[*Translating and the Computer 19*: Papers from the Aslib conference… 13 & 14 November 1997]

# Remote-Access Translation Services: Software Design with the User in Focus

**Klaus Schubert**
**Fachhochschule Flensburg**

**Abstract** The growing market of translation services, whether human or machine, accepting orders from customers via the Internet or similar means of communication has brought about a demand for suitable tailor-made service environment software to handle (1) customer access and task management, (2) access rights to linguistic resources (dictionaries, translation memories, proprietary terminology), (3) application of customer-specific linguistic quality assurance conditions, (4) teleworkplaces for remote linguistic resource maintenance. The study takes stock of the parameters involved, draws up a number of workflow scenarios and derives elements of a generic functional design of such a system from a user's point of view.

## 1 Remote-Access Translation Services

Teletranslation is a coming industry, to use Minako O'Hagan's words (1996). What is new in this? *Tele* means 'at a distance', but obviously translation services have always been provided by agencies situated at some distance from their customers. Traditionally, the gap was bridged by express letters, then by fax and now by the Internet. Is there a *qualitative* difference between communication through a computer network and more traditional means of ordering and delivering a translation? And if so, how is this to be reflected in the software designed for this new purpose?

As far as the shift towards teletranslation can already be observed or predicted, the development follows the three-stage model which has been typical for automation both in our present computer age and in much the same way in the days of the mechanical automation known as the industrial revolution. In the first stage there is a traditional solution, in the second a new medium of automation carries out the same function by other means, possibly quicker and cheaper but still copying the previous procedure quite closely, and only in the third stage are functions and capacities of the new solution discovered and systematically applied which were impossible without it. Teletranslation services are a coming, and novel, industry inasmuch as they are attaining the third stage, making use of information technology in a way which goes beyond that which could be done before.

This study takes stock of these new parameters of functionality with a view to the design of the software needed for remote access to translation services in a way which allows for a service profile customized to a broader and much more varied range of user needs and usage conditions than with traditional communication media. This is a study about putting human and machine translation to work in a telecommunication-mediated environment.

The starting point for all considerations of the lay-out and design of future remote-access translation services is the need experienced by its potential users. Yet, not every user's wish list is a useful requirements definition for an innovative development. Among the not seldom disappointed users of machine translation systems, for example, there is a tendency to word no medium-term, let alone long-term, vision at all but only the most modest wishes for close-at-hand improvements totally within the limits of available systems (cf. e.g. Schubert 1994). The present study therefore steps back a little and tries to look into the possibilities of this shift of medium, which has already begun, from a more principled point of view.

After a short review of user requirements in section 3, section 4 looks at professional skills needed and available instruments of automation. Section 5 draws up a series of workflow scenarios taking into account different degrees of automation and different ways and locations of resource management and maintenance. These overviews form the background for section 6 where design requirements for the envisaged remote-access translation service software are outlined. Section 7 then tries to look further into the future.

## 2 Some Available Services

Translation services which can be accessed remotely via a wide-area computer network have been available for some time. As far as an observer can judge from published reports, some of these services have been successful within the limits of what is usually called a success in the world of machine translation, though no company has been reported to have achieved the momentous breakthrough in user numbers or subscription revenues which can be expected from a high-quality translation service at every Internet user's fingertips. Nevertheless, market analysts (e.g. Lockwood/Leston/Lachal 1995: 142) predict an upsurge in Internet-based translation services within the next years.

Without giving an exhaustive overview of available services, this section looks into the types of service which are available at the time of writing.

Before turning to the services themselves, a few words on the communication medium are in order. When speaking of remote-access translation services or teletranslation, one normally has the Internet in mind. Indeed most of the services to be discussed below are accessible via the Internet, either by e-mail or in the World Wide Web. This is possibly the broadest available medium to reach the general public. At least in traditional translation services, however, the general public is by no means the largest customer. Non-literary translation is by and large a service offered by language professionals to professionals of other industries and thus the primary target group of a professional remote-access translation service is, in my view, not the general public but trade and industry, followed by government authorities. For these customers it is becoming less and less attractive to wait the unpredictably long response times in the Internet. Moreover, sending corporate documentation, law proposals, medical reports and other confidential and sensitive materials over a public network may deter these customers. A professional translation service will therefore have to consider other forms of computer network connection as described below (section 4.2).

This need is not only felt in translation services, so that the requirements of speed and confidentiality may in the next decade promote a development in which, alongside the Internet for the public, new corporate networks and networks for industrial subscribers are established. The translation services will have to be aware of these developments to make themselves readily accessible to the right type of customer.

Among the available machine translation systems, there are systems which lend themselves to professional translation work for trade and industry and others which seem more aimed at the general public. The difference is not primarily in the linguistic quality of the raw translation output of the untouched system. The main quality difference can only be measured when the translation quality is assessed which results from the combined efforts of the system and its maintainers. This does not mean that the post-edited machine translation output should be taken as a basis for an evaluation of the system. It means that those improvements in the linguistic performance of a system should be taken into account which the user makes not to the target text but to the resources of the system. These have a lasting effect in that all translation jobs carried out after such an improvement will profit from the work done only once, whereas post-editing needs to be carried out for every new text again. The linguistic quality of a machine translation system should therefore be evaluated on the basis of the raw translation output plus the effects of a skilled user's interaction with the system's resources. The resources in question are the dictionaries where words and word groups with their translation equivalents can be entered, but to some extent also rules of syntactic transfer (cf. Schubert 1996).

In view of this understanding of the quality of machine translation systems, it is possible to distinguish systems which may achieve a quality sufficient for professional services to trade and industry, and others which simply do not allow for the required interaction because they are not designed for expert users.

The services offered to the general Internet public include systems such as Globalink's Power Translator, Intergraph's Transcend and a new system called Diplomat. One user's experience with Power Translator is reported by Frank Berberich (1996, 1997), while Atoosa Savarnejad (1997) briefly reviews

Diplomat. The Transcend experience is better known. The system is used in CompuServe's network. Mary Flanagan earlier reported on it and does so again in this conference (Flanagan 1995, 1997; cf. also Berberich 1996, 1997).

Typically, the larger machine translation systems are not offered in the same way. The providers direct themselves more directly to trade and industry. An attempt at a subscriber access to services based on the METAL system was the DTS environment developed at Siemens Nixdorf Informationssysteme GmbH (Masion 1995)[1]. A service based on the LOGOS system is described at this conference (Hatley 1997).

Large organizations, both companies and authorities, are aware of the need to view translation services not as an isolated business, but as an activity integrated in the overall documentation and information processing which is so closely interwoven with present-day production, service business and administration. Networked translation services will therefore increasingly be included in document-processing environments of a broader scope than just translation. The European Union's EURAMIS project (Leick 1995,1997, Blatt 1996, 1997, Brace 1997) is a point in case, although corporate solutions certainly need a considerable amount of engineering to suit the specific needs to be met there.

## 3 The User Side: Types of Translation Need

The parameters of possible user requirements include linguistic conditions on the type of source text to be translated and the quality of the target text as well as technical conditions on the format of the source and the target text. Another series of conditions concern the management of the work and the required resources. These are the question of where which resources are kept and who maintains them, along with conditions on the time, speed and cost of the service, data security and confidentiality, copyright issues etc.

In somewhat more detail, the following questions may influence the choice of parameters in the solution to be chosen.

*Linguistic Parameters*

- Grammatical correctness of source text: well-written text by skilled native author; broadly correct but somewhat unidiomatic text by skilled non-native author; loose informal notes.
- Required grammatical correctness of target text: first-rate publishable quality; good quality with visible features of "translatedness"; raw translation.

