Computer-based translation tools, terminology and documentation in the organizational workflow: report from recent EAMT workshops

John Hutchins
(University of East Anglia, Norwich NR4 7TJ, U.K.)
[Email: WJHutchins@compuserve.com]

Introduction

The workshops held by the European Association for Machine Translation (EAMT) in the past three years have been devoted to exploring the practicalities of using machine translation software and other computer-based translation tools in organizations. In these environments, translation has to be seen by management as a positive enhancement to the promotion and sales of company products. For those involved in the provision of company translations the cost-effective exploitation of the most appropriate translation software and the careful design of operational workflows are of crucial importance. At two recent workshops in 1997 at Copenhagen and in 1998 at Geneva, the focus of discussion was the integration of MT and translation tools in the overall documentation workflow. Since this topic involved also the use and management of terminology resources, I shall attempt in this paper to relate the findings of these workshops to the interests of this conference. (It should go without saying that this will not be a full report of these workshops, which covered a range of topics and issues outside the focus of this presentation. See EAMT 1997, EAMT 1998.)

The first point to stress is that when talking about the use of computer software for professional translation work, there is no question of utilizing the cheap PC software that is becoming increasingly available in stores. These programs may be of value for the occasional non-professional (perhaps bilingual) user who wants a rough version of a letter or document, but who has no intention of producing from the results anything of publishable quality. There is, of course, a danger that some people may underestimate the difficulties of translation and may overestimate the capabilities of these cheap systems, but that is a problem of marketing – and, I must add, of the associations such as IAMT and EAMT which follow developments in this field. In this respect, I should add that the IAMT is proposing to set up a 'seal of approval' as guidance for consumers.

We may broadly distinguish four basic purposes for using translation software. First, for *assimilation* – to find out what an article, document, letter, etc. is about in general terms – where a rough, uneven version may be all that is necessary. Many MT systems are capable of providing such 'translations.' Secondly, for one-to-one *communication* – e.g. for translating electronic mail, for Internet discussion groups, and for spoken language, telephone communication, etc. – where high-level accuracy is not essential, but rapid delivery is crucial. Thirdly, the related area of *multilingual access* to databases and information resources in foreign languages for searching, retrieving and translating full texts of documents or for acquiring information (and including translation of Web pages), where the quality of translation in demand may vary from 'rough gists' to 'high quality' and where the speed of translation may or may not be a crucial factor. Fourthly, there is the most traditional purpose of all – translation for *dissemination*, the production of translations of

'publishable' quality, the traditional sphere of human translation, performed by the trained professional translator.

For the first three purposes, demand is growing rapidly and has not normally been provided by human translators. In fact, in the case of electronic communication and information access whenever translation has to be in near 'real time', the only viable 'solution' is automatic translation, where poor quality is acceptable. Conflict between human translation and computer-based translation is not an issue. In the fourth area (*dissemination*) there is an area of potential conflict, since the translation profession, since long before the advent of computers, has been devoted to the delivery of translations of good quality, to standards required by their customers require. In this area, computer tools and MT systems are still in the process of demonstrating their value and usefulness, and this has been a major focus of the EAMT workshops.

Translation systems and tools available

What computer tools for translators are available? At one extreme are the fully automatic translation (MT) systems, which take as input a text in one natural language and produce as output a text in another language. Since the quality of output is rarely good enough for publication, human assistance is inevitable – either before or after the machine processes. Assistance before processing may involve the standardization (or control) of the language of the input (grammar, terminology, or style) as well as preparation of the text by the temporary removal or material not to be treated (e.g. graphs and tables). Assistance after processing invariably includes the revision of the translation (correction or selection of terminology, correction of grammar and style) – traditionally called 'post-editing' – but it will also involve the restoration of tables and graphs and preparation for printing. In addition, human assistance covers the creation, maintenance and updating of glossaries and terminology databases.

Computer aids are equally familiar. They include software for the creation and maintenance of translator's own glossaries, for access and downloading from remote termbanks, for transmitting and receiving electronic texts, for desktop publishing, for producing concordances of texts, and above all for storing and accessing previously translated texts. The latter facility, in the form of the 'translation memories', comprises texts and their translations aligned sentence by sentence enabling the translator to locate examples of phrases (or sentences) already translated which may be re-usable in the current job. A translation memory is particularly valuable when translators are frequently given the task of translating versions of texts, such as manuals, that differ relatively little from one edition to the next. It is also valuable where the translator (or translation service) is translating many texts in the same specialized subject field. The important feature is that the translator retains the option to choose which aids to use, and to accept or reject the terminology or translation examples offered.

The most recent development is the translator's workstation which combines many of these aids, in particular translation memory and terminology management, at a single terminal. In addition, many workstations on the market offer automatic translation as a further option for translators. It means that translators can select the 'best' or most appropriate version from a multiplicity of sources.

However, this very variety can have profound effects upon work patterns. To make the optimal use of the resources available the translator must know when each is most appropriate for the particular environment. It is true to say, for example, that

there are spheres of traditional translation practice where many of these computer aids are still irrelevant. In the case of many free-lance translators undertaking work on a variety of subjects and in a variety of languages, it may be just too costly (in time and money) to create and maintain specialized glossaries: much of the terminology may be used for just one translation project. In other cases, however, much time (and ultimately money) may be saved by an initial effort to build up a domain-specific terminology resource.

