmac, we develop tools that can be best typified as grammar checkers. These are integrated in the authoring environment, and are able to give the author support for the lexical specifications, the grammatical specifications, and style. The author can use the functions of this tools in the document itself to check whether the text is in accordance with the specification. The author will get help and suggestions, if the text contains non-approved elements).

For this type of tool, it becomes more important to manage the resources properly. Again, in operational terms this often concerns primarily the lexicon, as this is subject to change, where the grammar typically is not. Authors should be able to continue there work, even if certain terms may be yet unknown (e.g. a new product name). But, when language control is used in an MT environment, any updates to the source language must first be verified against the possibilities of the MT system⁴.

Future developments

When we have augmented the author's work place with tools as described above, it is tempting to take a look at what kind of information is available electronically and might be used to improve or add functions to support the authors. Below are some examples we might want to consider.

Apart from text, documents contain more information, such as layout and/or structure. This structure information can be used to improve grammar checking. For example, headers generally do not have determiners, whereas you might prefer determiners (for disambiguation, as we have seen above) in running text. You can then trigger certain rules, depending on type position in the document (the type of text).

We might also consider to make the authoring environment more 'interactive'. Consider we want to use a term that is ambiguous, yet we cannot rewrite it because it is accepted and preferred by the audience. We could ask the author to indicate which interpretation is meant, store the information (annotation) and use that again when the document is processed in an MT systems. Another possibility is to work with pronoun resolution --what does a pronoun refer to in a text-- this way. The problem of PP-attachment mentioned earlier could also be handled along similar lines. In simple terms, this would change the author's job from writing to storing knowledge.

Finally, we can use information available in other automated systems (CAD/CAM, MIS, DIS, EIS, STEP ..).In these systems, we might find information that could help the author verify more pragmatic aspects. Suppose we are able to distill and model information in or from these systems that tells us a bit about the world. We could then start checking on pragmatic correctness of documents. For example, an author may write: 'Attach module XYZ to unit ABC.'. If we are able to check, in the CAD/CAM specification, that indeed this module should be attached to that print (by

⁴ This is not a trivial matter. Consider for example an MT system that cannot handle category switching, such as:

Dutch:	Hij wandelt snel	Hij wandelt graag
	(He walks fast)	(He walks 'likingly')
English:	He walks fast	He likes to walk

Although we want to be able to use constructions on the left (in Dutch), we want to block some adverbs (such as 'graag'). This means we always have to make sure that we can translate what we allow.

design), we can signal this to the author, or issue a warning when it does not match the design.

Conclusion

Controlled language is a useful instrument in optimizing the efficiency of the translation process, especially if it uses MT systems. For the specification of a controlled language, we must look very closely at the requirements of the MT system, and we must make sure the specification is complete in the sense that it must be consistent and must properly describe the approved language. The authoring environment can be upgraded with additional tools that help the author in writing in controlled languages. The resources used by these tools will introduce new resource management activities, mainly in the area of terminology or the general lexicon.

Organizations are not anxious to take on the whole issue of translating their documents. They prefer to have this done by subcontractors. Therefore we must look for solutions where translation bureaus provide translation services that are in key with the use of controlled language by their customers.

The adoption of controlled language will make this relationship very smooth since its major objective is to eliminate interpretation problems.

Machine Translation, Translation Memories and the Phrasal Lexicon: The Localisation Perspective

Reinhard Schäler

Abstract

Software localisation is the ideal application area for translation automation because of *what* is being translated, *how* it is being translated and *who* is involved in the translation. But existing translation tools and, more specifically, machine translation and translation memory technology cannot effectively deal with *all* the translation requirements of the localisation industry. This article examines these needs by highlighting characteristics of software localisation which have a direct impact on its translation requirements. It shows how and why localisation companies have used machine translation (MT) and translation memory (TM) based technology to address issues arising out of time and budget constraints. The article concludes with an assessment of the inherent limitations of both the MT and the TM approach and reports on research which could eventually lead to an enhancement of the currently available technology through the use of a Phrasal Lexicon.

Reinhard Schäler

Reinhard Schäler has been working with Irish and overseas companies in the area of software localisation since 1986. Together with Irish universities and industrial partners, he has organised several international conferences and workshops. He has given lectures on the subject at Irish universities, presented papers at international conferences on Natural Language Processing and has published a number of articles on NLP and software localisation.

He has been a researcher in the Department of Computer Science, University College Dublin, for a number of years and is the manager of the Localisation Resources Centre (UCD). He is a founder member of the Software Localisation Interest Group (SLIG) and its current chairperson.

The Localisation Resources Centre

Ireland is the number one location in Europe, and probably world-wide, for the localisation of software and its documentation. To further develop this important industry, the Localisation Resources Centre was established in December 1995 at UCD with support from the Irish Government and the European Regional Development Fund (ERDF). The Centre offers the following services to the localisation industry: Localisation Tools Library; Evaluation of Localisation Tools; Education and Training; Research and Development of Localisation Tools; Consultancy; Industry Watch; Regular Publications.

Mr Reinhard Schäler
Localisation Resources Centre
Roebuck Castle, UCD, Belfield
Dublin 4, Ireland

Tel: +353-1-2830644, Fax: +353-1-2830669

E-mail: Reinhard.Schaler@ucd.ie

Url: <http://lrc.ucd.ie>

1. Software Localisation

Approximately ten years ago, large North American software publishers, realising that international markets offered enormous potential for growth, decided to establish their first manufacturing sites in Europe. They also set up development teams to adapt the original US-English product to the requirements of European users. These European manufacturing and development sites were soon supported by a growing service industry offering, initially, translation and DTP services and, later on, complete 'turn-key' solutions. The whole of this new industry has become known as the Localisation Industry.

The process known as *software localisation* covers a wide range of activities, including:

- adapting and re-engineering of software according to the requirements of European and world markets;
- software testing (QA) of localised products;
- translating of documentation;
- "porting" of multimedia products into other languages and cultures, involving, among others, actors for voice-overs and graphic artists for the adaptation of the visual contents;
- printing of documentation;
- duplication of diskettes and CD-ROMs.

Despite its image as a high-tech industry, a great number of tasks in the localisation process are still carried out manually and are, as a result, very labour intensive and costly.

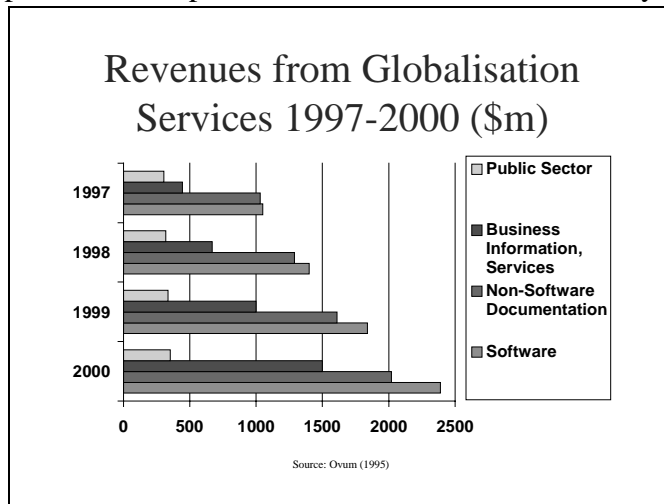
What makes software localisation different from other (even related) industries, e.g. in the translation or the I.T. sector, are its following characteristics:

1. It is an industry with phenomenal growth rates which will continue to expand into the next millennium.
2. Product life cycles are extremely short.
3. The source to be localised is never 'stable' (but always coming from the same restricted technical domain).
4. Localised products are supposed to become available at the same time as the base product.

1.1 A booming industry

The importance of a global perspective is no longer a disputed issue among the big North American software publishers. While the US internal market is stagnant and extremely competitive, Western Europe and especially the new markets in Eastern Europe and Asia are still 'underdeveloped'. Many publishers, therefore, are now investing considerable resources into their international business - an investment that is paying off: some of them already achieve more than half of their revenues from international markets. In this context, product internationalisation and localisation become more important than ever for company survival and growth.

In 1995, well known and respected British based marketing experts OVUM published a report on Globalisation in which they attempted to track the market for



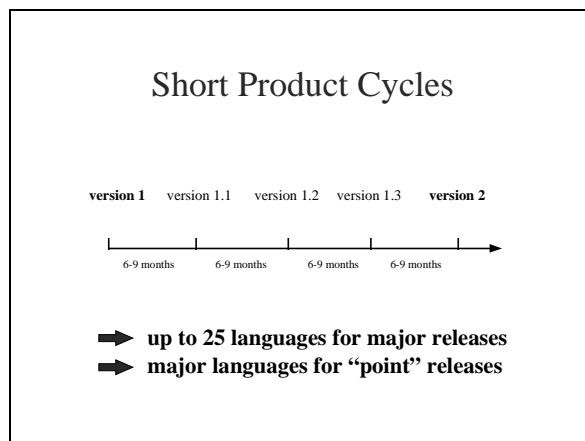
globalisation services and provide a forecast up to the year 2000. Among the four sectors surveyed were:

- Public Sector
- Business Information and Services
- Non-Software Documentation
- Software

According to their forecast, software localisation is one of the fastest growing industries world-wide and the most important sector within the globalisation-services market. The software sector will, at least over the coming 3 years, outperform all other sectors by more than doubling revenues between 1997 and 2000. In 1996, the world-wide translation products market was estimated to be worth around \$300 million and growing by nearly 50% per year, reaching \$1.5 billion in 2000.¹

1.2 Short product cycles

Product life cycles have become shorter in recent years, with some software publishers only leaving 6-9 months between 'point releases' of their products. Not all releases of all products are always being localised for all languages. While big publishers, among them Microsoft, Lotus and Corel, localise major releases of the



original product for approximately 20 languages, minor 'point releases' will only be published for major markets like Germany, France or Italy and, increasingly, some Asian markets including Japan.

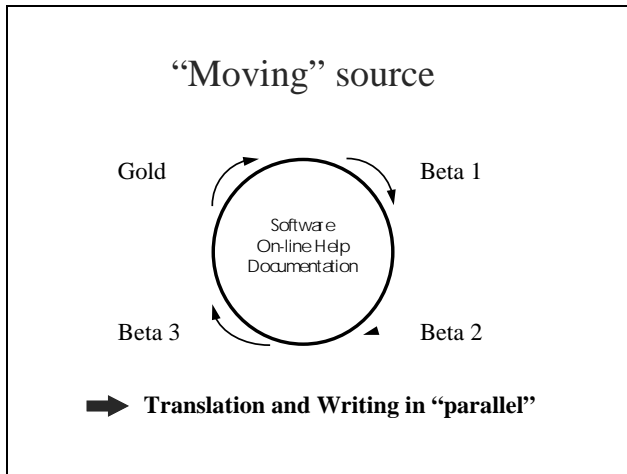
Obviously, the publishers would prefer to service all markets with localised versions of their current

¹ Ireland is now the number one location in Europe, and probably world-wide, for the localisation of software and its documentation. Currently, 40-50% of the PC based software sold in Europe originates in Ireland. It is expected that this figure will rise to 60% over the coming years. In Ireland, the localisation industry accounts for around 2 billion pounds worth in exports. There are approximately 4,000 people employed directly in the translation, engineering and manufacturing of localised products. It is estimated that for every person directly employed two others are employed in dependent industries. (cf. Murphy 1994)

products - not least because users constantly deprived of up-to-date software from one publisher will soon turn to other publishers with a more progressive update policy. The high costs currently involved in updating products, however, make this impossible for less significant markets. Software publishers are, therefore, actively searching for methods and tools which would allow them to update their current localised products with each update of the original product at minimum expense.

1.3 "Moving source"

Only in exceptional circumstances can localisers work on a stable, final version of a system (including software, on-line help and printed documentation). Localisation



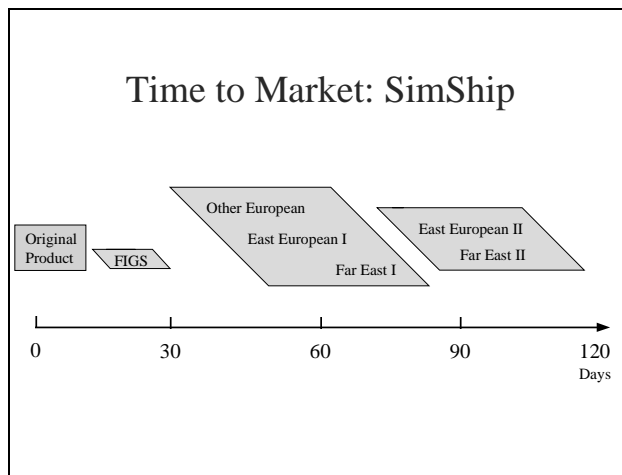
teams are involved from the very beginning in a development project by providing feed-back on the implementation of international features into the base product and by identifying those areas which will need to be localised, i.e. text, templates, defaults etc.

The first translation pass is usually performed on late alpha or very early

beta builds of products, at a stage when implementation problems regarding international features can still be addressed by the development team. Further builds are analysed as they come in and necessary modifications are implemented with each new beta build. This way the first localised versions of a base product are usually only very few days or weeks behind.

1.4 SimShip strategy

Simultaneous shipment ('SimShip') of the original and the localised version of an application became an issue in the early 1990s and was probably achieved for the first time with the German and French versions of 1-2-3 Release 2 by Lotus Development Ireland. The very costly exercise of simultaneous shipment was initially justified by the assumption that the company would lose revenue and valuable market share with every day that a product - although available in English - could not be sold in its localised version.



In 1996, 'SimShip' does not necessarily mean anymore that the localised product has to be shipped on exactly the same day as the original US version. Over the past few years, most companies have found that the enormous costs associated with the 'SimShip' strategy could not be justified. While they still aim for short

'deltas', i.e. the period between shipment of the original and the localised versions, 'SimShip' has now been redefined as shipping within the same financial quarter.

2. Translation and Software Localisation

The translation of the original software products, often into some 20 languages, is one of the key tasks in the localisation process (and still the single most expensive one). There are basically two types of 'texts' that have to be translated: the *software* itself (menu commands, dialogs, error messages etc.) and the *documentation* (on-line help and printed user manuals).

The translation of the *software* is technically quite challenging and presents translators with a number of issues that go far beyond those traditionally associated with translation. The nature of these issues depends largely on the design of the original software and varies according to the target language. Most software publishers have over the years developed their own proprietary in-house translation tools to help translators deal with these issues.²

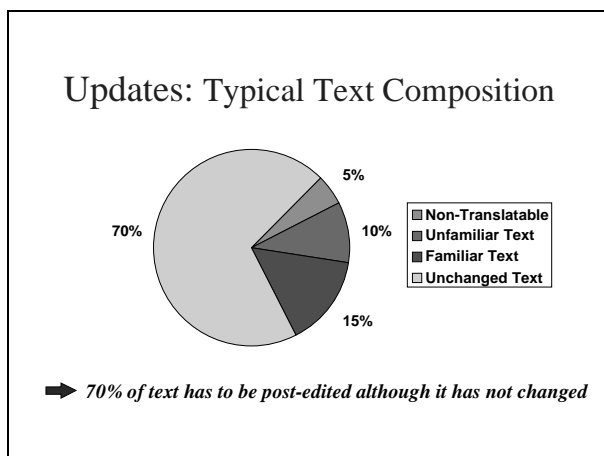
In terms of quantity, however, it is the *user documentation*, i.e. the on-line help system and the printed user manuals, which presents the biggest challenges. Projects of up to 1 million words cannot be managed anymore by individual translators. They require well managed and experienced teams able to guarantee a short turn around time while maintaining the required level of linguistic quality, the most important of which is consistency of style and terminology across:

- different areas of the documentation
- software and documentation
- different releases of the same product
- the same product family

The text to be translated is mainly composed of explanations and instructions outlining the features of an application and providing guidelines on how to use them,

² As Birch (1993) points out, software products should be designed from the outset with the multinational market in mind in order to achieve lower costs for localisation, faster time to market and higher customer satisfaction. He lists a number of potential issues associated with the translatability of user interfaces and provides examples for each of these to illustrate his points.

i.e. translators are generally dealing with a text of a consistent style and from a well defined and restricted domain. This requires, on the one hand, a certain period of familiarisation with the style and terminology - on the other hand it enables experienced translators who have become familiar with the product and the terminology to turn-around translations in sometimes amazing time frames.



As the software itself, the text to be translated very rarely is completely new. In fact, in very many cases it is updates of previous versions of the programme which are being translated and a high proportion of the text has already been translated and edited before. In a typical update situation up to 70% of the text has not changed at all, i.e. its original (or 'old')

translation can be 'copied' straight into the new target text.³ A further 15% is familiar text which has been modified but is not completely new, i.e. its translation has to be adapted to reflect the changes made in the source language.⁴ Only 10% of the text is completely new or 'unfamiliar' while 5% is text which for various reasons should not be translated.

It is not unusual, therefore, that translators are remain with a product over a number of years and translate all of its regular updates.⁵ Their translations almost become identified with the product - not unlike the identification of certain actor's voices with well-known American actors when dubbed.

Software localisation is one of the few industries which offers translators an acceptable level of income while allowing professionals to specialise exclusively in this area.⁶ This, however, requires translators to continuously update their knowledge and renew their equipment as they are, in effect, always working on tomorrow's software products. Professional translators in the localisation business see the need to keep up with modern technology not as a burden but as a welcome, interesting ingredient adding to the diversity of an otherwise quite dull and often routine job - far removed from the intellectual challenges and creative liberties aspired to during their college education.

All the above make translation in software localisation an ideal application for automatic translation: machine translation (MT), translation memory (TM) based systems and other computer assisted translation (CAT) systems such as on-line terminology databases etc. Translators, agencies and publishers - because of the very nature of the business they are in - are all open to experimenting with new

³ These are the *Exact Matches* in translation memory speak.

⁴ These are the *Fuzzy Matches* in translation memory speak.

⁵ All major software applications are being updated at regular intervals, often not exceeding 12 months.

⁶ In countries like Ireland the software localisation industry is in fact the only industry which can provide professional translators who wish to specialise in one area of expertise with an acceptable level of income.

technology, they are already equipped with the necessary hardware to run it and - last but not least - they want to be seen by the industry as operating on the leading edge of technology, as innovators in their field.

It is not surprising, therefore, that in the past the software localisation industry has probably been the biggest single user of translation technology and that its innovative, dynamic and technology driven approach has already contributed to the development and opening-up of markets for new language technology applications.

2.1 Machine Translation in Localisation

Issues related to increased competition, mounting pressure of bringing down the cost of translation (the single biggest cost in localisation), and the growing need to translate into more languages without a corresponding increase in the budget available, have prompted localisers since the early 90s to invest considerable amounts of money into either the implementation of MT systems or at least into long term evaluation and feasibility studies. By doing so, they have subscribed to the widely held view that MT had reached a level of maturity that would allow its successful operation in a commercial translation environment.

They were encouraged by reports from current and new users of MT technology within the localisation industry which had achieved significant savings in translation cost and a considerable increase in translation quality through the commercial use of available MT systems.

Success Stories

Daniel Grasmick from *SAP*, Europe's biggest indigenous software developer, reported at the MT Summit V in Luxembourg (July 1995) on his company's integration of MT into the translation process. He pointed out that at SAP

- MT is totally accepted by human translators;
- MT produces high quality translations at high speed;
- The operation has an excellent profitability rate;
- Its 13 employees (8 of whom are full-time) translated 3m words in 1994 and 1/2m words per month in 1995.

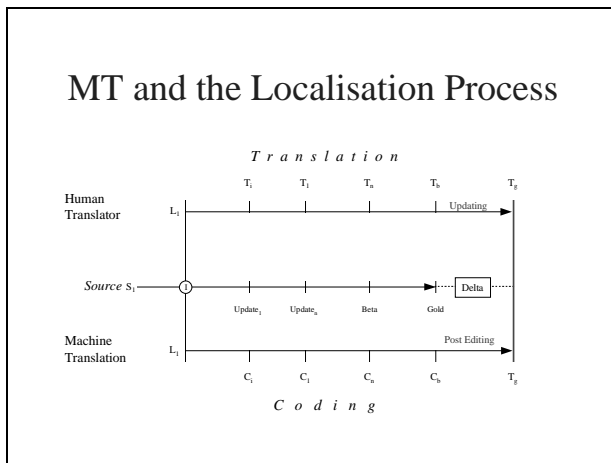
Gary Jaekel from *Ericsson Language Services* reported at the Localisation Industry Standards Association (LISA) Forum in Amsterdam (April 1995) that they were a user of MT since 1993. He pointed out that:

- By 1995, they had achieved a 50% productivity increase
- Experienced translators at Ericsson can produce 12 pages per day after working 3 months in production mode. Using MT on certain texts, the same work can now be done (on average) in an hour and a half.
- By 1995, they had achieved 100% increase in quality
 - permitted before (per 100 pages) : 16 minor, 1 major faults
 - achieved after (per 750 pages): no major, 4 minor faults

The problems which arose in connection with the introduction of MT at Ericsson were not, as originally expected, related to the quality of the translations but to the

(non-linguistic) limitations of the MT system and the format of documentation processed as well as the internal setup of the operation.

Among other localisation companies actively using or at present considering the use of MT technology are Berlitz, Corel, Gecap, Idoc, Lotus/IBM, and Oracle. To date, none of them have reported success stories similar to SAP and Ericsson. What they found was that if financial savings could, in fact, be achieved at all, they could probably only be expected after 2-3 years. Given the characteristics of the localisation process, an MT supported localisation project was also not likely to ship earlier than one run with human translators.



Their experience would indicate that, while *human* translators used the crucial period between the 'gold' date for the original source and the 'gold' date for the localised version to incorporate the final changes made to the last beta build in their translations, the translators working on the *MT* supported project spent that time post-editing the

machine translation output based on the source gold build.

Other limitations of the (exclusive) MT approach encountered in the context of localisation are:

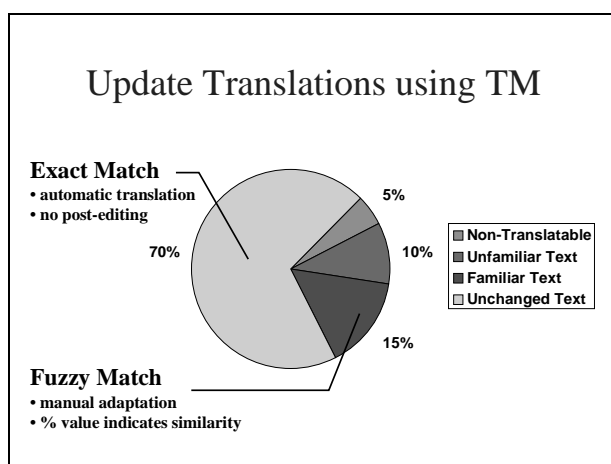
- *Translation Scenario*: Translations are often performed by freelance translators working from home in geographically diverse global locations without easy access to the central MT coding facilities.
- *Non-uniformity of material*: Many MT systems were not able to deal with the variety of text formats encountered in localisation. MT systems, for the most part, were also not able to re-use existing language resources (word lists etc.) and required an investment in development and maintenance that was difficult to justify.
- *Post-editing work is always lost*: While the performance of MT systems increases over time with the development of lexical and semantic dictionaries, output from current systems requires considerable post-editing to reach publishable standard. This necessary work is always lost - even if the text of the original has not changed. In other words, MT does not take advantage of the highly repetitive nature of the source text, it does not re-use the translation of text that has remained unchanged between the previous version and the current one.

2.2 Translation Memory Based Tools

Approximately 4 years ago, in 1992, the first translation memory (TM) based tools became commercially available. The basic idea behind these tools is simple: *Never translate or edit the same translation unit twice*. Translation memories are basically a

database of aligned sentence pairs, one being the translation of the other.⁷ From a technical point of view they are straight-forward, easy to use and maintain, and provide a high degree of ergonomics and portability. Financially, they do not require a substantial investment: the tools themselves are affordable with a list price of around 2,000 dollars⁸; they do not require special hardware equipment and require only limited training of translators and CAT administrators.

Translation Memory (TM) based tools do not require coding as MT systems do. TMs are created either during the translation process by translators as they enter the translation or by 'aligning' previously translated files using so called alignment utilities. When a newer version of a previously translated file has to be translated, the system automatically offers or even inserts the translation of all the sentences⁹ in the new file which were translated previously (and which have not changed), i.e. the system identifies 'exact matches'. If no exact matches can be found, the system attempts to find similar sentences in the TM to those in the new source file. If it finds similar sentences it displays them and the original source as so called 'fuzzy matches'. All the translator now has to do is to adapt the old translation to reflect the new source.¹⁰



In the context of localisation this means that in the typical update situation (be it between different beta builds or new releases of a product), around 70% of the text can be translated automatically using text segments previously translated and made available automatically through the translation memory.¹¹

Although TM systems do not provide a practical solution for all translation problems in localisation they are widely seen as a very pragmatic solution - the famous *80/20 solution* - potentially saving time and money, and providing a higher degree of consistency while they are, at the same time, easy to use, PC based, integrated in popular word processing environments and, above all, affordable.

⁷ Recent TM based systems allow the creation and use of multilingual translation memories.

⁸ Substantial discounts are offered to users purchasing a number of licences at a time.

⁹ Although the TM systems usually refer to 'segments' rather than 'sentences' to make clear that they do not just segment *text* into 'sentences' but also titles, headers etc., we will use the term 'sentence' for the purpose of this article.

¹⁰ With recent systems, the degree of similarity between the old and the new source segment which make it an admissible 'fuzzy match' can be defined by the translator.

¹¹ Text formatting information preserved as TM systems offer special filters for the most widely used word processors and desk top publishing systems.

However, there is one basic short-coming which results from the design of the TM based systems and the approach chosen by their developers: The basic translation unit in TM based systems are sentences ('segments'). Phrases which have been translated before are not recognized as exact matches but, and even then only in a 'best-case' scenario, as fuzzy matches:

Consider the following example:

TM Entry I

[ENG] The bullets move to the new paragraph.
[GER] Die Blickfangpunkte rücken in den neuen Abschnitt.

TM Entry II

[ENG] The title moves to the center of the slide.
[GER] Der Titel rückt in die Mitte des Dias.

New sentence

The bullets move to the center of the slide.

Although the two phrases, *The bullets move* and *to the center of the slide*, had already been translated before TM systems cannot combine phrases from different entries in the translation memory to form a new phrase, nor can they identify the 'best match', in this case probably *The title moves to the center of the slide*, and substitute the changed segment, *The title moves*, with the correct new phrase, *The bullets move*, to produce the required translation:

Die Blickfangpunkte rücken in die Mitte des Dias.

To solve this problem, translation memory systems need to be 'linguistically enhanced' using techniques which are based on those developed in the context of Example Based Machine Translation (EBMT).

3. Extending MT and CAT: The Phrasal Lexicon

In a basic implementation of Example Based Machine Translation (EBMT) a bi- (or multi-) lingual corpus of aligned source and target segments (not necessarily sentences) provides examples which during the translation are matched against the new source (analysis). This is done using sophisticated algorithms which measure the distance or similarity between the examples and the new source. In a second step the examples identified as appropriate are combined (transfer) and the new target is generated (synthesis). One of the best known representatives of the EBMT approach, Nagao, defined the concept in 1984 as that of implementing the human learner's technique of using examples as a guide to translation.

Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases (...),

then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence.

Almost 10 years before Nagao, Joseph Becker proposed the concept of the *Phrasal Lexicon* at the 1975 TINLAP conference.

I suspect that we speak mostly by stitching together swatches of text that we have heard before; productive processes have the secondary role of adapting the old phrases to the new situation.

The basic assumptions of the *Phrasal Lexicon* are:

- The use of language is based at least as much on *memorisation* as on any impromptu problem solving.
- The process of language production is *compositional*.
- The phrasal lexicon provides the *patterns* that can provide major elements of “new” expressions.

There are, as far as we could establish, few implementations of this concept. Sato and Nagao developed an experimental system and the Dutch company DLT developed what became known as the Bilingual Knowledge Bank (BKB) which is an aligned corpus of equivalent texts in two languages, structurally analysed by the same type of parser into translation units. There is also a research project at the Localisation Resources Centre (University College Dublin) and the University of Manchester Institute of Science and Technology (UMIST) which is exploring the application of modern computational linguistic techniques to this approach.¹² This project has developed a basic research prototype based on Becker’s idea about the phrasal lexicon. The aim of this project is to establish whether the approach could, if applied to the restricted domain of software localisation, overcome the inherent limitations of translation memories as outlined above.

The experiments with the phrasal lexicon carried out at the Localisation Resources Centre (University College Dublin) and the University of Manchester Institute of Science and Technology (UMIST) cover:

- The production of an aligned, bilingual *phrasal lexicon* (with linguistic analysis)
- The use of *one parser/grammar* for English and German to analyse source/target
- The “translation” of new source text by *combining/substituting* known phrases
- The use of a *dictionary* as an index into Phrasal Lexicon

These experiments will test two conceptually different strategies of re-using previously translated phrases for new translations:

(i) *Tree combining* (identify matches for phrases in new source in phrasal lexicon; combining known phrases to reflect the new source). It is expected that this approach will be computationally ‘cheap’ but that it will require a large phrasal lexicon to produce satisfactory results.

(ii) *Tree substitution* (identify ‘best match’ for new source in phrasal lexicon; substituting modified phrase(s) in the new sentence with phrases from the phrasal lexicon). Our expectations here are that this approach will be computationally

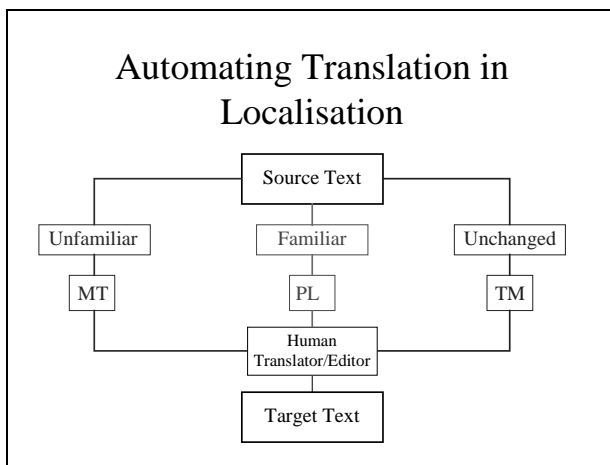
¹² This research was initiated by Prof. Alan Ramsay in the Department of Computer Science (UCD). Prof. Ramsay is now at UMIST.

'expensive' but will only require a smaller phrasal lexicon to produce satisfactory results.

The issues related to the development of the phrasal memory approach so far have been:

- Large development overheads (grammar, parser, lexicon)
- Large processing overheads (parsing)
- Large storage overheads (structural representation)

Although the initial experiments with the 'new' phrasal lexicon approach have not been concluded and, therefore, not yet produced results which could justify us to jump to premature conclusions, we feel that the phrasal lexicon could be the solution to the translation of text which with current TM based systems can - at best - only be identified as a fuzzy match.



The phrasal lexicon has the potential to offer translators all the advantages of the translation memory (speed, consistency, cost savings - no coding or post-editing as with MT systems) and more because it is not restricted to the matches found at sentence level but includes matches identified at the lower phrase level - a

strategy which will yield a higher percentage of exact matches as it is dealing with smaller translation units.

4. Conclusion

Translation in localisation is without a doubt one of the most suitable application areas for MT and CAT. Whether translation technology products will be used more extensively in the localisation industry depends on its economic viability, in other words: does the use of automatic translation systems bring down the cost of translation? Although there are some case studies available, some of which we briefly discussed, the question has yet to be answered.

However, there seems to be a consensus in the industry which suggests that, under average circumstances, stand-alone MT systems are not commercially viable. According to this view, only when combined with translation memory technology which overcomes some of its limitations can MT become useful. There are already attempts by some developers to offer this solution which has been received with great interest, although the translation and localisation agencies currently prefer to work with TM systems only, as they assess the implications and opportunities arising from a possible future integration of MT systems.

While definitely not yet offering an alternative to already available technology in a commercial translation environment, research into the phrasal memory approach will continue with the aim of assessing the potential of this technology to further enhance currently used MT and TM technology.

5. References

[Becker, 1975] Joseph D. Becker, Bolt Beranek and Newman. The Phrasal Lexicon. *Proceedings of Theoretical Issues in Natural Language Processing*, pages 70-73. Cambridge, Massachusetts (10-13 June 1975).

[Birch, 1993] Richard E. Birch. Making user interfaces translatable. *Proceedings of the First Irish Conference on Language Technology*. Dublin (12 May 1993.)

[Grasmick, 1995] Daniel Grasmick. 2 MT Systems and still hungry. *Proceedings MT Summit V*. Luxembourg (10-13 July 1995).

[Jaekel, 1995] Gary Jaekel. Machine Translation and being business-like. *Proceedings LISA Forum (Annual Meeting)*. Amsterdam (3-5 April 1995).

[Murphy 1994] Barry Murphy, National Software Director, Irish National Software Directorate, Forbairt (the Irish agency for science and industrial development), private communication included in: *Proposal for the Establishment of the Localisation Resources Centre*, Dublin 1994 (not published).

[Nagao, 1984], M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In: A. Elithorn and R. Banerji, editors, *Artificial and Human Intelligence*, pages 173-180. North Hollan, Amsterdam, 1984.

[Ovum 1995a] Ovum. *Globalisation - Creating New Markets with Translation Technology* (by Rose Lockwood, Jean Leston, Laurent Lachal). Ovum, London, 1995.

[Ovum 1995b] Ovum. *Translation Technology Products* (by Jane Mason and Adriane Rinsche). Ovum, London, 1995.

[Sadler, 1989] V. Sadler. (1989) *Working with analogical semantics: disambiguation techniques in DLT*. Distributed Language Translation, 5. Foris Pub., Dordrecht, 1989.

Session 2: Domain-Specific Lexica

Chair: Prof. Bente Maegaard

Introduction

Machine translation systems basically contain two types of linguistic knowledge: grammatical knowledge and lexical knowledge. In broad terms, the grammar contains the general rules for a language, and the lexica contain the particular or idiosyncratic rules for each word of the language. The lexica constitute an extremely important part of a translation system. They are often structured in a way that reflects the various domains, in order to keep different meanings of lexical items apart. In this session we will hear about domain-specific lexica which are to a large extent provided by the MT system producer, and about lexica which are almost exclusively built by the user.

