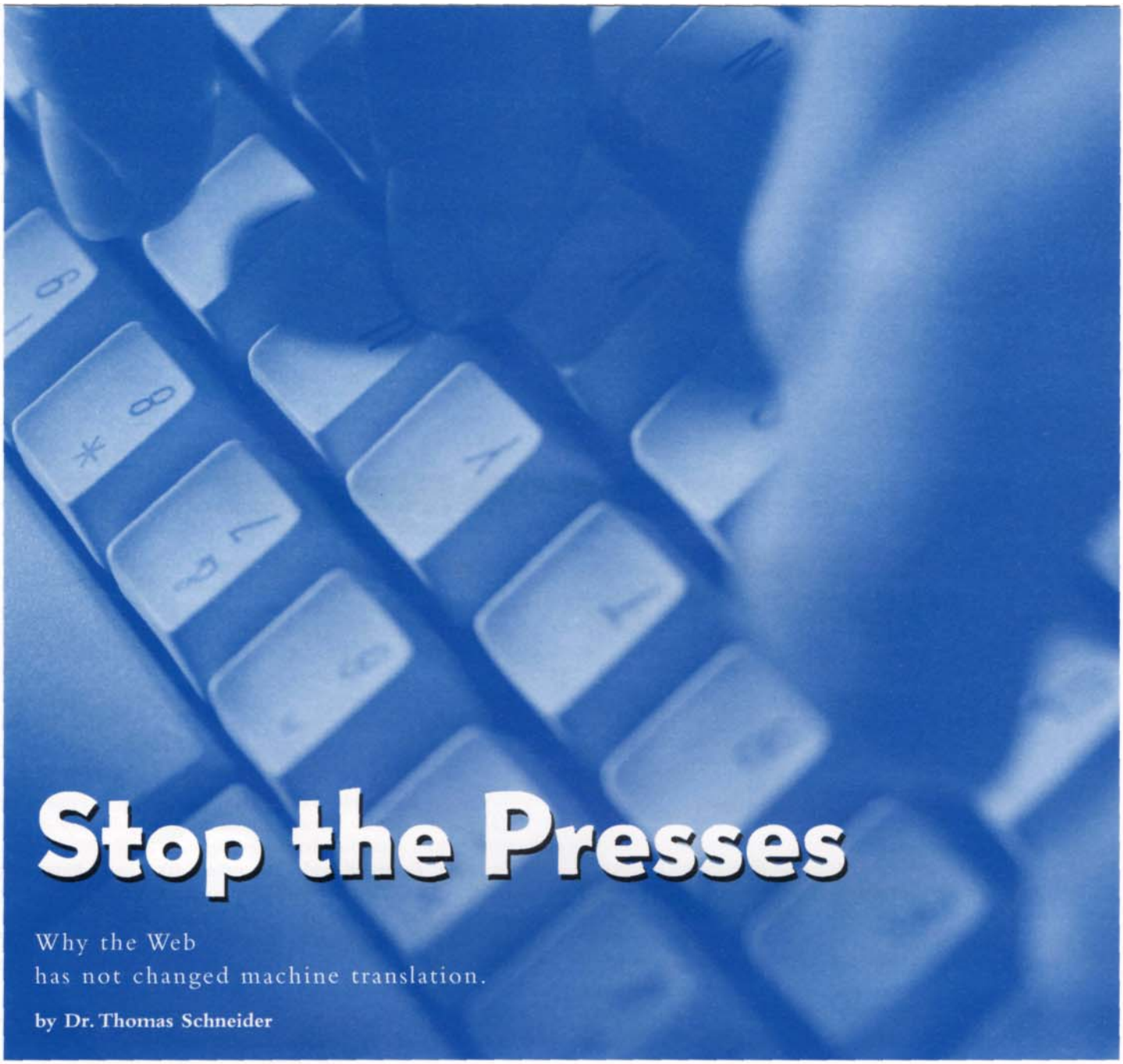

Stop the Presses

Why the Web has not changed machine translation.

by Dr. Thomas Schneider

Forget it. You cannot have what you want: high-speed machine translation (MT) for all your sundry texts at the push of a button, in publication quality, and without human intervention. And if you think that MT engines offered by providers on the Web can deliver better quality than stand-alone systems, you have probably been spared the experience of looking at Web-translated output and then adding the pages to your wastebasket contents. The fact that it is easy to send files via the Net does not change anything about the concepts, the possibilities, and the limitations of MT.