[1] After the major reorganizations in SNI in 1995-96 DTS is still available, though outsourced (personal communication, SNI Sprachendienst/Keith Roberts, Sept. 1997). It is my personal impression that the service is not being widely advertised.

- Syntactic variation: normal freely worded text making use of all syntactic constructions within the limits of the grammar vs. standardized messages in some restricted format; ordinary vs. controlled language.
- Lexical variation: broad variation of subjects in a single text vs. texts restricted to a narrow subject field; consistent vs. varied use of terminology in a single text.
- Linguistic quality requirements: prescribed or preferred corporate terminology; proprietary vocabulary; consistency with previous versions of the same or related documents or with corporate style.

*Technical Parameters*

- Formatting: format of the source text and formatting requirements of the target text: formats of word processors, desktop publishing and more professional document creation systems, generic mark-up formats such as SGML and HTML etc.
- Composition of the document: plain text vs. text with tables, figures, captions, headers, footers etc.
- Access software environment: user interface at the customer's side as a separate application vs. function integrated in available software.

*Management Parameters*

- Storage and maintenance of linguistic resources: dictionaries, documentation, translation memories etc. stored at customer's site or translation agency (or with translators); maintenance at or remote from any of these sites.
- Speed: maximal allowable time-span between request for and delivery of translation.
- Time: translation requests and deliveries during vs. outside normal office hours.
- Cost.

# 4 Required Skills and Instruments of Automation

Before trying to make the requirements and the possibilities meet, it is useful to have an overview of the professionals needed and the instruments of automation which are at their disposal. Further, it should be made clear which forms of automation are suitable under which conditions.

## 4.1 Degrees of Automation

For *manual translation,* the translation agency needs a translator and usually a reviser. Network communication can create new possibilities for an agency to resort to remote translators and revisers. This links the agency to free-lancers, external employees and partner companies in a fast way, and it may contribute to enhancing the linguistic quality of the translations, if the experience applies that translators deliver better translations when living in the community of the target language. The resources used in manual translation are normally dictionaries and documentation. Documentation can include texts from subject fields similar to the current translation task, in either the source or the target language or in both. It can further include previous versions of the current text, with or without previous translations.

It is easier to automate some of the resources than the translation process proper. This yields *computer-aided translation.* The dictionaries may be replaced (or, more realistically, complemented) with electronic dictionaries and terminology databases. These may be equipped with an active term recognition function which searches the text under translation for terms for which there is an entry in the database, offering the hits to the user. If the texts to be translated are terminologically and stylistically monotonous and repetitive, a translation memory may be used to good advantage. Especially in the case of documentation regularly published in new updated versions, this type of instrument is expedient. The professionals needed are a translator and a reviser as in fully manual translation, plus possibly a terminologist. In the case of large volumes of text handled by e.g. a team of translators, some maintenance of the translation memory is necessary.

If an attempt is made to automate not only the resources and possibly the process of consulting them, but also the translation process itself, a *machine translation* system can be used. In some very specific cases such as information skimming, a raw machine translation output quality may be sufficient to meet the user's requirements. Otherwise, the texts need to be post-edited. In practice, professional machine translation users have found that some pre-editing, in part itself automatable, may increase the productivity of a machine translation system. A machine translation system requires the source text to be computer-readable. It needs to be equipped with a dictionary and some systems can make use of a translation memory. The professionals needed are a machine translation operator, a reviser/post-editor, a dictionary coder and, if applicable, a translation memory maintainer. The role of translator has only apparently disappeared, since revision of translation is work for an experienced translator. There is an obvious difference, however, between revising human translation work and post-editing machine-translated raw output.

It should be noted that the list of resources includes documentation for the manual translation method but a translation memory for the machine translation method. The two things are not the same. It is well known that translators can exploit useful information found in monolingual documentation in either of the two languages and even in other languages. A translation memory, though sometimes labelled "documentation" by the manufacturers, does not contain this broad range of information sources, but only a very narrow portion of what a human translator can use: previous translations. Materials contained in a translation memory are helpful to (human or machine) translation only when there is a relatively large degree of similarity between the current sentence to be translated and a sentence stored in the translation memory. Available systems match task and memory by technical (rather than linguistic) means and get good results in some cases and poor ones in others.

A linguistically more meaningful matching technique is tried by the Kielikone team (Arnola et al. 1996) who seem to be implementing the theoretically long-demanded method of matching sentences in the form of syntactic dependency trees rather than plain strings. This is a first step towards a possible quantum leap in translation memory efficiency: what is repetitive in language is not sentence strings (as assumed in the matching algorithms underlying current products[2]) but, at a structural level, elementary syntactic dependency patterns (Schubert 1986: 44), representing syntagmata with slots for inserting syntactically relatively freely formed pieces of sentence (rather than fixed word strings).

## 4.2 Which Automation for Which Requirements?

Essentially, what is described in section 4.1 are the major components from which a translation service can be composed. It is interesting to see how the computer network technology can contribute to making an innovative service out of these elements. Before arranging the components into workflow scenarios in section 5, I discuss which of these instruments of automation meets which of the user requirement parameters.

From the user requirements listed in section 3, the linguistic parameters condition the decision whether or not a *translation memory* or a *machine translation system* will be profitable to the envisaged service.

*Grammatical correctness of the source text* is a prerequisite for machine translation systems available at present. Only rigorously correct texts lend themselves to handling by these systems. As long as all machine translation mechanisms have to rely on form features of the text (such as morphology and syntax) and most mechanisms labelled semantic in fact are more or less efficient circumventions of semantic or knowledge-based selections, the systems need every little cue to arrive at correct syntactic analyses and transfers and cannot allow for any kind of error-friendly variation in their rules and algorithms. Non-native writing may at times be less easy to machine-translate than genuine texts. Informal writing, such as e-mail or newsgroup chat, is even more problematic for this instrument.

When the parameter *of grammatical correctness of the target text* can be released, a machine translation solution can become very profitable. Translation work of first-rate publishable quality will need to be done entirely by human translators, and indeed it often involves a considerable amount of original text work or copywriting.  A flavour of translatedness is often acceptable where documents are

[2] As far as the matching algorithms are known, there appears to be a line of heritage from the old Translation Support System to present-day products such as Translation Manager/2 by IBM Deutschland Informationssysteme GmbH, Translator's Workbench by Trados GmbH and Transit by STAR GmbH (cf. Melby 1995: 225-226 note 73). To the best of my knowledge, Translation Support System by ALPS was the first system to realize Alan Melby's idea of a translation memory in the form of what was labelled repetition files. The company, now called ALPNET, ceased to market the system a couple of years ago, but a somewhat revamped version is reported to be used by the company's in-house translators (personal communications from ALPNET Stuttgart, Germany, and Ieper, Belgium, both 1997). Black-box analyses of the matching algorithms are reported among others by Bischoff (1994) and Moller (1996).

drafted, written and revised in an entirely multilingual environment (e.g. European Union documents, working documents in particular). This is a field where post-edited machine translation output may well meet the quality requirements, especially when quality is seen in a trade-off towards speed and cost. Unedited machine translation output is used for information skimming and related purposes, where speed and cost-effectiveness count more than linguistic quality (e.g. Nuutila 1996).

A *restricted syntactic variation* in source texts, as in standardized messages and notes and especially in controlled-language documentation, often allows for a machine translation solution with shortcuts, because the restricted grammar reduces to one the number of options from which to choose in cases that would be ambiguous in unrestricted language. In the case of large numbers of short, stylistically uniform messages, machine translation can be a very efficient instrument. The same holds for a translation memory or a combination of both instruments. Controlled language is an attempt to carry the idea of artificial non-ambiguity further to allow for longer and more varied texts. In the case of longer controlled-language texts the repetitivity may be lower than with many uniform short messages, so that a translation memory may become less efficient, while an adapted machine translation system will still give good results.