Prof. Bente Maegaard

Bente Maegaard holds a M.Sc. in Mathematics and French from University of Copenhagen, 1970. She was employed at the University of Copenhagen, Department of Applied and Mathematical Linguistics, 1971-90, being a research professor 1984-89, and visiting professor at the University of Geneva (ISSCO) 1981. She is currently director of the Center for Sprogteknologi (Centre for Language Technology) since its creation 1991.

Her main areas of expertise are machine translation, evaluation methodology, dictionaries, corpora. She has held and holds positions as officer in scientific and other associations, member of editorial boards, reviewer for journals and conferences.

Center for Sprogteknologi

The Center for Sprogteknologi, CST, is a research centre under the Danish Ministry of Research and Information Technology. The centre was established in 1991 with the purpose of promoting research and development in computational linguistics and language technology. CST has around 20 employees covering the subject range of computational and theoretical linguistics, lexicography, Danish and a number of other languages, computer science and artificial intelligence.

The centre participates in European and national research programs, and performs commercial development and consultancy under contract with Danish as well as foreign companies.

Prof. Bente Maegaard
Director, Center for Sprogteknologi, CST1
Njalsgade 80
DK-2300, Copenhagen S, Denmark
Tel: +45 35329074, Fax: +45 35329089
E-mail: bente@cst.ku.dk

Users' Experiences with a Set of Domain-Specific Dictionaries for the Stylus Machine Translation System

Svetlana Sokolova

Abstract

STYLUS is currently the most widely used machine translation software for the Russian language. There are more than 15,000 registered users of the STYLUS system to date. One of the most important features of STYLUS is that it allows the user to customise the system through the management and editing of the dictionaries. The linguistic database of the STYLUS system contains three different dictionary categories: general-purpose dictionaries, domain-specific dictionaries, and user-defined dictionaries. In the current version of STYLUS, the English-Russian and Russian-English systems contain a set of domain-specific dictionaries that are arranged into groups. This grouping of several different dictionaries gives a better overview of all dictionaries. Each of the domain-specific dictionaries can also be used individually.

The Technical group provides a good illustration. This group is made up of eleven domain-specific dictionaries entitled 'Telecommunications', 'Software', 'Automotive', 'Mining', 'Building and Construction', 'Military', 'Oil & Gas', 'Electrotechnical', 'Home Appliances' and 'Navy'. This article discusses the data we obtained from a questionnaire sent to registered users.

Dr Svetlana Sokolova, Ph.D.

Graduated from Leningrad State University in mathematics. Obtained a Ph.D. in computer science. Has been involved in several machine translation projects for state organisations as head of a software design group. In 1991, established the company "PROject MT" Ltd. of which she is currently the president.

AO PROject MT Ltd.

The PROject MT team consists of mathematicians, programmers and linguists. The company staff now includes 52 employees. The leading specialists of the company have extensive experience in designing machine translation systems. Several versions of STYLUS have been launched on the market since the company was set up five years ago. The first version, called PROMT- PROgrammers Machine Translation, could translate software documentation from English into Russian. The current version of STYLUS can translate from French into Russian, and from Russian into English, German and French. STYLUS runs on Macintoshes and on PCs under DOS, Windows and Windows 95.

Dr Svetlana Sokolova
AO PROject MT Ltd
PO Box 632
St Petersburg, 199053 Russia

Tel: +7 812 275 78 87, Fax: +7 812 2757893
E-mail: svetlana@prompt.spb.su

What is STYLUS?

By way of introduction I should like to explain what STYLUS is and how it works. STYLUS is a machine translation system that can translate from Russian, and into Russian, for a number of European languages (English, German and French). It is a commercial product designed for PCs and runs under DOS, Windows, and Windows 95. There is also a Macintosh version. STYLUS is currently the most widely used software in machine translation for the Russian language. We have more than 15,000 registered users of the program. They are individuals, small and big companies, universities, and state and government organisations. Our users include the Administration of the President of Russia, the Russian Space Agency, NASA, the Central Intelligence Agency, AT&T, the BBC World Service, Inmarsat, Siemens, Lockheed Corp., Chevron, NPO "ENERGIA", the US Air Force, the FBI, Volvo, and Ernst & Young.

The popularity of STYLUS can be ascribed to its user-friendly linguistic interface designed for end users, the high speed of the translation process and, indubitably, the intelligibility of the output. One of the most important features of STYLUS is the fact that the system can be customised to users' own needs by managing and updating the dictionaries.

Dictionary structure

The linguistic database of STYLUS contains three different types of dictionary: general-purpose dictionaries, domain-specific dictionaries and dictionaries created by users themselves.

As a rule, a general-purpose dictionary in one direction (English-Russian, for example) contains about 60,000 entries and is consulted as the system's basic dictionary during the translation process. It contains entries for the most frequent words and phrases of the source language. These entries have a functionally sophisticated collection of semantic and syntactic tags that are used for the translation algorithms.

The domain-specific dictionary contains not only terms specific to the corresponding domain but also general-purpose words, if they have some specific function in that domain.

For the correct translation of software documentation, for example, words such as "program" or "application" need specific information. The volume of a domain-specific dictionary varies from 10,000 to 40,000 stems.

A user dictionary is created by the user himself. There is no limitation in the system on the number of user dictionaries called up during the translation process. Normally, if a user translates texts in different domains, he organises his own terminology into different user dictionaries, and then selects the appropriate user dictionaries along with the general-purpose dictionary, and, where necessary, domain-specific dictionaries.

A user can add new words or phrases and change the meanings of existing entries in both the general-purpose and the domain-specific dictionaries. However, all these changes will be saved in the user dictionaries, because neither the general-purpose nor the domain-specific dictionaries can be accessed by users. They can contain

information which is hidden from the user, and to prevent basic linguistic algorithms from employing this hidden information, the system saves all user-customised information in the user dictionaries.

Dictionary Manager and Entry Editor

The various dictionaries are handled by a very flexible feature in the STYLUS interface - the Dictionary Manager. This makes it possible to connect or disconnect the domain-specific and user dictionaries when translating a text and to change priorities for dictionary interrogation. New user dictionaries can be created and user or domain-specific dictionaries can be opened for consultation (see Fig. 2).

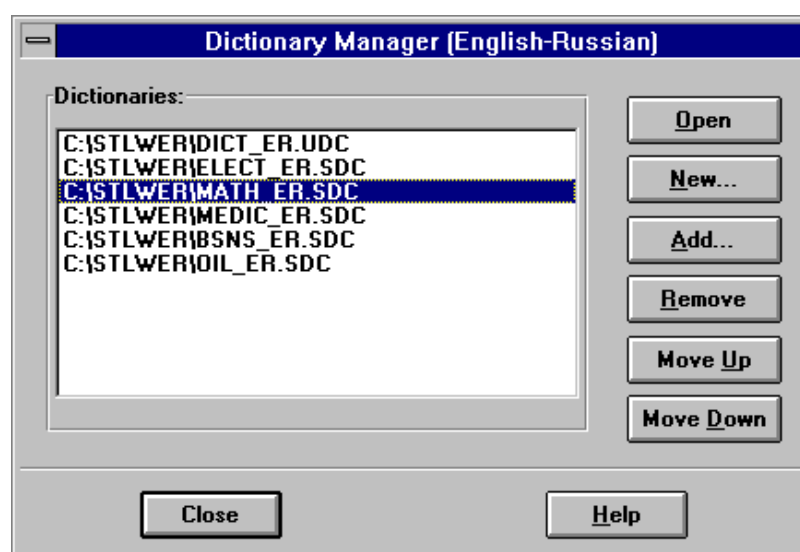


Fig.1 Dictionary Manager

When a dictionary is open, the user can copy an entry into his own dictionary by the "drag and drop" method, change the translations and grammar tags of the entry at his own discretion, and enter new words and phrases. It is possible to work simultaneously with several dictionaries.

These features offer a unique opportunity to customise the system to an individual topic and to the specific text being processed.

Another important item of the dictionary interface is the Entry Editor.

Each STYLUS dictionary is a bilingual dictionary, and all the dictionary entries have an identical structure. An entry includes the word stem, the grammatical description of this stem, and its equivalent translation in the target language. A machine translation dictionary contains highly specific information not included in a conventional dictionary. It is not easy for users to manipulate these dictionaries without a thorough knowledge of the linguistic methods employed. Hence, the STYLUS system includes the Entry Editor, which provides user-friendly access to the entry. This Editor is a kind of expert system that submits linguistic information to

a questionnaire, automatically forms a declension, assigns a set of patterns to input entries, etc. (see Fig. 2 and Fig. 3).

Fig.2 Dictionary Entry Questionnaire

The Entry Editor has two operating modes: Beginner and Expert. If the user chooses Beginner mode, his interaction with the system will be minimal. In Expert mode he can actively intervene in the updating process. For example, he can introduce “Government” by himself, change semantic information or correct automatically produced word forms. In this case the user should be familiar with both the source and the target language grammar. Automatic declension is a very important feature of the system, because STYLUS employs full morphology description for all the languages processed: 800 morphological types for Russian, 230 morphological types for English, more than 300 morphological types for German and French. Thanks to this automatic declension, the stems to be entered can be defined very quickly, cutting down the amount of routine work involved in dictionary adaptation.

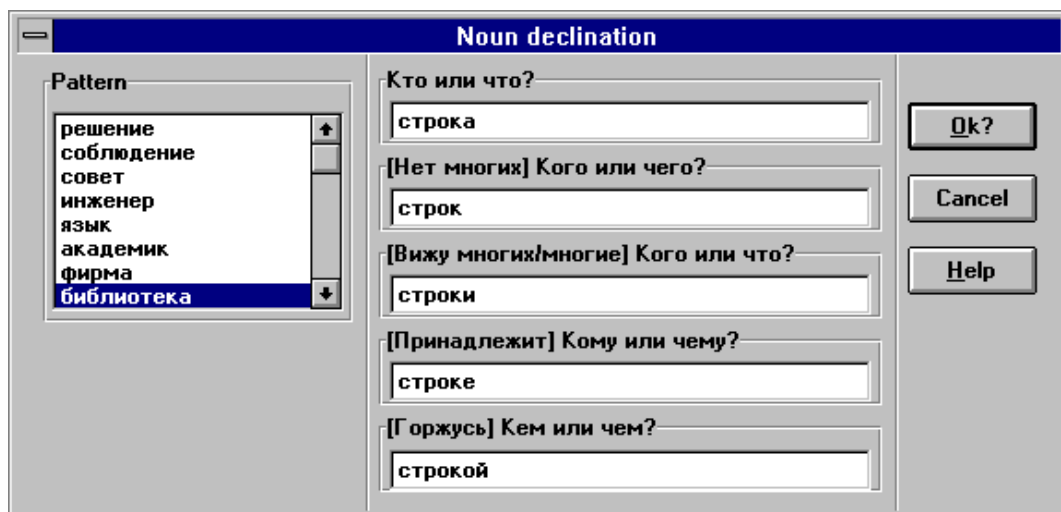


Fig.3 Automatically generated declension

Sets of Domain-specific Dictionaries

The current version of the STYLUS English-Russian and Russian-English systems has a large number of domain-specific dictionaries grouped together in sets. This grouping of several dictionaries in a set facilitates dictionary management. Obviously, each of the domain-specific dictionaries can also be used separately.

The dictionary sets cover Business, Polytechnic and Science. The Business dictionary set includes three domain-specific dictionaries: Business, Banking and Finance; Law; and Software.

The Science dictionary set includes five domain-specific dictionaries: Mathematics; Physics; Electrotechnics; Chemistry; and Computer Science.

The Polytechnic set is the most representative, comprising eight domain-specific dictionaries: Telecommunications; Software; Automotive; Mining; Building and Construction; Military; Oil & Gas Industry; and Electrotechnics.

There are also a number of individual dictionaries which are not included in any collection because of their very specific application: Military; Navy; Space Industry; Home appliances; and Medicine.

For the German-Russian and Russian-German systems there are only two domain-specific dictionaries: Business documentation and Software documentation. Dictionaries for Law and Medicine are under development.

For the French-Russian and Russian-French systems domain-specific dictionaries are under development.

How to use the set

A novice user will often start off by attaching as many domain-specific dictionaries as he can connect simultaneously. However, meaning is often lost by doing this, because the system interrogates the dictionaries according to the priorities that are fixed in the dictionary manager.

Take, for example, a text concerning a business agreement involving a software product for the management of a chemical process. Let us suppose that three dictionaries are connected during the translation session: the business dictionary (priority 1), the software dictionary (priority 2), and the chemical dictionary (priority 3). The general-purpose dictionary is employed in the background and has the lowest priority.

The program will always translate the word “file” as a business term, the word “exception” as a software term and the word “facility” as a chemical term. Hence, it is important to select priorities very carefully when connecting dictionaries to the translation process.

Sometimes, the only way to solve a dictionary conflict is to create a user dictionary and save the appropriate meanings for the conflicting words in that user dictionary, which is then given the highest priority.

User questionnaire

We have received very interesting feedback from our registered users on their experiences with the dictionary set in their translation work. This information is extracted from the questionnaire we sent to users. About 800 questionnaires were returned, which we consider to be quite a good result. The percentage figures below are in relation to the number of responses received.

According to our users' experiences, the best way to obtain a good translation is to use one of the domain-specific dictionaries combined with several user dictionaries.

Most users customise the system by creating their own user dictionaries:

- 29% of users create less than 5 entries a month
- 16% of users create 5-10 entries a month
- 30% of users create 10-20 entries a month
- 15% of users create 20-50 entries a month
- 9% of users create 50-100 entries a month
- 6% of users create more than 100 entries a month.

More than 52% of users consider the dictionary editing procedure to be simple enough and easy to use. About 17% of users find this work difficult, and the remainder did not answer this question.

More than 86% use the machine output as the basis for a final translation, 46% as an aid to understanding and classifying texts, and 6% use the system for other purposes, such as checking their texts in a foreign language by a process of reverse translation.

In response to the question “What type of text do you usually translate?”, the users answered as follows:

- about 70% translate technical documentation
- about 42% translate contracts and agreements
- about 65% translate business letters
- about 32% translate help information for software products

- about 44% translate scientific articles
- about 20% translate other types of documents.

The following question produced some interesting data:

In what way do you input your texts into the computer?

E-mail	16.5% of the users
Optical Character Recognition software	75% of the users
Typing	50% of the users
Files from diskettes	65% of the users.

The other questions were related to marketing issues, which are not considered in this paper.

Summary

The STYLUS translation system first appeared on the Russian market four years ago. There are now many users of STYLUS in a number of different countries. The experiences of users show that STYLUS suits both professional translators and those with a less specialized linguistic background. When separate domain-specific dictionaries are used in combination with user dictionaries to obtain a draft translation the results are very positive, making translation work more creative and productive.

Building Term Dictionaries for Machine Translation in Practice: A User Experience

Annelise Bech

Abstract

The paper starts with a brief outline of the machine translation set-up at Lingtech and the work-flow in the organisation. In-house the main task concerns preparing the patent texts for machine translation; the bulk of this work actually consists of identifying and coding technical terms and expressions. To this end we have two computational linguists, who are also in charge of creating and maintaining term databases for the various domains the texts concern. We intend to focus on our experiences and discuss problems and solutions from a practical point of view.

Annelise Bech

Ms Bech has been involved in language technology for more than ten years. She worked on the Eurotra project from 1986, partly as a member of the Danish language group and partly as a member of a designated specialist group with the task of designing and implementing the Eurotra formalism. Ms Bech also worked with knowledge representation and text understanding during her stay as research fellow at SRI International in California.

Since the creation of the Danish Centre for Language Technology, (CST), she has worked there as project manager of various language technological projects, most notably the development of the PaTrans machine translation system. Ms Bech has worked as a project manager of system development projects at Ramboll; and in November 1995 she took up an appointment as Director of Lingtech A/S.

Lingtech A/S

Lingtech A/S is a specialised translation company established by the patent agency Hofman-Bang & Boutard, Lehmann & Ree A/S. Lingtech has specialised in the translation of patent texts from English, German and French into Danish, and the company translates a total of about 8 million words per year. Lingtech has a core staff of 8 linguists and approximately 50 highly qualified technical experts working on a freelance basis. As a translation company, however, Lingtech is quite exceptional and a pioneer, in that since 1994 the PaTrans machine translation system has been used for English to Danish translation. PaTrans, which exploits Eurotra technology, was specifically designed and developed for Lingtech by CST. Ever since the system was applied in production, Lingtech has seen a steady increase in the number of words translated by the system every month and a level of reduction of more than 50 percent in the translation cost per word when compared to the costs for manual translation. Currently, it is expected that this year will see some 70 percent of the suitable English patent texts being machine translated. These results have led to the planning of further extensions of Lingtech's application of machine translation.

Annelise Bech, Lingtech A/S

Vesterbrogade 24, DK-1620 København V., Danmark

Tel. +45 33 25 71 71, Fax. +45 33 25 61 71

E-mail: lingtech@login.dknet.dk

(Outline of the Presentation)

- Lingtech and the MT scenario
 - PaTrans
 - Work-flow

- Term dictionaries
 - Definition
 - Organisation

- Building term dictionaries
 - Strategy
 - Problems

- Conclusion

Lingtech:

- Specialised in the translation of patent texts
 - English, German and French into Danish

- Bulk of English to Danish translation by MT
 - 70 - 75 % of texts (2 - 2.5 mio. words)
 - cost savings of more than 50 %

The PaTrans MT system

- purpose-designed system for Lingtech
- developed by CST exploiting Eurotra basis
 - translation kernel (Pok)
 - editor (PaEd)
 - term coding tool (PaTerm)

- used in production at Lingtech since 1994

The work-flow

- Registration and OCR scanning of text

- The pre-editing phase:
 - Format conversion (WP to PaEd)
 - Marking-up
 - Dictionary look-up
 - Alphabetised check-list
 - Simple and multi-word terms

- Term coding

- Batch translation of text
- Post-editing and language revision

Dictionaries

- The general dictionary
 - not maintained by Lingtech
- Term dictionaries
 - number and contents defined and maintained by Lingtech
 - the PaTerm coding tool; nouns, verbs and adjectives
- What is a term?
 1. A single or multi-word expression with a specific translation within a domain
 2. Any single or multi-word expression which we need to add to the lexicon
 - component*
 - abrasion resistant*
 - fox-shaped*

Organisation of term dictionaries

- Patent texts classified according to subject field codes
- The original idea:
 - many specialised dictionaries
 - very fine granularity
- In practice:
 - two 'main' term dictionaries: chem. and mech.
 - three 'supplementary' ones
- Why?
 - linguistically very hard to define
 - various system constraints
 - priority and interaction
 - nut*
 - locking_nut*
 - fox*
- ... 'then you only need to extend the dictionaries. No big deal!'

In practice it is rather a big deal, though

Why?

 - real life - size and coverage imperative
 - no off-the-shelf solution to be bought
 - a costly and time consuming task

- terminological homo- or heterogeneity
- the over-all cost-efficiency of MT
- the 'future value' of coded terms
- The applied strategy
 - on a text-by-text basis
 - identification and coding of new terms
 - critical selection of texts
 - hitting the right balance between coding and throughput of words
- Some reasons why it is a big deal
 - Coding a term is pretty fast, it's the finding which is hard
- The two main tasks:
 - identifying new source terms
 - target language translation
- The checklist and single word units
 - categorial ambiguity
space noun/verb
- The checklist and multi-word units
 - ball*
 - joint*
 -
 - ball_joint*

 - front_elevational_view*
 - side_elevational_schematic_view*

Interactive concordance facility; but one step

- Finding the translational equivalent
 - quality and consistency
 - expert knowledge; reference material
 - standardisation and corrections

A guiding principle

Work smarter, not harder!

Everyday pragmatics

- Keywords: structured extension of lexicon
 - maximise throughput
 - cost-efficiency

- prefer texts from well-covered fields
- prefer longer to shorter texts
 - from 2.5 to 1.2 - 3.5 to 1.3 percent [ratio *new terms/total words of text* for chem. and mech. respectively]
- for new subject-fields prefer short texts

Conclusion

- common format resources ?! In practice
- investment, pay-back, added value
- relative terminological exhaustivity
 - figures ?
- needs and market
 - linguistic strategies
 - quantitative measures
- need for invention and integration of tools
 - higher degree of automation
 - proactive/predictive tools
 - reactive/evaluative tools

Session 3: Experiences of a Large-scale User: The European Commission

Chair: Dimitri Theologitis

Introduction

The existence of two major translation tools, SYSTRAN and EURODICAUTOM, each with a 20-year history, makes for an interesting exercise in integration. Terminology and machine translation, until now conveniently classified as separate issues, are merging into an integrated linguistic resource. From term, to part of phrase, to sentence, today's technology makes it possible to reuse the same data in different systems.

Such is the success of translation memories and full text retrieval systems, that the usefulness of traditional machine translation has been challenged. A feasibility study will make clear what type of service is most appreciated by users throughout the Commission, and the Translation Service in particular.

Dimitri Theologitis

Born in Athens, civil engineer, specialised in integrated transportation systems. Opted for a major change of career in 1984 when he joined the Translation Service of the European Commission. Responsible for the "Rationalisation of Working Methods" from 1990. In 1994 became head of unit "Development of Multilingual Computer Aids", a multilingual team active in the technological modernisation of the Translation service of the European Commission.

The Translation Service of the European Commission

Situated in Brussels and Luxembourg, the SdT (for Service de traduction) houses today some 1 100 translators, 100 linguistic support staff, approximately 100 management staff and close to 500 secretaries and assistants. The total multilingual production of the Commission is some three million pages per year, in a combination of in-house, free-lance and machine production.

The SdT is currently undergoing major technological modernisation. The main translation aids being installed or overhauled include: the EURODICAUTOM termbase; the EURAMIS Linguistic Resources Database and search engines combined with translator's workbenches; SYSTRAN machine translation; and a document server with full-text search and retrieval possibilities.

Dimitri Theologitis

Translation Service, Development of Multilingual Computer Aids
CCE JMO B3/5, L-2920 Luxembourg

Tel: +352 4301 33632, fax: +352 4301 34069

E-mail: dimitrios.theologitis@sdt.cec.be

Generalised Language Resources: EURODICAUTOM, SYSTRAN and EURAMIS - a Case Study

Jean-Marie Leick

Abstract

One of the challenges of the EURAMIS project was to find data representation methods capable of covering the language resources of terminology databases, machine translation and translation memories. The importation of EURODICAUTOM terminology data into the SYSTRAN dictionaries gave an indication of the obstacles to be overcome, but it was the exportation of SYSTRAN dictionaries into EURAMIS which revealed the major difficulties.

The main problems arose, in the case of EURODICAUTOM, from the lack of strict encoding and, in the case of SYSTRAN, from the algorithmic representation of morphology. Translation memory entries are less problematic provided that logical representation and organic implementation are kept separate.

The advantage of common representation is that the various applications access data from different sources using the same mechanisms, thus creating unexpected synergies. The SGML-based approach for logical data description facilitates exchange in TIF-alike format.

Jean-Marie Leick

Graduated in electronic engineering, but moved onto informatics engineering in the seventies. Currently in charge of the EC-SYSTRAN machine translation project within Directorate General XIII of the European Commission. Since 1992, head of the former MLAP (Multilingual Action Plan) of the European Commission which ended in 1995. While EC-SYSTRAN development was the main project of the MLAP, the need to combine different tools (translation memories, term banks, machine translation and others) led to the design of the EURAMIS project, which has now been taken over by the Commission's Translation Service. Work is currently in a transitional phase as a new European Union Action Programme, MLIS (Multilingual Information Society) is put before the European Council.

DG XIII of the European Commission

DG XIII deals with: "Telecommunications, Information Market and Exploitation of Research". Directorate E "Information Market, Information Industry and Language Processing" is based in Luxembourg. Unit E5 (Language Processing and Applications) is mainly running the Language Engineering Programme within Telematics in the 4th Framework Programme for Research and Technological Development of the European Union.

Jean-Marie Leick

Principal Administrator, European Commission - DGXIII/E5
EUFO 0/168, L-2920 Luxembourg

Tel: +352 430134525, Fax: +352 430132354

E-mail: jean-marie.leick@lux.dg13.cec.be

Background

At the end of 1992 I took charge of the Multilingual Action Plan (MLAP) at the European Commission, which aimed at lowering language barriers by means of information technology. As the proposals came flooding in for NLP-tools or translation aids, it struck me right from the beginning that every tool, whether it be spell-checker, grammar-checker, term bank or machine translation system, had its own dictionary. While the Commission already had powerful dictionaries, both for its own version of the SYSTRAN machine translation system and for the EURODICAUTOM term bank, these resources could not be used for other tools. Another invaluable, but not easily accessible, language resource was the very large volume of parallel texts in the official languages stored in the archives of the Commission's Translation Service and in the Union's databases (e.g. CELEX). Translators need to have all this data at their fingertips. To this end, a comprehensible toolset working on a common linguistic resources database was designed in the framework of the EURAMIS project (EUROpean Advanced Multilingual Information System).

This paper discusses the first steps in this direction, which were the integration of the EURODICAUTOM and SYSTRAN dictionaries and their representation within the EURAMIS database.

Achievements

The main achievements, in chronological order, are:

- Importation of EURODICAUTOM into SYSTRAN
- Design of the EURAMIS Linguistic Resources Database (LRD)
- Exportation of SYSTRAN into the LRD.

Importation of EURODICAUTOM into SYSTRAN.

The question of importing EURODICAUTOM terminology data into the SYSTRAN machine translation system had been raised at intervals since machine translation was introduced at the Commission in 1977 and the conclusion had always been - mission impossible!

The problem was the fundamental difference in approach between an MT dictionary and a terminology database. MT entries necessarily reflect the most general translation of a term in a large variety of contexts, while terminology entries offer the most precise translation possible in a very specific context. Nevertheless, the wealth of terminology in the EURODICAUTOM base was tempting, and a feasibility study was conducted with Spanish as a source language. The results were so encouraging that they sparked off a revolution.

While the authors of the feasibility study were optimistic, there were plenty of pessimists to produce counter arguments and examples to show that the results of the feasibility study were biased and at any rate would not apply to other languages.

The software specialists feared that performance would deteriorate with the addition of millions of new entries, and the resulting volume explosion on the disks. If panic reigned, this was understandable.

The Challenge and the Strategy adopted

The nature of Eurodicautom entries presented a real challenge:

- 600 000 multilingual entries
- No formalisation of entries — only guidelines
- No grammatical information (such as word class or gender)
- No morphological information.

In short, Eurodicautom entries are meant to be understood by a human reader rather than a machine.

On the positive side, the project offered:

- Corresponding entries in up to 9 languages
- Large SYSTRAN dictionaries for cross-checks and for producing language-dependent morphological tables for automatic recognition of word classes.

Hence, we had at our disposal considerable language resources which could be exploited automatically.

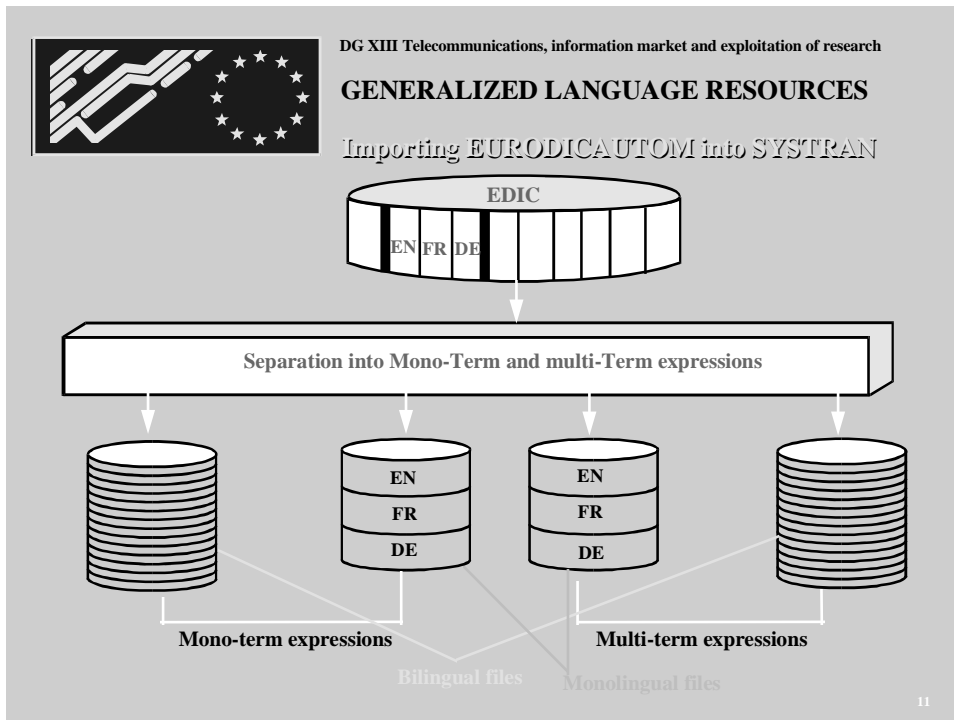
The following approach was adopted:

- Filtering of suitable entries
- Semi-automatic importation with plausibility checking at every stage
- Careful fine tuning of strategies.

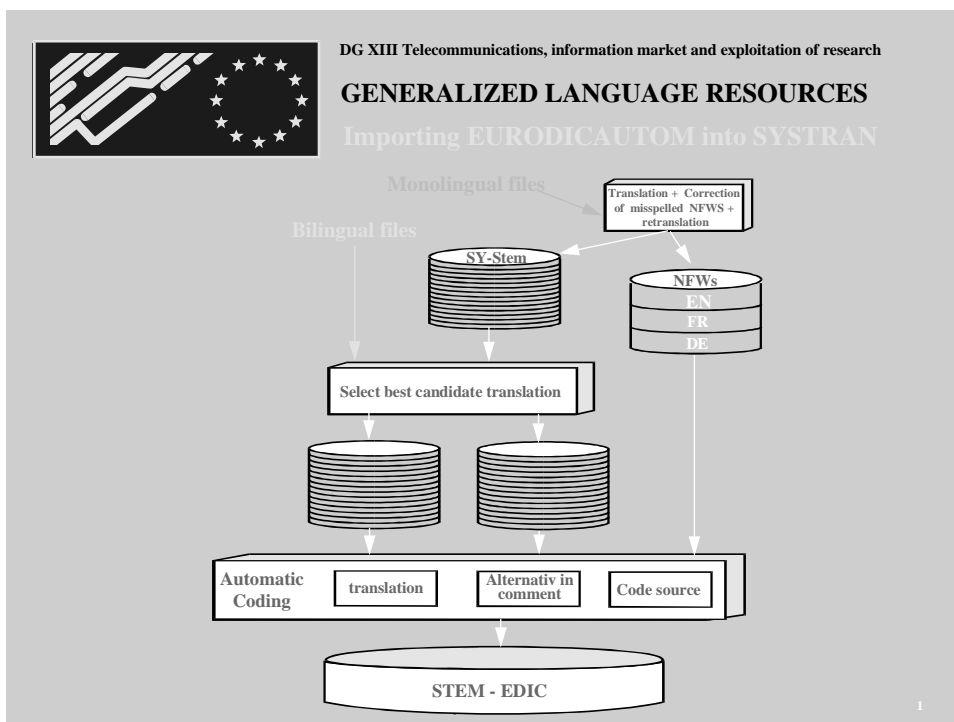
Filtering was done on noun groups. Verbs and adverbial groups were excluded.

Semi-automatic import was achieved by working out probabilities for corresponding entries in the various languages. Nouns could be identified in German (capitalised), verbs in English (preceded by "to"), other categories in Latin languages by sorting on endings, etc. The results were checked on sorted lists by the encoders.

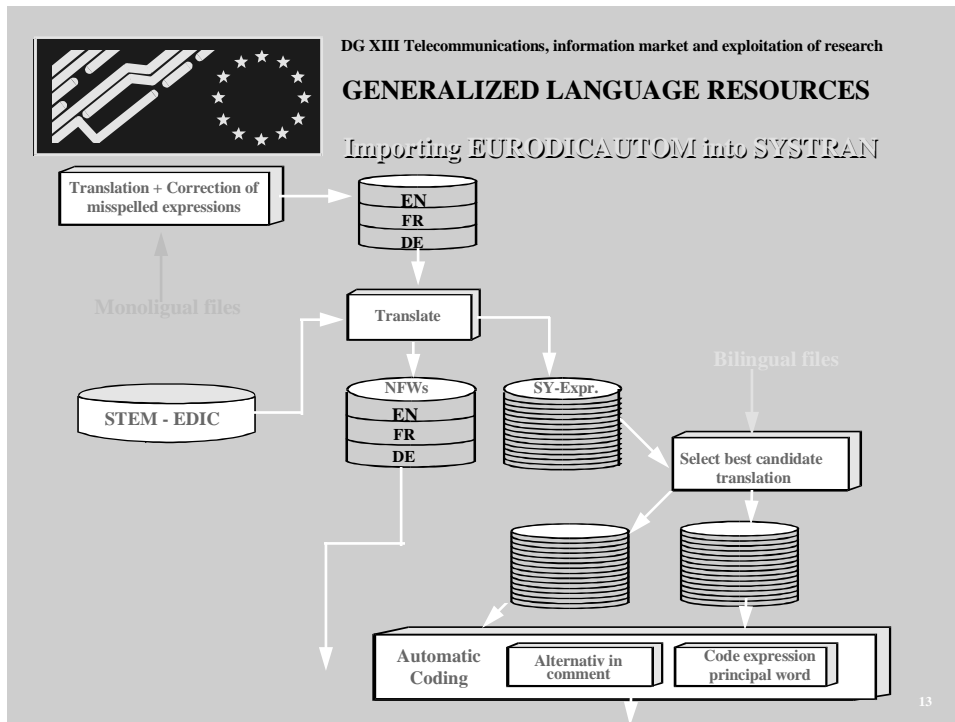
The following diagram shows the process in some detail.



