MACHINE TRANSLATION AS VALUE-ADDED SERVICE

by Prof. Harald H. Zimmermann, Saarbrücken (FRG)

draft: April 1990 (T6IBM1)

I. THE LANGUAGE BARRIER

There is a magic number common to the European Community (E.C.): "93" denoting the year when 12 countries in Western Europe will have reached a new level of cooperation and integration: the European Single Market. But there are other numbers representing one of the biggest problems not only in Europe but also in international communication: "72" denotes the translation directives of the 36 language pairs derived from the 9 official and working languages in the Commission of the European Communities (C.E.C.); if one more official language were added, this number would be "90" etc.. It is obvious that the language barrier is one of the greatest obstacles to European integration.

Theoretically, there are several ways to overcome this problem: One could concentrate on ONE or TWO selected and commonly agreed upon natural European communication languages. Fortunately, English as the common international communication language is within the languages in the E.C. In case one had to decide, there would be a good chance that English would be selected. This would not mean that, sooner or later, all the other languages would disappear, at least in communication environment. There is good reason for combining such a policy with another strategy: European citizens have to be or to become bilingual: speaking their native language and English as communication languages alike.

At first glance, this concept seems reasonable and also economical. In practice, there is indeed a real trend in this direction. Not to be misunderstood: This does not mean that there is no longer a chance to teach or learn other languages; but the difference to today's procedure would be that immense efforts should be undertaken to give more than 230 million inhabitants of the E.C. the chance and the motivation to concentrate on English as a second, fluently-spoken language. There are in fact examples of countries where bilinguality is the standard, even if this may have historical reasons, and English is not included (e.g. Luxemburg, Belgium and Switzerland).

But there are also reasons for or, at least, impediments to this concept. They are based on considerations involving more elements than the pure economic factors. One argument is that *language pluralism* also means cultural pluralism, and that language learning cannot be separated from the knowledge of the culture and society of that country where the language is used as natural instrument. On the other hand, one cannot divide a language which - apart from its pure informative role -has several functions (e.g. demonstrating social and cultural connections or feelings) into a "knowledge transfer instrument" and a "cultural and social indicator". The idea of developing more "neutral" pure communication and information languages like Esperanto has failed and one has to learn from this.

Things might be different if a (natural) language was chosen in an environment where the social and cultural (side) effects do not play an important role, e.g. in international trade or in scientific or technical communication and information. For example: It is a fact, that German scientists no longer have a chance to publish their scientific or research discoveries in German, at least in domains like Biology, Chemistry, Psychology and Medicine. So it is commonly "agreed" (as a result, without any political or economic influence), that at least in fields where basically international communication and information exchange is needed, where natural language plays a secondary role (where even specialised formal languages are invented or used to avoid ambiguities and vagueness), English plays the role of the *knowledge carrier.*

If we really take a look at this English, written or spoken by scientists and managers from France, Japan, Germany, Italy, Korea, Sweden etc., we have to consider (with exceptions) that they don't use "Shakespearean English" but more or less "basic" English, augmented with special or professional vocabulary. The only purpose for using English - normally - is to make sure that the communication partner will understand the problem, the facts and/or the solutions.

II. THE CONSEQUENCES

What are the consequences, with regard to the growing European Community? What can be done to fulfill the more or less contradictory aims within the E.C. - and especially the C.E.C. - in order to guarantee cultural and social pluralism in Europe and to promote communication via natural languages? The way it has been handled up to now within the C.E.C. - based more or less on decisions of 1952 where the E.C. had only a few member states - will lead to unrealistic and uneconomic results. Even today, it would be a utopia to expect that within the E.C., all natural languages would play an equal role. At the moment, the Luxembourg and the Gaelic languages are ruled out, but there are - as mentioned above - still 9 languages (Danish, Dutch, English, French, German, Greek, Italian, Portuguese and Spanish), in which, for instance, all official papers of the European Commission, published in the Official Journal, are translated.