No linguistic theory so far has been able to describe even a single language completely and unequivocally, and there is none in sight that promises to do better than the previous approaches. Advances in processing speed, storage media, and comfortable user interfaces have made MT easier and cheaper to use and have extended its range of applications considerably, but the inherent limitations imposed by analysis grammar and structural and lexical transfer remain.

MT can be used for two distinct purposes: for the dissemination of information to third parties and for the acquisition of information from foreign-language sources. A typical scenario for the first case is the rapid translation of technical documentation of complex systems for export purposes. A medium-sized telecommunications system requires more than 100,000 pages of documentation (installation, operating, and maintenance manuals) to be functional. By conventional translation methods, the task would use up more than 80 person-years of effort. Considering that it is extremely difficult (and expensive) to find translators with

the necessary expertise in the technology involved, chances for rapid delivery (and installation and operation of the system) are slim. (Would your average translator know what a "ram bang circuit" is?) Using a sophisticated MT system to produce a first draft, with consistent terminology throughout all documents, can increase the translators' productivity by several factors, decreasing the time-to-market for the product. However, there is no alternative to a post-editing phase.

Claims have been made that no post-editing would be necessary, if all ambiguities in the source document were resolved first. While it is true that a well-written original (ever seen one?) makes the task of translating, both by machine and by conventional means, much easier and faster, human intervention is still necessary. As a trivial example: the English verb "put" has to be translated into German as "legen," if the object is of horizontal shape and the intended location is a flat surface; as "stellen" if the object is upright and the location is a flat surface; as "setzen" if the object is either animate or very heavy and the intended location is a flat surface; as "stecken" if the object is light-weight and the intended location is a receptacle.

Even though the source text may be unambiguous, and the concept of placing an object is present in all examples, they still need to be translated differently. It would be ludicrous to describe each item in the world with such detailed semantic features. One would need to define the world several times over, and the system—even if it were theoretically feasible—would collapse under its own weight.

MT can only work well if grammar and lexicon are tuned to the types of text and the subject area treated. Technical manuals require imperatives in one language, either polite or casual form, and infinitives in another. The resolution of anaphora is often domain-specific, and the lexicon must be highly specialized to allow precise transfer and exclude unnecessary ambiguities. What, for example, is a "grain"? Is it an agriculturally relevant kernel of a plant, a measure, something of a piece of wood, of leather, stone, metal, a human trait? Each one definition requires, based on subject area, a different translation.

If one adds up the terminology of all subject fields, from coal mining to data processing or genetic engineering, one is likely to reach anywhere between 50 and 100 million entries. For a specific application, a user needs only a fraction of these, but no MT system will have them all available. Only the user knows his own technology and the correct terminology. Having MT on the Web will change none of this. On the contrary, tuning grammar and lexicon to the user's application is much easier on a stand-alone system rather than on a multipurpose and perhaps not-so-transparent MT engine run by a potentially anonymous provider.

It may be slightly different if MT is used solely for the purpose of information gathering. A typical application would be in international organizations such as the European Commission, where the content of foreign-language documents needs to be roughly understood. If the lexicon is geared to the subject area, post-editing can be dispensed with. Other organizations scan documents from open sources such as foreign newspapers to collect relevant information. According to the users' interest profiles (e.g., agriculture, military actions, crime, etc.) the lexicon needs to be built up differently. Still, for such specific purposes, Web MT would not provide any advantage.

A potential target group, however, would be the casual user, someone with not enough translation volume to warrant the purchase of a stand-alone system, and usually someone without the linguistic expertise to code his own lexical entries. To be sure, the Web-translated letter of a distant relative in Greece may read somewhat strangely, but at least the fact that the planned wedding will take place in September is discernible. On the other hand, the letter to a beloved person in Portugal may create more repercussions than the desired emotional response. Better to learn Portuguese instead.

To be sure, the Web-translated letter of a distant relative in Greece may read somewhat strangely, but at least the fact that the planned wedding will take place in September is discernible. On the other hand, the letter to a beloved person in Portugal may create more repercussions than the desired emotional response. Better to learn Portuguese instead.

Product localization, while benefitting from the rapid translation of relevant documentation, needs to go a step further and consider both technical and cultural aspects as well. If the target text has to fit into a form or into a table on the screen, some problems may arise if the translation turns out to be longer than the source. English "Press repeat" becomes German "Wiederholungstaste drücken" which is 14 characters longer. So sometimes a straightforward translation is not possible. One of the potential solutions is to abbreviate or reformulate. This, however, sometimes leads to semantic absurdities, as for example on airplane seats. Instead of "Keep seat belt fastened while seated," we read "Fasten seat belt while seated." Why would anyone get the idea to put the seat belt on while standing up?