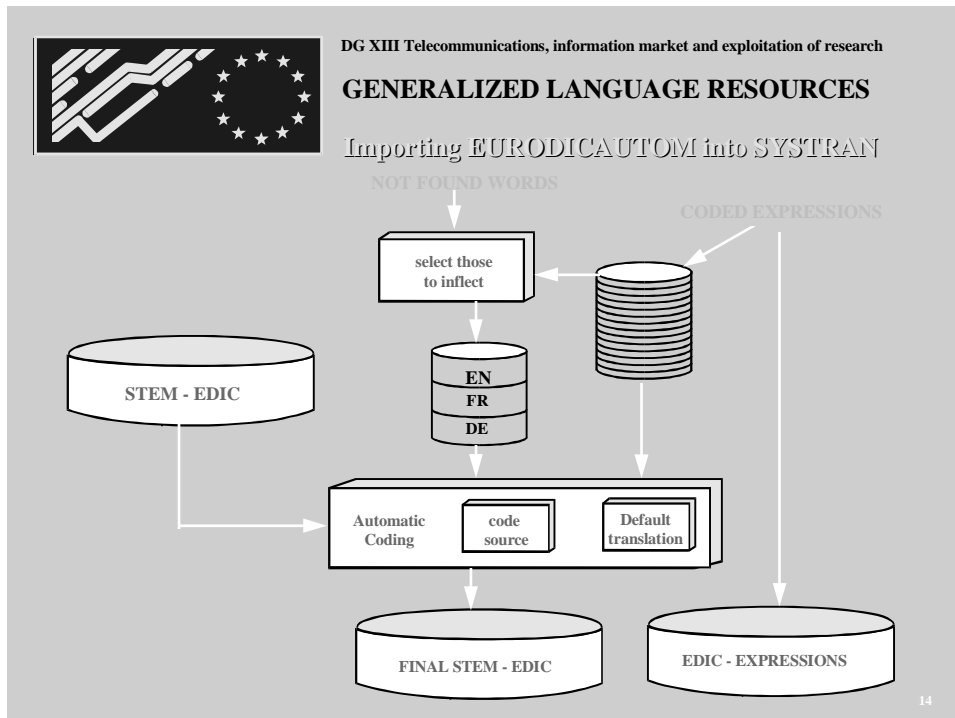
The general idea was to separate into monotermin and multitermin expressions and to deal first with monotermin.



The new monotermin were used to produce an MT dictionary, which in turn was used to translate the multitermin expressions. The selection of the best candidate translation via probabilistic methods then generated the new expressions dictionary.



The final EDIC-STEM dictionary was generated from information relating to the principal word within the expressions.



The Results

The final result? More than 3 million entries!

**GENERALIZED LANGUAGE RESOURCES**

Importing EURODICAUTOM into SYSTRAN

**Systran Dictionaries before and after
importing Eurodicautom data**

language pair	original entries	new entries
DE - EN	145.000	242.341
DE - FR	65.000	257.492
EN - DE	81.000	333.644
EN - EL	46.000	227.649
EN - ES	59.000	275.524
EN - FR	130.000	409.027
EN - IT	110.000	274.584
EN - NL	47.000	136.817
EN - PT	43.000	154.163
ES - EN	32.000	260.865
ES - FR	27.000	240.132
FR - DE	94.000	321.372
FR - EN	147.000	397.956
FR - ES	45.000	248.802
FR - IT	40.000	265.753
FR - NL	43.000	113.738
Total	1.154.000	4.159.859

15

The development of the whole procedure and the initial import took 180 man/months.

This exceeded initial estimates fivefold! (The original workload estimate was 34 man/months), but still resulted in a 600 entries per man day rate. The first two-yearly update only took 20 man-months for 750.000 added entries, giving a rate of 1.700 entries per man-day. It included the development of comparison functions allowing to identify and classify changes to existing entries.

Improvement of translation quality was:

- for technical documents: up to 60% or more
- for other documents: 2-10% or more, depending on the language pair and the type of document

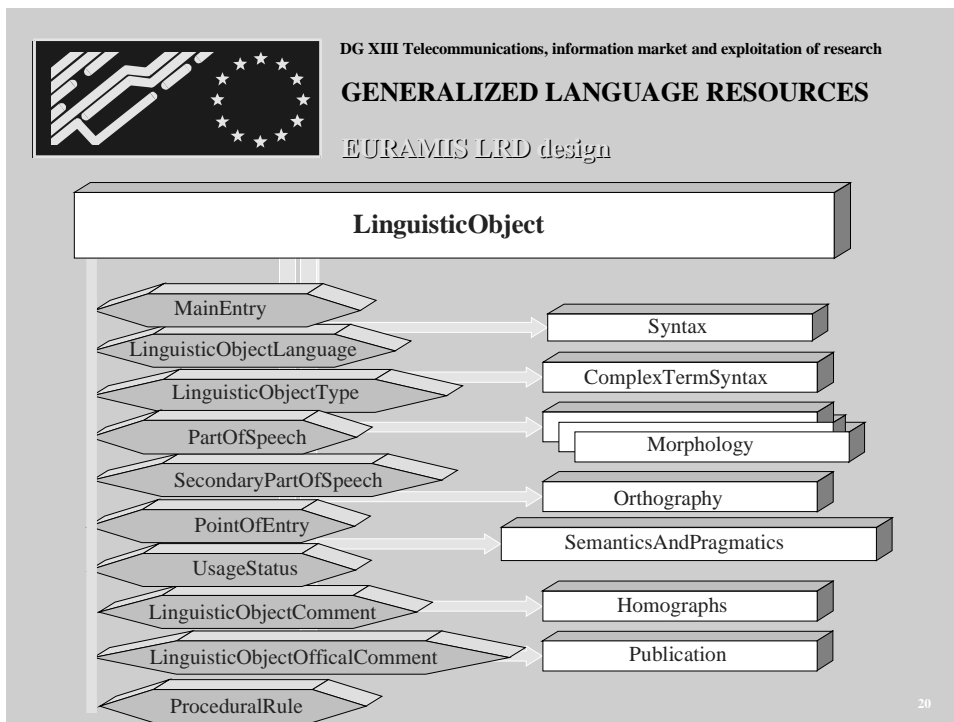
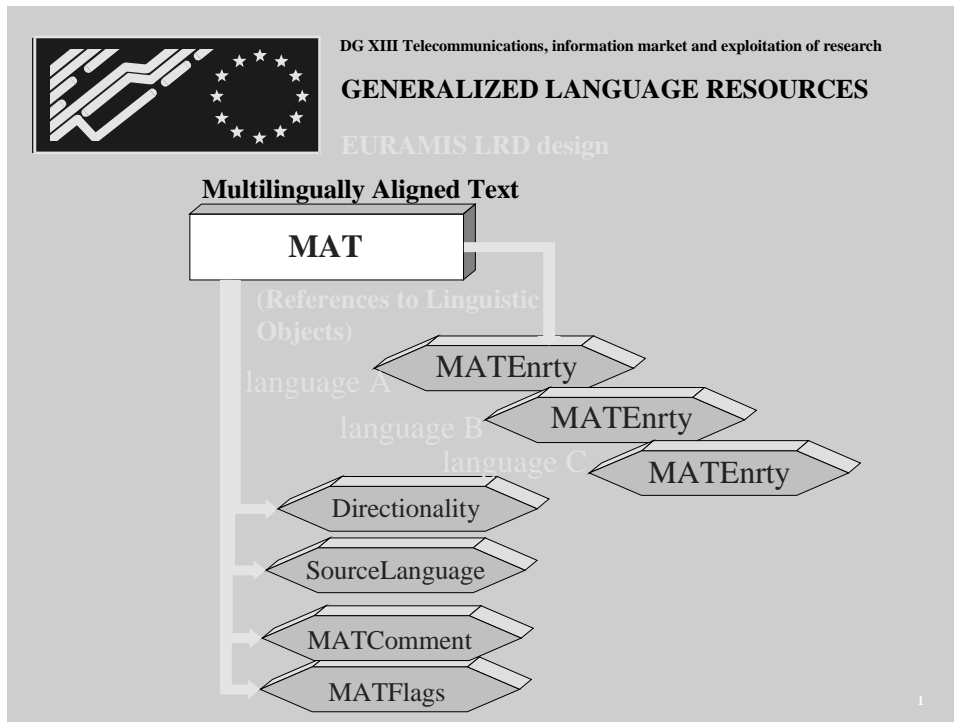
Quality enhancements are measured by translating a document twice, once with the EURODICAUTOM dictionaries and once without. An enhancement of 60% means that the number of sentences containing improvements minus the number of sentences containing deteriorations divided by the total number of sentences gives 0,6.

The fine tuning provided a better understanding of the Systran system and allowed to streamline coding strategies.

The EURAMIS Linguistic Resources Database (LRD)

The LRD has to cover the spectrum of multilingual entries from paragraph down to stem entries. In order to achieve this the multilingual aligned Text (MAT) - concept was defined as the generalized LRD-entry. Every MAT consists of corresponding

monolingual entries referring to linguistic objects with attribute-value pairs describing the details (see diagrams).



The structure is described by an SGML construct resulting in an SGML-database.

Exporting SYSTRAN dictionaries into EURAMIS-LRD.

After the design and initial implementation of the EURAMIS LRD stage, the natural progression was to consider the feasibility of exporting MT dictionaries into the LRD.

Human nature being as it is, there was again strong opposition. Of course the problems were very real:

- Contextual rules in the Systran translation engine
- Morphology, mainly for verbs (paradigms), is defined in Systran in order to optimise retrieval and is dependent on language
- Copyright considerations resulted in proprietary and public data sets
- The data volume explosion by the SGML tags was dramatic
- Character sets in Systran are quite antiquated.

The TIF-DTD was taken into consideration and the LRD-DTD adapted to contain SYSTRAN specific attributes and secrecy provisions. The export procedure from MT was kept separate from the import procedure into the LRD in order to take advantage of linguistic knowledge build-up specific to the LRD and to be able to represent morphology, for example, as an independent factor. Tag compression helped solve the problem of data volume explosion created by SGML-tagging, although the total volume is still estimated at some 20 GB of data for 5 million entries.

The main work was the definition of the trade-off between representing simple linguistic rules encrypted in the Systran specific coding by standard clauses understandable by linguists. This caused longish discussions and trials in order to isolate specific problems.

The import function for EURAMIS is still under construction. The fine tuning will have to be done after completion of the complete cycle.

Conclusions

The main benefits of this exercise have been the creation of synergies between developers of machine translation and terminology on one side and combined functions on the other:

- Scanning all concerned Eurodicautom entries allowed the correction of entries (spelling-errors, language clean-up, formal variants contradicting the guidelines a.s.o.)
- A combined function provides terminology look-up from text in combining the Systran retrieval with Eurodicautom lookup even for target languages not available in Systran.
- In using the domain identifiers in terminology entries a content identification function was developed allowing to prescan a document to identify by statistical methods the domain it belongs to.
- The use of Eurodicautom entries in Systran translations is of immediate use in machine translation especially for long expressions and lesser developed language pairs.

- In the future it is planned to use Euramis for entering terminology, so that the MT entries will be available to terminologists.

Some of the new products, such as terminology look-up from text, are available today. Others will be available in the EURAMIS system, with the first prototypes currently undergoing tests. We have already achieved a great deal, but there is still more to come!

Machine Translation Feasibility Study at the European Commission

Dorothy Senez

Abstract

Since the introduction of machine translation (MT) to the Commission 20 years ago the project has been funded from a research budget; but now that MT is becoming an operational concern, at least for certain language pairs, the Commission is obliged to review its policy. A five-point study is therefore examining how MT might be used in the best and most economical way.

- a) An in-house survey of Commission MT users ascertains their MT needs in regard to languages, speed and quality.
- b) In parallel, practical experiments are being conducted in the Translation Service to provide an evaluation, as objective as possible, of the effects of machine translation on the Translation Service's production line.
- c) An investigation is being conducted on the legal issues which dictate the use, by third parties, of the Commission's multilingual tools and linguistic resources.
- d) The Commission wishes to find out if there are alternatives on the market to its current machine translation system and what costs would be involved. A technological survey of MT providers and users is underway and will give a clear picture of the state of the market for MT products and services.
- e) Finally, out of the data from the first four studies a cost-benefit analysis will be made. At the time of writing only preliminary results can be presented. The final conclusions, however, will enable decisions to be taken regarding the future orientation of MT at the European Commission.

Dorothy Senez

Born in Aberdeen in the UK. Graduate of the University of Edinburgh (French and Philosophy) and of the University of Paris (Maîtrise in Applied Linguistics and Translation Diploma). Has worked in the Translation Service of the European Commission since 1980. Staff translator for 10 years in the English Division. In 1989 took part in the SYSTRAN evaluation project. From 1990 responsible for the promotion of machine translation at the Commission. Now based in the "Translation Workshop" where new tools for translators are being tested in a production environment. Runs a service for MT users in the institution offering rapid post-editing of raw machine translation output.

Dorothy Senez
European Commission Translation Service, Translation Workshop
IMCO 5/5A
200, rue de la Loi, B - 1049 Brussels
Tel: +32 2 29 56436, Fax: +32 2 29 62993
E-mail: dorothy.senez@sdt.cec.be

Background

Twenty years ago two departments of the European Commission, the Translation Service and DG XIII (which is responsible for the exploitation of telecommunications research), decided to work together on the development of the Systran machine translation system. Since then, funding for both linguistic and technical development has been provided by DG XIII under the Multilingual Action Plans, while the Translation Service has played an active and joint role in matters relating to language. This proved to be a satisfactory solution, particularly in the early days, when the use of MT was limited as an operational tool in the institution's daily work.

However, since the beginning of this decade the use of MT in the Commission has made giant strides, mainly as a result of the general adoption of electronic mail, but also of promotion within the institution. In 1995 a total of 170 000 pages were put through the system, with the Translation Service accounting for slightly less than one third of this demand. Machine translation could now no longer be treated as a matter for research only; indeed in the case of certain language pairs, MT had already moved from a research to an operational environment. The Commission was therefore obliged to review its policy. It hardly came as a surprise, when in 1995 DG XIII announced its decision to wind down its participation with a view to withdrawal from the project. Although the last Multilingual Action Plan came to an end in December 1995, financing would nonetheless be provided by DG XIII until 1998, giving the Translation Service sufficient time to devise a new strategy.

Feasibility Study

The Translation Service is thus faced with the decision whether or not to allocate funds from its own budget to support the use of MT in the institution. To this end, a feasibility study is examining the conditions under which all or a part of the machine translation service might be taken over. On its completion it should be possible to determine to what extent it would be appropriate to continue machine translation at the European Commission and under what conditions such a service could be managed by the Translation Service.

The study comprises five distinct, but interdependent tasks:

1. A survey of machine translation users, and non-users, in all Commission departments.
2. Practical experiments with in-house translators.
3. An investigation of the legal issues involved.
4. A market study.
5. A cost-benefit analysis.

1. The User Survey

1.1 *The Objectives*

It was felt that the experiences of users would be of fundamental importance in the Commission's decision to continue, or not to continue, with MT. Indeed it is on the strength of the findings of the user survey that any future action on the remaining

four points will be determined, and the questionnaires were designed to provide both factual information and strategic guidance. A Coordination Group, consisting of 10 members representing different backgrounds, fields and languages, was set up to monitor progress and provide an objective viewpoint.

1.2 Methodology

MT is freely available to all Commission officials via the internal electronic mail network and the majority of users are non-linguist staff who help themselves to machine translation as and when they need it. Consequently, there is no direct feedback. We set out to discover why, how, and how much MT is used, and who the users are. The survey was an ideal opportunity for users to voice their opinion on the service offered. But it was hoped that the remarks of those who had never used MT would indicate the reasons for this and help to remove some of the obstacles to the adoption of MT. It became clear that there were three groups to be considered:

1. in-house Commission translators
2. administrators in the operational departments of the Commission
3. non-users.

A different version of the questionnaire was sent to each group, tailored to their specific context.

In the case of the Translation Service the questionnaire was distributed to all members of staff, users and non-users alike. In the case of the administrative departments questionnaires were sent to all users of the MT system over the previous 16 months. For non-users a proportional sampling method covering 4.5% of the remaining Commission staff was adopted.

1.3 The Questions

Objectivity has been the overriding concern throughout the whole survey; for this is an opportunity to dispel a few myths and establish once and for all a number of hard facts on the subject of machine translation.

Ten basic issues are addressed in the questionnaires:

1. Who the users are.
2. Why they use MT.
3. How they use it.
4. The volume of MT usage.
5. How useful it is to the Translation Service **directly** (texts post-edited by translators).
6. How useful it is to the Translation Service **indirectly** (texts submitted for MT by administrators that would otherwise have found their way to the Translation Service).
7. How useful MT is to the Commission departments as a linguistic tool (for browsing, translating, drafting).
8. The service's strengths and weaknesses (customer satisfaction).

9. The reasons why MT is not used.

10. Assuming that it is of use, how the service can be improved.

1.4 Response Rates

Of the 1,700 questionnaires distributed to Translation Service staff, 520 of whom were identified as MT users, 773 responses have been received. A total of 2,600 users in the administrative departments were surveyed, with 735 responses received. In the case of non-users a proportional sampling method covering 5.5% of the remaining Commission staff was adopted and 270 responses have been received.

1.5 Results

The findings presented here are preliminary. The final results will be published in the autumn.

1.5.1 Translation Service - Users

Translators like to use MT because of its speed, the typing time it saves, and the fact that the raw MT is returned with its original format. MT is also sometimes used to provide assistance with terminology. As nearly all those translators who use MT do so to produce a final, polished translation, they tend to be somewhat negative about the quality of the MT output. They do not like the heavy post-editing involved, although a certain number do appreciate the system's sense of humour. The majority find less than half the documents useful, and a small percentage find no texts useful. Over 50% of users in the Translation Service request MT only occasionally, and although they do save some time, a significant number say they save very little time. Everyone at least agrees that the system's response time is very good. As regards the assessment of language pairs, the best marks are attributed to French-English, French-Spanish, French-Italian, and English-French. About half the users in the Translation Service would like to be able to create and manage their own personal dictionaries. Although only 25% say they would not have met deadlines if MT had not been available, the majority (67%) of MT users feel it is a tool worth having at their disposal.

1.5.2 Translation Service - Non-users

Most non-users do not know how to use the system, or tried once or twice and did not like the result. Some consider that the texts they have to translate are not suitable. Fears are expressed that it might dehumanise the work of translating. In many cases the relevant language pair is not covered.

1.5.3 Administrative Departments - Users

Administrators tend to request MT on an occasional basis for the translation of urgent documents they would have preferred to send to the Translation Service. They also use MT for information scanning in languages unknown to the reader and for drafting in a foreign language. They like MT for its speed, ease of use, the lack of bureaucratic procedures and the fact that it is available round the clock. They do not, however, like having to correct the texts, the fact that the system is slow to learn, and that some language pairs are missing. Texts are revised in most cases, normally by a native speaker. More than half of the respondents in this group do not indicate on the

text that it is revised MT output. Those who rely on the post-editing service are happy with it. On the assessment of language pairs, administrators tend to be more lenient than their translator colleagues. The majority find half or more of the output useful. In stark contrast to the Translation Service figures, 74% consider MT saves them a considerable amount of time. More decisively, over 50% think that some of their documents would have been late had it not been for MT. Also interesting, since it proves that MT is at least saving the Translation Service some time indirectly, 59% say they would otherwise have sent their texts to the Translation Service for translation. A resounding 94.8% feel that MT is worth having at their disposal.

1.5.4 Administrative Departments - Non-users

About 20% of non-users have no real need for translation. Of the others, most do not know how to use MT and those who do complain about poor quality, determined from direct experience in the past or from the comments of colleagues. As with non-users in the Translation Service, many judge that their texts would not be suitable. In some cases people solve their day-to-day translation problems with the help of colleagues. The majority of non-users requested more information about MT, which is still relatively unknown.

1.6 Preliminary Conclusions

There are two quite different pictures emerging here. On the one hand there is a lukewarm, but by no means entirely negative, reaction from professional translators. Although results vary from one language group to another, a significant number of in-house translators have been able to make MT work for them, particularly in the case of targeted development, where the Systran dictionaries have been programmed for specific text types. On the other hand, there is a more enthusiastic reaction from the administrative departments. This group has perceived needs for urgent translations, browsing and drafting, which MT is already meeting to some extent. Nevertheless, the initial findings show quite clearly that increased efforts are required here to provide better support and information for this user population.

2. Practical Experiments in the Translation Service

As the user survey confirms, a number of Commission translators are making good use of MT in their daily work, particularly where the system's dictionaries have been prepared with the appropriate terminology, and it was decided to back up the survey with data provided by these translators. The aim is to provide an evaluation, as objective as possible, of the ways in which MT helps in-house translators, if not to increase their page count, at least to speed up the workflow. The experiments are to be conducted from June until October. Initial conclusions will be drawn this autumn in the context of the cost-benefit analysis, where the results obtained from these volunteer translators will be produced as vital evidence.

2.1 Language pairs

The experiments are limited to those language combinations that have reached a sufficient quality to make post-editing worthwhile:

- English into Spanish, French, Greek, Italian and German.
- French into English, Spanish, Italian.

2.2 Translators taking part in the experiments

In line with this pragmatic approach, we initially approached those translators who are already experienced in the use and the post-editing of MT. Since staff translators are often understandably wary of MT, a modest response was expected to the call for volunteers. In the event, over a hundred translators have come forward to take part in the experiment.

2.3 Selection of texts

Volunteers are under no obligation. They are not required to use machine translation systematically, but only when they feel they have a suitable job in hand. The choice of text type and subject matter is left to the volunteer post-editor, but, naturally, text types for which terminology has already been coded yield better results.

2.4 Evaluation

Since the aim of the exercise is to produce quantifiable data, post-editors fill in an evaluation form each time they post-edit a machine translation in the course of their daily work in which they are asked to calculate how much time, if any, they have gained in relation to traditional translation, and this may include time saved in typing and lay-out as well as terminology searches - difficult questions to answer.

We tried to avoid a pitfall encountered in previous exercises of a similar nature, where the evaluation of the linguistic content of the MT output tended to become confused with the surrounding organisational issues. It is important to try to keep the two issues separate if a clear and objective picture of the actual performance of the MT system is to be obtained. Hence, we divided the form into two sections: the first (compulsory) asking for an appreciation of the linguistic content; and the second (optional) aimed at gaining an insight into any difficulties experienced in getting an electronic version of the source document to the translator's PC.

2.5 The results so far

The initial results show that although, generally speaking, translators do not find it easy to quantify the amount of time gained, a number do report time-savings of between 30% and 60%. The remarks concerning the quality of the MT output argue in favour of stepping up the linguistic development of the system. The striking improvements to the French-Spanish language pair which have been recorded recently are an illustration of just what can be achieved when translators and linguistic developers collaborate actively on the development of an MT system.

3. Legal Issues

This part of the feasibility study involves an examination of the legal aspects which dictate the use of the Systran system. A number of related questions, covering the policy of distribution of the Commission's multilingual tools and linguistic resources and their use by third parties, will also have to be treated.

The investigation so far shows that the Commission owns a licence which gives it the right to use its system without restrictions as to territory or sector. However, not wishing to distort conditions of competition, the Commission has decided to restrict itself to the territory of the European Union and to the public sector (governmental bodies and Community institutions).

4. Market Study

The Commission is seeking to establish whether there are alternatives on the market to its current system and what their costs would be. This market study covers MT only and not, at this juncture, other translation aids and other computer-assisted tools. The study, financed by DG XIII and conducted by external experts, takes the form of a questionnaire, distributed to MT vendors, aimed at identifying all machine translation systems which are available for use in production environments and any service organisations that provide access to such MT systems. At this stage it is a fact-finding rather than an evaluation exercise, and does not cover products which are still under development or the subject of research.

In particular, the market study sets out to identify:

1. what fully automatic MT **systems**, compatible with the Commission's informatics environment, are currently available on the market and what languages they cover.
2. what MT **services** can be offered, in what form, e.g. raw MT, post-edited MT; and via which access methods (e-mail, Internet or other communication channels).

Depending on the results of the study, which is due to be completed by September, a call for tenders might be held with a view to meeting the Commission's needs for machine translation and for post-editing services.

4.1 The questions

Questionnaires were sent out to 70 MT vendors. In the case of MT systems, the information required by the Commission includes:

- technical specifications
- linguistic processing techniques
- linguistic resources provided
- text handling features
- performance and quality
- product training and support
- costs and licensing schemes.

In the case of companies providing access to their MT technology as an outside service, some additional information is required including:

- means of access
- peripheral services (dictionary encoding, pre- and post-editing).

4.2 Preliminary results

Although it is too early to draw specific conclusions, some general trends are beginning to appear from the 30 responses received so far. The sector in general is still very shaky, with the exception of three or four stable players. Language coverage is patchy. On the new PC-based market one vendor has an 80% market share, but its success lies more in the home systems market and is no alternative to the current EC system.

As regards MT offered as a service, both raw and post-edited, the questionnaire elicited a very poor response. The message seems to be that the time is not yet ripe. In this embryonic market a consensus needs to be reached on the question of prices. For the moment, then, the overall picture shows little demand for raw MT via teleservices and very few new services coming to the fore.

The picture would not be complete without a view of the experience of actual users of MT products and services, and a parallel research project is being carried out with the MT user community. It would appear that the successful users of MT are those who have built up extensive dictionaries covering their specific fields, and users express a clear preference for maintaining control over their MT dictionaries.

5. Cost-Benefit Analysis

Once all the data from the first four studies has been gathered, a cost-benefit analysis should enable some conclusions to be drawn. Based on forecasts regarding the budgetary impact of any new policies, a decision will then be taken about the advisability of certain activities being taken over by the Translation Service.

There are no easy answers, particularly as MT at the Commission continues in some respects to be a hybrid creature, hovering somewhere between a development issue and an operational issue. A specific feature of MT in the institution is its wide user base and, hence, extremely varied text types and subject fields. We have at our disposal considerable resources in the form of twenty years of development on the system's dictionaries. We have a core population of over 2,000 regular users, whose needs cannot be ignored. In addition, the study clearly shows that more information about MT is required throughout the institution. It is probably safe to say at this stage that machine translation at the Commission is here to stay, in some form. Looking to the future, there are a large number of officials in administrative departments throughout the member states with a latent need for MT. Whatever the final outcome, our assets need to be exploited wisely in the best interests of current and potential users.

Acknowledgements

The feasibility study described in this paper is the result of the combined efforts of Dimitri Theologitis (Translation Service), Jean-Marie Leick (DG XIII) and the Luxembourg and Brussels machine translation team - Francine Braun-Chen, Cameron Ross, Rosemarie Sauer, Angeliki Petrits and Dorothy Senez.

Session 4: Economics of Using Machine Translation

Chair: Viggo Hansen

Introduction

It is often claimed that the use of MT-systems does not represent any significant cost savings, but merely is a way to produce huge quantities of raw translated documents. Not many private companies have actual experiences with MT. But it appears that some key factors are instrumental to obtain economics of using MT. Besides a high degree of functionality of the MT-system itself the procedures for creation and maintenance of domain based dictionaries and for the preparation work, the pre-editing and the post-editing seems to be some of the more important factors.

Viggo Hansen

Since November 1995 Viggo Hansen is the Managing Director of Hofman-Bang & Boutard, Lehmann & Ree A/S. In May 1993 he became Manager of a new established company Lingtech A/S, a translation company owned by Hofman-Bang & Boutard A/S and Lehmann & Ree A/S. Viggo Hansen's main task was to build up the activities of the company and to implement the machine translation system PaTrans. Up to May 1993 he was an independent Management Consultant working primarily with management information and control systems. As a management consultant he worked for Hofman-Bang & Boutard A/S and Lehmann & Ree A/S as project manager for their part of the development of the PaTrans machine translation system.

Hofman-Bang & Boutard, Lehmann & Ree A/S

is one of the leading patent- and trademark attorney companies in Scandinavia, and is the result of the January 1996 merger between Hofman-Bang & Boutard A/S and Lehmann & Ree A/S. The company has about 50 specialists in intellectual property rights and a total staff of 125 employees; the customers reflect a broad variety of large and medium-sized Danish and international companies. The main activities include obtaining and maintaining of patent, trademark and design rights. To obtain patent rights in Denmark, it is mandatory to supply the Danish authorities with a detailed description of the product/process in Danish, and consequently the translation workload of the company is substantial. About 8 million words per annum are translated from English, German or French into Danish.

To separate the core activities of the company (i.e. counselling and advising on intellectual rights) from associated activities (such as translations) it was decided to form a separate company Lingtech A/S in which most translation activities are handled. Lingtech started to operate in 1993.

Viggo Hansen

Hofman-Bang & Boutard, Lehman & Ree A/S
Adelgade 15, DK-1304 København K, Danmark

Tel: +45 33 150585, Fax: +45 33 157585

E-mail: vha@hofman.dk

Kielikone Machine Translation Technology and Its Perspective on the Economics of Machine Translation

Dr Harri Arnola

Abstract

This paper describes the machine translation technology of Kielikone Ltd. and gives an outline of TranSmart, a Finnish-English system which is a commercial application of that base technology. We argue that MT is fundamentally empirical research. Product development is a slow and strenuous process, and MT systems will remain incomplete not only vis-à-vis human translations but also with respect to the system's own potential translation quality. An evaluation method is described which measures the progress in MT development. This method can also be used for system and technology evaluation. The paper concludes with the claim that the real contribution of MT will be seen in the long term in applications that do not compete with human translations but in which MT is the only option.

Dr Harri Arnola

Harri Arnola teaches Artificial Intelligence in Helsinki University of Technology. He was the leader of the Kielikone project that was funded by the Sitra Foundation, a public Finnish fund whose objective was to support risky R&D that carries obvious national interest. The aim of the project was to design and implement computational algorithms that would be widely applicable language technology applications in Finland.

Kielikone Ltd.

The Kielikone project resulted in the setting-up of Kielikone Ltd., a Finnish language technology company that currently has 13 employees. Harri Arnola is the managing director of the company. Kielikone Ltd. is the market leader in Finland as a supplier of proofreading software and electronic dictionaries. The main R&D activity of Kielikone is Machine Translation technology and MT systems. The **TranSmart** Finnish/English system is the first MT system that is based on Kielikone's own MT technology. It is used by two major Finnish companies. The system has been used also to provide MT services to various companies and individuals. Kielikone's MT work has been supported by the Technology Research Centre of the Ministry of Trade and Industry in Finland.

Kielikone Group

Kielikone Ltd. is the parent company of a group of small companies. The other companies in the group are **Transwise Oy**, a Machine Translation services company, and **Wordmark Oy**, a marketing company, located in Jyväskylä, Finland. In addition, Kielikone Ltd. is a shareholder in **Käännöskone Oy**, a company that specializes in hand-held electronic dictionaries, located in Vihti, Finland.

Harri Arnola

Kielikone Ltd., P.O. Box 126, 00211 Helsinki

Tel: +358 0 6820 211, Fax: +358 0 6820 167

E-mail harri@kielikone.fi

(Authors: Arnola¹, Harri; Hyvönen, Kaarina; Juntunen, Jukka-Pekka; Linnanvirta, Tim; and Suoranta, Petteri of Kielikone Ltd.)

1. KIELIKONE MT

This section describes the base MT technology developed by Kielikone. The technology is language independent and can be used for building MT systems for various language pairs. Kielikone has built a commercial MT system, TranSmart, applying the base technology. The TranSmart system translates from Finnish into English.

1.1 Base technology

In terms of machine translation techniques, Kielikone MT technology (KMTech) uses the transfer approach; as to linguistic theories, KMTech makes a commitment to the dependency theory. Transfer, then, consists of a sequence of meaning-preserving transformations of source dependency trees into target dependency trees.

An **MT Engine** is the basic process for transforming dependency trees for a specific goal. An MT Engine instance is a virtual machine which consists of two parts: the generic code (the engine proper) and a task-specific rule base (Figure 1).

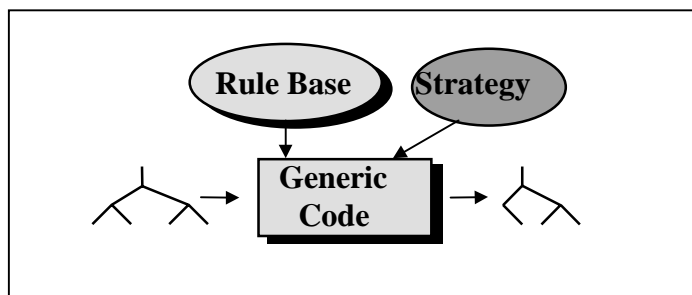


Figure 1: MT Engine

An MT system which uses KMTech has an extremely simple but flexible linear architecture. The transfer is a chain of the MT Engine applications (Figure 2). Each engine performs a certain transfer subtask. In KMTech also lexical transfer is performed by MT Engine applications. The number of MT Engine applications used depends on how the transfer problem is divided into subproblems. The architecture can be changed at any time by defining new subtasks and inserting new MT Engine applications in appropriate places in the chain. For example, an MT system may comprise several distinct lexical transfer phases and their priority is assigned simply by changing their order. A full MT system consists of an input analyzer and a dependency parser, which are not part of KMTech, a KMTech MT Engine transfer chain, a linearization module, and a postprocessing module. The two synthesis modules can also apply the MT Engine (Figure 3).

¹ Formerly Jäppinen

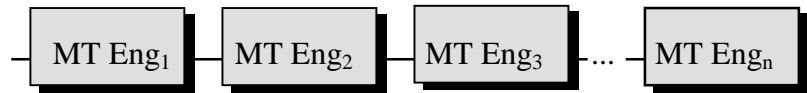


Figure 2: Linear transfer architecture

Fully automatic machine translation has distinct quality limits. One way to overcome such limits is to access and adapt high quality human translations (or corrected MT translations) stored in a translation memory. Obviously, the extent to which translation quality is improved in such a hybrid system depends first and foremost on the amount of recorded data. Since in practice an exactly matching sentence is only rarely found in the memory, the added value of a translation memory depends also on how intelligently the translations of nearly matching sentences can be used. Translation memories usually perform fuzzy matches on **strings** of words. KMTEch offers a linguistically intelligent translation memory which stores the dependency **trees** of source sentences. Matching trees rather than strings offers several benefits. To name one, dependency trees neutralize variations in constituent order in a natural way thus making the adaptation of near-matches easier (Juntunen, 1996).