Before Greece, Spain and Portugal joined the E.C., it was common practice that at official meetings, every spoken word was translated into the native language of the participants. This lead to the fact that more than one third of the employees at the E.C. are translators and/or interpreters. After the entry of Greece, Spain and Portugal, it became more and more difficult to organise official meetings where all the languages of the participants were allowed because it was practically impossible to get the full interpreting service.

Whereas, at the beginning, translations (i.e. interpreting) were done directly by an interpreter understanding the language of the speaker and the listener (e.g. Italian and Danish), nowadays the strategy has changed in case there is no interpreter available for direct translation. To give an example: For translating Greek into Portugese, English is normally used as a "switching language". This means that the Greek is translated into English by the first interpreter and the English translation is translated into the target language Portugese. Because this sometimes leads to misunderstandings and wastes time, especially in a real time situation, at "business meetings" within the Commission, the attendants agree, more or less, on English as the communication language, or to have only interpretation between French and English.

Until now, the member states of the European Community have hung onto the concept (or should I say: fantasy) that within the E.C. - at least for official events or declarations - every language is the same. And to say it again: It is quite clear that any natural language has the *possibility* to play the same role in communication. And there is also a true feeling that if one would agree to the concept of having only one (or two or three, in any case a strictly reduced list of) officially usable communication language(s), the other languages would lose capacity, starting with the missing of technical vocabulary, etc.

III. MACHINE TRANSLATION AS A POSSIBILITY?

There has always been - at least within the last 30 or 40 years - a dream or even hope that, one day, technical tools would allow a speaker or writer to use his native language and to be understood by his listener or reader in his own native language via machine translation or interpreting. But what are the future possibilities? And when will they have realistic impact? - The chances are growing - even if it takes several decades - that functions are available where - in some limited situations - automatic translation (with "full understanding" within a restricted communication situation) is available. For example: standard situations like buying a ticket or planning a flight; room reservation; emergency calls etc. But still, the real question remains: Is the computer able to contribute *effectively* to the problem of overcoming language barriers in the E.C.?

Before discussing today's practical possibilities and the future opportunities, one has to mention the actual situation and activities. To start with, one has to distinguish the following cases:

> Computers, especially the Personal Computer (PC) and the so-called Workstation, nowadays play an important role in word processing. The "magic" term here is "desk top publishing" which means that everybody is now able to write texts which look like they were published. Within this environment, most of the word processing software has functions to give access to word lists to select synonyms; but for specialists, even bilingual electronic dictionaries should be given access to during the process of text production. Using the word processing tools, the cost of word processing (and in our case: in human translation) can be reduced. In cases where standard elements of text occur frequently such sequences can be marked and automatically inserted. By using these facilities, the costs of translation can be drastically reduced. On the one side, this might lead to a lower price for the translation; on the other side, this pure technical function allows the production of more translations in the same time span (between 20 and 40 %).
>
> Today, Machine Translation resulting in quality translations which are - in all cases - comparable to translations of a human specialist (so-called Fully Automatic High Quality Translation, FAHQT) is not available. The situation would be a little bit better if the text type and/or the domains were highly restricted, e.g. in the field of weather forecasting, where in Canada a special system to translate between French and English, TAUM-METEO, has been developed; and in France where in Textile Documentation the MT-System TITUS is in use to translate abstracts between English, German, French, and Spanish in restricted natural language with a restricted special vocabulary.
>
> There are already systems on the market, where practical application is possible. To mention SYSTRAN, LOGOS and METAL, they all have, with regard to translation quality, more or less the same level: The results are not perfect, and especially not reliable in the sense that one could use them without supervision. The output of these MT systems is normally called a "raw" or "informative" translation. The process of "polishing" the results to get high quality translation comparable with the human translation is called post-editing (normally done by human translators; at least, competence in both languages, the source language and the target language, is *required).*

The possibilities for using one of these MT systems depend on a variety of parameters. It is not possible to list all the criteria, but at least some of them have to be mentioned:

> The availability of the language pair, the quality and the volume of the (discipline-oriented) electronic dictionaries. Most of the existing M.T. systems differ in that point. SYSTRAN has, for the moment, the biggest list (FRENCH-ENGLISH, GERMAN-ENGLISH, RUSSIAN to ENGLISH, ENGLISH to SPAIN, ITALIAN, PORTUGUESE, ENGLISH to ARABIC ...). METAL is available for GERMAN-ENGLISH and ENGLISH to SPANISH etc.