While a restriction of the syntactic variation is a way to avoid syntactic selections in the automated translation process and thereby facilitate the use of machine translation, *restriction of the lexical variation* addresses in a similar way a much more notorious issue in machine translation. This is the problem of semantic, pragmatic and knowledge-based selections for which essentially no feasible general-purpose mechanism is known to have left the research state. As these selections are particularly frequent in lexical transfer, a very common workaround in available machine translation systems is the classification of source-target word equivalencies into subject fields. When starting a translation job, the user is prompted for a subject field (or a set of hierarchically ordered subject fields) from which translation equivalents are then preferred in lexical transfer. Much can be said about this method, but what is immediately obvious is that if a subject field parameter can be set only once to apply for an entire document, texts with consistent use of terminology from a single subject field can be machine-translated better than thematically varied polytechnic texts, overview articles etc. (cf. e.g. Lukas 1994).

The combination of restricted syntactic and lexical variation is the "secret" behind the world's only fully automatic high-quality machine translation system, Météo, which translates nothing but weather bulletins.[3] The sublanguage approach tries to generalize this use of translation-enabling restrictions.

When a customer has explicit *linguistic quality requirements,* such as a prescribed terminology, corporate vocabulary or the like, it may be quite laborious for a translation agency to enforce these with their translators and revisers because a human mind cannot be simply switched to consult a certain list or database before using its own knowledge. In suitable cases, terminology databases, translation memories

[3] At a closer look, not even Météo is fully automatic, though often referred to as such. It is reported to include a recognizer function which singles out sentences the system cannot translate automatically which are then submitted to human translators (Slocum 1989: 637).

and, other requirements permitting, machine translation systems are much easier to tune to this type of requirement.

Among the *technical parameters,* the *formatting* of source and target documents is of utmost practical importance to all professional users of partially automated translation services. In a situation in which a translation memory or a machine translation system could be used to good advantage, the benefit may be outweighed by the effort required for manually reformatting the target text unless the systems can produce output in the right format. The need for a cross-vendor document formatting standard has been urgently needed in this field for many years. Maybe the general striving to the Internet and the transfer of Internet techniques to Intranet solutions will promote an HTML-type de facto standard.

In a similar way, the *composition of the source document* plays a role. A translation memory or machine translation system which handles running text only but cannot deal with headers, footers, tables, captions, text in figures etc. or cannot preserve them through format conversion, achieves a lower degree of automation and thus needs more manual post-editing than a system which can. As long as the document contains these pieces of text in some text format, the problem should be resolvable. When text is included in illustrations in a graphical format, however, the instruments and efforts needed to have it handled by these systems are prohibitive, so that manual translation is the most feasible solution. This is one of many factors in the trade-off between quality, speed and cost, mentioned below.

A technical parameter of immediate importance to the design of the translation service software is the question of the *access software environment.* The options are a separate application to be installed at the customer's workplace or an additional function to be included in some application which the customer has in daily use. The former solution would be easier to implement for the software developer, but it has several drawbacks. The need to familiarize oneself with a new piece of software may deter customers, particularly occasional users. It would also mean that a customer would have to contact the translation service provider and buy or download and install some software before being able to use the service. An access functionality for the translation service included in a number of standard software packages would make it easier for customers to use the service. This may be an important competitive factor when a service provider wishes to attract spontaneous customers. If it is unrealistic for a translation service provider to have their special access functionality included in widely marketed standard software such as word processors and Internet browsers, a solution with anonymously downloadable plug-in or add-on access software is in line with present marketing tendencies. How and at what price, if any, to spread the access plug-in depends on the overall pricing and accounting strategy.

At present, a remote-access translation service of the kind described here should allow for a variety of different forms of access including Euro-ISDN, direct login by modem, WWW browser software and e-mail for the public networks – with additional solutions for corporate and Intranet environments (e.g. Lotus Notes or SAP R/3 clients).

Another set of parameters concerns the *management* of the translation service and the business relation between the customer and the service provider.

One of these is the question of *storage and maintenance of linguistic resources.* The linguistic resources involved are mainly dictionaries (in paper or software form),  documentation (paper or software) and translation memories (software). Especially in the case of the software resources, the options are keeping resources at the translator's and the reviser's, the translation agency's or the customer's site, letting a translator or a machine translation system access a resource locally or remotely and having the required maintenance work carried out locally or remotely. The translation service software will have to allow for all combinations of these options.

The required *speed* and the allowable *cost* of a translation job are in a direct trade-off relation to each other and to the *quality* parameter. When while-you-wait translations are needed, only raw machine translation output can meet the speed requirement and the customer has to make do with an unedited quality. The same holds for the *time-of-day* parameter: if translations are needed during the night or at odd times of the weekend, a fully automatic machine translation service which receives and returns the translation job may be a feasible solution. (A through link to a partner company in another time zone may be another good option.)

# 5 The Service Provider Side: Workflow Scenarios

A translation agency wishing to meet the broadest possible range of customer requirements will derive their service profile from an inventory of parameters similar to the (incomplete) overview in section 3. From the requirement parameters, a set of basic service functions can be derived which can then be arranged to make up a specific workflow scenario for each of the different types of expectable translation order.

In this section, I present a number of such workflow scenarios. They all show the relationships of workflow and dataflow between a customer and a translation agency and between the translation agency and its, possibly external, language professionals. These scenarios are but a basic set. Variations are possible which I do not show here. In particular, the scenarios are based on the assumption that the customer, the agency and the professionals are all in different places so that some structured form of communication is needed between them. This parameter can be varied: the agency can have all its professionals in house, or the customer and the agency could be at the same place (the agency thus possibly being a translation or documentation department), or all three parties could be situated under the same roof. But even so, it is not uncommon to connect departments or individual professionals by an in-house network in a fashion similar to the scenarios shown below. The central issue here is the new dimension added to translation services by the computer network technology - in other words, the shift from automation doing the same work as before by other means to automation exploiting the specific, innovative capacity of its new medium (cf. section 1).

The basic service functions from which the workflow scenarios below are composed are the following. They derive directly from the requirement parameters in section 3.

- Dynamic automation: automation of the translation process proper.

  Is a machine translation system available for the required language pair? Is the quality of the source text suitable for machine translation? Is the required target quality low enough for fully automatic translation? Does the trade-off between quality requirements and post-editing effort favour machine translation?

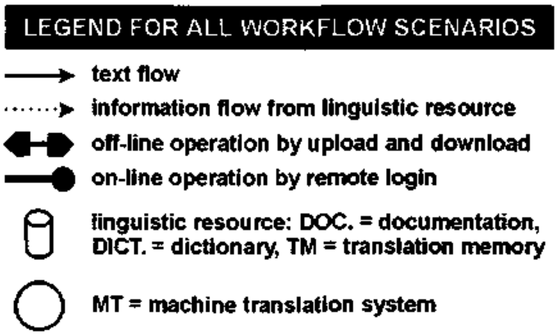
- Static automation: automation of the resources.

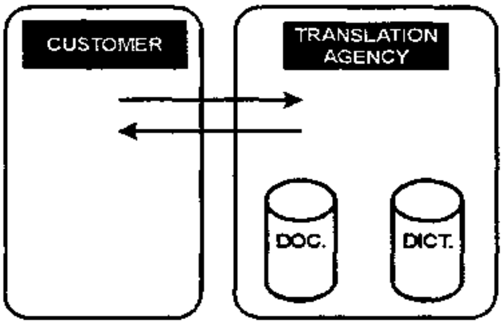
  Are dictionaries available for the relevant language pair and subject area? Is documentation available? Are previous versions of the text to be translated available in the source or the target language or both? Does the customer require consistency with previous versions, previous translations, similar documents, prescribed vocabulary lists, standards or the like?