1.2 A case study: the TranSmart Finnish-English system

Kielikone's TranSmart Finnish-English system is a fully implemented MT system which relies on KMTEch (Jäppinen et al., 1991, Jäppinen et al., 1993). Its basic architecture is shown in Figure 3. The shaded squares represent MT Engine applications. The analysis has two parts: morphological analysis of word forms, and dependency analysis of sentence structures. Deterministic parsing in linear time is the main theoretical attraction of the parser (Arnola, 1996). As the figure shows, transfer activates the MT Engine seven times (lexical transfer three times). As a product, TranSmart has two versions: a workstation solution (Unix), and a client-server solution (Unix).

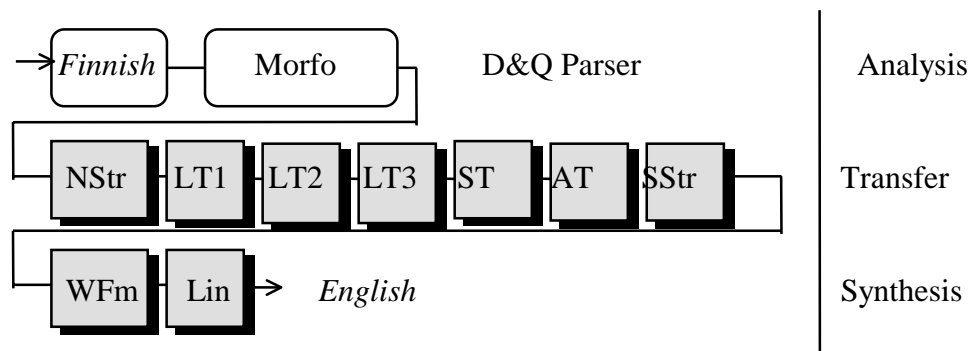


Figure 3: Architecture of the TranSmart Finnish-English system

1.3 TranSmart in practical use

Kielikone MT has been under development for several years. The project was initially supported by the Sitra Foundation, and later on, in the product development phase, by the Technology Development Center of the Ministry of Trade and Industry. From the outset there has been strong user participation in the work. The companies listed below have participated in the development work both financially and concretely. Although systems have already been sold to customers outside this consortium, Kielikone has so far mainly concentrated on catering to the needs of the members of the consortium. Major marketing efforts have been deferred.

The consortium

Nokia Telecommunications, net sales over FIM 10 billion, a subsidiary of the Nokia Group, is one of the pilot customers who have participated in the development work from the beginning. The original intent of the company was to use TranSmart as a workstation to support customer documentation. In this role the translation memory feature plays an important role as documents often have common sections. During the current year (1996) a new application has surfaced spontaneously at Nokia. The company has plenty of employees whose native tongue is not Finnish stationed in various countries. Occasionally they face documents which are written in Finnish. Thanks to the TranSmart server version installed in the company net they are now able to obtain quick rough translations which often satisfy their needs.

Rautaruukki Oy, a Finnish steel company, net sales over FIM 9 billion, is another pilot customer. The TranSmart system is installed as a server in the company net, and employees can access it through the internal electronic mail system.

Trantex Ltd., a Finnish translation services company, net sales over FIM 28 million, is a third pilot customer. The company specializes in localizing software products but produces translation services in other fields as well. Text types and domains tend to vary greatly in a translation services company and since MT requires domain specific lexical tuning, the benefits of MT are not so clear in this area.

MT services

Jointly with Trantex Ltd., Kielikone has established a translation services company, Transwise Oy, to offer machine translation services fast and at a reasonable price. This service function has two goals. The first and paramount goal is to offer customers a new form of attractively priced translation services. The second, subsidiary goal is to collect texts from different sources for purposes of system tuning. Since the texts are used also for tuning, translation speed has not yet been as high as machine translation technology would ordinarily allow. Transwise Ltd. intends to install the TranSmart system in the WWW in the near future and start offering instantaneous translation services.

Next we want to share our experience in the strenuous work of building MT systems. MT systems have idiosyncratic characteristics which often make the work seem frustrating and evasive. In fact, MT systems are never complete. Nevertheless, they do have promising applications.

2. MT Systems Are Incomplete

It is no surprise to anyone that MT systems are incomplete. This section attempts to argue briefly why incompleteness is an essential characteristic of MT systems. This discussion serves as a stepping stone to the discussion on evaluation of incomplete systems.

2.1 *There is no theory of MT*

Incompleteness is a result of the lack of rigorous theories of MT. We might start by asking if there exists a verifiable theory of translation of sentences between two languages in the sense of hard sciences. By such a theory we mean a detailed and explicit method or a set of formulae which relates source language sentences to their perfect translations in the target language, **without resort to human intuition**. We believe that such a theory does not and will not exist, at least in a near future. (Of course there are other forms of theories of translation which teach and study the trade of translation, but these theories are not rigorous in the sense we are looking for.)

One of the reasons why there are no scientifically precise theories of translation is that translation is about meaning-preserving transformations between utterances in two languages, and there is no rigorous theory of meaning available. It is unlikely that anything except fragmentary theories of meaning of natural languages (say, Montague semantics or situation semantics) will surface in the foreseeable future. Furthermore, natural languages abound idiomatic expressions and conventions whose correlations with other languages constitute an empirical and not a theoretical inquiry. In any given natural language there are dozens of idiomatic expressions which resist any precise semantic analysis which would unequivocally link the expressions to another language. Such expressions can be only approximately projected into another language and the projections are language specific.

If our argument is true, it follows that there neither is nor will be a computational theory of translation. A computational theory presumes a conceptual theory; if the latter is missing, so is the former.

2.2 *Yet, MT can and should be theoretical*

However, from the argument it does not follow that MT cannot be **theoretical**. It can (and should) be theoretical, in at least two different ways. First, MT should use linguistic theories of word and sentence structure and text cohesion. There are several morphological theories which relate word forms with their morphological structure and syntactic theories which relate sentences with their syntactic structure. Moreover, there are computational algorithms which produce such theoretical structures more or less reliably. As translation is about meaning-preserving transformation of sentences, it is advantageous, we believe, to choose from among alternative syntactic theories one that indicates functional rather than constituent structures of sentences. Our choice is the dependency theory.

Second, even if there is no comprehensive computational theory of translation worthy of the name, MT should be theoretical in the sense that translations should capture as many linguistic generalizations between two languages as possible. Even if such a set of generalizations falls short of a comprehensive theory, it takes steps towards that direction. And, applying Occam's razor, we might then say that one MT system for a given language pair is more "theoretical" than another if the former applies more

salient generalizations about the language pair than the latter (and presumably needs a smaller number of contrastive rules). From this viewpoint, even word for word translation systems are “theoretical” to an extent (since probably they implement at least some rudimentary generalizations between the language pair), but they are less “theoretical” than more advanced systems.

2.3 Hence, MT systems are incomplete

If there is no comprehensive computational theory of translation, MT technologies, as embodiments of certain theoretical principles for the purpose of producing translations, are incomplete. Thus, if we are right, there cannot be an MT technology which would produce a closed solution for the problem of translating sentences between a given language pair in the sense that one would be able to conceive of a set of formulae that, once written down, would produce perfect translations between the pair. The problem is and remains open. Theoretical research is able to produce a partial solution but there remains a considerable empirical work load. Consequently, the quality of an MT system is not a step function but improves only slowly as new empirical associations are added to the theoretical base (Figure 4). The more “theoretical” a system is, the faster its quality improves and the higher is its asymptotic, potential translation quality. MT systems are not only incomplete vis-à-vis human translation quality; they are also incomplete with respect to their own, system dependent potential translation quality.

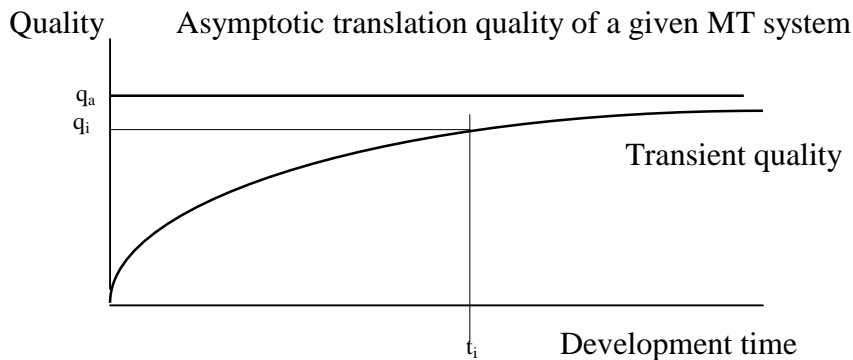


Figure 4: Gradual improvement of the quality of an MT system

MT system development is slow and labor intensive work and its goals are evasive. MT is therefore not only a theoretical endeavor, but it has also clear engineering goals. MT technologies should support both the implementation of linguistic theories and the unavoidable empirical work. KMTech, we believe, is such a technology.

3. Evaluation of MT Systems

Usually artifacts are so designed that they serve the intended purposes fully. MT systems, being inherently incomplete, are therefore odd commodities. The issue of the economics of imperfect commodities, per se, is outside the scope of this paper and we shall content ourselves with only a few remarks on the question.

Incomplete systems have no absolute value; they have only relative (or functional) value. Usually a product, say, a boat, has absolute value in that you can use it for crossing any waterway. An MT system is like a slightly leaking boat in that it cannot be used for all translation needs but it can, nevertheless, be used for some specific purposes. It has therefore only relative value, and one needs evaluation to secure that the quality of a given MT system is high enough for a specific purpose.

A specific purpose is often a certain functional role, say, the production of rough translations for a human translator. An incomplete system qualifies for a functional role only if its use creates added value in the whole process. For example, if an MT system is used for producing rough translations for a human user, the positive value of the system (rough translations) must be greater than the negative value which its use generates (postediting cost). The best way to ensure that an MT system qualifies for a certain functional role is, again, to perform evaluation. The rest of this section discusses evaluation methods of MT systems.

3.1 The purposes of MT evaluation

If incompleteness and far from perfect performance is inherent in a system, evaluation becomes very important. It is not acceptable to market an MT system either by claiming that it is perfect, which is false, or by just claiming that it is imperfect. The latter claim, albeit honest, leaves the customer puzzled. He or she rightly asks, but **how** good is it? To answer such a natural question one needs reliable evaluation.

There are several different purposes one might want to evaluate MT systems for. At least the following purposes have been discussed.

System comparison. A potential user of an MT system for a given language pair wants to compare different systems in order to choose the right one (e.g. Flanagan, 1994).

System evaluation. A potential user of an MT system for a given language pair wants to evaluate a particular system in order to see if it is good enough (in terms of quality and economy) for the intended purpose.

Technology comparison. A funding agency or an MT system developer wants to compare different MT technologies in order to see which one is worth funding or choosing, if any (e.g. White et al., 1994).

Technology evaluation. An MT system developer wants to evaluate a specific MT technology in order to see if it provides a good enough base technology for MT system development.

Progress evaluation. An MT system developer wants to evaluate progress in MT system development work in order to see when the system has reached its limits and the cost of additional work no longer pays off (e.g. Ishara et al., 1994).

3.2 Black box evaluation and glass box evaluation

Of these, system evaluation is by far the most frequently performed one although results are rarely reported. Whenever a potential user thinks seriously of using an MT system in production environment some kind of system evaluation is performed. The

most frequently discussed evaluation issues seem to deal with comparing systems or technologies. Two methods have often been mentioned.

In **black box evaluation** test systems are given identical inputs and their outputs are compared and ranked. If the systems are mature such a strictly behavioristic evaluation method is fair and equitable. For a practical working system it should not matter what internal processing takes place; all that counts is what the system delivers.

If the systems are not mature and it is permitted to correct errors located during evaluation, it makes a difference how the corrections are made. If an error results from a missing piece of linguistic knowledge or a missing or incomplete lexical entry and if the fault can be fixed following the standard procedure used for implementing linguistic knowledge in the system, the error does not indicate a theoretical or structural weakness in the system. If, on the other hand, the fault can be corrected only in an ad hoc manner by associating hand-coded translation with the input or a part of it, the error reveals a fundamental weakness in the system.

In **glass box evaluation** test systems are given identical inputs and their outputs are compared and ranked as in black box evaluation. Furthermore, corrections are allowed and the theoretical principles and linguistic generalizations employed by the systems are observed and ranked. Glass box evaluation is fair if the systems are not mature or if they always require customer specific tuning before production.

Black box and glass box evaluations are plausible choices for system comparison. When technologies are compared, glass box evaluation is preferable to black box evaluation. Next we discuss an evaluation method used by Kielikone. This method, which is in the spirit of glass box, has applications in progress evaluation and system evaluation, and it may offer material for technology evaluation as well.

3.3 Window evaluation

Kielikone developed the method of window evaluation originally for the purpose of monitoring progress in system development work. As argued earlier, the quality of an MT system approaches its final quality extremely slowly. At any give moment t_i there is a “distance” $|t_a - t_i|$ between the current quality of the system (t_i) and its potential quality (t_a) (Figure 4). The distance should decrease monotonously while the work progresses. Window evaluation attempts to give a measure for such a distance and to monitor the monotony of the progress.

For the sake of this discussion, let us assume the concept of a **translation space** generated by various linguistic attributes. A point in the space is said to be **covered** if the system can translate the source structure represented by that point at the asymptotic level. Figure 5 shows a hypothetical translation space drawn in two dimensions. (In practice, such a space would, of course, have many dimensions.) This crude visualization attempts to separate areas covered by structural transfer rules representing linguistic generalizations (shaded irregular areas) and lexical transfer rules representing word-specific translations (shaded circles). Due to the empirical nature of the work, the space will always have uncovered areas (white areas in Figure 5). To evaluate the current quality against a certain text type one takes a representative sample of the text. The piece of text covers a certain fixed area in the translation space. (Represented by a rectangular in Figure 5. In reality, the subspace covered by any piece of text would not be continuous.) The piece of text offers a kind

of a “window” into the translation space, and the finite subspace “seen” through the window can be fully covered. White areas (missing or incorrect structural and lexical translations) are revealed in this subspace and the system can be tuned to its asymptotic level.

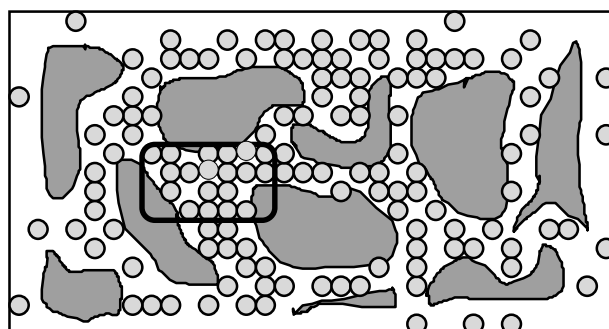


Figure 5: A translation space covered by generic and specific rules

Let us call the initial translation (no corrections) of a piece of text its **rough rough** translation and the final translation (all possible corrections made) its **polished rough** translation. The rough rough translation indicates the current quality, the polished rough translation shows the potential quality, and the corrections made represent the “distance” between the two. Notice that in this context “polished” does not indicate post-editing.

Table 1 gives an example. The text is a light news item printed in a Finnish newspaper. The original text was 56 sentences long. To save space, the table shows only the first 30 sentences. The text has not been pre-edited. Sentence #7 (marked by a star) is parsed incorrectly in two ways: the highly elliptic structure is parsed incorrectly, and a genitive attribute gets an incorrect structure (*Nesteen aurinkoenergiaan...* / *solar energy of Neste* should read *Nesteen... tutkimusosasto / research department of Neste*). The translations of this sentence are therefore quite wrong. Whenever TranSmart encounters a source word for which it has no translation it copies the original source word. These are marked in bold face in the rough rough column. (Headings are also in bold face.) If the missing word is a compound and there exist translations for the parts, TranSmart performs a part-for-part translation as a default. Such translations appear in italics in the rough rough column. The polished rough column indicates systematic corrections between rough rough and polished translations: lexical corrections are in italics and structural corrections are in bold face. Polished rough translations have **not** been post-edited in any way.

#	Source	Rough rough target	Polished rough target
1	AURINKOENERGIAN VOIMALLA LIKKU-VASTA VENEESTÄ PUUTTUU MOOTTORIN	THE SÄKSÄTYS OF THE MOTOR IS MISSING FROM THE BOAT WHICH MOVES WITH THE POWER OF	THE <i>CLATTER</i> OF THE MOTOR IS MISSING FROM THE BOAT WHICH MOVES WITH THE POWER OF

	SÄKSÄTYKS	SOLAR ENERGY	SOLAR ENERGY
2	Kolmen vuoden kehittälyllä yhdistettiin suomalainen puuvene ja aurinkoenergia	With the three year development the Finnish wooden boat and solar energy were connected	With the three year development the Finnish wooden boat and solar energy were combined
3	Tavallisen soutuveneen perässä on pikkuinen moottori.	There is a little motor behind an ordinary rowing-boat.	There is a little motor behind an ordinary rowing-boat.
4	Vene liukuu vedessä kuin unelma, tasaisesti, äänettä-mättä ja saasteettomasti.	The boat slides in water like a dream, evenly, silently and saasteettomasti .	The boat slides in water like a dream, evenly, silently and <i>without pollution</i> .
5	Sähkömoottori saa voimansa auringosta.	The electric motor gets its power from the sun.	The electric motor gets its power from the sun.
6	Suomalaisen soutuvene on työstänyt aurinkoveneeksi Juha Nyman Särkisalosta.	The Finnish rowing-boat has been worked to be a <i>sun boat</i> by Juha Nyman from Särkisalo.	The Finnish rowing-boat has been modified into a <i>solar boat</i> by Juha Nyman from Särkisalo.
7*	Hanketta on ollut kehittä-mässä myös Albican-verkos-toyryitys sekä tukemassa Nes-teen aurinkoenergiaan keskit-tyntyt tutkimusosasto.	The project has also been developed by the Albican <i>network company</i> and has been supporting the <i>research department</i> which has concentrated on the solar energy of Neste.	The project has also been developed by the Albican network company and has been supporting the research department which has concentrated on the solar energy of Neste.
8	"Tämä on kolmen vuoden kehittelyn tulos.	"This is the result of the three year development.	"This is the result of the three year development.
9	Hanke on edennyt pikkuhiljaa ja keväätalvella idea toteutetiin käytännössä.	The project has proceeded little by little and in late winter the idea was carried out in practice.	The project has proceeded little by little and in late winter the idea was carried out in practice.
10	Pääajatus oli yhdistää vanha suomalainen puuvene ja aurinkoenergia", selittää Nyman.	The main idea was to connect an old Finnish wooden boat and solar energy", Nyman explains.	The main idea was to <i>combine</i> an old Finnish wooden boat and solar energy", Nyman explains.
11	Paneeli piiloon penkin alle	Panel to hiding place under bench	Panel to hiding place under bench
12	Teknisesti aurinkovene vaikuttaa sangen yksin-kertaiselta.	The <i>sun boat</i> seems very simple technically.	The <i>solar boat</i> seems very simple technically.
13	Aurinkopaneeli lataa akun, josta sähköinen perämoottori saa voimansa.	The solar panel loads the accumulator from which the electric outboard motor gets its power.	The solar panel loads the accumulator from which the electric outboard motor gets its power.
14	Paneelin voi jättää näkyviin tai työntää piiloon istuimen alle. Vene saavuttaa noin kymmenen kilometrin tuntinopeuden ja yksi lataus riittää reilun tunnin ajomatkaan täydellä vauhdilla.	The panel can be left in sight or can be pushed to the hiding place under the seat.	The panel can be left in sight or can be <i>hidden</i> under the seat.
15	Nymanin aurinkovenevers-taaseen on kantautunut uutisia, joiden mukaan Ruotsiin on syntymässä hänelle kilpailija. Muista pohjoismaisista kilpailijoista ei ole tietoa.	The boat will reach about the ten kilometre speed per hour and one charging is enough for the good hour drive on a full speed.	The boat will reach <i>the speed of about ten kilometres per hour</i> and one charging is enough for the good hour drive <i>at full speed</i> .
16	Nymanin aurinkovenevers-taaseen on kantautunut uutisia, joiden mukaan Ruotsiin on syntymässä hänelle kilpailija. Muista pohjoismaisista kilpailijoista ei ole tietoa.	News according to which the competitor being born to Sweden to him have been carried in Nyman's <i>sun boat workshop</i> .	News according to which the competitor <i>is being born in</i> Sweden to him has <i>reached</i> Nyman's <i>solar boat workshop</i> .
17	"Ruotsin TV4 esitteli taannoin ruotsalaisyrityksen valmistaman aurinkoveneen.	There is no information about other Nordic competitors.	There is no information about other Nordic competitors.
18		"Sweden's TV4 demonstrated	"Sweden's TV4 demonstrated

19	Sen sähköratkaisut vaikuttivat tosin kovin alkeellisilta. Myös Saksassa päin on tietävästi muutama tämän-	recently the <i>sun boat</i> made by the Swedish company. However, its electric solutions seemed very elementary.	recently the <i>solar boat</i> made by the Swedish company. However, its electric solutions seemed very elementary.
20	suuntainen hanke", sanoo Nyman. Tavoitteena Keski-Eurooppa	There probably are a few such projects also somewhere in Germany", Nyman says.	Also somewhere in Germany there probably are a few such projects", Nyman says.
21	Aurinkovene on ollut esillä muutamilla venemessuilla.	As objective Central Europe	As objective Central Europe
22	Suomalaiset messuvieraat ovat jonkin verran vieroksu-neet	A few <i>boat exhibitions</i> have had a <i>sun boat</i> up for discussion.	A <i>solar boat</i> has been <i>shown at</i> a few <i>boat exhibitions</i>
23	Nymanin venettä. "Perussuomalainen	The Finnish <i>exhibition guests</i> have shunned Nyman's boat a little.	The Finnish exhibition guests have shunned Nyman's boat a little.
24	soutuveneen ostaja on aika konservatiivinen. Yksi kommentti on kuulunut, että veneemme on digitaalilla pilattu puuvene."	"The <i>basic Finnish</i> rowing-boat buyer is quite conservative.	"The basic Finnish rowing-boat buyer is quite conservative.
25	Sen sijaan keskieurooppalaiset ovat olleet kovasti innostuneita. "Etenkin saksalaiset ovat olleet ihastuneita aurinkoveneen ympäristöystävällisyyteen."	One comment has been that our boat is a wooden boat that has been spoiled with digitaali ." Instead the Central Europeans have been very excited.	One comment has been that our boat is a wooden boat that has been spoiled with <i>digitals</i> ." Instead the Central Europeans have been very excited.
26	Nykyisellään Nymanin vene ei kuitenkaan kelpaa Keski-Eurooppaan.	"Especially the Germans have been attracted to the kindness to the environment of the <i>sun boat</i> ."	"Especially the Germans have been attracted to the kindness to the environment of the <i>solar boat</i> ."
27	Siellä päin veneitä säilytetään pääasiassa telillä, koska vedessä olevia venepaikkoja on niukasti. "Jos nykyisen veneen nostaa aina vesireissun jälkeen telille, puu kuivuu ja seuraa-vassa vesillelaskussa vene täyttyy vedellä." ...(cont.)	As it is now Nyman's boat does not suit to Central Europe, however. There the boats are retained mainly on the spindle because there are scantily <i>boat places</i> in water. "If it lifts the present boat always after the vesireissu to the spindle, the tree will dry and in the following launch the boat will become full of water."(cont.)	As it is now Nyman's <i>boat is not suitable</i> , however, <i>for</i> Central Europe. There the boats are <i>kept</i> mainly on the spindle because there are scantily <i>places for boats</i> in water. "If one lifts the present boat always after the <i>boat trip</i> to the spindle, the tree will dry and in the following launch the boat will become full of water." ...(cont.)

Table 1: Example text in window evaluation

The polished rough translation in Table 1 does not yet represent the final word of TranSmart, since certain phenomena (such as the proper assignment of articles or the proper ordering of adverbials) have not been handled conclusively yet. There are several such errors in the polished rough translations. We already mentioned that translation errors of sentence #7 are mainly due to the parsing errors. Notice how elliptic headings may get awkward translations (#11, #21). In Finnish the word *puu* means tree, wood, or timber, depending on context. In sentence #30 a wrong translation is chosen (*the tree will dry...* should read *wood will dry...*). This error cannot be corrected by the general method and the error is therefore left as its is.

Corrected word specific translations are shown in Tables 2-4. Table 4 does not show semantic information. Nouns can be semantically classified, and this typology can be used in selectional restrictions for example in choosing correct verb configurations.

#	Source	New target
1	säksätys (informal)	clatter
2,10	yhdistää	combine
6,...	aurinkovene	solar boat
15	vauhti(Advl, Ad)	at speed
25	digitaali (rare as a noun)	digital
29	venepaikka	place for boat
30	vesireissu (informal)	boat trip

Table 2: Added or corrected domain specific word or preposition translations

#	Source	New target
4	saasteettomasti	without pollution
27	ympäristöystävällisyys	kindness to the environment

Table 3: Added or corrected generic word or preposition translations

#	Source	New target
6	työstää smth1(Obj) smth2(Advl,Transl)	modify smth1(Obj) into smth2(PComp)
14	työntää smth(Obj) piilo (Advl, Ill)	hide smth(Obj)
15	QuantNoun(QuantAttr) tuntinopeus	the speed of QuantNoun(PComp) per hour
16	olla syntyä (Advl,IIIinf,In) smth(Advl,III)	be (Progr) born in smth(PComp)
16	kantautua smth(Advl,III)	reach smth(Obj)
22	olla esillä (Advl) smth(Advl,All)	show (Pass) in smth(PComp)
28	kelpaa smth(Advl,III)	be suitable for smth(PComp)
29	säilyttää smth1(Obj) smth2(Advl,Ad)	keep smth1(Obj) on smth2(PComp)
32	smth(Advl,All) olla hinta (Subj,Part)	smth(Subj) be expensive (Compl)

Table 4: New or corrected generic configurations

The text revealed also two structural errors. In sentence #20 the surface ordering of adverbials was corrected. In Finnish the passive voice indicates an unknown actor. The passive is usually indicated morphologically with a passive morph in the main verb. An unknown actor can be indicated syntactically using a 3rd person singular finite verb without a subject. Both structures are translated into English using either an English passive structure or using one of the formal subjects *it*, *one*, or *they*. The rough rough translation of sentence #30 opts for the incorrect formal subject *it*.

The distances between the rough rough and the polished rough translations are shown in Table 5. A distance is a real number between 0 and 1. It is calculated simply by dividing the number of sentences requiring corrections by the total number of sentences.

Correction type	Domain specific	Generic	Combined
New words or prepos	4	2	6
Corrected words or prepos	3	0	3
New configurations	0	3	3
Corrected configurations	0	6	6
New structural rules	0	1	1
Corrected structural rules	0	1	1
Lexical distance	0.23	0.33	0.5
Structural distance	0	0.07	0.07

Table 5: Quality distance of the example text

The figures in Table 5 are not informative for the casual reader. For a system engineer they provide relatively straight-forward data on progress. When quality distances are measured regularly, using similar, general text type, the numerical generic distances should demonstrate a decreasing tendency. Since structural distances are further divided into subclasses at Kielikone (see Figure 3), the numbers tell how mature the different transfer parts are and where attention should be focused.

Window evaluation can also serve a potential customer's needs. Both rough rough and polished rough translations are made for a representative piece of the customer's text. When the customer sees both translations and the domain specific distance figures, he or she gets at least an approximate idea about the maturity of the system and its potential performance for his/her text type. The domain specific lexical distance gives an estimate about the amount of lexical work required by the text type before the system can be used in production.

Window evaluation may give data also for technology evaluation. Assume that an MT technology has been used for building an MT system for the language pair a-b and the question is raised how good a quality that technology would provide for another language pair c-d. Polished rough translations show the final quality for the pair a-b for a given text type. If a and b are structurally more distant from each other than c and d, window evaluation tells indirectly how good a quality is achievable for the pair c-d for a similar text type. If, on the other hand, c and d are more distant from each other than a and b, window evaluation remains silent.

Window evaluation does not tell anything about the absolute translation quality of the system. This evaluation method provides only relative figures about the maturity of a given system. The method yields also samples of the final quality, and these samples can be subjected to any of the evaluation methods proposed in the literature of the field in order to get an estimate of the absolute final quality. Informally, these samples tell an informed reader something definite about the quality of the system.

4. Niche Economics of MT

Our discussion has had a slightly negative bend so far since it has mainly dealt with the limitations of MT. This final section straightens things up and ends the discussion with an upbeat note. In search for a better term, we call sudden new economic opportunities opened by the use of a totally new technology **niche economics** (Church and Hovy use the term for a different purpose). A new technology may pave the way for commodities of a totally new kind which consumers have not even

dreamt of. If ordinary product development is driven by demand pull, the products in niche economics, if they take customers by surprise, manifest technology push.

We believe that MT may yield technologies which create their own niche economics. Translation services, as they are currently offered, are greatly limited by the physical limitations of the human body and its central nervous system. Human translations are slow and often inconsistent. A human translator gets easily fatigued and his or her memory is imperfect. Human translation requires the presence of the whole bulky human body and functions only under restricted environmental conditions. Moreover, privacy is breached when a text is translated by another human. Of course this last problem can be legally solved by writing binding agreements, but nevertheless the fact that somebody to whom the text was not intended reads it may in some cases be so great an obstacle that translations are avoided. If, on the other hand, translations are performed by machines, people probably would not feel their privacy threatened.

MT is free of all these drawbacks. It is fast and consistent, a machine never gets fatigued and it has perfect memory. LSI circuits fit in a small space and function in a great variety of environmental conditions. And machines do not breach privacy. But, of course, quality has been quite low at least so far. We believe that the MT technologies already available may be good enough to produce MT products for niche economics. In particular, we believe that KMTech is such a technology.

There are certain products and services which for different reasons are beyond the capabilities of human translators but which MT might deliver in the near future. Interpreting telephone, portable interpreting aid, translation of e-mail messages, translated newspaper, translating fax machine, and translating copying machine are examples of such products or services. As is well known some of these products or services are already under R&D in many countries.

There are probably many more to come that we cannot yet even think of. It seems that discussions of MT economics have often been too preoccupied with short term economics and comparison with human translation. The real strength of MT will be seen in the longer run, in its niche economics, where MT does not compete with humans but turns out to be the only choice.

References

- Arnola, H., On parsing linear dependency structures deterministically in linear time. Manuscript, 1996.
- Church, K., and Hovy, E., Good applications for crummy machine translations.
- Dowty, D., Wall, R., and Peters, S., *Introduction to Montague Semantics*. D. Reidel Publishing Company, 1981.
- Falkedal, K. (Ed.), Proc. of the Evaluators' Forum. ISSCO, 1991.
- Flanagan, M., Error classification for MT system evaluation. Proc. of the First Conference of the Association for Machine Translation in the Americas, 1994.
- Ishara, H., Uchino, H., Ogino, S., Okunishi, T., Kinoshita, S., Shibata, S., Sugio, T., Takayama, Y., Doi, S., Nagano, T., Narita, M., and Nomura, H., Technical evaluation of MT systems from the developer's point of view: exploiting test sets for quality evaluation. Proc. of the First Conference of the Association for Machine Translation in the Americas, 1994.

Juntunen, J-P., Intelligent translation memory. M.Sc. Thesis, Helsinki University of Technology, 1996 (in Finnish).

Jäppinen, H., Kulikov, L., and Ylä-Rotiala, A., KIELIKONE Machine Translation Workstation. Proc. MT Summit III, Washington, DC, 1991.

Jäppinen, H., Hartonen, K., Kulikov, L., Nykänen, A., and Ylä-Rotiala, A., KIELIKONE Machine Translation Workstation. In Nirenburg, S. (Ed.), Progress in Machine Translation. IOS Press, 1993.

White, J., O'Connell, T., and O'Mara, F., The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. Proc. of the First Conference of the Association for Machine Translation in the Americas, 1994.

Use and Value of Computer-Assisted Translation in the Central Translation Service of Coop Switzerland, Basle

Martha Ebermann

Abstract

Since the beginning of the 1990s, the translation department of Coop Switzerland has had to cope with rapid growth in the volume of orders. This has given rise to two groups of problems: growth of "corporate language", and the need to rationalise and control costs. Coop's solution to these quantitative and qualitative difficulties was to set up its own data base of Coop terminology and resort to the computer-supported processing of certain categories of text.

These adjustments in working methods led to a "cultural change" in the translation department. This phenomenon is the subject of this paper.

Martha-Mariette Ebermann

Is the Director of the Translation Service of Coop Schweiz. Studied Germanic and Romanic languages at Salzburg University. MA in philosophy. For 9 years, German lecturer at the Universities of Reims and Mulhouse. Graduated in Germanistic Linguistics at the University of Besançon. Taught German as a Foreign Language and French for Finance at a private school in Germany. Became a translator at Coop Schweiz in 1981 and director in 1989.

Coop Schweiz

Coop Schweiz is the services branch of the Swiss Coop Group which is made up of 17 regional associations and 24 production, commercial and service companies. It is the second largest distributor in Switzerland, with a market share of 13,9%, a turnover of 11 billion, 1 000 shops of various sizes and about 46 000 employees.

The **Translation Service** of Coop Schweiz translates into the various languages of the Swiss cantons. Organised as a cost centre, it houses 8 French translators, 3 Italian translators and contracts several internal and external free-lance translators. It has a terminology service and uses translator's workbenches as a computer aided translation service.

It has a Multiterm-based termbank, uses various dictionaries and encyclopaedias on CD-ROM, as well as on-line termbanks. It has its own library and documentation service.