> The access availability and the speed of the machine translation. Most of the systems have to be used inhouse, that means the customer has to install a copy on a special computer or on his facilities and to instruct his staff (e.g. technicians and human translators) in handling the system. Normally, such systems are installed in combination with a translating department or a translation company. In such a case where machine translation normally will be post-edited speed plays not an important role, because the human translator is the "bottle neck".

But now there is a first possibility given to professionals (or even to everybody) to use Machine Translation via telecommunication. This means that a translation system can be used combining PC (and texts written with different word processing systems) via a standard telecommunication network (telephone line or packet switching network) with a host computer where the system itself is installed. The C.E.C. has now developed a function to give direct access to their translators to the use of SYSTRAN installed on an internal computer within their inhouse office network. In such a case, automatic translation speed plays an important part: the raw translation has to be back within a few minutes.

> System development, especially dictionary development, maintenance and service are still important. It is possible today to get relatively good translations (in the above-mentioned sense), especially within domains and text types where research and applications have existed for several years, where post-editing saves time (and money) or the raw translations is sufficient for pure information purposes (see, for example, the picture where a result of a raw translation of German to English with SYSTRAN is displayed). But on the other hand, a great effort and investments are necessary to reach a higher quality level in all important domains and language pairs.

IV. MACHINE TRANSLATION AS VALUE-ADDED SERVICE IN EUROPE

The term "value added service" today normally is used in connection with the use of new techniques in telecommunication. Especially the "Integrated Services Digital Network" (ISDN) with higher speed and digitisation of the data is involved. But even today, many possibilities are available to get or to distribute information or to communicate (outside the standard telephone function). Telefax services, access to databases, electronic mail and mailboxes, videotex systems may be mentioned as the main examples.

The way machine translation will be integrated as a value-added service in such an environment is quite simple: M.T. can be handled similar to the access to or communication with databases. After a contract with the translation center (as the host) has been signed, a client is able to send the source text to the host where the text is translated automatically and sent back to the client. As an - even augmented - alternative, a M.T. function within a mailbox system can be used: In this case, the mailbox provider has a contract with the host and the user has a general contract with the mailbox provider which includes the possibility of using the M.T. function.

One of the most interesting scenarios - which is very realistic, because all components nowadays exist and have only to be combined - is the following: The user sends his text to the mailbox which automatically selects the right (or desired) translation host computer to get the raw machine translation. If the user only needs an informative good-enough translation, he (or his) communication partner) will get the machine output directly. Otherwise, the raw translation is automatically sent to the "box" of a translation office where human translators (possibly specialised within the domain of the text) will "polish" the raw translation to get the right quality and send it back to the client or his partner - via file transfer or even telefax.

If one takes a look at the labs of the computer industry, one has to consider that only a few companies - especially in Japan - are aware of that. In Europe up to now the greatest effort has been done within the C.E.C. in developing SYSTRAN as an application tool and starting a big project on basic research in M.T., with the lab system EUROTRA. Indeed, a lot of basic research still remains so that universities and industry laboratories will have to cooperate in the 90ies in order to improve existing concepts.

But there is an extreme high probability that the next decade - even before the year 2000 - will bring *practical M. T.* up to the European or even worldwide market via using existing facilities. The technical bases and instruments as well as the framework (word processing and telecommunication) are available, more investigation has to be done in realising customer-friendly environments and finding the right marketing policy including pricing and services.

Even if one has to have in mind that overcoming language barriers in Europe means a lot more than just using machines, there now is a chance to integrate M.T. step by step as a "normal" function - like word processing - in the international communication process. Several partners - system developers, network providers, software developers, translating houses, and - last but not least - the users have to build a group of common interest for using and developing M.T. functions in their daily business. So, machine translation will no longer be a fiction or a pure toy for language research, but become a practical tool.