- Growing degree of automation: maintenance of the resources.

  Who owns and who manages the available linguistic resources? Who maintains them? Who owns and who manages possible updates created in connection with the translation job? In which form can (relevant parts of) the resources be made available to the professionals?

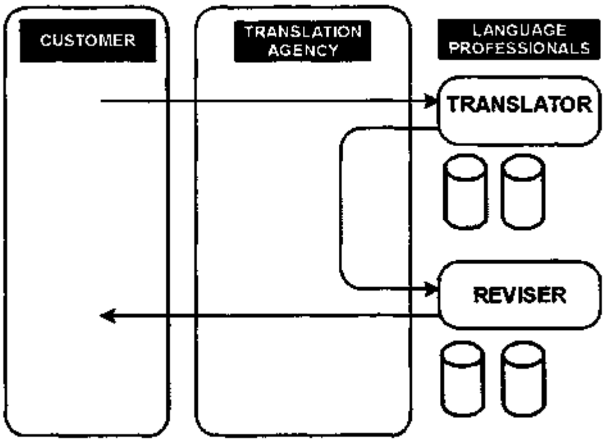
From these considerations, the following workflow scenarios can be derived.


LEGEND FOR ALL WORKFLOW SCENARIOS
text flow
information flow from linguistic resource
off-line operation by upload and download
on-line operation by remote login
linguistic resource: DOC. = documentation, DICT. = dictionary, TM = translation memory
MT = machine translation system



Workflow Scenario 1

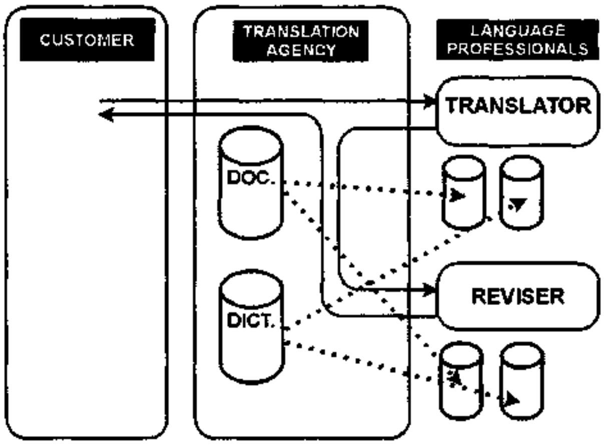
Workflow Scenario 1 shows the traditional situation, in which the customer sends only the text to be translated to the translation agency and receives from the agency only the target text. The agency holds their own linguistic resources, especially dictionaries and documentation. This scenario does not show the workflow within the agency.



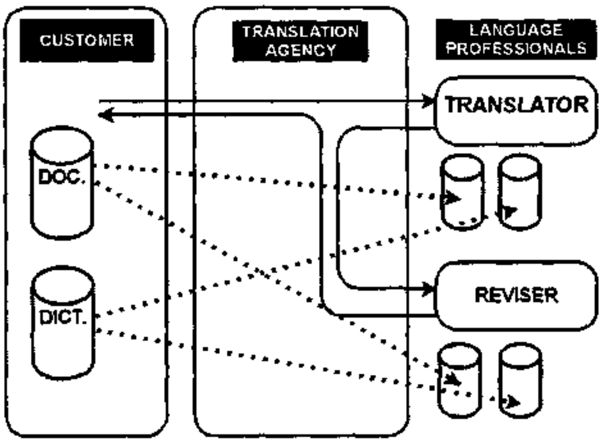
Workflow Scenario 2

Many translation agencies resort to external translators or revisers, so that in these cases a scenario arises in which linguistic resources are held at two different external places. These resources need not be congruent or even consistent with each other, and the translation agency need not necessarily own any linguistic resources of its own. This is *Workflow Scenario 2.*

Quality requirements may lead the translation agency to a policy of quality assurance, enforced among other things by ensuring consistency among the linguistic resources used by the translators and revisers. A frequently used method is sending the relevant word lists together with a translation job. This often means that a specific part from the agency's larger database is extracted and sent to the translator or reviser in print-out, text file or database form. This is *Workflow Scenario 3.* It is a task for the agency's quality management to make sure that the extracts sent to the translator and the reviser are consistent, that the translators and revisers understand whether or not they are supposed to keep the extracts and whether they should use them for the next assignment or expect the agency to send new dictionary extracts with each new job.



Workflow Scenario 3

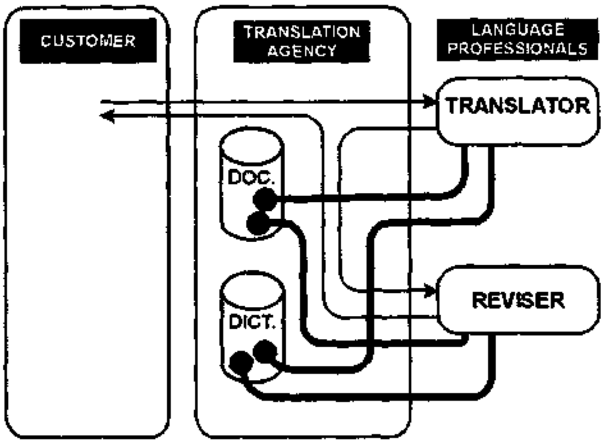


Workflow Scenario 4

Thus far, the workflow scenarios assume that the customer does not contribute to the overall translation process beyond sending off the translation job and receiving the outcome of the work. However, quality assurance measures are not only taken by translation agencies. The requirements equally often originate from the customer, and this frequently entails an obligation for the agency to use vocabulary prescribed and made available by the customer. Obviously, as in Workflow Scenario 3, these materials should be forwarded to the translators and revisers. This is *Workflow Scenario 4.*

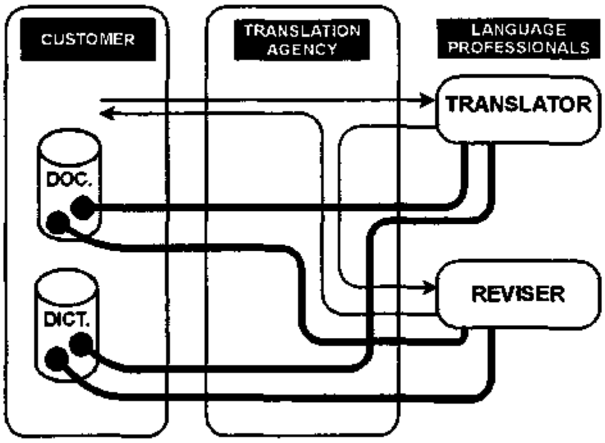
Various combinations of Workflow Scenarios 3 and 4 are possible. The translators could be given resources both from the customer's and from the agency's archives, or the customer might provide only documentation but no dictionary data or vice versa. The overview of scenarios shown and discussed here is not intended to be exhaustive.