Martha-Mariette Ebermann

Dornbachstrasse 23

4053 - Basel, Schweiz

Tel: +41 61 336 6964, Fax: +41 61 336 69 55

E-mail: martha.ebermann@coop.ch, 100625.2021@compuserve.com (private)

(Outline of the Presentation)

Switzerland - a country of institutionalised translation

- Multilingual nature of public life guaranteed by the constitution
 - hence
 - product information
 - advertising
 - communication
 - business documents
- In three national languages: German, French and Italian

The Coop Group in Switzerland

- Turnover of CHF 11 billion, approx. 1 100 sales outlets
- 1.3 million member households
- 17 regional cooperative societies
- 24 affiliates (manufacturing, trade and services)
- Coop Switzerland is the umbrella organisation of the Swiss coop societies

Translation at Coop Switzerland

- Central Translation Service: profile and goals
 - 10 translators and 1 terminologist
 - secretariat and management
- Specifics of Coop Switzerland
 - oriented to the Swiss market
 - active in trade, manufacturing and services
- Organisation as cost centre
 - economic significance
 - value of work provided
 - cost-to-performance ratio

Requirements of translation

- Speed (shortest possible time)
- Efficiency (methods commensurate with desired result)
- Consistency (terminological, textual etc)
- Readability (communicative quality of texts)
- Accuracy (e.g. requirements relating to product liability)

Work analysis 1993

Conducted by Coop Switzerland's Internal Controlling

Investigation

- Organisation of work
- Quality
- Quantity
- Speed
- Flexibility
- EDP compatibility
- In-house/outsourced ratio

Measures

- Organisation as cost centre
- Streamlining of procedures
- Orders to be processed via LAN
- Creation of terminologist's post
- Creation of Coop terminological database
- Introduction of computer-assisted translation at all workplaces

Development of work methods

- Typewriter, paper, pens, dictionaries (until 1983)
- Word processing, dictionaries (until 1993)
- Word processing, terminology database, computer-assisted translation, CD-ROM, on-line services (since 1993)

Change in work methods (resource management)

Efficient use of available resources and capacity, i.e. avoiding

- duplication of translation work
- wasteful use of time and resources

Change in work methods (providing customer satisfaction)

Inconsistencies in repetitive text (packaging, advertising etc), existence of different translations for the same text

- impairs understanding (intertextual references are lost in the translation process (example: product lines, NP)
- undermines the creation of a "textual" image of the organisation

Change in working methods (quality management)

Quality assurance

- accuracy and consistency of terminology
- intertextual consistency
- the means used must be appropriate to the text to be translated (quality at a reasonable price)
- text appropriate to its purpose (product liability)

New tools

- Multiterm 95 Professional Plus from Trados GmbH
- Translator's Workbench 1.07 from Trados GmbH
- CD-ROMs, on-line services

Efficiency in translation I

Method must be adapted to the type of text

Efficiency in translation II

Criteria for using computer-assisted translation:

- repetitive texts (significant amount of repetition within one and the same text)
 - manuals (IT, etc)
 - instructions (e.g. for different items in a line of electrical appliances)
- multi-version texts
 - administrative planning documents subject to periodic revision
 - manuals subject to periodic revision
 - collective agreements
 - statutes etc.

Efficiency in translation III

- standardised texts (recipes, packaging texts, manuals, instructions, management documents, plans, diagrams)
- texts with frequently recurring terms (do-it-yourself/building centres)

Reshaping of the translation process

I Preparation and follow-up

- Selecting the translation method according to the nature and type of text involved
- Preparing the document for computer-assisted translation (formats, abbreviations, standard text, analysis and/or Translate function)

- Post-translation editing of documents and updating databases

II Work method: The translator

Selecting the most effective work method according to the principle of “leaving all repetitive work to the machine”

- interactive (manual or semi-automatic)
- automatic adaptation during computer-assisted translation (especially with long texts)
- traditional

General principles

Working principle I: suitability of text

Only suitable texts are selected for computer-assisted translation

Working principle II: Approach orders from a long-term viewpoint

Losing time in order to gain time = cutting costs!

Working principle III: Consistency in translation memory

Ensuring the consistency of texts in the translation memory = creating a shared knowledge base for future translations. This means avoiding any inconsistencies arising from translators' personal ambitions.

Results as per 30.6.96:

Costs/performance

- Achievement of departmental target: 75%
- More than a 30% improvement in output while personnel costs and billing rates remain unchanged
- Substantial increase in number of translation orders (especially in the IT field) because of efficient approach used

Translator response

- High level of acceptance of new CAT tools after initial scepticism
- Terminology database (15 000 entries) and translation memories (French: 107 000 entries and Italian: 50 000 entries) are treated as a shared activity that merits careful attention
- No inappropriate “creativity”

Quality of texts

- High level of consistency in CAT texts
- Translation memories as terminology resource

- Long texts easy to divide up among several translators
- Preparation of texts for external translators

Infrastructure requirements

- Strengthening of IDP support for translators
- Organisation of support for network and database
- Organisation of more intensive Windows, Word and network training
- Organisation of comprehensive servicing of tools, translation memories and terminology databases

Conclusions

- Reduction in relative translation costs
- Establishment of quality standards in the terminological and the intertextual domain
- Less duplication of effort
- Additional benefits for the translation service's customers
- Reduced loss of know-how on departure of staff

Machine Translation, Terminology and the African Languages in South Africa: An Overview

Milde Jordaen-Weiss

Abstract

The South African language policy changed in 1994, increasing the number of official languages from two to eleven. These are Afrikaans, English, isiZulu, isiXhosa, isiNdebele, siSwati, Sesotho, Sepedi, Setswana, Tshivenda and Xitsonga. IsiZulu tops the list by being the home language of more than 20% of the population.

In South Africa, where many opportunities for translators exist, terminology forms an important link in the translation chain. The National Terminology Services (NTS) is the only national office for terminology work in this country. Co-operation between MT developers and the NTS is probable in future, concerning the verification of terms in the African languages.

The company EPI-USE Systems is the only local developer of MT in South Africa. It has close ties with the University of Pretoria and offers inter alia a full translation service. Their product Translator Professional contains ten language pairs, four of which are African languages officially used in South Africa. All of these can be linked to optional domain dictionaries containing various terminologies. The product will shortly be available in Europe.

Milde Jordaen-Weiss

Milde Jordaen-Weiss graduated from the University of Pretoria, where she is presently completing her doctorate in History. She started working at the NTS in 1992 in the Section for Systems Development and Research and is currently an Assistant Director and head of the Section for Computer Facilities.

National Terminology Services (NTS)

The NTS is part of the South African Central Government's Department of Arts, Culture, Science and Technology. It facilitates the development and modernisation of technical and scientific terminology in all the official languages. Twenty-eight terminologists handle the excerpting, creation, documentation, standardisation and dissemination of terminology. The management of the National Termbank, which contains approximately 600 000 bilingual entries, is also the responsibility of the NTS.

Ms Milde Jordaen-Weiss
Head: Computer Facilities, National Terminology Services
Department of Arts, Culture, Science and Technology
Private Bag X894
0001 Pretoria
Republic of South Africa
Tel +27 12 314 6165 Fax +27 12 325 4943
EPI-USE's Web page: <http://www.epiuse.co.za>
NTS e-mail: mj-weiss@acts2.pwv.gov.za

1 Introduction: South African Languages

1.1 History of languages

Two official languages were recognised when the Union of South Africa was proclaimed in 1910. These two, English and Dutch, reflected both the composition and the history of the white part of the population. Afrikaans, the only Western European language indigenous to Africa, which replaced Dutch as an official language in 1925, developed from 17th century Dutch, influenced by inter alia Khoisan, French, Portuguese, English and Malay. English came to the country at the beginning of the 19th Century when the British took over the Cape. As a world language and the language of the conqueror, it quickly took root in this country.

In 1994 a new political dispensation changed the language policy in South Africa. The nine Bantu or African languages indigenous to South Africa also became official. They can be divided into four different families. In the Nguni family we find isiZulu, isiXhosa, isiNdebele and siSwati. The Sotho family includes Sesotho, Sepedi, and Setswana. The two other languages, Tshivenda and Xitsonga, form families on their own.

1.2 Statistics

IsiZulu is the home language of the largest group of people in South Africa. It is spoken by 8,5m people or 21.6% of the population. Second is IsiXhosa with 6,8m mother tongue speakers (17.4%) and third Afrikaans, which is the home language of approximately 6,1m people (15.6%). Next in line is Setswana with 3,6m (9.1%) and Sepedi with 3,4m (8.7%). English is spoken at home by less than 3,4m (8.6%), Sesotho by 2,6m (6.7%), Xitsonga by 1,3m (3.5%), Siswati by 926,000 (2,3%), IsiNdebele by 799,000 (2%) and Thivenda by 763,000 (1.9% of the population).

1.3 Other languages

There are other languages spoken in South Africa in an unofficial capacity. There are various communities using their own languages, like Germans, Portuguese and Greeks, while languages like French, Spanish and Dutch are spoken as well. There is also a strong Asian presence in the country. Some of these languages are taught at schools and universities.

2 The National Terminology Services as facilitator of terminology development in South Africa

The availability of terminology in various subject fields comes to mind when one talks about translation and Machine Translation (MT). Since the fifties, terminology development in South Africa's official languages has been the mission of the National Terminology Services (NTS). This office falls under the jurisdiction of the Department of Arts, Culture, Science and Technology of the Central Government. Until 1994 the NTS only attended to the two official languages, Afrikaans and English. Since then, the nine African languages mentioned above were added.

The NTS excerpts, documents, creates, standardises and disseminates terminology. By assisting terminology projects from outside collaborators as well, it facilitates terminology development on a national level. Although it is geared to handle any subject field, topics related to education, training, health, building, etc. receive

priority attention at present. Of the 23 terminologists working on projects, ten are mother tongue speakers of the different African languages.

3 The role of the NTS in MT development

The primary function of the NTS is terminology development and dissemination. However, as national body in this field it is logical that terminology verified and disseminated by the NTS will be used when MT programs involving technical and scientific domains are developed.

As MT development in South Africa is fairly recent, the exact extent to which the NTS will be involved in this development still has to be determined. As part of the Civil Service, giving free information and assistance to the public, terminology for MT development will have to be made available to all interested parties. At present, only one company has shown an interest in making use of this service.

4 EPI-USE Systems as developer of Machine Translation (MT) Software

At the University of Pretoria a research group existed at the Department of Computer Science. One of its projects was MT. About 15 years ago, a private company, called EPI-USE Systems, was formed by some of the lecturers involved in this research.. Contact between the NTS and this company dates back many years, when the NTS developed its own database program for the documentation of terminology in the official languages, which were then Afrikaans and English. This program, called Lexikon, was distributed free of charge to clients, and EPI-USE Systems considered using this database for its MT software.

This company offers various services, including translation, Network Management and SAP (R3) consulting. The component handling translation is called the Lexica Division and the software they use was developed by the company itself. It was called Lexica at first, but recently changed its name to Translator Professional. Various Windows packages are being marketed. Translator Professional has a stand-alone version using ASCII text and an integrated version to be used with MSWord 6. A scaled down version of the program is available (Translator), as well as a bilingual glossary for Afrikaans and English, called Glossit.

The minimum hardware requirement for using the program is a 486DX2-50 with 8Mb of RAM, but a Pentium with 16Mb RAM is recommended.

Translator offers the following language pairs in the European languages:

English to Afrikaans/ Afrikaans to English (50 000 concepts)

English to German/ German to English (60 000 concepts)

English to French/ French to English (90 000 concepts)

Portuguese to English (70 000 concepts)

Spanish to English (25 000 concepts)

Italian to English (35 000 concepts)

Of special interest are the African language pairs that have been developed. Swahili, although not an official language in South Africa, is one of the better developed Central African languages. Translator offers Swahili to English with 25 000 concepts. Of the official South African languages, the following are available:

Setswana to English (20 000 concepts)
Sepedi to English (20 000 concepts)
English to IsiZulu (45 000 concepts)
English to IsiXhosa (40 000 concepts)

The development of the language pair Sesotho to English is in progress, and English to Sesotho will probably follow.

Several bilingual domain dictionaries (Afrikaans and English) are available optionally for use with the MT software. These domains are Industrial, Office, Military and Academic. Within each domain, several subject fields are covered, for example the Industrial domain includes a Technical (30 000 concepts), Electrical (3 000 concepts), Chemical (7 000 concepts) and Engineering (9 000 concepts).

Although the cost of development is R100 000 (approximately \$25 000) per language direction, the company markets its products at a competitive price. Translator Professional will be marketed in Europe shortly, where it will probably sell for about \$100. The lighter Translator version, which is only available in South Africa, sells for about \$76. Training for these products is done in a 5 hour session at additional cost.

5 Proposed co-operation between the National Terminology Services (NTS) and EPI-USE Systems

In its quest for a new Terminology Management System (TMS) able to handle eleven languages, the NTS realised that such a TMS would probably have to be imported. As it turned out, the German program MultiTerm, developed by Trados of Stuttgart, satisfied most of the NTS requirements. Local vendor support was an important factor in the final decision. As Trados had no local representative, the NTS considered EPI-USE Systems to fulfil this role. EPI-USE Systems imported the software (which the NTS as a Civil Service organisation was not allowed to do directly), installed it and provided training for all staff members.

On an MT level, EPI-USE Systems would like to co-operate with the NTS as well. As the leader in terminology work in South Africa, the NTS will probably in future verify African language terms documented by the EPI-USE Systems team. In return, the NTS will have the use of these documented terms for further processing.

6 The future

The demand for translations into the African languages is increasing. The NTS, with its limited staff resources, is currently working on a large number of terminology projects in all the official languages. These terms will eventually be accessible via electronic medium. However, if the standardised terms could also be incorporated into domain dictionaries linked to MT software, the dissemination of African language terminology to translators and other users will be facilitated. We all realise that MT is not an instant solution to all translation problems. However, in the current South African language environment MT into the African languages could be a tremendous help in promoting multilingualism.

Bibliography

WEBB, V (red) Afrikaans na Apartheid. Pretoria, 1992
1VN Webb (red). Afrikaans na Apartheid, p 28

Session 5: Machine Translation vs. Translation Memories: Rivals or Partners?

Chair: Dr. Jörg Schütz

Dr. Jörg Schütz

Jörg Schütz studied Computer Science, Mathematics and Medicine (Diplom-Informatiker) at the University of the Saarland from which he also received his doctoral degree (MT/CL and AI). Since 1985 he is working for the Institute of Applied Information Sciences (IAI) in Saarbrücken where he is responsible for the institute's R&D as deputy director. He was/is the project leader or supervisor of several research and development projects in the field of multilingual language technology and information technology. He also teaches at the University of the Saarland and acts as a consultant for industrial companies. He is a member of the EAGLES group on Evaluation and Assessment. His current scientific interest is information engineering, in particular the integration of language technology in network-based applications.

IAI, the Institute of the Society for the Promotion of Applied Information Sciences at the University of the Saarland

The IAI, the Institute of the Society for the Promotion of Applied Information Sciences at the University of the Saarland, is a non-profit research and development organisation founded in 1985 as a university related research institution by the Society for the Promotion of Applied Information Sciences (GFAI) in Saarbrücken, Germany. Its charter requires the institute "*to work in the public interest, to promote and foster the application of information and communication science in the development of commerce, trade and industry, ...*". The IAI was established to carry out R&D in multilingual natural language processing, machine translation, and other areas of information sciences and information technologies. In this context, IAI promotes the application of findings from the scientific research area in industry, scientific research institutes and authorities within application projects.

Dr. Jörg Schütz

IAI, Martin-Luther-Strasse 14, D-66111 Saarbrücken, Germany

Tel: +49 681 38951 32, Fax: +49 681 38951 40

E-mail: joerg@iai.uni-sb.de

Url: <http://www.iai.uni-sb.de>

Use of Linguistic Resources like Translation Memories in Machine Translation Systems

Lee Humphreys

Abstract

The context and general technical strategy for MAHT and MT at ERLI is presented. After presenting the commercial context of translation, we outline the architecture of AlethTR, ERLI's assisted-translation platform.

Special attention is paid to the question of linguistic resources : general lexicon, terminology, translation memory. We consider briefly the integration of translation memory with MT and its extension to the sub-sentence level.

Lee Humphreys

Lee Humphreys currently heads the Grammar group in the Research and Development Department of ERLI. He is also involved in the design of translation systems. He was previously responsible for French-English MT development at SITE-EUROLANG (Paris) (1992-1994). Prior to this he was part of the CL/MT Group in the Department of Language and Linguistics at the University of Essex (UK), working on MT Evaluation and linguistic specifications in the EUROTRA project.

GSI-Erli

Created in 1977, ERLI is Europe's leading language and document engineering company. With a turnover of MFF 31 in 94, the company has a team of 60 computational linguists and software engineers. The principal areas of activity are natural-language based indexation and document retrieval (AlethIR), text generation (AlethGen), controlled-language checking (AlethCL), translation (AlethTR), linguistic knowledge manipulation (AlethKES), terminology management (AlethGT) and general lexical management (AlethGD).

Lee Humphreys

ERLI

1 place des Marseillais

F-94227 Charenton-le-Pont CEDEX

France

Tel: +33 (01) 48 93 81 21, Fax: +33 (01) 43 75 79 79

E-mail: info@erli.fr

WWW: <http://www.erli.fr>

Introduction

AlethTR™ is a translation support tool from ERLI that integrates

- bilingual terminology management
- identification and default translation of candidate new terms
- translation memory (TM)
- light machine translation (MT)

together with text handling, project costing software, translation utilities and user interfaces.

As has become standard in translation engineering, the MT engine is used to translate - if required - text which is not found in translation memory. The grammar of the current engine can be customised by ERLI to suit the particular type of text routinely handled by a customer.

Translation Context

ERLI's translation tool engineering effort has historically concentrated on supporting high-volume professional technical translation. It is worthwhile for the translator to invest a considerable amount of time and effort in a project preparation phase since this is recouped by improved quality and speed during the translation phase.

A project typically involves from one hundred to several thousand pages, where all the texts in the project belong to the same domain and text type. The French texts we have looked at have syntactic and lexical restrictions characteristic of sub-languages:

restrictions on verb form

Verbs restricted to 3rd person, mostly present tense, limited use of past and future, perhaps complete exclusion of the French passé simple. Interrogatives are restricted to very specific document parts such as fault-finding e.g.

Is the main circuit stop valve closed?

anaphora

Pronouns almost always represent things rather than persons

restrictions on relatives

Object relatives are rare

determination

Greater use of zero determination than in standard French e.g. before a deverbal noun e.g.

Avant montage du joint, nettoyer ...

Sentence and part sentence repetition rates are high, often because they correspond to standard warnings or advice, or because they describe frequently repeated operations / states. For example, we have found in a series of related projects for the same customer (same domain, same text type) that typically half the text is covered by exact matches in the TM.

General vocabulary is nearly closed and polysemy reduced. For example, whilst a French dictionary may give several senses to the pronominal verb *s'afficher*, the sense meaning *flaunt* as in

Elle s'affiche partout

is most unlikely in a technical text.

Translating with AlethTR™

In this section we rapidly sketch the various steps involved in a translation project with AlethTR™. The object is to highlight the tools available and their contribution to the translation process.

Project Preparation Activities

Validate terminology

One of the first thing a professional translator does before tackling a translation project - be it machine-assisted or otherwise - is to draw up a list of the terms in the source text and their translations.

Although the lexicon of AlethTR™ may contain some appropriate terms at the start of the project, their translation must be validated. AlethTR™ lemmatises and syntactically tags the source text, identifies known terms and presents their translations for validation.

Identify potential new terms

Even for a familiar domain, the texts in a new translation project are likely to contain new terms. AlethTR performs a syntactic analysis of the text and uses a template-based search algorithm on the noun phrases in the resultant structure to allow identify instances of candidate multiword terms. For example, a template for French might be

N de/à N

as in *bac de récupération* or *filtre à huile*. Instances of candidate terms found in the text e.g. *bacs de récupération*, *filtres à huile*, are normalised to standard term form and statistically processed, taking into account such factors as the frequency of the candidate term in the text(s). The translator validates the resultant list.

Find translations for new terms

default term translation rules

Many new terms entering into a domain tend to translate compositionally e.g.

$N1 \text{ de/a } N2 \Rightarrow N2 \text{ } N1$

AlethTR™ exploits such rules to propose translations for the candidate terms it has found in the source text. The proposed translation can be modified before being entered into the term translation dictionary.

statistical identification of candidate translations

In another approach currently under development we align source and translation in the customer's legacy corpus. Statistical techniques find segments in the translations which correspond to terms in the source text. Although computationally quite heavy,

this approach has the advantage that the target terms identified need not be compositional translations. (Not yet available in the standard AlethTR™ product.)

Identify repeating text

AlethTR™ identifies repeating text segments in source text. It can provide a default translation for repeating sequences, using existing translation memory, terminology translation and the translation engine. The sequences with their translations are made available to an editing interface, allowing the translator to select appropriate sequences and to edit the proposed translation. The result is used to create a project Translation Memory.

Identify unknown words

AlethTR™ signals the presence of unknown words and attempts to predict their category. The translator can enter the words in the dictionary and supply an appropriate translation.

Identify possible translations for general language words

The general language dictionary can be organised to allow the user to partition translations on semantic grounds. A nice example of this is the French word *fraise*, which can be translated as strawberry (fruit) or reamer (engineering tool). Since it is possible to arrange for (say) fruit and vegetables to be grouped into a sublexicon, inappropriate references to strawberries can be avoided in engineering text translations by selection of the appropriate sublexicon.

However, no amount of chopping and changing of sublexica will entirely eliminate the problem of multiple translations for general language content words. Rather than trying to calculate very elaborate lexical selection constraints in the translation engine, AlethTR™ encourages direct control by the translator during the preparation phase. Working from a lemmatised and tagged internal representation, AlethTR generates a list of all the general language word types in the text, together with examples of source text context and their possible translations as given by the bilingual dictionary. It is very easy to rapidly scan through this list and indicate to AlethTR™ the preferred translations.

Of course, the preferred translation of a general language word can easily vary from one context to another. However, recall that in technical texts the number of senses (and hence translations) of a word is usually limited with respect to general language, and that often a potentially polysemous word turns out to have only one sense in a particular text(s). Hence this translation preselection step turns out to be surprisingly cost-effective. (Surprising, that is, for MT professionals, who are used to devoting a great deal of effort to the lexical selection problems posed by general purpose translation).

Costing

Using a sample text(s), and various user definable cost parameters, AlethTR generates a fully detailed quotation with estimates for the contribution of Translation Memory to the projected project. This allows an immediate assessment of the cost effectiveness (which is closely related to the repetitiveness of the documents).

Translation and Revision Phases

After preparation, translation can involve

- Translation of terminology alone
- Translation using exact-match or exact- and fuzzy-match translation memory
- Translation using MT

The result is made available in a revision/translation interface, which displays the source text and its translation (line by line). Colour codes and underlining highlight client terms, base terms, terms with several translations, terms with usage notes, unknowns, phraseology, exact match TM, fuzzy-match TM and MT. This interface allows entirely free navigation and editing i.e. unlike with some other products, the user is not required to translate/revise in a strict sequence from the top to the bottom of the file.

Update of TM

When the translation has been revised and corrected, AlethTR creates an extension to the Translation Memory. The compiled TM can be used for other texts in the project or in other projects.

Billing

At the end of the project, AlethTRTM generates a customer bill.

Directions in Linguistic Resource Management

AlethTRTM has a linguistic knowledge base (monolingual and bilingual vocabulary together with terminology information) which can be directly inspected and modified by the user.

The AlethTRTM linguistic knowledge base fits into a high-level approach to lexicographic resources at ERLI. Our general strategy involves the creation of very rich lexical resources in a generic lexicon model - GENELEX. A powerful lexicon manager and lexicographer workbench - AlethGD - provides

- OO management of large lexical resources according to the GENELEX model
- graphical display and navigation of lexical information e.g. semantic networks
- special facilities (including a linguist-usable OO dictionary programming language) for
 - importing dictionary data in virtually any external format
 - exporting dictionary data to
 - * ERLI applications
 - * other end users or dictionary managers

All general language lexical resources are created, validated and maintained in the GENELEX format by the Dictionary team. ERLI applications such as AlethTR, AlethCL (a Controlled Language checker), AlethGEN (a language generator) and AlethIR (a natural language front end to document management systems) call upon the same GENELEX resource via specific export programs. The approach allows us

this one source to supply linguistic processors which use completely different underlying grammatical theories such as

- LFG
- Meaning-Text Theory Dependency Grammar
- Simple Constituency Grammars

and so on

GENELEX

The GENELEX model (MENON 94) is extensively described in the three reports (Consortium 93a; Consortium 93b; Consortium 94) and will not be presented in detail here. With relatively minor modifications, this model forms the basis of the EC's LE PAROLE project, which involves the construction of lexica for 20,000 or more Ums in 12 European languages. Given the level of activity based on this model, it can be considered to be a de facto European standard.

The model is intended to be multi-theoretical, providing a descriptive formalism which allows linguistic facts to be drawn from different linguistic theories. The GENELEX model provides for a very high level of factorisation of linguistic information.

There are three layers of representation : morphology, syntax, and semantics. The morphological layer relates a given morphological unit (a UM) to its variant written forms (UMGs) e.g. *bosun* and *boatswain* are variant spellings of the form *boatswain* in English. A given UM is associated with a inflection paradigm.

A given UM has one or more USYNs - syntactic units. Each such unit describes a single syntactic behaviour of the UM. Typically a verb UM e.g. *tackle* might have several USYNs, each USYN describing a particular syntactic complementation pattern for that verb. (The internal structure of a USYN is quite rich, and allows full description of local trees. Interestingly, the described element does not necessarily have to be the root of the local tree.)

Finally, a given USYN can have one or more semantic units (USEMs), where each semantic unit corresponds to the notion of a word sense. Since a given sense can have more than one syntactic realisation, a USEM can itself point back to more than one USYN. A USEM can be linked to a predicate-argument structure. USEM-USEM links allow the construction of a classic semantic network.

The multilingual part of the GENELEX model (Consortium 95) establishes reversible bilingual links between two monolingual models : these links can be at the usyn level, at the usem level, at both usyn and usem level, or at the predicate level.

Terminology - TRANSTERM

The GENELEX model was particularly intended for the representation of general language. An extension of GENELEX which adds further information for representing the usage of terms was developed in the context of the TRANSTERM project (EC LRE project TRANSTERM - Creation, reuse, normalisation and integration of terminologies in natural language processing systems). The TRANSTERM approach allows a terminologist or ordinary user - who does not

necessarily have the theoretical linguistic expertise of an GENELEX / AlethGD user - to enter terminological data without entering into the full linguistic richness of the GENELEX model.

Just as the GENELEX model has its associated lexical DBMS - AlethGD - intended for specialists, so also TRANSTERM has a corresponding DBMS - AlethGTTM - intended for use by customers. Like AlethGD, AlethGTTM is based on an OO DBMS with specialised navigation and lexicographic facilities.

AlethGTTM also stocks an entire GENELEX-conformant general language dictionary. When the user creates or modifies his/her terminology, AlethGTTM automatically hunts down in the GENELEX part of the dictionary all the morphological, syntactic and semantic information associated with the component parts of the term. This information is essential for the operation of linguistic processors such as AlethTRTM. The AlethGTTM terminology manager hides all this complexity from the user.

Lifecycle aspects

In the design of the TRANSTERM model and the resultant DBMS AlethGTTM, particular attention was paid to the problem of ensuring consistency between customer supplied and managed terminology, on the one hand, and ERLI supplied and maintained general language dictionary information on the other. The customer manages his/her terminology from day to day: at periodic intervals, ERLI or other third-parties might issue new and extended versions of the GENELEX general lexicon. Using the AlethGTTM terminology manager, the customer can download this new dictionary version - thus improving the performance of the linguistic processor - with minimal disruption to his/her terminology.

AlethGTTM is just coming on stream and will be progressively integrated with ERLI products.

Linguistic Knowledge Extraction - KES

As we have seen, AlethTRTM provides a number of tools to aid the translator in the preparation of the lexicon for a project e.g. tools for the identification of new terms. These tools are frequently used on a historical corpus for a given client.

Extraction of lexicographic and grammatical information from corpora is an activity carried out in the context of all types of NLP applications and products, not just translation (OGONOWSKI 94). For example, the construction of an appropriate thesaurus is an important prelude to optimised domain-specific information retrieval with AlethIR.

ERLI's AlethKESTM responds to this need, integrating facilities for corpus analysis and the manipulation of hypotheses. One typical use is to allow a terminologist to progressively construct a conceptual network appropriate for a particular domain. Starting from initial hypotheses, the structure of this network can be updated, modified or completely restructured in the light of additional information or further insight. The resultant network could be used as a basis for lexicographic coding in AlethGTTM. (AlethKESTM is very closely coupled to AlethGTTM and shares the same underlying software architecture.)

It is intended to progressively integrate in AlethKESTM a variety of specific corpus exploration tools, including, for example, the repetition analysis tools currently found in AlethTRTM.

Directions in Machine Translation and Translation Memories

ERLI is developing a new robust general purpose translation engine based on a new analyser. The first component in the analyser is a constraint-based morpho-syntactic analyser (KARLSSON 90; KARLSSON 95; ZAYSSER 96). This analyser attempts to assign appropriate categories and basic morpho-syntactic features (e.g. number, tense, mood etc) to words in a sentence by intersecting an initial word automaton produced by lexical lookup with a finite-state automaton compiled from a set of linguistic rules in the form of regular expressions.

An additional ruleset in this analyser allows the calculation of a *presyntax* - the assignment of a surface syntactic function such as **main-verb**, **subject**, **object**, **preposition-complement**, **noun-premodifier** and so on to each word in the sentence. In the case of complex sentences, further rules indicate the location of clause boundaries.

A special lifting algorithm parses the presyntax and creates a dependency graph. This graph is similar to a classic dependency tree structure except that a given node may have more than one potential governor, reflecting residual functional ambiguities and attachment ambiguities characteristic of these surface functions. Subsequent operations establish which of the component trees in this graph maximises the satisfaction of syntactic, semantic and general heuristic constraints. The selected optimal surface dependency is the result of the analysis (GRAAL project carried out in conjunction with Aérospatiale).

A relatively classic transfer phase transforms the source language surface tree into its target language equivalent. Thereafter a generation component - based on the linguistic part of AlethGEN - creates the corresponding surface string.

Integration with Translation Memory

Translation Memory techniques do not have to respect natural linguistic boundaries. It is perfectly appropriate to stock paragraphs, sentences, or fragments of sentences in TM provided that these correspond to a certain level of repetition.

However, most existing integrations of MT and TM only pass the MT engine segments delimited by strong punctuation when these are unrecognised by the TM: the MT engine is not asked to translate arbitrary fragments of sentences. If it was, having no idea what sort of fragments it is supposed to be looking at, it is likely to produce poor results. We are exploring two approaches to this problem.

Repetition-driven segment identification

The first approach is to go ahead and identify TM candidate segments by carrying out the classical repetition analysis. Once these segments are obtained, we then try to automatically identify an appropriate category for the segment. For example, the segment

In a scalar context

as in

In a scalar context, function blah returns the number of items in the list
might be classified as an Adverb.

We would then supply to the MT engine something like

proADV, function blah returns ...

where *proADV* is a generic adverb placeholder word in the MT lexicon. It is replaced in the translated output by the TM translation.

One simple approach to automatic categorisation is to try to analyse - or at least tag - the segment in one (or perhaps more) of its simplest use contexts and then to compare the result with a set of templates e.g

START_SENT PREP DET? ADJ? N --> ADV

This, of course, is just an extension of familiar term candidate recognition techniques.

The features in placeholders may need to be quite rich. Thus

allows users

as in

The cancel command allows users to cancel print requests ...

requires a placeholder with features *verb, 3p, singular, indicative, active, SUBJ+TO-INF*. The success of the approach depends on the reliability with which we can carry out the automatic classification.

Clause Boundary identification

Another approach that we are starting to investigate uses clause boundary information. Recall that our constraint-based morpho-syntactic analyser has a rule set which identifies clause boundaries in complex sentences. For example

The Prime Minister advised people † not to work

As is well known † detergents can damage the skin

Smoking cigarettes † is dangerous

The watched him † smoking cigarettes

It is said † that detergents can damage the skin

where † indicates a clause boundary. When repetitions are found, it is possible to search back in their sentential contexts to see whether they correspond to clause boundaries.

As before, we replace the clause in the matrix sentence by a placeholder. Since analysers expect clauses to be more complex than single words, we use multiword skeletal clause structures where important clause elements are replaced by placeholder words.

proVing proN † is dangerous

It is said † proThat proN proVfin

(If we used single words as placeholders rather than multiwords, we would have to change our analyser grammars.)

After passing through the translation engine, the skeleton is replaced by the TM translation. If for some reason the skeleton contains insertions induced by the translation process, the whole MT translation is rejected.

The approach depends on an underlying assumption that clauses typically have a translation which is not affected by context i.e. by the matrix sentence in which they are found. This is a question we are only just starting to look at.

Conclusion

AlethTRTM provides powerful facilities for the translation professional. In the context of a global strategy for linguistic resource management based on generic dictionaries, we intend to integrate it with a new-generation terminology manager (AlethGTTM) and linguistic knowledge management tool (AlethKESTM).

In two further lines of development we are

- preparing a completely new translation engine
- exploring ways to improve MT-TM integration.

Acknowledgements

The techniques and tools described in this paper are the result of work by many people at ERLI. My particular thanks to:

Sophie Corbel, Sylvie Greverend, Marie-Claude Guérin, Béatrice Marchand,
Dominique Maret, Laurent Roussarie, Simon Sabbagh

and to Veronika Lux of Aérospatiale for information on the linguistics of Aircraft Maintenance manuals.

References

(Consortium 93a) GENELEX Consortium. Report on the morphology layer. Technical report, 1993.

(Consortium 93b) GENELEX Consortium. Report on the syntactic layer. Technical report, 1993.

(Consortium 94) GENELEX Consortium. Report on the semantic layer. Technical report, 1994.

(Consortium 95) GENELEX Consortium. Rapport sur le multilinguisme. Technical report, 1995.

(KARLSSON 90) F. KARLSSON. Constraint grammar as a framework for parsing running text. In *COLING-90. 13th International Conference on Computational Linguistics, vol. 3* Kargren, H. (ed.) Helsinki, Finlande, 1990.

(KARLSSON 95) F. KARLSSON. Constraint grammar: a language-independent system for parsing unrestricted text. 1995.