For Arabic countries, including Iran and perhaps Malaysia, the sequence of illustrations needs to be rearranged from right to left, in keeping with the direction of the Arabic script. Use of the color green is restricted, and in some countries no women may be shown in training material or marketing brochures. Translation may indeed form an important aspect of localization, but it does not solve all the problems. So where is the Web? It isn't. You're on your own.

MT, if used sensibly, can be a powerful tool. It can be supplemented by others, also with their own application range. Automatic document version comparison can identify text portions that occurred in previously written documents. Case studies in technical documentation have shown that the manual of the subsequent version of a product on the average repeats about 30 percent of the original text. If that document had been translated before, a translation memory (TM) would be an excellent choice to minimize the translation effort. The previously translated portions are inserted into the new text and only the rest is channeled through the MT system. TM is especially suited for repetitive texts like error messages, EU calls for tender, etc.

For information acquisition, any one of the more than 3,000 languages of the world may be relevant. It is understood that from a

So sometimes a straightforward translation is not possible. One of the potential solutions is to abbreviate or reformulate. This, however, sometimes leads to semantic absurdities, as for example on airplane seats. Instead of "Keep seat belt fastened while seated," we read "Fasten seat belt while seated." Why would anyone get the idea to put the seat belt on while standing up?

cost/benefit point of view, commercial developers of MT systems concentrate on the major languages and not on "minor" ones like Kazakh or Karen. The development of a full-blown MT system for languages which have not been described formally may cost more than 80 person-years—if linguists with expertise in these languages can be found at all. There is simply no way to recoup the investment.

But sometimes, for political reasons, information search in "exotic" languages is imperative. In these cases, developing lemmatas and specialized lexicons can lead to results which may well be adequate for the intended purpose. A term-substitution system checks the foreign-language text against the lexicon, translates the known terms, inserts them and highlights them on the screen. Thus the user is able to identify the topic and the general content of the text and can decide if the text is important and deserves closer attention.

Tools like these were developed within the Aventinus/Sensus project, partially funded by the European Commission under the Language Engineering program. The purpose was to support law-enforcement agencies in their fight against organized crime, especially in the area of drug trafficking. Relevant information resides on a variety of different databases, in different formats, and in different languages, but rapid access to the information is crucial and Boolean search is too slow and not precise enough. A formal query like "drug AND deal AND Netherlands" is likely to produce a lot of white noise, e.g., thousands of presumed hits like "Company X concluded a deal with company Y on the joint marketing of antihistamine drugs in the Netherlands." At the same time, it would miss items such as "heroin smuggled via NL" or "cocaine sold by organization Z in NL," etc.

Now, an officer may formulate a query in his own native language, something like "Give me a list of all persons who have been suspected of dealing hard drugs in the Netherlands during May and June 2000." This natural language query is automatically converted into a formal query; all (multilingual) databases are searched, and the result—the factual content—is presented to the officer in his own language again. An integrated thesaurus links concepts and points out synonyms ("snow" means cocaine, heroine is a drug; Mercedes is a means of transport) and a domain model (drug smuggling involves persons, locations, means of transport, times, route, and substances).

The overall system contains an MT system (T1, marketed by Langenscheidt), a translation memory and term substitution, e.g., for Chinese, Arabic, and Russian. Additional modules identify topics and compare different spellings of names as to their similarity. Especially with non-Roman script languages which need to be transliterated, there is a great discrepancy between different spellings for the same name: is it "Al-Bab," "Al-Beb," "Al Beb," "El-Bab," or "El Bab"? Even in European languages there is the problem: is it "Meier," "Meyer," "Mayer," or "Mayr"? Phonetic similarities must be filtered and weighted.

The system has now been implemented at Europol and is considered a major step forward in facilitating communication throughout Europe. However, due to the sensitivity of the transmitted data, links to the Web cannot be tolerated.

It seems the only way in which the Web may change MT is that on account of considerable marketing hype fewer people will buy stand-alone systems, hoping for the great leap ahead on the Net. The corresponding loss of revenue for the developers will reduce the available investment into badly needed further development of MT systems—hardly a bright prospect.

After teaching at universities in the US and the South Pacific, Thomas Schneider, MA, PhD, DPhil, joined Siemens to head the development of a machine translation project (METAL), the administration of a large terminology database, and, later on, all of software development in the area of natural language processing. Since 1995 he has been a freelance consultant. Contact him at tomasi.schneider@t-online.de