Another parameter can be varied. It does not affect the customer, at least not directly. Workflow Scenario 3 shows how a translation agency can make its resources available to external translators and revisers by sending them relevant extracts from a possibly larger collection of data held by the agency. Instead of extracting materials and sending them, the agency can as well choose to give the professionals access to the archives at the agency, and the computer networks offer an opportunity to do so by remote access, as shown in *Workflow Scenario 5.*



Workflow Scenario 5

Virtually the same solution is possible with resources at the customer's site, as shown in *Workflow Scenario 6.*



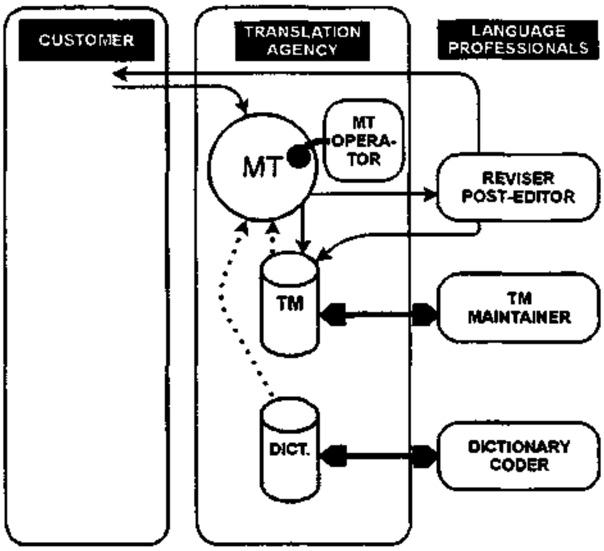
Workflow Scenario 6

Workflow Scenarios 1 to 4 can also be realized by traditional means of communication. If a computer network solution is used, it does not function in any essentially different way to the traditional solutions. This is stage-two automation in the sense of section 1. For the first time in this discussion, scenarios 5 and 6 show a functionality which was impossible until a computerized remote-access solution became available.

The six scenarios shown above vary the parameter of the availability of linguistic resources. In principle, the maintenance of the resources could be an issue as well, but normally resource maintenance is not seen as an urgent task to which a highly skilled external workforce would be attracted. This attitude is possible as long as the overall translation process relies on human work using software systems only for the communication task and maybe for *static* linguistic resources.

The distinction between static and dynamic systems is meaningful only when the linguistic resources are computerized. I distinguish systems which provide information to a human or to other software systems from systems which carry out an essential part of the translation process, labelling the former static and the latter dynamic. Static linguistic resources are dictionaries, terminology databases, translation memories and similar reference tools, whereas machine translation systems are dynamic. Some terminology databases and translation memories include functions by which they could qualify for the status of a dynamic resource. These functions are applied when a text is being translated by a human translator. The relevant functions contained in some terminology database systems scan the current sentence or piece of text for terms contained in their database and signal their hits to the user who can consult them and cut and paste a chosen target language equivalent into the target text. Translation memories present suggestions for target language equivalents of sentences in a similar way, selecting from their database the stored source-target sentence pair most similar to the current translation task. The involved mechanisms carry out an active function; I nevertheless do not label them as dynamic, because their functionality does not bring about any new piece of translated text. They only retrieve and select translation equivalents previously obtained from a translator or a machine translation system.

The role of the linguistic resources changes drastically, however, as soon as the translation process involves a dynamic component, i.e. a machine translation system. Whilst a human translator receives useful help for instance from a terminology database without being confined to using only the words found there, a machine translation system cannot translate to anything but the words and multi-word units contained in its dictionaries.  The same holds for translation memories: a translator can take a vaguely similar translation suggestion as a starting point for further adaptation, while a machine translation system, at least with the present-day state of the art, cannot use sentences from a translation memory unless they can be taken over without further processing. Machine translation dictionaries therefore need regular updating work, and experience shows that in practice every single translation job requires immediate dictionary coding. This makes resource maintenance much more important and time-sensitive than in a fully manual translation process.



Workflow Scenario 7

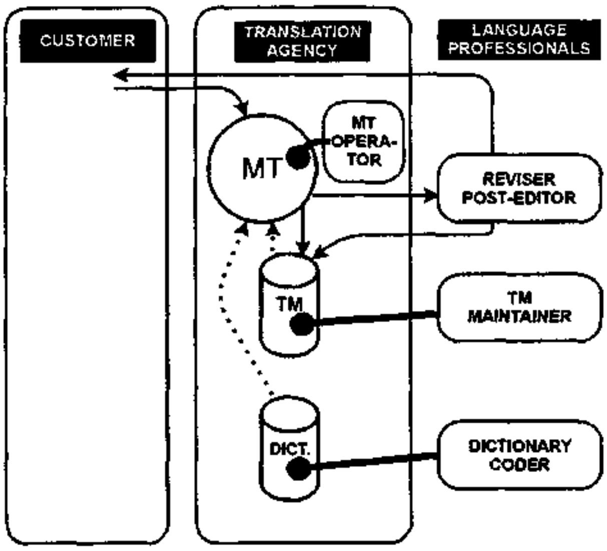
*Workflow Scenario 7* shows a relatively straightforward workflow with a machine translation system situated in the translation agency. The system is run by a local operator but the post-editing is done by an external reviser. The machine translation system makes use of a dictionary and a translation memory. The dictionary needs updating for virtually every translation job. In the scenario, this is done by an external dictionary coder with an upload/download connection to the machine translation system.

As for the translation memory, Workflow Scenario 7 shows a solution in which the raw machine translation output goes directly into the translation memory and is overwritten when the post-editor makes changes or corrections. This can be organized differently, depending on the needs of speed and efficiency. The solution shown in Workflow Scenario 7 assumes a mechanism in the translation memory which can be set to overwrite a raw target language equivalent with a corrected one. In variants of this scenario it may become useful for other functions (e.g. team-work of several post-editors) to choose a translation memory parameter setting where a new translation equivalent for a previously stored source language sentence is stored in parallel rather than overwriting the old equivalent. In this case, only the post-edited version should be stored in the translation memory, or the machine-translated version should be marked for lower preference[4]. Regardless of which option is chosen, applying a translation memory over a longer period or with a larger volume of translation work done requires a certain amount of maintenance work. In the scenario, this is carried out externally by upload and download.

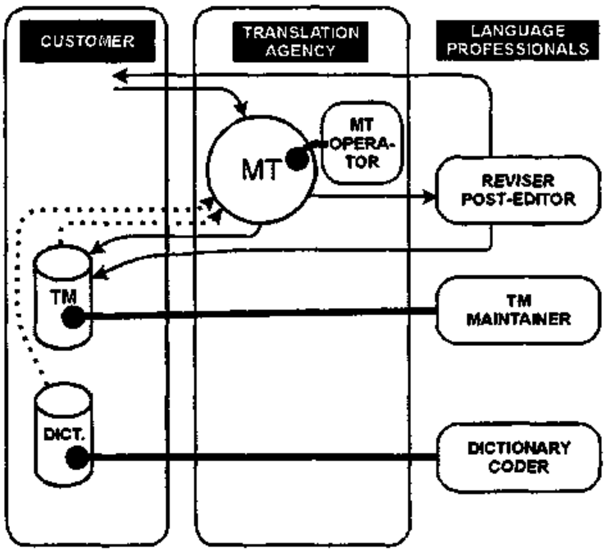
[4] Translator's Workbench by Trados GmbH, for example, contains a feature for automatically assigning a lower preference figure to machine-translated sentences than to human translations.

A faster solution can be achieved by direct on-line dictionary coding. *Workflow Scenario 8* shows this solution together with on-line translation memory maintenance.

Similar to the distinction between Workflow Scenarios 3 and 4 or 5 and 6, respectively, it is possible that customers wish to keep linguistic resources of their own and let the machine translation system receive the required data via the network connection. As before, the maintenance work can in this case be carried out either by upload and download or on-line. *Workflow Scenario 9* shows only the on-line solution.