(MENON 94) B. MENON. Eureka project GENELEX: an overview. In *Journées du Génie linguistique: actes, Paris - La Défense*, 1994.

(OGONOWSKI 94) A. OGONOWSKI. Constraint grammar as a framework for parsing running text. In *COLING-94. 15th International Conference on Computational Linguistics, vol2, Kyoto, 1994.*

(ZAYSSER 96) L. ZAYSSER. Representing morpho-syntactic ambiguity. In *MIDDIM-96 Seminar, Le Col de Porte, France, 1996.*

Integrating Machine Translation into Translation Memory Systems

Matthias Heyn

Abstract

Within the last few years, there had been a remarkable change within the use of tools at the desktop of professional translators. Whereas, traditionally the keyword for the automation of the translation process had been *machine translation* (MT) this has significantly changed in the last few years towards the usage of *translation memory systems*. On the other side, MT-systems are more and more targeting their genuine market. Non-professional users' main interest lies in "quick information translation".

This general development doesn't mean that MT is not used any more at translator's desktops. It rather means, that the role of MT in a professional environment has significantly changed. MT for professional translators means that MT is one software component among other ones within a central translation memory system. MT is reduced to a "proposal machine" for worse case situations: if no information at all is accessible or if the source is simple enough.

Integration of MT into translation memory systems can be done by several architectures and this paper will investigate the different possibilities and their pros and cons. Existing integrations with the TRADOS Translator's Workbench for Windows will be discussed.

Matthias Heyn

Matthias Heyn studied at the universities of Heidelberg and Stuttgart *linguistics and information science*. He worked as a lecturer at the University of Heidelberg and published in the fields of lexicology and computational linguistics. He specialised in corpus research, alignment strategies and computational lexicology / terminology before he started in 1992 to work with TRADOS GmbH in Stuttgart. In 1994 he got manager of the research and development department of TRADOS and founded in 1995 TRADOS Benelux S.A. in Brussels. He is currently managing director of TRADOS Benelux S.A.

Trados GmbH

TRADOS is widely acknowledged in developing software products in the field of translation tools. TRADOS exists now 12 years and therefore is one of the most experienced companies in the field. With more than 8000 users of TRADOS products - to a large extent terminology database systems and translation memory systems - TRADOS has in-depth knowledge of user needs and requirements, the current market situation and linguistic knowledge within specialised languages. TRADOS is recognised to be one of the most successful commercially driven companies in the field of CAT-tools. TRADOS International Network has offices in Germany, Belgium, Great Britain, Spain, Sweden and Switzerland and resellers all over the world.

Matthias Heyn

Trados Benelux SA/NV

303, Avenue de Tervuren, B-1150 Brussels, Belgium

Tel: +32 2 775 84 70, Fax: +32 2 775 84 80
E-mail: matthias@trados.com

1. Introduction

Over the last few years, the use of tools at the professional translator's desktop has significantly changed. Whereas the keyword for the automation of the translation process used to be *machine translation (MT)*, the dominant notion for **language professionals** is nowadays *translation memory systems (TMS)*. This development is due to several factors: generally speaking, language professionals are experts, dealing with semantics and pragmatics much better than any machine can do. They do not need to bother with imperfect machine translations, but they do need substantial aid in the organization of their work concerning terminology and retrieval of existing human translations. A TMS now takes over this part: the machine does what a machine can do best.

On the other hand, MT has proven its worth in informative translation, helping non-professionals to understand the rough meaning of documents. Therefore, MT vendors start now targeting the market of the standard application user by means of marketing and pricing.

This general development does not necessarily mean that MT has no future at the translator's desktop. It rather means that the role of MT in a professional environment has changed significantly and hence must be redefined.

In this paper, we will have a closer look at that new role of MT within a language professional's environment and see that the integration of MT into TMS can sometimes lead to fruitful synergies. We will investigate the conditions for the use of MT and discuss several integration architectures. Furthermore, we will give a few examples of existing integrations in the TRADOS Translator's Workbench for Windows.

2. Shifting towards Translation Memory Systems

There are some general reasons for the recent success of translation oriented software applications:

- The general tendency towards computerisation of text flow, gives translators more and more access to machine readable source documents.
- The processing power of modern (desktop) computers enables functionalities that were not available in the past and that are crucial to the successful implementation of translation tools on standard machines. Generally speaking, all improvements to the hardware are very welcome in this specific application area.
- The integration of translation software into the translator's software environment has improved considerably.
- Translation software has met the overall quality standards of the industry software with regard to user friendliness and software ergonomics.
- The knowledge of translators about the benefits of computerising their work is steadily growing.

Within this general trend towards the use of translation software, we can distinguish between two approaches: MT and TMS. Both approaches are defined as follows:

A TMS stores in a computer all translations made by a translator. In case of re-translation, these translations are retrieved automatically.

An MT system applies grammatical rules and information from dictionaries to a given source sentence in order to translate it.

These two approaches to translation are quite contradictory. MT tries to model the translation process, so to speak replace the translator; whereas a TMS supports the translator by making the individual translation process reproducible.

2.1 What comes out of the system?

Whereas a TMS can be described as a system where *all output is based on human input*; an MT system can be described as a system where *all output is performed by a machine process*. TMS avoids generative capacities whereas MT relies on them. In a professional environment, this means that a translator can blindly rely on any TMS, if he or she trusts the translator who has previously worked with this system. In contrast to working with a TMS, an MT translator can never trust the output and has to proceed to a time-consuming and boring revision- (or better: repair-) phase.

2.2 What does the system learn?

Another interesting feature of TMS is the “learning”-factor: a sentence has to be translated once and never to be translated again, whereas with an MT system, a translation, with its possible errors, is always *re-generated*. In other words: repetitions are only learned once within a TMS, where MT always re-generates them.

2.3 How to get a better profit from the tool?

A very important aspect for language specialists is the “tuning” of a system. In the case of TMS, this is very simple. The only thing a translator has to do is translating with this system. Since a TMS “learns” in the background the introduced translations, it improves automatically. There is no other specialist knowledge required but good professional translation skills.

Improving MT output is a rather tedious work. Documents can be preprocessed by using controlled authoring or controlled language mechanisms; output has to be postprocessed and revised; the dictionary component of the MT system can be updated; grammar rules can be adapted etc.

Updating the MT dictionaries is always very time consuming and requires specialized knowledge at different linguistic levels. Updating the generative core component of an MT system (the “grammar base”) is difficult and may yield side-effects that are almost always uncontrollable. In short, tuning an MT system is rather complicated and time-consuming and requires skills beyond standard application user knowledge and beside the standard skills of a language professional.

2.4 Psychology of a tool

MT tries to *replace* the translator, a TMS is doing the opposite: it tries to *support* the translator. A translator who works with MT looks like working most of his time against the machine because of operations like error-prevention and error-repair. A translator working with a TMS is feeling even more responsible about her or his translations because their work is going to be “re-used”.

To summarise: a TMS frees translators from boring and repetitive tasks and lets them concentrate on what they do better over machines, i.e.: handling the semantics and the pragmatics. This generally leads to a broader acceptance of TMS by language professionals.

2.5 What is needed by language professionals?

Professional translators do not have problems with morphology and syntax but with semantics and pragmatics. In most cases this has to do with lack of knowledge about the subject area or, in other words, with lack of terminology and specialised language collocations. A professional translator does not need a system that handles syntax and morphology, but a reliable term bank or a translation memory covering the subject field.

Maybe an analogy can help to clarify the relationship of MT and TMS as software products: there is software for the benefit of professional users and software for simulating the professional. We can think of a software that tries to replace partly an accountant and software that is used by accountants; or else: software that tries to simulate the skills of an architect and software that is used by architects to facilitate their job. MT tries to simulate a translator and a TMS software is used by a translator to do a better job.

3. MT as an “add-on”

From the above mentioned discussion we can redefine the role of MT in the field of professional translation. First of all, we can narrow down the scope of conditions for a successful use of MT. If a translator is confronted with a sentence and:

- this sentence or a sufficiently similar sentence cannot be retrieved by a TMS;
- the sentence lies syntactically more or less within the scope of the capabilities of the MT system;
- there is a certain coverage of the MT dictionaries of the required subject area;
- the MT system is capable to preserve the formatting;
- the MT system is a keystroke away and responds quickly (or has already prepared a translation over a previous batch process);
- the MT system uses the terminology of the private term bank system of the translator,

then probably a good proposal of the MT component can speed up editing time. The proposal can be corrected and next time the memory is in charge of the sentence!

Therefore, we can describe MT in a professional context as "a proposal machine" that can be switched on and off - dependent on the conditions of the text to be translated. MT is not a core component, but plays a subordinate role as part of a set of useful tools a translator can choose from, like spell-checkers, electronic dictionaries etc..

4. Integrating MT into TMS

In principle there are two possible architectures to integrate MT into TMS. It can be integrated using batch processing or interactive integration. Already in the earliest

development phase of the object-oriented class system of the TRADOS Translator's Workbench for Windows, was decided for the implementation of a neutral layer for machine translation integration. This enables to integrate an existing MT system seamlessly either in batch or as an interactive component.

4.1 Batch integration

Batch integration means that a text passes first through an analysis process of a TMS which sorts out all sentences (translation units) that are unknown to the TMS. These sentences are then passed on to the MT whereafter the results are reimported into the TMS. The access to the TMS yields now results that are probably MT output and not former human input. It is self-evident, that these entries have to be marked and treated separately.

Batch integration can be implemented easily: MT and TMS only have to communicate over a common file exchange format. The disadvantages of this approach are the following:

- the required file processing adds an additional preprocessing phase to the translation process;
- it disables the translator from making interactive decisions such as expanding or shrinking phrases, correcting on the fly errors in the source text (typos!), adding terms to the termbank, switching to a different MT lexicon (or changing the access sequence of MT lexicons), adding an abbreviation in order to avoid segmentation faults, etc.. This kind of changes can only be respected by the MT system after manually restarting the complete batch process and are therefore in practice left out most of the time.

On the other hand, even slow MT systems can be integrated over batch, since the proposals are quickly accessed over the TMS.

Batch integration is often the only possibility to integrate an MT into a TMS. This is the case e.g. if the MT is not running on the same platform as the TMS or if the MT is too slow for interactive integration or if the exchange protocols are not fit for adaptation.

4.2 Example batch integration

4.2.1 LOGOS Machine Translation combined with the TRADOS Translator's Workbench for Windows

The TRADOS Translator's Workbench for Windows integrates the LOGOS MT system using a typical batch processing environment.

A given document (RTF or WordPerfect format) is analysed in order to detect all segments (sentences) that are unknown to the current translation memory of the TRADOS Translator's Workbench for Windows. Fig 1. shows the menu where this operation will be performed.

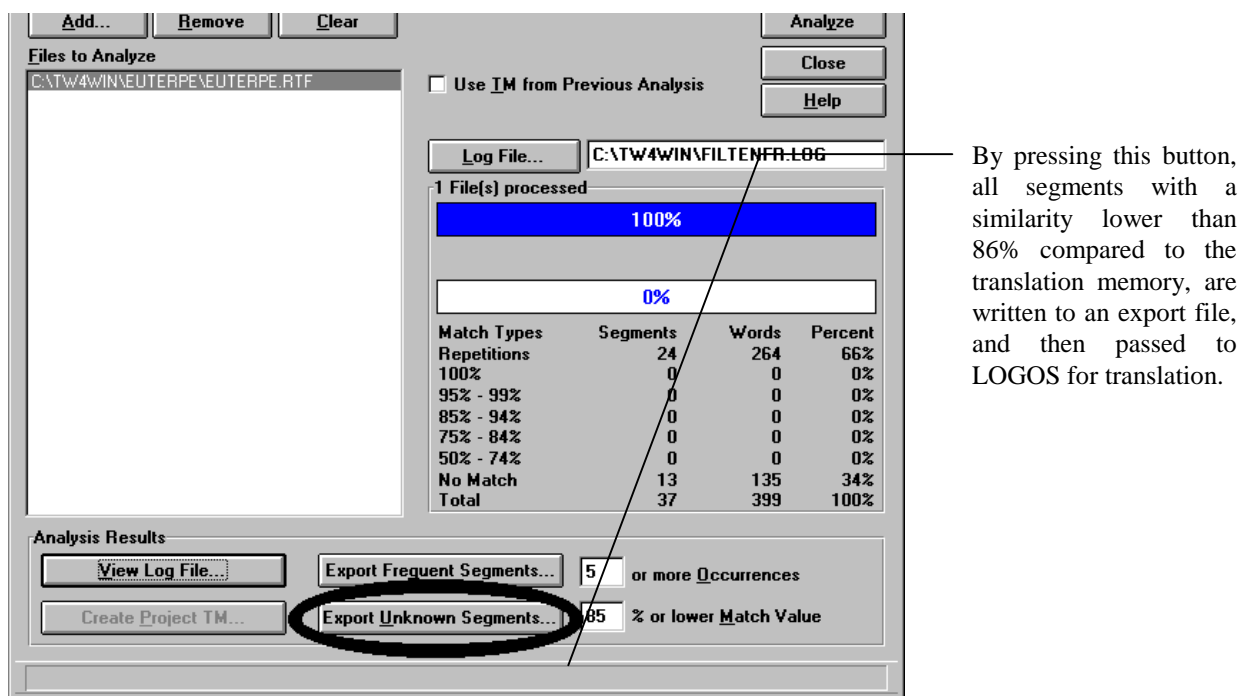


Fig. 1: The “Analyse dialog” of the TRADOS Translator’s Workbench for Windows

This file is then passed to the LOGOS machine translation, which translates the segments. The result of the machine translation is then reimported into the TRADOS Translator’s Workbench for Windows. During this process all formatting information of the source text will be respected by LOGOS and the formatting is later stored within the TMS.

This approach has the advantage that the segmentation process is performed by the TMS, which ensures uniform integration into the later translation process while preserving the same segmentation.

After this preparatory work, the translator proceeds in the familiar way: from the well-known word processing environment (e.g. WinWord or WordPerfect) the TMS is interactively consulted. All proposals from the machine translation are marked as machine translations. In addition, the user has the option to set penalty values to “punish” machine translation entries (see Fig. 2 of the TRADOS Translator’s Workbench for Windows “Translation Memory Options” dialog window).

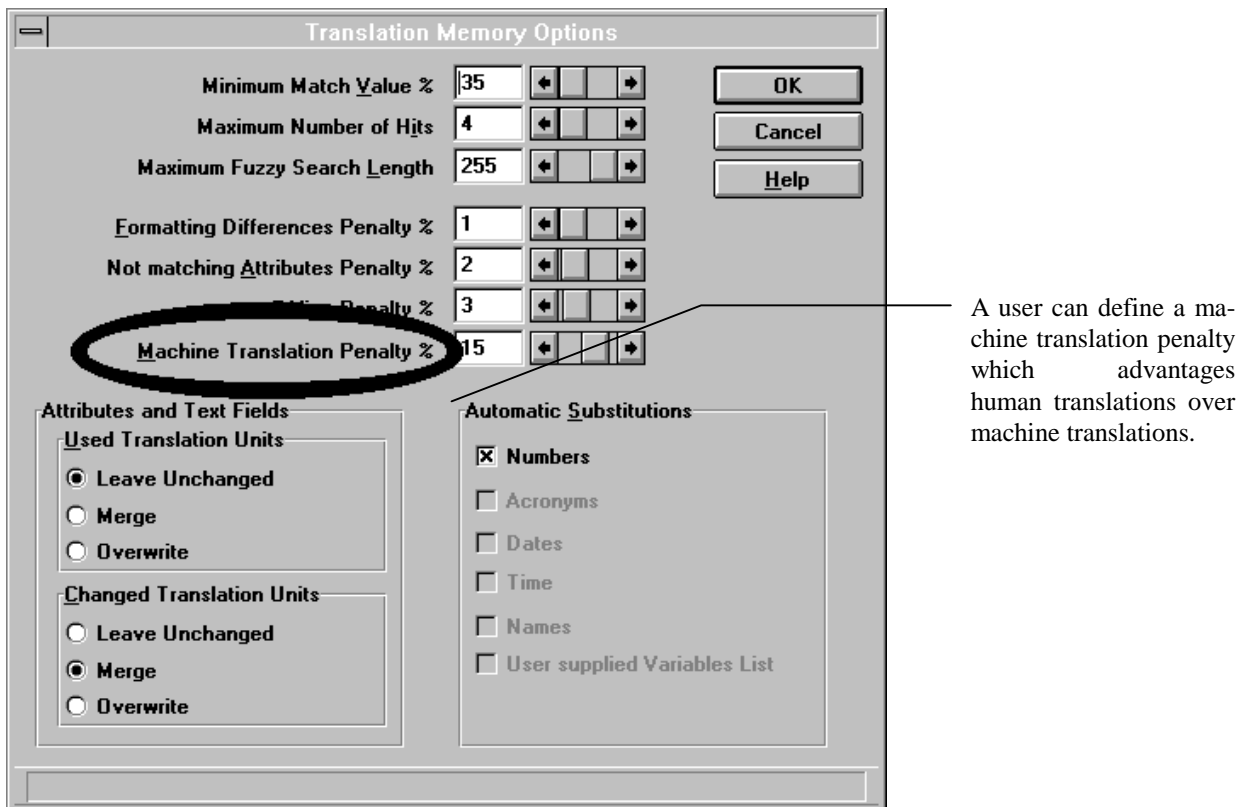


Fig. 2: The “Translation Memory Options” dialog

This mechanism ensures that human translations are proposed before machine translations to the user.

4.3 Interactive Integration

Interactive integration means that the user can interactively decide whether he sends a given segment from the source document to the MT and that the MT thereupon sends back a result. After correcting the errors of the MT system, the segment, as usual, is then stored in the TMS and retrieved in case of similar or equal sentences.

Interactive integration is more difficult to implement and requires the same platforms as MT and TMS (or sufficiently powerful exchange protocols) and a quick response time from the MT.

The big advantage of this solution lies in the flexibility for the user. Interactive decision-making like changing the segment sizes, correcting source text errors etc. can be performed without interfering with the MT.

On the other hand, interactive access is by nature slower than the batch access and must therefore rely on appropriate hardware and efficient MT systems.

4.4 Example Interactive Integration

4.4.1 Intergraph TRANSCEND and TRADOS Translator's Workbench for Windows

The interface to TRANSCEND is an extension to the TRADOS Translator's Workbench for Windows, which is automatically installed (See Fig. 3: TRADOS Translator's Workbench "About" dialog).

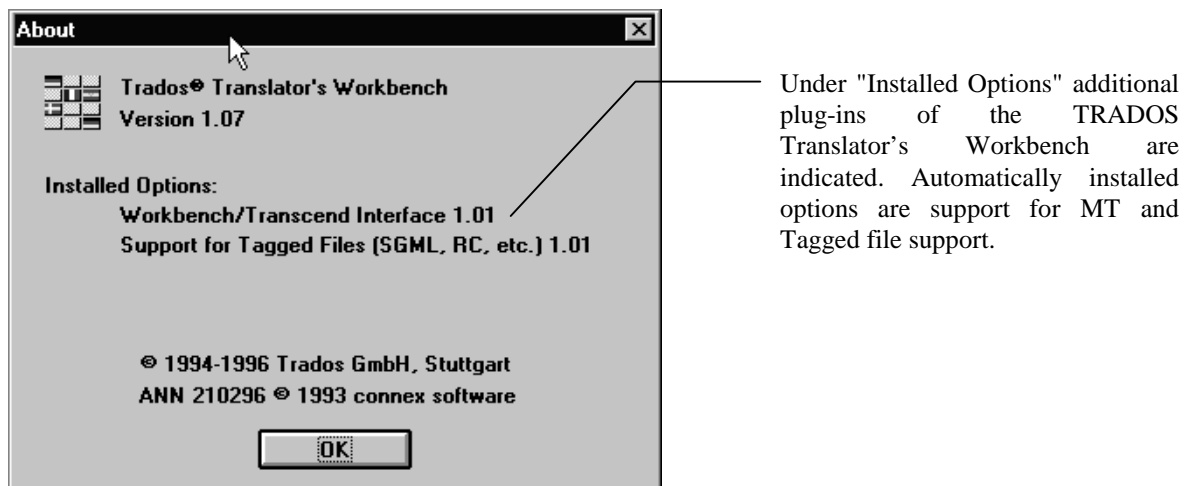


Fig. 3: TRADOS Translator's Workbench "About Menu"

The TRANSCEND MT has to be loaded into memory in order to be accessible over the TRADOS Translator's Workbench for Windows. One keystroke (see Fig. 4) activates an option that makes all unknown sentences pass from the TRADOS Translator's Workbench to TRANSCEND.

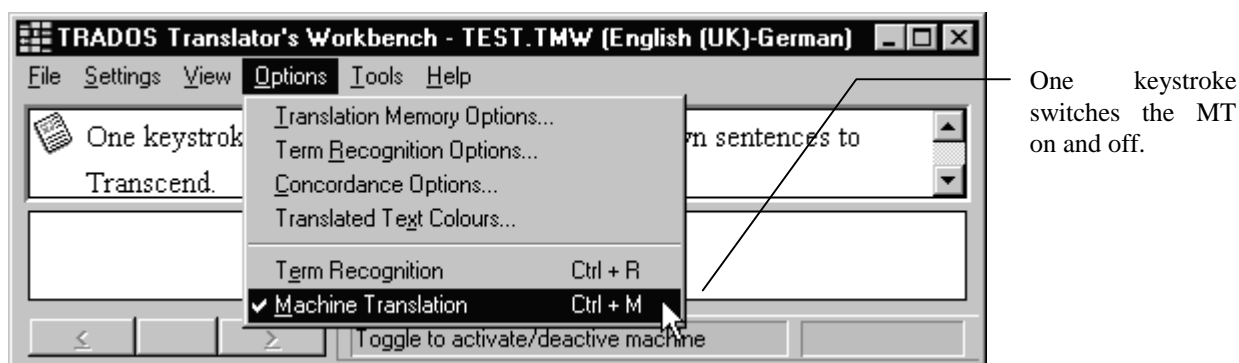


Fig. 4: Activating Machine Translation within the TRADOS Translator's Workbench for Windows

An example is given by Fig. 5, where the heading *After the wash* was taken from an instruction manual of a washing machine and could not be retrieved from the translation memory. Now, TRANSCEND English/French machine translation comes up with the proposal *Après le se laver*. The proposal of the machine translation is clearly distinguished by colours (a grey frame). If the translator now corrects the wrong proposal of the MT system and stores it in the TM, the next time the

translation is needed, the corrected version of the sentence will automatically be presented by the translation memory.

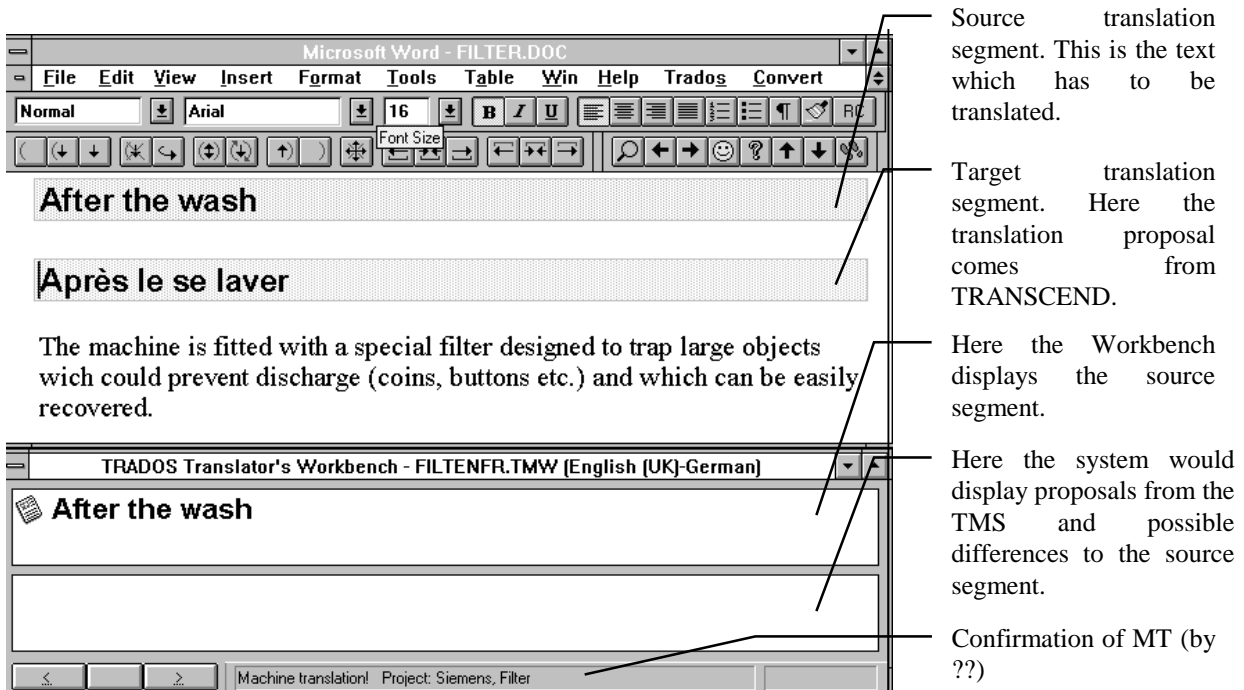


Fig. 5: TRANSCEND has been activated after unsuccessful search in the TM

5. MT lexicon versus TMS term bank

A general problem with the combination of a MT system and a TMS is that both systems provide lexicons. That means that probably two competitive sources have to work together. A possible solution could be the storage of the TMS information in the MT or vice versa or the passing on of information from the MT to the TMS or vice versa. In the context of MT used by language professionals, there seems one solution preferable.

MT dictionaries are specialised dictionaries for the explicit storage in a formal way of information from different linguistic levels. The more sophisticated the MT the more elaborate the dictionary structure. Scientific prototypes of MT are confirming this tendency with complex feature structure lexicons and specialised editors for these dictionaries.

The more advanced the system the higher has to be the (theoretical) language and information science competence of the user coding an MT dictionary. But, for all MT systems it is true that the coding time for one dictionary entry is rather high.

On the other hand, TMS are offering professional translators the possibility to encode their own terminology. This is necessary since one of the biggest problems for professional translators is to find wordings that are **not found** in standard dictionaries. Good TMS are offering sophisticated term bank systems that can be freely configured for appropriate terminographical work. In practice, considering

time constraints and production stress in a professional translator's environment, the available time for coding terminology is rather limited. That means often that under production circumstances only quick term-list equations are possible.

Now, let's look to both dictionaries from two angles:

5.1 Use of the MT lexicon by the TMS

MT lexicons are normally constructed round a core lexicon that covers the standard lexicon of a given language. This is not at all of interest for professional translator's. If there are specialised dictionaries available in MT systems these could be interesting for manual consultation by the translator. Precondition is a suitable access for "human" users or even better an import into the TMS term bank system.

5.2 Use of the TMS lexicon by the MT

In this perspective, it is very important that the MT respects the terminology of the translator. There are again two possibilities:

1. The contents of the term bank are transferred to the MT dictionary. In general, this involves the manual adaptation of the lacking linguistic information - which consumes a lot of time and effort.
2. The TMS is not only passing to the MT a particular sentence that has to be translated but also all known terminology of that sentence found in the term bank of the TMS.

The second possibility is certainly more appropriate, since a terminology database is more frequently updated and more easily maintained than an MT lexicon.

One can argue that the MT produces more errors if there is not enough linguistic information found in the lexicon, but in the environment of professional translators it has to be stressed that **the syntactical correctness is not important compared to semantic and pragmatic correctness**. The only reason for maintaining minimal linguistic information within term banks that have to be passed on to MT is that frequent morphosyntactical deviations are slowing down the editing process.

5.3 MultiTerm and TRANSCEND

An example solution for the passing on of terminological information from the term bank system of a TMS to an MT has been implemented within the TRADOS Translator's Workbench. The term bank component of the TRADOS Translator's Workbench - MultiTerm - is a fully fledged term bank system. MultiTerm allows for free database definitions and can be used in rather sophisticated terminology driven environments or within pragmatic production driven environments.

The TRADOS Translator's Workbench performs the automatic detection of terminology stored within MultiTerm ("term recognition") and points out this information to the user. Term recognition is in itself a rather complicated function which involves non-trivial tasks like the decomposition of complex compound phrases or the handling of separable verb-prefix constructions etc. Recognised terms are then passed to TRANSCEND. Besides the passing on of the terminology, the TRADOS Translator's Workbench also provides a protocol for passing on a few basic morphosyntactic features to TRANSCEND.

The complete MultiTerm entry of Fig. 6 can be written with 14 keystrokes. After adding this entry to MultiTerm and reopening the heading of the washing machine operation manual again, TRANSCEND produces the translation given with Fig. 7.

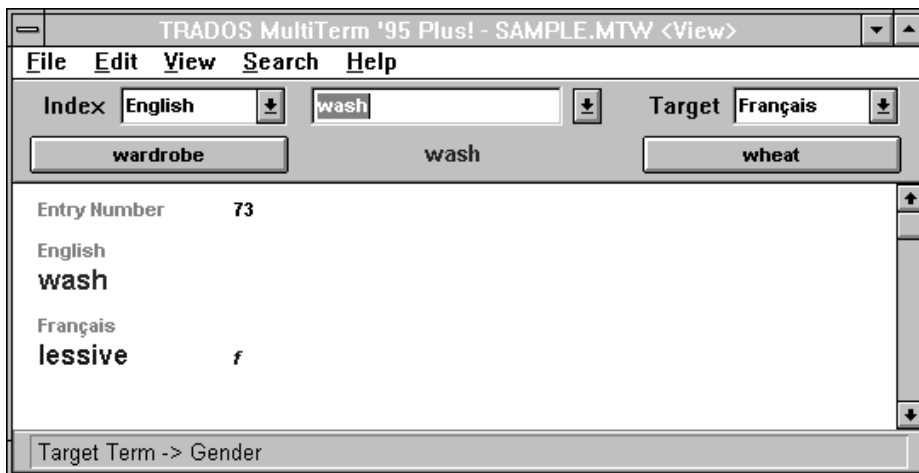


Fig. 6. MultiTerm database entry for English “wash”

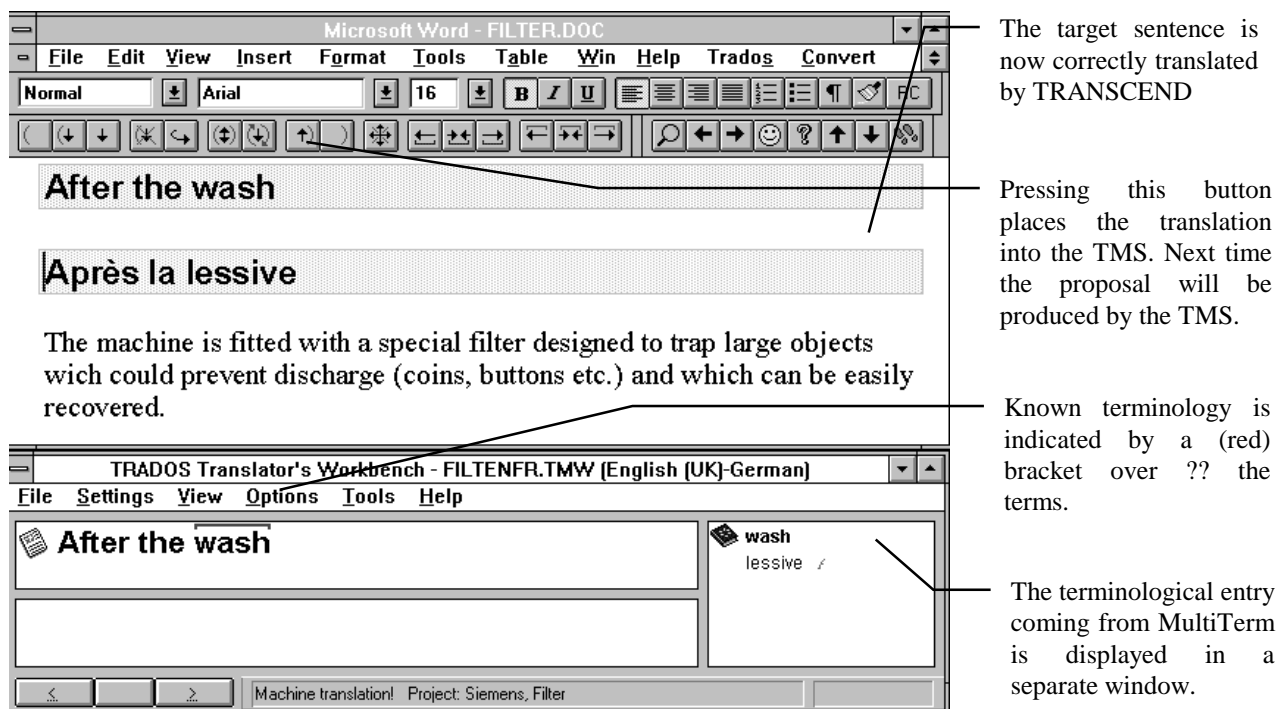


Fig. 7. Retranslation of TRANSCEND using a term bank entry of MultiTerm

When the translator now confirms the translation, it will be stored in the translation memory.

The next sentence in the text (*The machine is fitted ...*) is translated by TRANSCEND with: *“La machine est ajustée avec un filtre spécial a conçu pour prendre au piège de grand wich d’objet pourrait empêcher la décharge (les pièces, les boutons etc.) et*

qui peut facilement retrouvé.” The translation is disturbed by a typographical error in the source (...large objects could...). The translator can correct the source error and restart TRANSCEND with two keystrokes. The result changes only slightly: “*La machine est ajustée avec un filtre spécial a conçu pour prendre au piège de grand d’objet pourrait empêcher la décharge (les pièces, les boutons etc.) et qui peut facilement retrouvé.*” In this case the translator does not benefit from the MT (or can even be hindered by it) and will certainly propose a different translation, such as e.g. the one given in Fig. 8.

Fig. 8. gives an example where the TMS retrieves a former manual translation in the case of a retranslation.

