

[From *The Sunday Times* (Malta), 25 April 2005]

Machine Translation and Human Translators

by Michael Rosner

WHEN Maltese became an official EU language in May 2002, the decision came as something of a surprise. We had all assumed that one of Malta's selling points, linguistically speaking, was the fact that the nation's second language was English, thus avoiding the need to augment the EU translation budget to handle another official language. Currently there are 11. After May 1, there will be 20, including Maltese.

Despite the cost, the granting of official status to Maltese has to be a good thing - for the language and for the speakers. A native speaker of an official EU language enjoys certain privileges: the right to speak in Maltese during official EU meetings, to correspond with the EU in Maltese, and to receive legal and other official documentation in Maltese.

Secondly, in an age when linguistic diversity tends to be under threat, official status, like listed status for a building, offers a certain level of protection against further deterioration.

Although there are some who would claim that Maltese is alive, kicking and in no need of protection, our ears and eyes tell a different story, to judge from reports about the steadily deteriorating quality of Maltese in the spoken and written media, not to mention the incredible bilingual salad exchanged between individuals on the street and especially through computer and mobile messaging systems.

The privileges, however, do not come automatically. Neither do they come for free. Official status is costly since, not surprisingly, it concerns translation services, and thus, the most expensive resources: highly specialised humans.

It is important to stress that the translation problem will not go away when the *acquis communautaire* is available in acceptable Maltese, nor when the Maltese version of the EU Website is finally up, running, and correct.

For other legal documents will follow, and as our involvement with the EU progresses from the official to the informal, the demand for translation services will expand to include other kinds of documents, other kinds of content, language pairs other than English and Maltese, and translation in both directions.

The translation problem, then, is real, large, and probably getting larger. How is it going to be addressed? A recent editorial in *The Times* of Malta concludes that the government and the EU Commission should be prepared to bring appropriately qualified human resources to bear.

This is obviously a necessary part of any solution, but the costs are unlikely to be sustainable given the expected increase in translation volume.

The only hope of a solution has to lie with technology. However, there are not only different technologies, but also different ways of involving them in the translation process, and the choice is far from simple. Below we discuss a few home truths about Machine Translation (MT).

Most of the points have been previously made by well-established practitioners in the field, notably Martin Kay and Jonathan Slocum in the 1980s.

The field of MT has a long history, being the first non-numerical application proposed for the emerging technology of digital computers in 1947 by Warren Weaver, the pioneer of information theory.

A very large amount of money was poured into MT research by the Americans during the 1950s and 1960s, the primary goal being Fully Automatic Machine Translation (FAMT), in the sense of an autonomous procedure capable of carrying an arbitrary text from one language to another with human intervention only in the final revision.

But the earliest efforts wildly underestimated the complexity of the task. Many factors contribute to the difficulty: ambiguity (words with multiple meanings, sentences with multiple structures) is one of them. Yet an essentially word-by-word approach had been adopted with little regard for grammatical or semantic issues.

This not only led to obvious mistranslations, some of them amusing (e.g. *les soldats sont dans le café* - "the soldiers are in the coffee"), but to results that were so unacceptable that the funding authorities called for an official report (the well-known ALPAC Report) that in 1966 resulted in a disillusionment with the field as a whole. This lasted well over a decade.

In passing, we should note that despite this lesson from the past, word-for-word, literal translation was the approach that prospective translators of the *acquis* were instructed to use by an official of the Office of the Attorney General.

Since the early days, our understanding of the computational techniques associated with syntactic and

semantic processing, and the speed and capacity of the hardware, have increased enormously. So has our respect for the complexity of the translation problem. We can now accept that high quality FAMT is, and will probably remain, elusive, if applied to text of arbitrary complexity.

However, in restricted domains, where a controlled sub-language of some kind is used, high quality FAMT is entirely feasible, as demonstrated by the very successful METEO system, developed at the University of Montreal, which still translates Canadian weather reports between English and French.

Conversely, when quality of translation is not an issue, but throughput is, as for example, in information-gathering tasks where the requirement is mainly to determine the general subject matter of a document rather than the exact translation, FAMT can profitably be exploited.

We now recognise that FAMT lies at one extreme of a continuum of ways in which technology can be brought to bear upon the translation problem. At the other extreme there are the mundane tools of the modern office: word processing software, fax machines, and even mobile phones. Between these two extremes there are other points of interest where technology can radically affect the productivity of the individual translator.