Workflow Scenario 8



Workflow Scenario 9

As in the cases discussed above, the on-line solution is an innovative element which cannot be achieved without the computer network technology. Before proceeding to more complicated scenarios, it may be worthwhile asking what is so important about the place where a certain resource is kept. Practice shows that customers who maintain large resources wish to protect their intellectual property and prefer to give away only extracts needed for a specific task. In some cases even word lists are rigorously guarded, for example when they contain the maintenance terminology for the newest, yet to be marketed, model of a car, which, as the manufacturers are afraid, might give away technical details not yet publicly known. When the resources are more than the many word lists, semi-privately kept in many documentation and translation departments, customers are increasingly reluctant to give them away. If the resources become more sophisticated, for example a regularly maintained and updated corporate terminology database, a regular upload to the translation agency may become more costly than an on-line access.

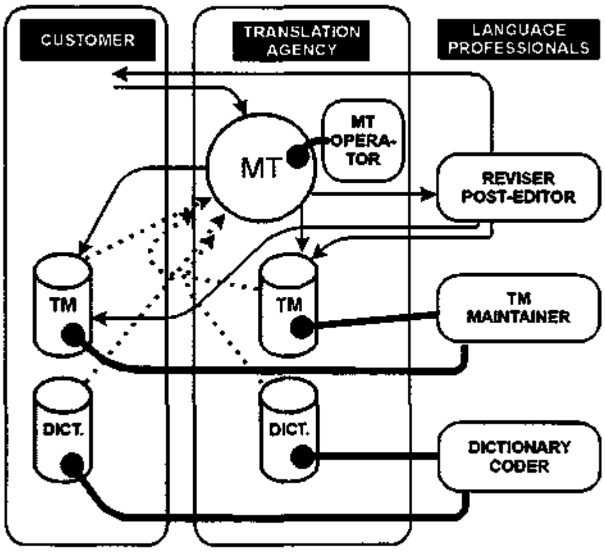
An as yet largely unresolved question concerns conditions under which a corporate terminology database designed for human users can at the same time function as a machine translation dictionary.

A direct on-line access from the machine translation system to the terminology database as depicted in Workflow Scenario 9 (and others below) would require the database to contain the codes needed for the machine translation system and to generate an output structure in which all required information is expressed in these codes. The larger, professionally applicable machine translation systems and, to a lesser extent, the database systems suitable for terminology work, are still far from any form of cross-vendor standardization, so that generic exchange formats or the like are still to come.[5]

While the awareness of intellectual-property issues around lexical materials is still relatively weak, the question of a copyright to translation memories has been around for a longer period. Does a customer tacitly accept a translation agency to keep the full text of completed translation jobs stored in the form of a translation memory? Does the customer allow the translation agency to use this translation memory for other customer's assignments? Could corporate terminology leak out this way? A translation memory is a tool designed for saving money by avoiding to repeat work previously done. Is this the customer's or the translation agency's money? Or the translator's?

The present study does not aim to resolve these questions. Rather, it aims to investigate the possible scenarios for remote-access translation services and derive from them the requirements for the software system to be used. This means that the questions of where which resource should be kept and from where and by whom it should be maintained should not be decided in one or another way, but where it is purposeful, the translation service software should allow for all reasonably viable variants of workflow.

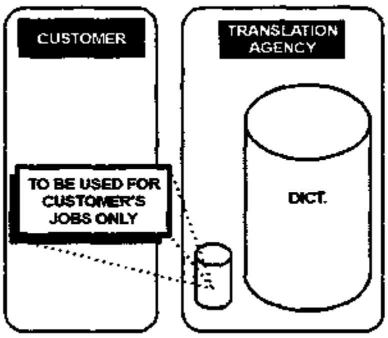
As for the question of proprietary linguistic resources, a solution which can become common in step with on-line access connections becoming cheaper and easier to use, is *Workflow Scenario 10* in which the translation agency holds a large general-purpose vocabulary and translation memory for their machine translation system, while the customer keeps his own. The dotted arrows which show the information flow from the linguistic resource to the machine translation system in the illustration of Workflow Scenario 10 are meant to illustrate the fact that the customer's lexical materials should be given priority. Prioritizing of specific portions



Workflow Scenario 10

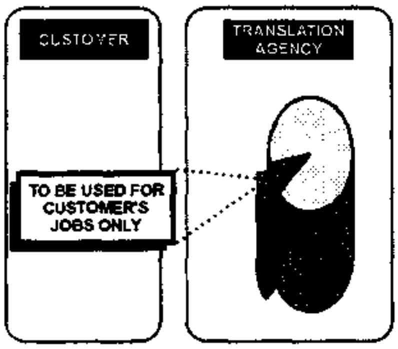
[5] A small R&D project with the objective of developing the prototype of a terminology database tailor-made for the purpose of both fulfilling the needs of a human translator and providing the required upload formats for the METAL and the LOGOS machine translation systems has been completed at Fachhochschule Flensburg under my supervision with funding from the German Ministry of Education, Science, Research and Technology (BMBF). Half-way reports have been published (Lass 1996, forthc.), while at the time of writing the final report is still to be completed.

of the overall vocabulary is a mechanism commonly used in machine translation systems for performing lexical transfer in context to the extent possible in present-day systems which almost completely lack semantic or knowledge-based capacities.

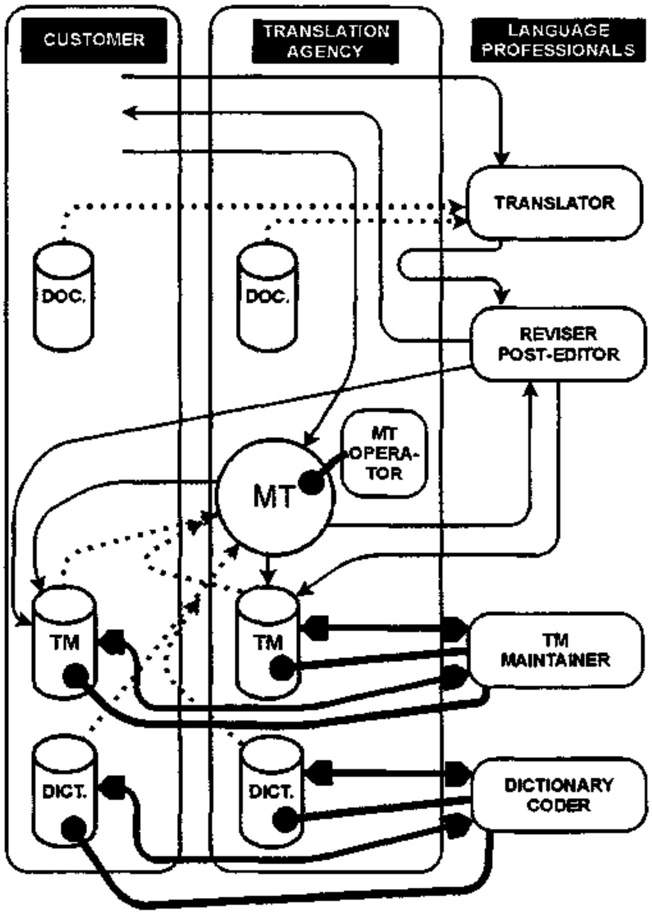


Workflow Scenario 11: Detail

The management of proprietary linguistic resources can be resolved in different ways. In the case of the dictionary, for example, Workflow Scenario 10 shows two distinct dictionaries, one at the customer's site and one in the translation agency. If the customer only wishes to keep the two resources apart without necessarily holding his own part physically within his own premises, the same functionality with a prioritized and a secondary resource can be achieved by keeping two separate databases in the translation agency. (Only this detail of *Workflow Scenario 11 is* shown in the illustration.) This in turn would be functionally equivalent to a solution in which the customer's proprietary materials are included in the translation agency's larger database, though labelled as accessible for this customer only. (This is *Workflow Scenario 12;* only the relevant detail is shown.) Current machine translation systems provide options for setting attributes in dictionary entries either at the monolingual or at the transfer level and to filter for these attributes when carrying out a specific translation job. These instruments were not originally designed, however, for protecting proprietary lexical material from being used by unauthorized translation jobs but for selecting the preferred out of a number of concurrent translation equivalents. The METAL system, for instance, will first search the dictionary areas chosen by the operator, but if nothing is found, it will carry on to the other dictionary areas, searching eventually the entire database. For this functionality, therefore, a solution as in Workflow Scenario 11 is the best which can be achieved without modifying the machine translation system itself.