V. CRITERIA FOR EVALUATION AND USAGE OF MT FUNCTIONS

It is quite clear that most of the readers of this article do not have experience with the use of machine translation. But there might be an interest now to start with the (experimental) usage of machine translation. Trying to avoid negative effects or even frustration when introducing MT functions within the information and communication process of a company or an authority, some central "rules" for decision makers are formulated as a brief checklist. Computer-Aided Translation (CAT) where more or less only electronic dictionaries are used can be excluded because it seems quite clear - at least for the author - that they will be accepted (if they are available) in any case. So one can concentrate on "full" Machine Translation (MT) with or without the post editing component (which - in any case - is needed if one wants to get full reliability and high translation quality)

- The most important advantage of the application of today's practical MT systems is the translation speed, even if one takes into consideration the post editing part. Even if one has to consider - at least in standard situations - that the costs will today be more or less the same compared to traditional translations (including high quality post edition), this advantage may lead to a high acceptability of MT systems in the near future.

- The decision for introducing an MT system should be made *in close contact with the translating staff* (if it exists). But one should not discuss only the problem of quality of the raw translation. There is a broad area of new functions for the human translator. The most important is the feedback function between the system and the user, especially in dictionary improvement: The (specialised) translator will play an interesting part as a kind of knowledge engineer of the ("intelligent") MT expert system. Up to now, there are functions established on the MT provider's side which give the users a chance to influence the results of the translation.

- The decision cannot be limited to the aspect of using MT functions for every text type or language pair. The quality of the machine translation normally is influenced by the complexity of the text ("sentences"), but also by the volume of the internal dictionaries. The human translator (or - more generally - the user) should therefore be free in if the MT has to be used or not within a concrete situation.

- There are many situations where raw translations are "good enough" even for the partner. They are - and this seems to be an important advantage - much cheaper than the human translations (HT) or the "combination" of MT and HT. Up to now, this variant of a "quick and dirty information" given by automatic translation of (written) texts could not be used because there was no technical possibility. At least during the testing phase of MT applications one should integrate experiments in which partners have to be informed about this type of written communication. Moreover, one has to be very careful about using this variant in order to preserve the "image profile" of a company.

- If one looks at the existing systems, the biggest problem besides the problem of the availability of a specific language pair is the deficit in technical vocabularies. There are some areas like Electronics, Computer Science, Building Construction where (in language pairs as English / French or English / German) the basic technical vocabulary of such a discipline generally has been integrated today. But normally, the client has to contribute (directly or indirectly) to the updating of these vocabularies at least by giving some commentary as feedback. A more constructive way would therefore be to send files containing lists of words and their translations. When installing an MT function in a closer cooperation with one of the MT system providers, one has also to consider the case that one day one wants to use another system. That is why one should make sure that the dictionary data integrated via the cooperation are also available for further use outside the original system.

The usage of an MT system should be combined with the "right" environment of text or word processing. One has to remember that a spell check has to be done before the translation. Some MT systems allow - more and more - to save the structural information (at least the most important markers like centralisation, left margin, tabulator ...) and to give them back in the target text. CAT functions, like core-resident dictionary access to lexical alternatives during the post editing phase, can be combined with MT functions.

To start using MT nearly without any risk the use of "teletranslation" via telephone line or - much cheaper - via packet switching networks is recommended. SYSTRAN - as one example - can be accessed as mentioned before by using a PC, a modem and a small software tool called EXPRESS. Most MT providers have a POST service or at least a testing possibility so that nobody has to buy a pig in a poke.

***

Prof. Dr. Harald H. Zimmermann, * 1941, University professor for Information Science at the University of Saarbrücken, FRG; director of the "Institute of Applied Information Research (IAI)"; also manager of SOFTEX Software Institute for Automatic Text processing Ltd. Fields of research and development: natural language processing (automatic indexing, machine translation, mono- and bilingual dictionary development, spelling checkers), application of expert systems; interfaces to databases; data base development and information broking.