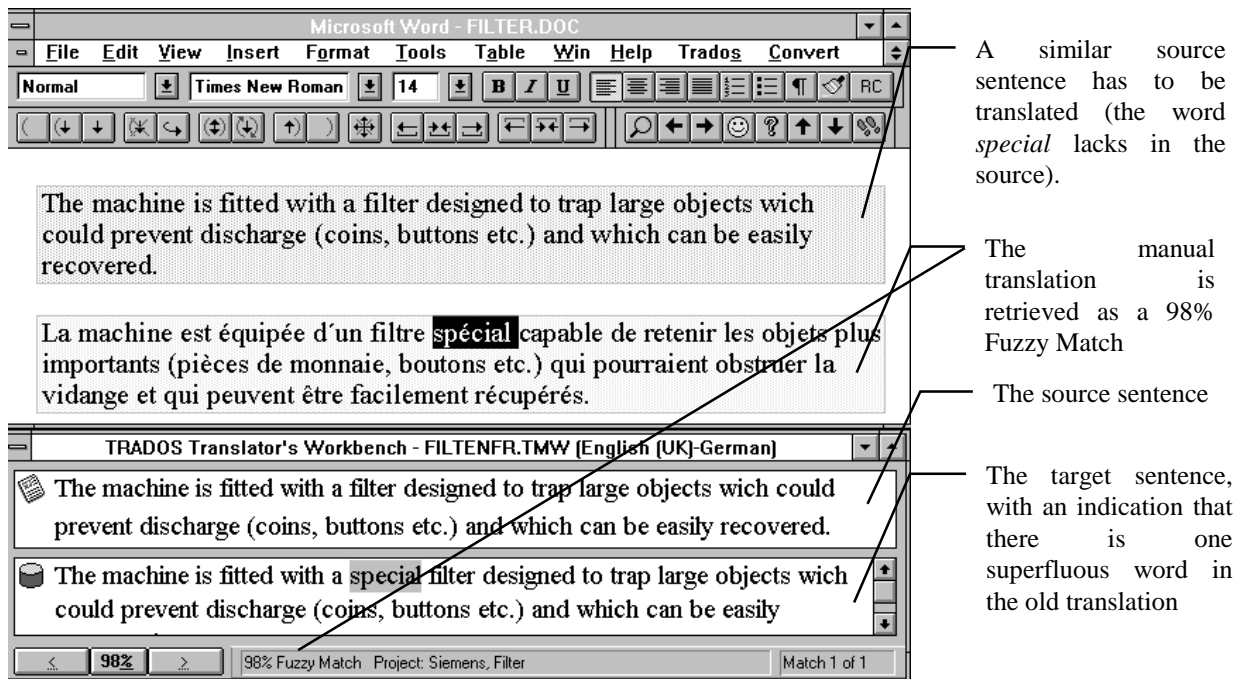


Fig. 8 Retrieval of a human translation within the TRADOS Translator's Workbench

A side effect of working with the TMS is that all translations are immediately retrievable in form of concordance searches. If a translator for example searches for *discharge prevention*, she or he will get the results shown in Fig. 9.

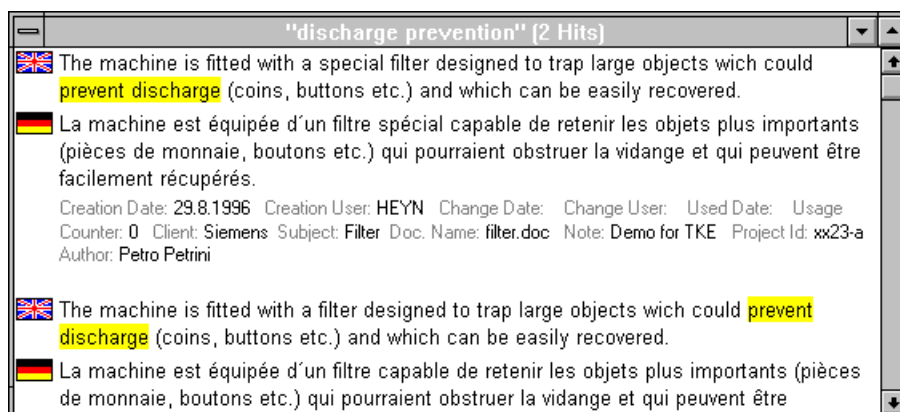


Fig. 9. Concordance search for *discharge prevention*.

Concordance searches are in many cases more important for reconstructing the semantics and pragmatics of a given translation task than using MT. At least this is true for the language professional.

6. Summary

By investigating the relationship of machine translation systems (MT) and translation memory systems (TMS) in the context of (technical) translations by language professionals, we first had to redefine the role of machine translation. The MT's main application lies within information translation and consists in helping non-language specialists to overcome language barriers. MT tries to simulate the skills of a translator and this is sufficiently successful for certain application fields in the mass market.

On the other hand, the professional translator does not need an "less skilled" electronic colleague, but reliable professional software helping to do a professional job. Specialised software for translators today is TMS, which takes over the parts in the translation process that can be successfully delegated to a machine. Therefore, it is evident that the role of MT in the context of professional translations has to be redefined as an optional "add-on" tool within the TMS. If certain conditions prevail, MT can speed up the editing process. Preconditions are: a seamless integration - preferably an interactive integration - and sufficiently powerful links of the term bank system to the MT.

Successful translation process automation as one part of an overall document production flow means for the future a better harmonisation of the involved technical solutions. We are faced with the problem to integrate authoring tools, document retrieval software, workflow solutions, translation memory tools and last but not least machine translation. A key role for satisfying solutions will be the "interconnectivity of software modules" combinable to holistic solutions.

Tendencies in information science towards distributed objects and in general towards object orientation are very important in this respect.

Translator's Workbenches: A Practical Application

Adriane Rinsche

Abstract

The paper presents an example of a solution designed for a multinational client involving translator's workbenches.

As a first stage, the user requirements were assessed, as concerns contents and form of documents, the technological base, human resources and volume and cost of translation.

The range of possible solutions was then evaluated, in terms of budget, technological base again, language combinations, user friendliness, customisability and linguistic features.

Recommendations were then made in terms of either setting up an in-house language business unit or an outsourcing solution. The client opted for an outsourcing solution. We converted, aligned and imported existing information into the workbench solution and resolved organisational issues.

We have offered MT as an integral part of Workbench solutions to a variety of clients. We will discuss the financial viability and our own views of integrating MT into a translation technology environment.

Dr. Adriane Rinsche

Dr. Adriane Rinsche is currently managing director of The Language Technology Centre Ltd. She studied English and Philosophy at a German University, was a Research Assistant for three years, has a PhD in Computational Linguistics, and worked as a language and language technology consultant with major multinational clients. She has more than 10 years experience in the language and IT industries.

The Language Technology Centre Ltd.

LTC specialises in:

- evaluation of major language technology developments
- software consultancy in multilingual documentation and computer-assisted language training
- delivering and implementing software solutions
- coding and converting language resources
- providing and managing outsourcing solutions in multilingual documentation

Dr. Adriane Rinsche

The Language Technology Centre Ltd.

27 Cotswold Close, GB-Kingston, Surrey KT2 7JN

Tel: +44 181 549 2359, Fax: +44 181 974 6994

E-mail: xe_s412@kingston.ac.uk, or 101365.676@compuserve.com

www: http://ourworld.compuserve.com/homepages/Language_Technology_Centre

Introduction

Translation technology offers solutions to a variety of multilingual documentation problems which may vary considerably from one prospective user or user organisation to the other.

The increasing number of tools and systems becoming available on the market place requires a careful evaluation as a basis for any far-reaching decision, because the investment envisaged is not restricted to the cost of the software solution chosen. In order to provide the best result, any off the shelf product or product configuration requires considerable customisation. The time and cost involved in designing and implementing an appropriate solution should not be underestimated. Very well defined and well organised procedures need to be introduced in order to achieve a cost-effective, high quality and high speed solution to the translation problem. Today, many international corporations do not organise the processes involved in the most appropriate way yet.

An example of a solution we designed for one of our multinational clients will be discussed below. It illustrates the most crucial steps, embracing

- an analysis of the user requirements
- an evaluation of the relevant technology
- recommendations and an implementation plan.

As our evaluation was based on the requirements of a specific client, it cannot be generalised. Although we developed global evaluation criteria for translation technology, which were, for instance, applied as a basis of our recent "OVUM Evaluates" report, the choice of a particular solution depends on the requirements of a given company. There is no global "best solution" or "best product" on the market.

Towards integrated multilingual document processing

Our client is an American multinational manufacturer of machine tools. The European Headquarters are based in the UK. The management within one department of this organisation asked us to provide a technology based solution to their translation requirements. We gathered the following information as a basis for further research and our final recommendations:

Step 1: Evaluating User Requirements (May 1995)

1. Contents and form of documents

1.1. Text types:

The company handles, in principle, four text types within the technical/engineering domain, three of them highly repetitive, the last has a more informal and chatty style.

- Repair Instructions, repetitive
- Consumer Manuals, repetitive
- Machinery manuals
- Training materials, non-repetitive

1.2. Information available in machine readable form:

A considerable amount of terminology (parts lists) and parallel text in seven languages is stored in machine readable form on a mainframe computer, all in capital letters. The information stored on the mainframe will not be supported beyond the end of 1995. It is therefore vital to retrieve the information and convert, store and reuse it in an appropriate environment.

1.3. Languages:

English is the source language, at the time of the evaluation 6 target languages need to be covered, to be extended to 12 target languages later.

1.4. Formatting and tagging:

Formatting and tagging requirements are practically non-existent in the text type "Repair Instruction", with the exception of occasional tables and statistics. Machinery and consumer manuals, on the other hand, are produced in Pagemaker.

2. The technology base

Several departments use Ami Pro as their word processing environment throughout, on networked PCs and workstations.

The graphics department uses one Macintosh powerstation and a number of smaller MACs with Pagemaker as their major input and output medium for linguistic data and QuarkXpress as their major medium for pictorial data. Some employees use other packages such as Word for Windows.

3. Human resources

A group of technical authors, illustrators, project managers is responsible for documentation.

Translations are carried out:

- partially in-house by trainees or technical staff with language skills
- partially by various translation agencies
- partially by low level bilingual staff in the various national markets.

This is a fragmented and unorganised way of organising multilingual documentation, with a high degree of inconsistency, at low speed, resulting in low quality output.

4. Volume and cost of translation

The volume of text material translated in the past in one department was comparatively low. The source text material consisted of approximately 20 000 words a year and was translated into 6 target languages, with an overall annual cost of £14,000. If additional requirements are met without technical support the cost of translation will rise to £250,000 for traditional external translation services in one department alone (calculated on the basis of £120 per 1000 words translation).

Other departments within the organisation have similar translation requirements in the same subject domains.

A limited budget for introducing translation technology is available to the department with very urgent translation requirements, and although we would have preferred a top to bottom approach where we would have aimed to introduce an all-embracing solution, we decided to adopt an initial "pilot" approach in order to help our client with their immediate needs and in order to support their unusual interest and motivation to dedicate themselves to a technologically oriented approach to multilingual documentation. The department is obviously aware of the fact that a centralised and unified approach to multilingual document production and management is not only more cost-effective and consistent, but that it provides them with critical influence on the company's international corporate image.

Step 2: Evaluating the range of possible solutions (June 1995)

Based on the information obtained from the client, we used the following strategy for evaluating the translation technology products available on the market:

1. Budget

In calculating the cost of supplying and implementing a computer-assisted translation solution, the following costs had to be taken into account:

- a) Purchasing price for hardware and software
- b) Customisation cost (e.g. converting data)

We considered the following systems:

MT systems considered: METAL, LOGOS, DP/TRANSLATOR, SYSTRAN
all too expensive, therefore ruled out

Power Translator Professional
cheap, not ruled out for financial reasons

LMT
not ruled out for financial reasons

Translation Workbenches considered: IBM Translation Manager
Eurolang Optimizer
Trados Translation Workbench
Star Transit
XL8
investment in principle acceptable

2. Language combinations

The very advanced IBM terminology system Translexis is multilingual, but the translation system LMT with which it is integrated offers only two commercially available language pairs at present: English to German and German to English. The solution was not further evaluated for this reason.

The Power Translator Professional covers only 4 language pairs altogether, only 5 out of 8 language pairs offered by DP/Translator were relevant to the required solution.

Due to the high estimated degree of repetitiveness of 60% applicable to two out of three text types covering approximately two thirds of the documents involved, it was therefore appropriate to consider a translation memory based solution. All translation workbenches available on the marketplace offered the required language combinations, except the Eurolang Optimizer, which does not cover Greek. We retained the other four systems for further evaluation.

3. Formatting requirements

At the time of the evaluation, it was impossible to resolve all issues regarding layout and formatting, which ideally should be retained throughout the multilingual production process. Pagemaker and QuarkXpress were not supported by any of the translation products. The client requires further consultancy at a later stage in order to adopt an integrated and considerably more cost-effective strategy at a later stage.

4. Overall design

We distinguished between two categories of Workbench solutions:

Type A systems: database structure

Type B systems: flat file - reference material approach

Type A products: IBM Translation Manager,
Trados Workbench,

Type B products: Star Transit, XL8

We considered the Type A products more suitable in this context, the Type B products were therefore ruled out.

5. User-friendliness

Our client employs no linguistic staff in-house and does not intend to do so in the future. For the application planned it was therefore not feasible for the corporation to employ translators. It was our intention, however, to make the documentation process as transparent as possible to the client. We wanted them to be able to control the process from their premises to a certain extent by keeping one copy of the software on site. In order to do so, the system introduced had to be particularly user-friendly, with little training required, an adequate help system and a practical user-interface. Some features of the two systems which “survived” the evaluation to this final stage were:

IBM Translation Manager: System specific editor (at the time of the evaluation)

Project oriented design

Automatic translation memory update within project file

Maximum of 3 fuzzy matches displayed

List of not found words

no terminology database, flat file SGML dictionaries
instead

Trados Workbench: Integrated with Word 6
Integrated working environment for terminology and translation memory databases
Fuzzy matches up to user definable matching percentage
Access to bilingual concordance

6. Customisability

In order to meet the requirements of the specific application, a product chosen should allow for maximum customisability. Some features of the remaining products were:

IBM Translation Manager: Many WP and DTP systems supported

Flexible creation of dictionaries

User definable number and sequence of up to 10 dictionaries

alignment tool to build translation memories from parallel text

Trados Workbench:

Central translation memory database, user definable structure

(system, text and attribute fields)

User definable range of translation memory options

User definable automatic substitution of numbers, dates, times, names

Alignment tool to build translation memories from parallel text

7. Linguistic features

Translation memory should have identical and fuzzy matching capability, existing terminology and translated material should be easy to import.

Both solutions (Trados and IBM) had fuzzy matching capability and routines for morphological reduction. The Trados Workbench had the following additional features:

Trados Workbench:

Concept oriented terminology database design

Fuzzy matching at terminological level

Bilingual concordance

Neural network design (not a linguistic feature, but supports linguistic processing)

8. Conclusion

Our decision in this specific context was in favour of the Trados Translation Workbench. We expected a cost reduction for translation of approximately 60% for the repetitive text material, with considerably increased quality, consistency and

speed. The client requires further consultancy in terms of designing source document production with a view to multilingual documentation.

Step 3: Recommendations and implementation plan (August 1995)

1. Because of the highly repetitive structure of three out of four text types we recommended the above translation memory based approach with a terminology database both for terminology coding and update processing.
2. One multilingual terminology database and one copy of all bilingual translation memory databases required are held on site by our client. The reasons for this are several:
 - a) Data security is increased and risk management improved by having the data on more than one site. Although LTC holds copies of the databases on two powerful PCs and on tape, fire or burglary could result in the loss of the databases.
 - b) Any new text is submitted to the Workbench at the client's site to estimate the cost of a given translation cycle beforehand.
 - c) The percentage of identical matches is estimated to be 60%. The client is, in principle, able to increase this percentage by modifying very similar matches, using the database contents as their standard, and eliminating stylistic, syntactic and terminological variants. By this means an even higher percentage of the new source text becomes identical with source text already stored. This leads to a highly effective and standardised use of both language and storage capacities. In practise, with one year operational experience (August 1996) the client has relied on us for any source text standardisation.
3. The structure of the multilingual database and the translation memories is designed in such a way that a more comprehensive and integrated approach to language engineering within the overall corporate business processes can be implemented step by step.
4. LTC converts the contents of the existing mainframe database into Workbench format which provides instant reusable text and terminology in 7 languages and adds the relevant information for the additional 5 target languages in a second phase.
5. LTC manages the translation process on behalf of the client, as in-house linguistic staff is not available. LTC trains specialist translators in the use of the workbench and sends one fully updated copy of the databases to the client after each translation cycle. The client pays only for the translation of new text.
6. During a later stage in 1997 source document control will be increased by providing an English to English translation memory database. Standard text is stored in the database, new segments are accepted only if they are not variants of existing translation units.

After one year...

the solution is fully operational.

Statistics show that our estimate of an average reduction of 60% of the volume of translation and cost is correct. The example below was taken from our August '96 repair instruction update, German. There are 49% of identical text (cf. 100% below).

Repetitions are repeated untranslated elements, 50% of repeated words are charged to the client.

```
Start Analyze: Mon Aug 19 17:01:08 1996
Translation Memory: C:\PROJECTS\B&D\GERMAN\EN_DE1.TMW
C:\PROJECTS\B&D\GERMAN\R_ALL_EN.RTF
Match Types  Segments      Words  Percent  Placeables
Repetitions      250      1,063     11         0
100%             631      4,536     49         0
95% - 99%        28        290        3         0
85% - 94%        31        140        1         0
75% - 84%        20        150        1         0
50% - 74%        46        247        2         0
No Match         334      2,678     33         0
Total            1,340     9,104    100         0
Chars/Wo         4.28
```

We have an open strategy in terms of read and write access to the translation memory database. All translation work is, at present, carried out on site, therefore each translator can take advantage of maximum information, resulting in high motivation. Each translator is responsible for the status of his/her database, which increases motivation further. We are in the fortunate position to be able to agree any deadlines several weeks beforehand with the client to ensure that the same specialised and trained translators are available for each translation cycle.

The possible role of MT

We believe that machine translation plays an overall important role in information scanning and multilingual information retrieval, and will be increasingly relevant in value added and customised internet and intranet applications. In the past, considerable success was achieved with Machine translation and controlled language in technical documentation (cf. Caterpillar and Rank Xerox, to name only two examples).

With the emergence of translation workbenches, machine translation loses relevance for repetitive documentation in limited domains where a considerable amount of 100% matches can be expected through translation memory. However, machine translation could support the initial process of populating a translation memory database where parallel text is not available. Once a workbench is operational, it seems more practical to give translators terminology and translation memory information rather than having them post-edit machine output.

If machine translation is integrated into a workbench solution, design, implementation and production procedures should be as follows:

The first step is obviously to create or update the terminology database. Then all terms from the term bank are copied and imported into machine translation system dictionaries, at the same time adding the required linguistic coding. For the above client, this would have been possible, but expensive. The cost would have been GBP 2 per term = GBP 4000 per 2000 terms = GBP 48 000 for 12 languages, and the benefit would have been limited and restricted to only some target languages. The

client is therefore at present not interested in machine translation integrated into the current solution.

After updating the system dictionaries, the relevant source text can be submitted to a batch translation routine, with all identical segments automatically replaced by translation memory database content and all non-100% matches are machine translated.

In an interactive revision process the translator checks and post-edits the flagged machine translated text material and enters the correct sentences into translation memory. We would recommend this strategy to companies with a comparatively low repetition rate in their documentation.

Session 6: Integration of Machine Translation in Information Management

Chair: Colin Brace

Introduction

As an object of study, MT has for decades, with a few notable exceptions, been largely conceived as an isolated entity, and hence singularly failed to adequately address real-world needs. However, the past few years has seen a major shift in this regard. This session will address issues involved in implementing MT in document management systems, particularly within the context of heterogeneous networks environments.

Colin Brace

Colin Brace is a writer and consultant specialising in language technology. Since 1991, he has published Language Industry Monitor, a bimonthly newsletter dedicated to developments in the field of language technology. Mr Brace has given presentations at a variety of conferences and has participated in several EU-funded initiatives.

Colin Brace
Language Industry Monitor
Eerste Helmerstraat 183, NL-1054 DT Amsterdam, The Netherlands
Tel: +31 20 685 0462, Fax: +31 20 685 4300
E-mail: cbrace@lim.nl

“Translation and the Internet” A Sample Application Based on the Logos MT System

Joachim Meyer

Abstract

The workshop aims to give an overview of the current situation of the translation industry within the internet. As is the case in many other areas, the internet has considerable influence on this industry. It will be pointed out which kind of changes have already been observed (e.g. presence of numerous translation agencies, networks of freelance translators etc. in the internet) or are to be expected respectively.

The Logos Internet Client is to be presented as an instance of those kinds of tools which even today permit utilising machine translation technologies via internet. In addition, the various aspects of the Client that are also of interest to the user will be demonstrated:

- glossary functions with automatic coding of user terminology
- pattern matcher routines
- off-line lexicographic coding (ALEX for Windows)
- utilisation of user-coded off-line dictionaries in connection with the internet client.

Joachim Meyer

Born 20.12.63. Has an MA in Economics and Information Science. Previously freelance consultant for several small and medium size enterprises. He currently holds a position as Business Development Project Manager at Logos.

Joachim Meyer is at the moment working on a PhD in Information Science (Thesis: Information Management in Service Companies).

Logos GmbH

Logos Corporation was established in the US in 1969 as a privately owned company. The first mayor projects were carried out in collaboration with the US government, to develop English-to-Vietnamese and English-to-Russian machine translation systems. Between 1977 and 1983, the company concentrated on developing the Logos linguistic process using German and English as the base languages.

Logos opened its European office in Germany 1982. At the end of 1992, ownership passed from US investors to a group of private and institutional investors in Germany. Logos conducts its development efforts in the US and in Europe. The company also has a site in Silicon Valley. The European market is serviced from Logos's Eschborn office, near Frankfurt.

Joachim Meyer

Logos GmbH

Mergenthaler Allee 79-81 , 65760 Eschborn, Germany

Tel: +49 6196 59030; fax: +49 6196 590 315
E-mail: jmeyer@logos.de

(Outline of the presentation)

Impact of the Internet for the Translation Industry

- 24 hours access to translation services
- opportunity for new marketing strategies
- new communication mechanism between freelancer and translation provider
- distributed working environment

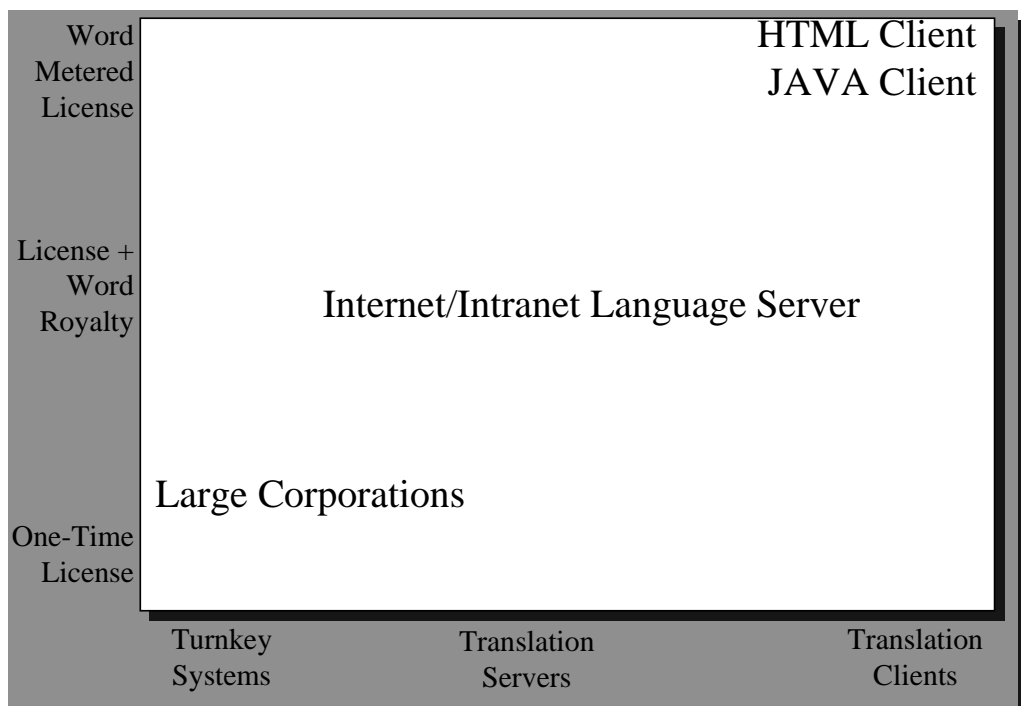
Today's use of the Internet in the Translation Industry

- web pages as a new access medium for customers
- different MT developers offer translation via the Internet
- world-wide networks of translators have been formed
- different lexical resources are available

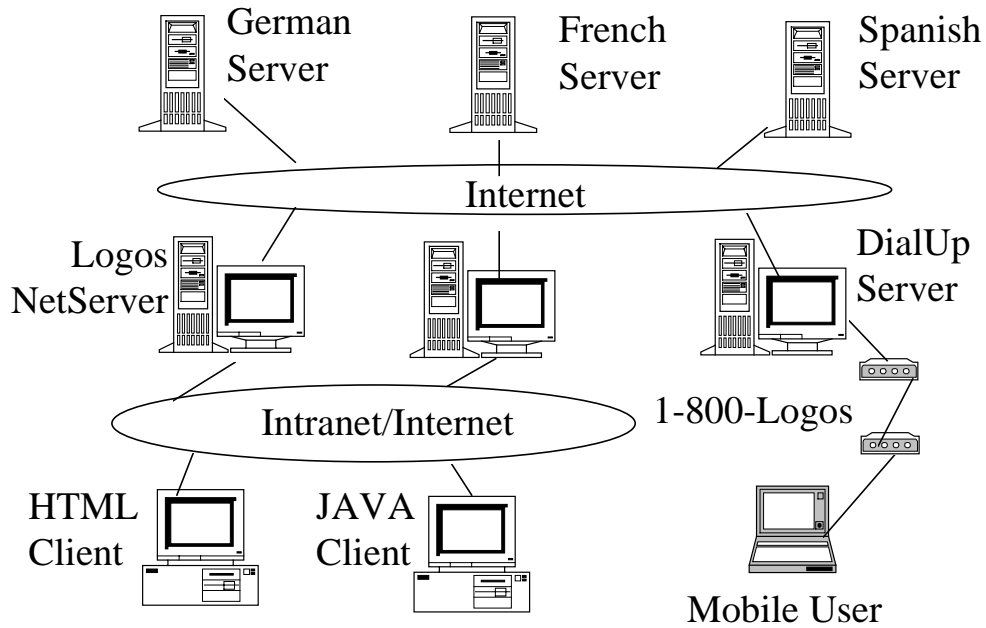
Existing Problems

- security of the documents
- accounting and billing mechanism
- bandwidth of the Internet
- incompatibility of the different lexical resources (no exchange mechanism between the different language applications)

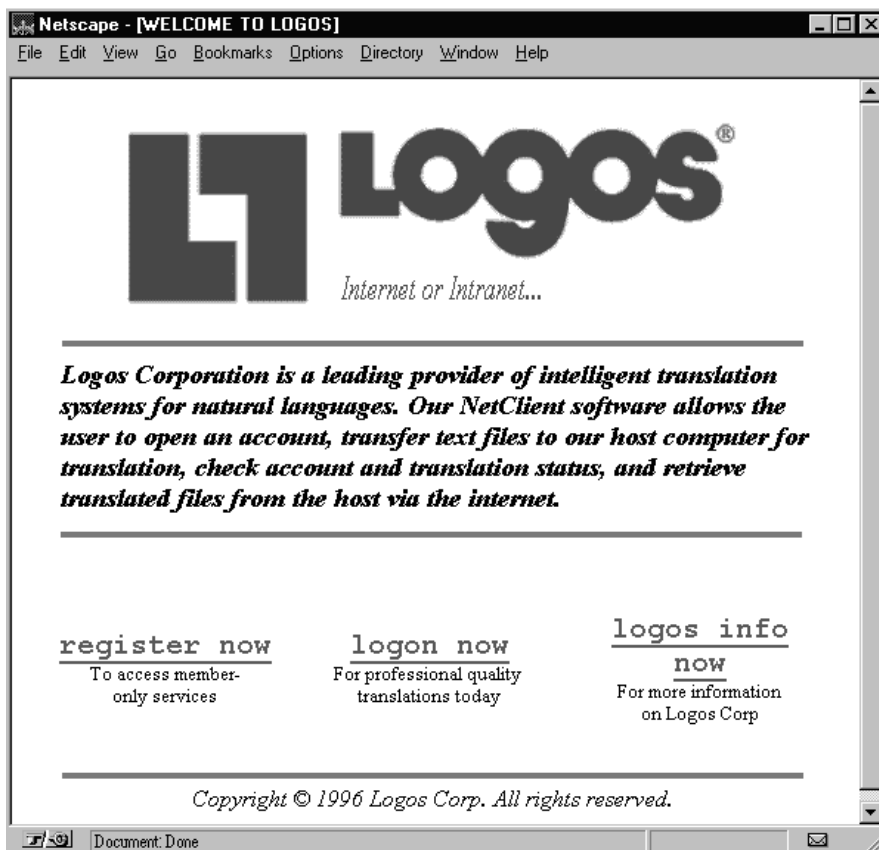
The Logos Picture



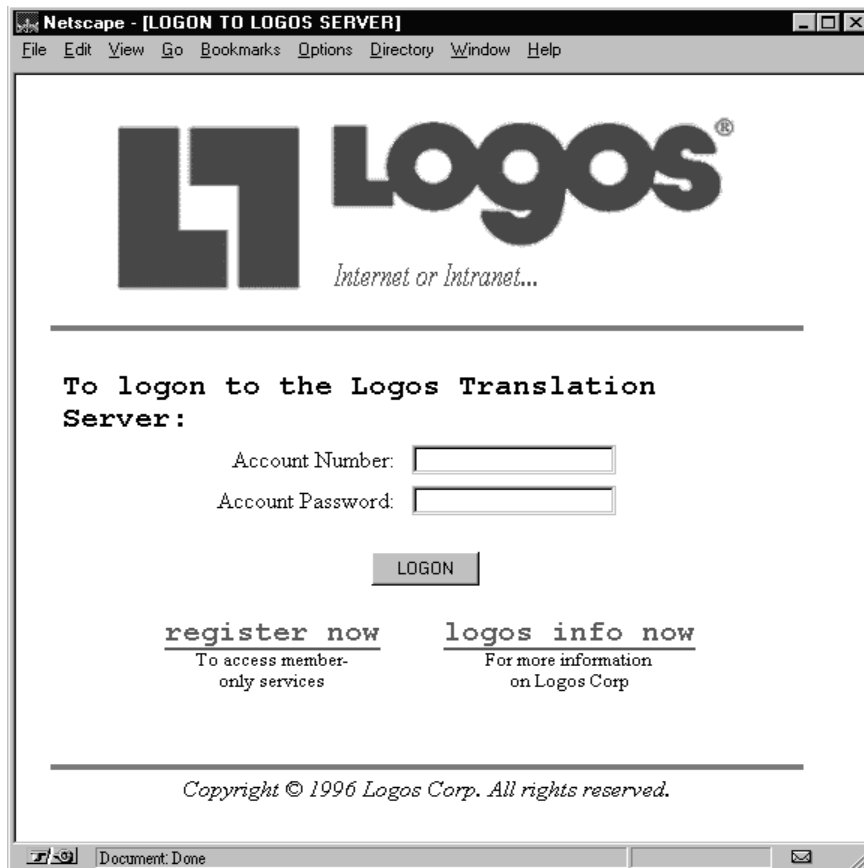
Logos: Product Strategy “AnyNet” Connectivity



The Logos Internet Client (1)



The Logos Internet Client (2)



The screenshot shows a Netscape browser window titled "Netscape - [LOGON TO LOGOS SERVER]". The menu bar includes "File", "Edit", "View", "Go", "Bookmarks", "Options", "Directory", "Window", and "Help". The main content area features the Logos logo, which consists of a stylized "L" and "G" followed by the word "LOGOS" in a bold, sans-serif font. Below the logo is the tagline "Internet or Intranet...". A horizontal line separates the logo from the login section. The login section is titled "To logon to the Logos Translation Server:" and contains two input fields: "Account Number:" and "Account Password:". Below these fields is a "LOGON" button. At the bottom of the login section, there are two links: "register now" with the subtext "To access member-only services" and "logos info now" with the subtext "For more information on Logos Corp". A horizontal line is placed below the links, and at the very bottom of the page is the copyright notice: "Copyright © 1996 Logos Corp. All rights reserved." The browser's status bar at the bottom shows "Document Done" and a mail icon.

LOGON TO LOGOS SERVER

File Edit View Go Bookmarks Options Directory Window Help

LOGOS
Internet or Intranet...

To logon to the Logos Translation Server:

Account Number:

Account Password:

LOGON

register now
To access member-only services

logos info now
For more information on Logos Corp

Copyright © 1996 Logos Corp. All rights reserved.

Document Done

The Logos Internet Client (3)



The Logos Internet Client (4)

The screenshot shows the Logos logo and tagline "Internet or Intranet..." at the top. Below a horizontal line, the text "Select the file to translate:" is displayed. The form contains several input fields and dropdown menus: "File to Translate:" with a text box and a "Browse..." button; "File format is:" with a dropdown menu set to "Rich Text Format"; "Source Language is:" with a dropdown menu set to "English"; "Target Language is:" with a dropdown menu set to "English"; "Brief Description:" with a text box; "Glossary File:" with a text box and a "Browse..." button; "User Dictionary File:" with a text box and a "Browse..." button; "Pattern Matcher Rules:" with a text box and a "Browse..." button; "Processing Priority is:" with a dropdown menu set to "Same-Day"; and "When translation is done:" with a dropdown menu set to "Send via E-mail". At the bottom center, there is a button labeled "Begin New Translation".

The Logos Internet Client (5)

Translation Status Report

There are no jobs that are currently queued for processing

The output queue currently contains 3 job(s).