Workflow Scenario 12: Detail



Workflow Scenario 13

In summary, an agency operating a remote-access translation service for a variety of customers with different requirements is ultimately in a position in which it becomes necessary to cater for all possible combinations of translation job requirements, workflow routings in the agency and to its external professionals. Eventually, the designer of the relevant software system will end up with a scenario incorporating all the possibilities and options side by side, as, for instance, in *Workflow Scenario 13.*

But even Workflow Scenario 13 is not the ultimate in possible variations. For example, I do not illustrate here scenarios in which the translator and reviser work on-line on the customer's computer, making use of tools and resources held there.

Moreover, workflow scenarios are being experimented with (cf. Pyne/ Grasmick 1997 in this conference) in which the dictionaries and translation memories held by the customer and the translation agency are mirror-images of each other, so that machine translation systems used at the two sites can produce identical results. This is another step into automation making use of the networking technology in an innovative way. Other new variations will emerge.

# 6 Design Requirements for an Integrated Service Solution

In view of the user requirements and the workflow scenarios sketched in the above sections, the design requirements for an integrated remote-access translation service solution can be outlined.

Given commercially available applications for the linguistic resources and the machine translation, the software to be tailor-made for the translation service should include

- a customer access environment,
- task management facilities,
- system maintenance and remote operating facilities.

The customer access environment is the outer shell of the software and the point of entry for all customers. At the telecommunications level, it should allow for a variety of different forms of access including direct dial-in, Internet and Intranet options (cf. section 4.2). The customer access environment channels these into a uniform set of options (presented in a form and with a user interface tailored to the relevant type of access software at the customer's end) through which the customers can detail the service they wish to make use of. The functionality of the customer access environment should allow for a broader range of variations to suit different degrees of customer involvement in the overall translation process. In particular, it should cater for different skills at the customer's end and for the possibility of customers wishing to upload resources or to open them for on-line access. The more advanced forms of interaction between the customer and the agency may require well-negotiated and technically carefully established and tested arrangements concerning exchange formats, file access rights etc. so that these forms of on-line cooperation are less likely to be opened for new or spontaneous customers. The translation service provider will therefore need to differentiate customers who can be offered the more straightforward services from those who are sufficiently skilled and equipped for advanced forms of cooperation. For frequent users, customized parameter settings should be available to speed up and possibly fully automate the service request procedure for routine jobs. (For the flatter access formats such as e-mail, there should be the option of attaching the relevant parameter file to the document to be translated.)

The task management facilities should ensure a well-organized workflow between the operators and the systems as well as between the components of the system. Basically, two sets of tasks need to be covered:

- Task management and scheduling. Identification of incoming job types: mapping of user specifications to service elements and degrees of automation. Proper distribution of jobs in the local network, processing of jobs on the different machines. Task queue manipulation. Human-machine workflow integration. Channelling of jobs to remote language professionals and teleworkers.

- Accounting and billing. Word count and text comparison functions. Interface to standard accounting and billing facilities and customer database maintenance packages.

The system maintenance and remote operation facilities should allow for the activities to be carried out by the professionals in the translation agency. Apart from standard system maintenance (machinery, back-ups etc.), the main issue is the maintenance of the dictionaries and the translation memories.

The overall translation service software, composed of these elements, should allow for the following features of functionality (non-exhaustive list). Especially for the customer, some elements are a necessity, while others should be offered to expert users only.

*Customer* (when submitting a translation job)

- Selection of source and target language.
- Selection of subject field.
- Selection of proprietary resource to be used: resource to be sent to translation agency, proprietary resource to be used at translation agency's site, access to resource at customer's site to be opened for translation agency or translator.
- Selection of machine translation parameter settings or attachment of parameter file.
- Selection of pre-editing command file.

*Translation Agent*

- Routing of incoming translation job: human or machine translation, which translator/reviser, which machine translation system/post-editor/dictionary coder,
- For job routed to translator: Selection of resources to send to the translator or to instruct the translator to use. (Same for revisers.)
- For job routed to machine translation: Selection of resources: dictionaries, proprietary dictionaries, translation memories, proprietary translation memories.
- For job routed to machine translation: Selection of parameters: subject fields, system-specific parameters[6].
- For job routed to machine translation with parameters set by customer: Onward routing of parameters, possibly automatic mapping to parameters of the system.

[6] Some machine translation systems allow for or require specific parameters to be set such as the form of address (e.g. EN *you* → DE *du* vs. *Sie),* equivalent of ambiguous pronouns *(*EN *you* → DE *du/Sie* vs. *man),* translation of capitalized words, protection codes for strings not to be translated, maximum processing time or ambiguity breadth etc. etc.

*Translator*

- Selection of resources to download or to access remotely.
- Selection of parameter files if directly importable into tools such as translation memory.

*Reviser*

- Selection of resources to download or to access remotely.
- Selection of parameter files if directly importable into tools such as translation memory.

*Reviser/Post-Editor*

- Selection of resources (dictionary, translation memory) to be used.
- Selection of translation memory to which to download the corrected text.
- Selection of storage option for translation memory (overwrite vs. add, cf. discussion of Workflow Scenario 7).

*Dictionary Coder*

- Selection of connection (upload/download vs. on-line access).
- Selection of session attributes to be set (marking entries as customer's proprietary, regional, product-specific or the like).
- Selection of possible upload facility for importation of data generated in external software (e.g. from terminology management systems into the machine translation dictionary or from an alignment tool into the translation memory).

*Translation Memory Maintainer*

- Selection of connection (upload/download vs. on-line access).

# 7 Outlook

The present study looks into the future. Remote-access translation services have begun to appear on the market. Some of them are cases of stage-two automation in the sense of section 1, whilst a few may have begun to proceed towards stage three. This study goes a step further. It describes the user and operator functionality of a generic solution for a remote-access translation service which is variable in several dimensions, allowing for service providers to adapt their activity to the customer's translation skills and resources. This solution can range from an all-inclusive translation service for non-professionals in language matters to a graded specialist service offered to copywriting and documentation departments skilled in some parts of the documentation management process.

This is a look into the near future. Looking further ahead is beyond the scope of this contribution. However, some of the foreseeable directions of development may well be identified already now. In the present study I do not say much about the functionality and performance of machine translation systems, translation memories, terminology databases and other instruments of automation. I take the available systems for granted and describe how to build them into an integrated service environment which makes use of computer network technology to combine manual and automated work. If the aim of continued development is to increase the degree of automation, then a series of improvements in available systems should be addressed:

- Machine translation quality.

  The main obstacle to a tangible progress in machine translation quality is the semantic barrier (cf. Melby 1995: 151). Selection mechanisms based on meaning (i.e. semantics, pragmatics and knowledge of the world) are urgently needed for lexical transfer, for syntactic analysis and transfer and for text coherence mechanisms (Papegaaij/Schubert 1988: 190), to name only a few of the tasks in the machine translation process.