Below is a discussion of two such points: Machine Aided Human Translation (MAHT) and Human Aided Machine Translation (HAMT). The essential difference between these two lies not only in the way in which the person is involved but also in the extent of their involvement.

MAHT refers to translation systems in which (i) all initiative resides with the human and (ii) the basic program is essentially a text editor which has been souped-up with certain translation-specific functionalities.

Minimally these are likely to comprise simultaneous access to source and target texts, and online access to dictionaries, thesauri, terminological databases, and word concordance tools. These basic functions can be supplemented by other facilities.

Communication among translators, for example, which is of extreme importance when sharing the translation of a large document, can be improved enormously by integrated e-mail, instant messaging and the like.

Another issue is identification of and access to secondary materials. For instance, one of the most valuable resources a translator has for solving difficult problems is the text he is working on and other texts like it, in both source and target forms.

In the classical set-up of a physical translation bureau, access to these is often haphazard - yet it is clear that with the right mixture of database technology, smart indexing, and networking, improvements can be achieved that are well within the scope of current technology.

A good example of this is Translation Memory (TM) software, which stores matching source and target language segments that were translated by translator in a database for future reuse. Newly encountered segments are compared to the database content, and the resulting output (exact, fuzzy or no match) is reviewed and completed by the translator.

Human Assisted Machine Translation (HAMT) is another interesting point on the continuum and refers to systems in which the machine retains the initiative, but works in collaboration with a human consultant.

The central idea is that the system translates autonomously until it recognises that a linguistic difficulty of a certain type has arisen which it is unable to translate. When this happens, it seeks help from the consultant, communicating the nature of the difficulty in such a way as to elicit a quick and unambiguous response.

When designing such systems, the main technical issue is this ability to reliably and automatically identify, classify and communicate the nature of the translation difficulty. A valuable characteristic of HAMT systems is that if this issue can be successfully addressed, there is a guarantee of high quality output - even in less restricted domains.

A second important advantage of HAMT systems is that the human consultant need not necessarily have the skills of an expert translator. The minimum qualification is that of being a native speaker, who can be drawn from a significantly larger segment of the labour pool than fully-fledged expert translators. HAMT can thus have a significant impact upon the problem of skill shortage.

From this brief expose it should be clear that MT is capable of increasing the productivity of the individual translator or translation bureau, but to be successful, the choice of MT paradigm is crucial, depending upon a subtle interplay between factors such as the human cost, the machine cost, the quality of the result, and the nature of the translation task under consideration.

This distinction between paradigms is to some extent orthogonal to the underlying technology of MT. This remains in a state of constant evolution. Currently, there are two broad approaches, one of them knowledge-based, the other statistical. Knowledge-based approaches attempt to encode, using formal, rules understandable to a machine, the linguistic, semantic and contextual knowledge that is used, more or less unconsciously, by expert translators.

The system must then deploy that knowledge appropriately in order to carry out a successful translation. Systems designed on these principles tend to work well in restricted domains, where the number of rules applicable to a given situation is kept within reasonable limits. The greater the number of rules, the harder it is

to choose between them, and thus the greater the potential for wrong choices and poor quality results.

Statistical approaches, on the other hand, are not based upon rules at all, but upon the statistical regularities that can be automatically learned from existing translations. The starting point is typically a large body of parallel text which has been aligned, at paragraph and at sentence levels, before statistical analysis can be applied.

Typically this yields likely equivalences between units of different sizes: words, phrases, or even complete sentences, and these can then be employed for the translation of unseen material.

The success of this method depends on the availability of training texts. When these are easy to align a priori, this approach can be extremely valuable, not only for translation of sentences, but also for related tasks such as the acquisition of terminological equivalents.

It is too early to say which approach is best. Each has its merits, and given that MT is a subject area which has to grapple with the inherent messiness of real-world data we have to be pragmatic not dogmatic. So for some time to come, we are likely to see research into both kinds of system.

Some of the pertinent issues will be presented and discussed at the forthcoming workshop of the European Machine Translation Association, (EAMT), which is being hosted in collaboration with the Department of Computer Science and AI, University of Malta, tomorrow and on Tuesday at the Foundation for International Studies, St Paul Street, Valletta.

The special themes for this workshop are machine-translation-related issues concerning Semitic languages, and the languages of the newly accessioned states of the European Union.

For further details of the workshop please contact the local organiser, Michael Rosner on mike.rosner@um.edu.mt.
Mr Rosner is head of the Department of Computer Science and Artificial Intelligence of the University of Malta.