JobNumber	Date	Time	Comment	Priority	Delivery	Input File	Output File
0	6/17/96	11:28:52 AM		Same-Day	Send via E-m	logostest.rtf	543370.rtf
1	6/17/96	12:56:51 PM		Same-Day	Send via E-m	logostest.rtf	543371.rtf
2	6/18/96	8:29:53 AM		Same-Day	Send via E-m	logostest.rtf	543372.rtf

You may download the translated file for any of these jobs, including those that may have already been delivered via email.

<u>translate</u> <u>file</u> To translate a new document	<u>account</u> <u>status</u> For translation status report	<u>get translated</u> <u>files</u> Get translated documents
--	--	---

Copyright © 1996 Logos Corp. All rights reserved.

Alex for Windows, Sample Coding

The screenshot shows the 'Entry' window of the Alex for Windows software. The window has a menu bar with 'File', 'Edit', 'Window', and 'Help'. Below the menu bar are four buttons: 'Add', 'Delete', 'Previous', and 'Next'. The main area contains two text boxes: 'Source - English:' with the text 'computer scientists' and 'Target - German:' with the text 'Computerwissenschaftler'. Below these are two columns of settings for 'Source' and 'Target'. The 'Source' column has 'Entry Type: Word/Phrase', 'Number: Singular & Plural', and 'Category: Human - Profession/Designation'. The 'Target' column has 'Entry Type: Word/Phrase', 'Number: Singular & Plural', 'Gender: Masculine', and 'Inflection: [-s/-]'. At the bottom, there is an 'Alternate:' field, a 'Part of Speech:' dropdown set to 'Noun', a 'Subject Matter Dictionary:' dropdown set to 'General', and a 'Company Dictionary:' dropdown. 'OK' and 'Cancel' buttons are at the bottom right.

Sample LEF-file

```
#LEF1
FLD_SEP = ; VAL_SEP = ,
S_LANG = (EN)
T_LANG = (GE)
L.GIsCC = (ABC)
L.GIsSMC= (000)

ENTRY Word = S_Word T_Word S_PartOfSpeech T_Gender
#
Bernard;Bernard;N;M
costreduction;Kostensenkung;N;F
Scott;Scott;N;
semantic feature;Semantikfunktion;N;F
translation tool;U4bersetzungsprogramm;N;N
trauma;Trauma;N;N
unedited;nichteditiert;AJ;
```

Network-based Machine Translation Services

Dr. Jörg Schütz

Abstract

The emerging network technologies such as Intranets and the Internet, particularly the World Wide Web (WWW), will create a new model of computing which I will call *Network Computing (NC)*; others, e.g. Forrester Research Inc. call it Internet Computing with the obvious stress on the Internet. Recent analyses of this emerging market conclude that Network Computing:

- Extends and improves client/server technology.
- Will be used by companies to build new customer connections.
- Will change the structure of the industry and the role of information technology as well as the role of language technology.

The focus of my presentation will be multilingual language technology and includes the following aspects:

- What is Network Computing?
- How Network Computing will emerge.
- How Network Computing will change the industry.
- What this means to vendors.
- Impact on users.

I will limit these aspects primarily to network-based translation services, which comprise translation on demand, multilingual communication, multilingual information search and retrieval, as well as customisable translation engines.

For details on Dr. Jörg Schütz and the IAI, please refer to Session 5 cover page.

0. Introduction

In order to thrive in the next millennium industries will have to overcome the confusing revolutions introduced by business and technological areas. On the one hand, the business revolution is characterized by increased competitiveness, the pressure to reduce costs, increased productivity and an increase of responsiveness; on the other hand, the technological revolution concerns the step-wise deployment of distributed systems, PCs, client/server technology, graphical user interfaces (GUI) and open systems. Currently, significant reengineering efforts are underway to taming the complexity of both business processes and computer information systems. The information systems that are needed for tomorrow's global markets must be far more robust, intelligent and user-centered than the data processing systems of today. With the increasing global competition of trade and industry, multilinguality is an additional asset of future information systems, particularly in combination with the ubiquitous information networks (or information highways) that form the digital foundation of the information society (personally I prefer to talk about the knowledge society, cf. below). Just as quantum physics and the relativity theory have changed the thinking of how to describe nature (particularly the recent works of Hawking and Penrose summarized in [Hawking & Penrose, 1996]), the Internet and its multimedial extension, the World Wide Web (WWW or Web), have changed the way of thinking about human communication and the world's globalisation in general (and of course we have the same problems of what this is all about).

Today, we are still faced with the Babel threat, especially when we are surfing on the Internet. Therefore the myth of Star Trek's universal translator is alive more than ever, especially because past experiences, in particular the various evaluation and validation projects, have shown that current machine translation (MT) approaches are not flexible enough to solve the diversity of the multilinguality problems of today's global business and communication situations. The time is ripe for the introduction of a new MT paradigm that is able to combine existing and emerging information technologies and language technologies in an integrated way, thus MT can be seen as a kind of specialized information system (as described above) that is able to contribute to overcoming the existing language and cultural barriers of the world-wide network community. For this, however, a better understanding of the concept of network-based MT is needed. This shall be the first dimension of our investigation.

The combination of Web, information and language technologies will also create a new model of computing, which will revolutionize the existing hardware and software markets. I will call this new model of computing *Networked Computing (NC)*¹. Recent analyses of these emerging markets conclude that Networked Computing:

- Extends and improves current client/server technology.
- Will be used by companies to build new customer connections.
- Will change the structure of the industry and the role of information technology as well as the role of language technology.

¹ Although I have used *NC* as the abbreviation for *Networked Computing*, there is not an intended relationship with the Network Computer (also abbreviated NC) which has been specified by Sun Microsystems, Oracle and others as a low cost Internet PC.

In this model MT will obviously have its place, but its exact locus has yet to be defined. This constitutes the second dimension of our investigation.

When talking about a new MT paradigm founded on multilingual network-based communication capabilities we are not aiming at monolithic local machine translation systems, as they have been on the market for several decades, but at systems that deploy the potentials of the international networks in terms of language resources and software capabilities. Examples of the former type of systems are Logos, Metal and Systran (the real dinosaurs of MT in terms of the employed language and information technology), the huge amount of available PC-based translation systems such as Globalink's Professional Translator, IBM's Personal Translator and Langenscheidt's T1, and the new class of translation support systems known as example-based translation (EBT) or as translation memories (TM) and offered for example by Trados, IBM and others.

However, currently the network scenario with embedded machine translation functionality (MT plug-ins) is the vision which might turn MT (hitherto I will use this term to subsume fully automatic MT as well as machine aided translation and translation memories) to challenging new application domains such as interlingua-based email transmission, the translation of Web pages on demand, interactive machine translation and speech translation in multiparty teleconferencing situations.

Today, existing MT applications are often faced with serious problems when they are employed in a network environment. This can be exemplified by the following description which I received recently by email:

In short, the most efficient way to use [... the TM system ...] is by sharing a translation memory (TM) between several translators on a network. That way translators can benefit from repetition across manuals and they can also control consistency in both style and terminology. However, if your network software is either not state of the art or has too much traffic on it, then the response time for the Workbench is seriously reduced and any increase in throughput figures, thanks to leveraging from the TM, is canceled out. You're faced with even more trouble if your network crashes regularly (something which happens quite a lot in large companies!). One solution is to have a separate network server just for translation memories, but that's quite costly.

In this example, the bottleneck of the application can be manifold ranging from the MT system proper, especially its network capabilities (e.g. concurrent requests), to the surrounding technical infrastructure, i.e. the local area network (LAN) in which it is operating. The MT infrastructure aspects then define the third dimension of our investigation.

To summarize, the specifications of a new MT paradigm which is capable of being deployed as a network-based MT service (local and wide area) have to comply with three dimensions:

1. Concept of network-based MT.
2. Locus of the MT engine and the MT resources (local vs. distributed).
3. Technical infrastructure of MT (hardware and software aspects including costs).

The first dimension has to take into account the application domain of an envisaged translation task, i.e. the purpose of the MT employment; this is also a serious aspect for the evaluation of MT systems. The second dimension is crucial for planning the future of an application and relies heavily on the existing and planned workflow cycles, e.g. the extensibility of the service to other tasks and domains or the treatment of multilingual documents within a company where the recent trends are characterized by systems that allow content-based access to documents (document databases). The third dimension interplays with the second dimension because it is concerned with the overall technical infrastructure in which the MT system is embedded.

The remainder of this presentation is organized as follows: in the first section we discuss Networked Computing in general, we will contrast it with current client/server technology and we will develop a calendar for its evolution within the next three years; in the second section we will extend the Networked Computing scenario with the concept of intelligent software agents which finally results in the definition of an intelligent translation assistant. The third section reports on how we can benefit from these developments yet to come into operation in existing machine translation applications such as outlined in the above example. The paper closes with some prospects of ongoing research and development activities.

1. Networked Computing and Machine Translation

1.1 Networked Computing: Listening vs. Conversation

Networked Computing (NC) can be defined as:

"Remote servers and clients cooperate over Intranets and/or the Internet to fulfill a certain task."

Intranets are corporate networks that are based on the Internet protocol and Web technology; this means the look and feel of an Intranet application is similar to a Web application embedded in a browser such as Netscape, Mosaic, etc. In a NC application the user on the client end will connect to a Web site, but instead of asking for a file (the usual way today) the user will request a session and will receive client code. Once this code is loaded, the client computer and the (network) service server will cooperate, exchange data and communicate. While standard Web applications are based on listening, NC applications are based on conversation.

On the one hand, NC will extend the classical client/server applications which are local, prearranged and limited to a set group of users, because it will have global reach and a potentially massive scale, and on the other hand, it will also enhance Web applications, which handle static documents, by allowing clients and servers to carry on rich discussions based on Web technology, they will use LANs, the Internet, Web browsers and Web-enabled servers.

NC has the advantage that it is:

- **global**, this is facilitated by the standardized communication protocols;
- **frictionless** because maintenance and customization will be drastically reduced by the possibility to download and install client code from remote servers on demand;

- **modular** because NC will rely on software objects to minimize bandwidth and allow for flexible application design; and
- **scaleable**, this again is facilitated by the Internet technology.

The time seems to be ripe for NC because of the existing global standards such as the Internet protocols, browsers, HTML, etc., and the tens of millions of people that are on the Internet. However, NC certainly will be introduced step-by-step:

1996: experimentation; no applications with real customers; building of the basic technology and competences (Java, JavaScript, Tcl/Tk, Virtual Reality, etc.).

1997: deployment of NC applications with focus on Intranets (corporate users).

1998-1999: deployment of NC applications with focus on the Internet.

Obviously, NC will change industry:

- New systems will emerge that are not entirely under the control of current hardware and software leaders, e.g. systems based on VRML (Virtual Reality Modeling Language).
- New and different pricing models will appear, e.g. freeware.
- No software development cycles in the traditional sense will exist.
- New distribution channels will exist, e.g. downloading from the Internet, interactive system upgrades, etc.
- The industrial market will be driven by split-offs, venture-funded start-ups and new private companies, i.e. by a creative divergence.

For software vendors the impact of NC depends on their move towards NC applications which demands new user interfaces, especially multimodal and multimedia based user interfaces. Software will be customized according to predefined application scenarios, updates of software vendors will be done overnight. And last but not least new database players will emerge, in particular those who are concerned with object-oriented databases and those who optimize their products for Java and the emerging Scripting Languages such as Java-Script, Visual Basic Script, Lotus Script, etc.

Users will have the opportunity to run traditional client/server applications in parallel to NC applications. This certainly will demand new skills, both on vendor and on customer sites. Like in every application where time plays a critical role (cf. the example in the introduction section) speed will be a crucial factor. Therefore there is a need for organizational transition, which then also permits new mechanisms for client support.

1.2 *NC environments for MT*

Today, MT systems with acceptable translation results are still very costly and mostly available for high-end workstation platforms. The existing PC-based systems do not accomplish comparable translation results and they are only available for a limited number of language pairs; we particularly start getting into trouble when our application focuses on Asian and Arabic languages. However, similar problems exist

when we look for systems dealing with French as source language (one of the universal languages). This situation is clearly demonstrated on the Internet: multilinguality does simply not exist; the English language is the cyber-lingua of today's information highways.

Obviously, the idea for an MT service available as an Intranet/Internet browser plug-in is fascinating and is becoming more and more feasible with the emerging NC technology. Some MT vendors have already started to offer MT services via the Internet, e.g. the off-line translation of HTML documents (Systran and MTSU) or form/email-based translation services (e.g. CAT2). However, online translation is only provided by the Web site of Rivendell International Communications on a word-to-word basis (this Web site is a very useful resource of language related data; for this Rivendell maintains world-wide links to different resources and services). NC will not only provide online translation services but will also be a basis for other interactive services. Today, the Web is only mildly interactive with just hyperlinks to take the user from location to location.

This situation has changed a bit with the introduction of forms, for example, users can specify search queries or parameters and texts for off-line services. A form consists of two parts: the form itself, which is rendered in the browser, and a Common Gateway Interface (CGI) script or program located on the server. This script processes the user's input mainly to validate the correctness of the user data (e.g. online subscription). This form mechanism is also employed by most of the above mentioned Internet translation services. However, the server executes only one program (CGI script). Since the interaction is immediately direct over the network, this approach causes a higher load of the network. Incorrect and incomplete input can only be perceived on the server side, and the possibility to implement appropriate user interfaces is very limited. This situation led to the idea to execute some tasks on the client side similar to client/server programming. For this the Web browser must be able to run a program which has been implemented by the service provider and which is accessible via the offered HTML page.

Thus, the next step to interactivity is the use of so-called *browser plug-ins*. The employment of plug-ins also reduces the network traffic, which today still is a major problem. However, with the introduction of high-speed networks based on the Asynchronous Transfer Mode (ATM) and mobile communication capabilities via satellites, this will no longer be a real bottleneck, especially because the costs are going down in this technical area. Plug-ins can be installed by running a setup program supplied with the plug-in. Plug-ins reside on the local hard drive and are detected by the browser when it starts up. When the browser encounters data handled by a plug-in (either embedded in an HTML page or in a separate file), it loads the appropriate plug-in and gives access to all or a part of a window. The plug-in remains active until the associated page or file is closed. Currently the programming language Java is mostly used for this application, because it is platform independent. A Java program is compiled into Java bytecode, which can be embedded in a HTML page. A Java enabled browser on the client side loads this bytecode over the network and executes the code. This sort of program is called *Applet*. Other programming languages or scripting languages used for plug-in development are for example SafeTcl/SafeTk, Phantom and TeleScript.

Such a plug-in program can also use additional resources needed by the application and located somewhere on the Intranet/Internet. Thus, plug-ins are the ideal mechanism for handling the trade-off between local and distributed resources and the overall network infrastructure. Today's plug-ins can be seen as the first operational examples of NC.

2. Intelligent Software Agents

2.1 Current and future situation of network environments

As outlined in the previous section, the interactivity of current network-based applications and services is limited, i.e. mildly interactive. Current Intranet/Internet applications provide static Web documents, and the user can use, on the one hand, presentation services provided by Web browsers, and on the other hand, search services as maintained by Lycos, Yahoo, Inktomi, etc. This situation is presented in Figure 1.

With the evolution of NC a new type of software is emerging which will operate on behalf of a user or another program. Thus, it is called intelligent software agent. The Intranet/Internet scenario of tomorrow then might look like it is shown in Figure 2. In this scenario the intelligent software agent acts as the personal navigator of a user. Therefore, we talk about either a personal net assistant or a softbot (software robot).

This softbot cooperates with other software agents for the fulfilment of a specific task by the technical means of NC. The cooperating agents constitute the intelligent network infrastructure, which is the knowledge layer above the information entities, e.g. multimedial Web documents, of an Intranet and the Internet.

A softbot of a user is able to execute different tasks such as

- Information filtering according to user defined parameters, e.g. technical watch in terms of selective dissemination of information and competitive intelligence of large companies.
- Information condensing, e.g. for the storage in digital libraries and in multimedial databases.
- Information brokering, e.g. for online services and emergency applications.
- Telematics assistance, e.g. in telecooperation situations with working cooperations across time zones including the synchronous exchange of working materials and the asynchronous development.
- Translation services, e.g. machine translation and online speech translation for teleconferencing.
- Data analysis, e.g. automatic validation and evaluation of service results.
- Knowledge acquisition, e.g. distance service operations, and distance learning and training.

In order to do this the softbot must have information about the available resources on the Intranet/Internet, either on the basis of local knowledge or through the communication and co-operation with other software agents located on the intelligent infrastructure layer, and information about his master (user), for example, predefined and learned usage patterns, a user model or a specific consumer/customer profile.

Such an intelligent infrastructure service scenario is shown in Figure 3. In this scenario the personal softbot gets an order from the user for whom it is navigating, and according to the user's order and its knowledge about the user, the softbot negotiates with the software agent of a selected supplier who offers the appropriate information, goods or services the user is looking for.

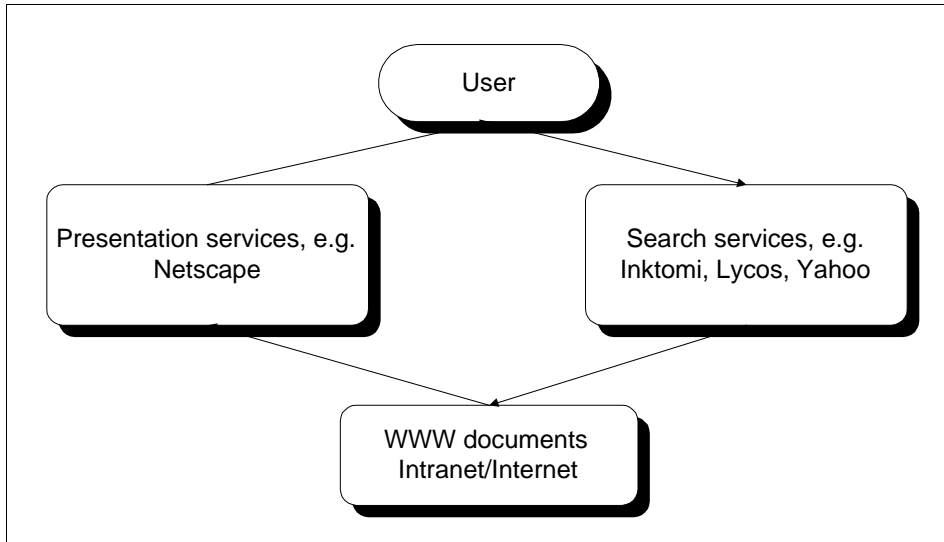


Figure 1: Intranet/Internet situation today

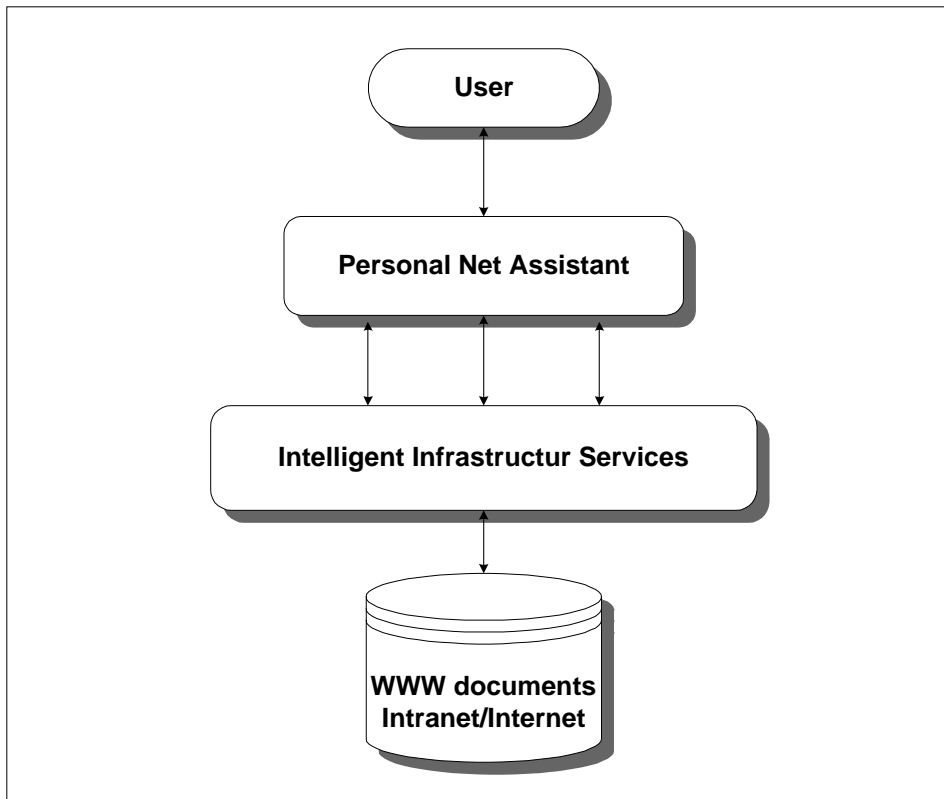


Figure 2: Intranet/Internet situation tomorrow

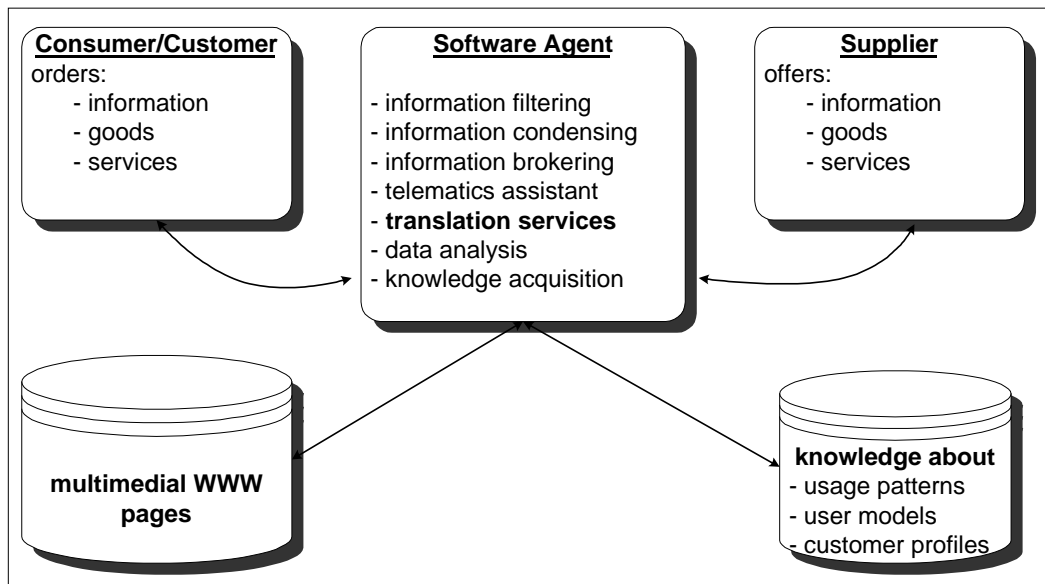


Figure 3: Intelligent infrastructure services

2.2 Intelligent translation assistants

Up to now we have described the technical basis of current and future network-based services, and we have demonstrated how the traditional MT system approaches could be an integrated service within this scenario. The feasibility of the integration in today's network infrastructure is also proven by the recent efforts of some of the international MT system vendors.

However, our envisaged concept of network-based MT aims at the design and specification of an intelligent translation agent, the translation broker, that on the one hand is able to identify and select translation services on the network infrastructure (Intranet as well as the Internet) for a specific application domain, and on the other hand provides a validation and evaluation of the translation results, which will be used in future decision making processes (MT service selection). In addition the MT service agent can inform the personal MT brokers about new features and enhancements of their services according to the principles of NC, i.e. communication and cooperation. This global scenario is shown in Figure 4. In the next section we will discuss this scenario and its realization on the basis of our three investigation dimensions.

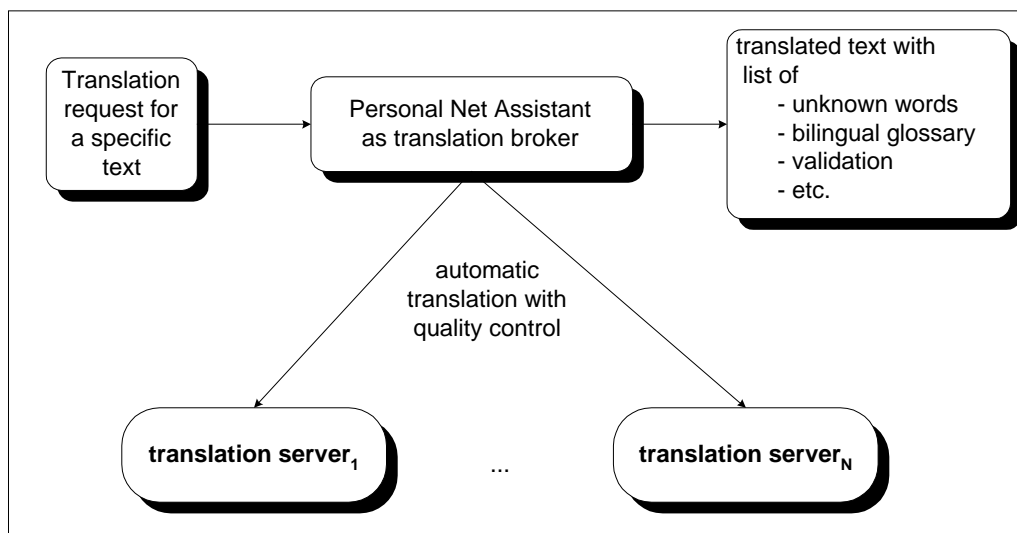


Figure 4: Intelligent translation assistant

3. A New Machine Translation Paradigm

To achieve the goal of an intelligent translation broker we will now develop the theoretical and functional foundation of a new MT paradigm. The need for a translation broker that operates on Intranets and the Internet is obvious, and such a softbot will certainly contribute to overcoming the language barriers of the international trade and commerce, as well as world-wide personal communications.

3.1 Technical infrastructure

The third dimension of our investigation, i.e. the technical infrastructure, can be achieved by the deployment of NC and high-speed network connections (ATM, satellites, etc.). Since the costs in this technical area are dropping drastically, the new investments of a company will be subsumed by their general investments in the field of Web technology. Today nearly every company is investigating or actually investing in the employment of Intranet technology, and most of these companies are already present on the Internet with Web pages or limited services. However, this implies detailed planning, and must be integrated in the general reengineering efforts of the industrial companies. Since the investments in Intranet technology already show an increase in corporate communication and information exchange productivity, as recently reported in market surveys and market researches, especially for the hardware and software market, as well as the pharmaceutical market (cf. [Bowen & Wong, 1996]), the extension to other Intranet/Internet services is prepared.

3.2 Locus of MT engines and MT resources

A remaining open question within this technical scenario is the openness of current MT system vendors with respect to Intranet/Internet based translation services, which go beyond their present efforts, and is of course a matter of competition. This leads us directly to the second dimension of our investigation, i.e. the locus of MT engines and MT resources.

MT users would certainly benefit if different systems and approaches would cooperate to achieve better translation throughput in terms of quality and speed

instead of competing with different approaches. Such a cooperation of MT system vendors could start with a standardization effort for an exchange format of lexical and terminological resources together with national and international projects in this area. This well-defined standard will have also an impact on other areas of network-based services such as multilingual search engines and multilingual text retrieval in general.

A second area of standardization could be the field of distributed environments, for example, the modularization of language resources and technical resources. The idea is to start with general lexical information and to enhance this source of information by additional terminological knowledge or general world knowledge based on shared ontologies (cf. the DARPA knowledge sharing initiative; [Gruber, 1993]) according to the specific translation task of a personal translation broker.

Additional standardization areas could include (automatic) evaluation procedures and processes, and the customization of a brokering task in a way similar to the customization of an intelligent software agent (cf. above). To some extent, these efforts can be compared with the ongoing initiatives in the field of distributed language engineering; there the focus is particularly in the area of modular grammar engineering. A cooperation of MT vendors and these initiatives would be of mutual benefit for each party.

Last but not least, the standardization and the sharing of the technical MT infrastructure and the linguistic MT infrastructure contributes to reducing the costs of both MT system development and MT system maintenance. This certainly will also allow for better translation results and thus contributes to the further development and evolution of the information society.

3.3 Networked Machine Translation

The overall concept of network-based MT, our first dimension, can now be defined as:

"Remote translation servers and translation brokers (clients) cooperate over Intranets and the Internet to fulfill a translation task."

This definition implies the relationship between network-based MT and Networked Computing on the basis of intelligent software agents (cf. Chapter 2). Actually, in this scenario a translation broker will request a translation session from a MT server, which will provide a basic client code for a specified translation task. Once this client code is loaded, the translation broker and the MT server will cooperate (e.g. specific requirements for the translation result), exchange information (e.g. local lexical resources) and communicate (e.g. for the identification of additional resources).

To summarize, the technology for network-based MT is emerging with the recent efforts and developments in Networked Computing and language technology. I will call this new MT paradigm *Networked Machine Translation (NMT)*. The success of this new MT paradigm depends on the steps the MT industry will take in this direction. Traditional MT approaches, including translation memory approaches, are integratable in this scenario. However, the actual effort for the design and the implementation of appropriate interfaces and APIs (Application Programming Interface) will differ from system to system, and will depend on their already existing network capabilities.

It should be noted that search engine developers have already initiated standardization efforts in the field of language resources together with language technology companies. Some of these efforts are spin-offs of former R&D projects funded by the European Commission.

In this context the on-going European project OTELO may also contribute to the further development and the realization of the presented new MT paradigm. The OTELO consortium is coordinated by Lotus Development Ltd. in Ireland and consists of several language technology related European companies, including MT vendors. According to the project description ([SAP, 1996]) the aims and goals of the OTELO consortium comprise:

- Defining standardized common lexical resource and text-handling formats.
- Providing an array of tools which will increase the quality and efficiency of machine translation.
- Developing a network infrastructure that makes NLP tools widely accessible.
- Integrating access to the OTELO network into the groupware framework.

I am looking forward to the first results of this project which certainly can contribute to the further evolution of Networked MT.

4. Conclusions and Prospects

In this presentation we have developed the initial prerequisites of a new machine translation paradigm based on the recent and emerging developments in the fields of information technology, Web technology and language technology. This new MT paradigm aims at combining the different technologies to achieve a better translation throughput in terms of translation quality and translation processing speed by means of cooperating and communicating network-based intelligent software agents, which make use of local and distributed language resources that are based on agreed standards. The language resources of this new kind of MT software is intended to be sharable with other Intranet/Internet based language technology applications and services, such as multilingual search engines, text retrieval, information brokering and telematics services.

Security and privacy issues have not been addressed in this presentation, although this topic is of crucial importance for both service providers and service customers. This is mainly due to space limits; an in-depth discussion of this topic would certainly be beyond the scope of this presentation. However, the available Intranet/Internet security measures, such as Firewalls and Phil Zimmerman's PGP (Pretty Good Privacy) freeware, would be sufficient for the investigated new MT application.

With the further evolution of the information society and the need to overcome the language barriers of global trade and commerce, the time is ripe for further R&D work towards the actual realization of this new MT paradigm.

In closing this presentation, I would like to come back to the network problem cited in the introduction chapter. Is there any direct answer to their problem?

The described application is a strict LAN (Local Area Network) application and the network failures seem to be based on the used communication protocols but not on

the used hardware (old Ethernet connections are still very reliable). Therefore I would suggest the analysis of the network infrastructure on the protocol basis including the network capabilities of the employed TM software and its use of language resources. This analysis should also include the investigation of the portability of the TM to an Intranet application with the support of the TM vendor.

5. References and Suggested Further Readings

- [**Andreessen et al., 1996**] Marc Andreessen & The Netscape Product Team, 1996. *The Netscape Intranet Vision and Product Roadmap*. Netscape Communications Corporation, Mountain View, CA.
- [**Bowen & Wong, 1996**] Barry D. Bowen and Carolyn W.C. Wong, 1996. *Spinning the internal Web*. Sun World Online 4/96.
- [**Dean et al., 1996**] Drew Dean, Edward W. Felton and Dan S. Wallach, 1996. *Java Security: From HotJava to Netscape and Beyond*. In: Proceedings of the IEEE Symposium on Security and Privacy, Oakland, CA.
- [**Fingar et al., 1996**] Peter Fingar, Dennis Read & Jim Stikeleather, 1996. *Next Generation Computing: Distributed Objects for Business*. SIGS Books & Multimedia, SIGS Publications Inc., New York.
- [**Hawking & Penrose, 1996**] Stephen W. Hawking and Roger Penrose, 1996. *The Nature of Space and Time*. Princeton University Press, Princeton.
- [**Gruber, 1993**] Thomas R. Gruber, 1993. *A Translation Approach to Portable Ontology Specifications*. Knowledge Systems Laboratory Technical Report KSL92-71, Stanford University, Stanford, CA.
- [**SAP, 1996**] SAP, 1996. *OTELO - EU Project LE-2703*. SAP AG, Walldorf, Germany.
- [**Stoll, 1995**] Clifford Stoll, 1995. *Silicon Snake Oil - Second Thoughts on the Information Highway*. Macmillan, London.

Conclusion

Viggo Hansen, EAMT Secretary

To make a long story short, it has proven to be beneficial to use Machine Translation provided that:

- the user has substantial quantities of machine-readable, domain specific text to be translated,
- the MT process includes a well planned work-flow with easy term identification and coding facilities, an effective pre- and post-editing tool and a reliable full text translation system,
- the MT supplier has user experiences and preferably is an institution/company with commercial objectives.



Membership Application Form

Name

Department

Company/Organisation.....

Address.....

.....

City.....Postcode

Country.....

Telephone.....

Fax.....

E-mail.....

Please check here if you would prefer that your name is not listed in the EAMT membership directory.

Membership fees (per calendar year)

Individual (SFR 35) Non-profit (SFR 175) Corporate (SFR 350)

Method of Payment:

Cheque, payable to EAMT, enclosed

Bank transfer to account number 351.091.40L, Union Bank of Switzerland, Bahnhofstrasse 45, CH-8021 Zurich, Switzerland (Note: all bank transfer charges must be borne by the applicant).

Please send completed form to:
EAMT Secretariat
ISSCO, 54 route des Acacias
CH-1227 Carouge (Geneva), Switzerland

Fax +41 22 300 1086