  Cheaper platforms, even nicer interfaces, integration in everyday software packages - all these are welcome improvements of available systems, but they will not bring about the necessary breakthrough in translation quality. It is only through translation quality that substantially larger numbers of users and satisfying sales figures can be achieved. The prerequisite for this is a huge, long-term investment in a semantic engine.

- Availability of machine translation language pairs.

  The number of language pairs with professionally applicable machine translation systems can be increased by enabling cross-vendor coupling of systems not only at the language pair level but also at the module level: not only from English into German by system A and from German into Russian by system B, but also system A's English analysis with system B's Russian synthesis. In view of the possibilities of networking technology, this leads to the question of distributed translation and a cross-vendor standard of transfer functionality and transfer representations which in turn revives the issue of a fully expressive intermediate language (cf. Schubert 1992).

- Perspicuous documentation of machine translation functionality for dictionary coders.

  Professionals who code dictionary entries for available machine translation systems normally have only superficial information about exactly what the rule mechanism of the system does. (This applies to dictionary coding by users, not by system developers.) Coders are often left to guesswork and trial-and-error coding to find out whether a certain syntactic transfer does not work because of the lack of an adequate transfer entry (which in many cases the coder can provide) or because of an erroneous analysis (which the coder cannot repair). I have elsewhere briefly sketched the need for professional documentation and incremental testing facilities at the levels of analysis, structural transfer, lexical transfer and synthesis for this purpose (Schubert 1996).

- A standardized document exchange format.

  Format conversions and missing converters often prevent a possible step in automation from being taken (cf. section 4.2).

- A standardized resource format.

  The idea of re-usable, multifunctional linguistic resources, especially when accessed by translation systems rather than translators, requires standardized formats. The standardization will in some specific cases have to extend beyond defining the representation of the data. Specific systems need specific pieces of information content to be provided. Machine translation systems in particular are still much too little standardized to be able to deal with, for instance, subject area codings in dictionary entries which are not their own.

Translating human languages is an enormous task which is not susceptible to full automation. But the achievable degree of automation is far from reached.

## References

Arnola, Harri; Kaarina Hyvönen, Jukka-Pekka Juntunen, Tim Linnanvirta, Petteri Suoranta (1996): Kielikone Finnish-English MT System *TranSmart* in Practical Use. *Translating and the Computer 18.* London: Aslib

Berberich, Frank (1996): Die digitalen Dolmetscher bieten ihre Dienste mittlerweile in Echtzeit an. *Computer-Zeitung* 18 January 1996, 8

Berberich, Frank (1997): Online-Übersetzungen kommen nicht ohne humane Nachbesserung aus. *Computer-Zeitung* 27 February 1997, 8

Bischoff, Heiko (1994): Vergleich von Translation Memory-Ansätzen. Hildesheim: Universität Hildesheim [unpublished diploma thesis]

Blatt, Achim (1996): The Euramis Project. Angelika Lauer, Heidrun Gerzymisch-Arbogast, Johann Haller, Erich Steiner (eds.): *Übersetzungswissenschaft im Umbruch.* Tübingen: Narr, 131-134

Blatt, Achim (1997): Euramis - Entwicklungsstand. Protokoll der 12. Europäischen Tagung des Arbeitskreises "Maschinelle Übersetzung" am 6./7. Feb. 1997 bei der Europäischen Kommission, Luxemburg [unpublished minutes] [copies of viewcharts, no text]

Brace, Colin (1997): Future Tense for Euro Systran? *Language International* 9[3]: 14-17

Flanagan, Mary (1995): MT in the Online Environment. MT Summit V. Luxembourg: European Commission [unpublished participants' proceedings]

Flanagan, Mary (1997): Online Translation: MT's New Frontier. *Translating and the Computer 19.* London: Aslib

Hatley, John (1997): LOGOS as an Internet and Intranet Application. *Translating and the Computer 19.* London: Aslib

Lass, Jan (1996): Terminologiedatenbank für Mensch und Maschine. Eine Datenbankprogrammierung aus Anwendersicht. R. Walczak (ed.), *Forschungsbericht 1995/96.* Flensburg: Fachhochschule Flensburg, 81-88

Lass, Jan (forthc.): Terminologiedatenbank T42. Die Äquivalenzbeziehung im Zentrum eines hierarchisch strukturierten multifunktionalen Terminologieverwaltungssystems. [Jahrestagung der Gesellschaft für Linguistische Datenverarbeitung, Leipzig 1997; forthcoming in the conference proceedings]

Leick, Jean-Marie (1995): Euramis: Integrated Multilingual Services for a Large Multilingual Community. MT Summit V. Luxembourg: European Commission [unpublished participants' proceedings]

Leick, Jean-Marie (1997): Euramis – Grundgedanken. Protokoll der 12. Europäischen Tagung des Arbeitskreises "Maschinelle Übersetzung" am 6/7. Feb. 1997 bei der Europäischen Kommission, Luxemburg [unpublished minutes] [copies of viewcharts, no text]

Lockwood, Rose; Jean Leston, Laurent Lachal (1995): *Globalisation: Creating New Markets with Translation Technology.* London: Ovum

Lukas, Mirjam (1994): *Merkmale der maschinellen Übersetzbarkeit von Fachtexten.* Arbeitspapiere der GMD 855. Sankt Augustin: Gesellschaft für Mathematik und Datenverarbeitung / Flensburg: Fachhochschule Flensburg [diploma thesis]

Masion, B. (1995): Offering MT as a Service from a Profit Center Point of View. MT Summit V. Luxembourg: European Commission [unpublished participants' proceedings]

Möller, Anja (1996): Der Abgleichsalgorithmus zwischen Text und Übersetzungsspeicher. Flensburg: Fachhochschule Flensburg [unpublished diploma thesis]

Nuutila, Pertti (1996): Roughlate Service for In-House Customers. *Translating and the Computer 18.* London: Aslib

O'Hagan, Minako (1996): *The Coming Industry of Teletranslation.* Clevedon/Philadelphia/Adelaide: Multilingual Matters

Papegaaij, Bart; Klaus Schubert (1988): *Text Coherence in Translation.* Dordrecht/Providence: Foris

Pyne, Chris; Daniel Grasmick (1997): DoVer: MT Information Translation. *Translating and the Computer 19.* London: Aslib

Savarnejad, Atoosa (1997): Dolmetscher für den Online-Chat. *Computer-Zeitung* 21 August 1997, 24

Schubert, Klaus (1986): *Syntactic Tree Structures in DLT.* Utrecht: BSO/Research

Schubert, Klaus (1992): Esperanto as an Intermediate Language for Machine Translation. John Newton (ed.): *Computers in Translation.* London/New York: Routledge, 78-95

Schubert, Klaus (1994): Sietec Listens to Metal Users. *Language Industry Monitor* [24]: 10

Schubert, Klaus (1996): Lexikon und grammatischer Regelmechanismus in der maschinellen Übersetzung. R. Walczak (ed.), *Forschungsbericht 1995/96.* Flensburg: Fachhochschule Flensburg, 97-107

Slocum, Jonathan (1989): Machine Translation: A Survey of Active Systems. István S. Bátori, Winfried Lenders, Wolfgang Putschke(eds.): *Computational Linguistics//Computerlinguistik.* Berlin/New York: de Gruyter, 629-645 + 799-900

## Author

Klaus Schubert is professor of computational linguistics and technical translation.

Address: Studiengang Technikübersetzen, Fachhochschule Flensburg, Kanzleistraße 91-93,
D-24943 Flensburg, Germany. E-mail: <schubert@fh-flensburg.de>