Each computer aid has its own advantages and disadvantages (Lewis 1997). An MT system produces a full sentence translation and it does not rely on the presence of previously translated texts in its memory; on the other hand, the results may require so much post-editing that the translator might have been more efficient and quicker without the system. A further disadvantage is that for each new subject domain the dictionary of an MT system will probably have to be updated. In the case of a translation memory (TM), the principal advantages are that there are recognizably good quality versions that can be re-used, that repeat texts do not have to be re-translated, and that exact matches are not essential; on the other hand, the translation memory has to be created and sentences have to be aligned: but, there may be insufficient electronically available documents in appropriate fields, and sentence alignment does not always work accurately. In addition there is the question of quality assurance: a translation service needs to be sure that its database of translations is up to date and of high quality and makes use of the current terminology appropriate for the particular domain and/or client. The advantages of a well-constructed and well-maintained terminology database are clear enough: even though it provides information only for individual lexical items and gives no help with text translation as such, it can save considerable time and it aids consistency; the disadvantages are the amount of time required for maintenance and, as with TM, the question of quality assurance.

Integration of systems and tools

Increasingly it is the experience of organizations making extensive use of computer aids for translation that MT, TM and terminology databases can lead to definite economic benefits (Marty-Albisser 1997). However, full efficiency can be achieved only when the tools are made available in an integrated system, and ideally from a uniform interface. It is recognized that technical documentation is ideally suited to integrated workflow solutions, in view of the range of facilities to be handled (writing, editing, graphics work, translation into one or more languages, multilingual DTP) and the large volumes of documentation, the tight deadlines and the variety and repetitiveness of document versions. This situation is particularly apparent in the localization industry, where software and the associated manuals and documentation are translated for simultaneous product launches in a number of languages (O'Donoghue 1997). But, it is equally the case in many other multi-national companies and organizations.

One of the most impressive examples of integration is the system still under development, but already operative in large part, at the European Union's Translation Service (Theologitis 1997, Blatt 1998). The core of the system is EURAMIS, which combines: (a) access to full-text databases, particularly the CELEX multilingual database of EU official publications; (b) access to centralized translation memories (for each of the Service's seven thematic groups), topical TMs (e.g. for EU legislation, Court rulings), and personal TMs for each individual translator (accessible by colleagues); (c) facilities for creating TMs (alignment programs); (d) access to the

Commission's MT systems (Systran now offers 17 language pairs); (e) automatic extraction of terminology from the EURODICAUTOM database, conversion of terms into MultiTerm format, and access to the MT dictionaries. Translators can choose two different working environments: the TRADOS workbench or from a familiar word processor. In either case, retrieval of data is available from central TMs, from MT systems, and terminology resources in MultiTerm format. (We may note that MT systems are also available directly to administrators outside the translation service, who use it for the purposes of assimilation (information scanning) or document drafting, i.e. for writing in non-native languages.)

Although the Commission has gone a long way to easing the translators' working environment, it is recognized that there are still problems: with receiving documents for translation in electronic form (an easy-to-use interface for translation requests has been developed), and with downloading and aligning reference documents. Above all, a good deal of experience is necessary for translators to determine which applications should be used under which circumstances.

Terminology maintenance and control

The centrality of terminology maintenance is clearly recognized: consistency of terminology translation is just as important for computer-based systems as it is for non-automated translation services. Where there is a variety of different computer tools, the problems of maintenance and integrity are compounded.

SAP (one of the world's largest software companies) has been using MT systems for many years (both Logos and METAL), and has built up substantial dictionaries for them (Luecke 1997). At the same time, it has continued to use both in-house and external translators working in traditional manner. For these translators, SAP has maintained its own separate terminology database (SAPterm, currently with 70,000 entries). To facilitate outsourcing of translations, SAP is making these data available in an improved database (STERM), which will offer more terminological information, improved language coverage, transparent links to other glossaries, guidelines for updating, etc. (McCormick 1998). At the same time, SAP is a partner in the OTELO project, an EU-supported project for developing a uniform interface for translators. The main feature of OTELO is a central lexical database for the interchange and management of terminology using an agreed interchange format (OLIF). It is designed to accommodate traditional terminology and lexicographic data as well as MT dictionary entries. The integration of entries from MT systems into a terminology database raises inevitable problems: the presence and/or absence of semantic, syntactic and morphological information; the presence or absence of definitions and examples; the presence or absence of equivalents in other languages (and different target languages for different sources), the presence or absence of intra-language semantic relationships (synonymy, taxonomy, part-whole, etc.). Other problems encountered are that SAPterm (and STERM) entries are concept-oriented, whereas OTELO is lemma-based; in OTELO entries part-of-speech indications are obligatory, in SAPterm they are optional; in MT systems the gender of nouns has to be marked, in terminology databases it is rarely coded.

Documentation workflow

The full integration of MT and TM tools in the total documentation process (from authorship to printing and distribution) has been stressed by many speakers. With growing demand for high-quality multilingual documentation at ever-shorter deadlines and at ever-larger volumes, it is clear that a cost-effective well-organized

documentation and translation workflow is essential. This is the experience of an increasing number of organizations. Océ Technologies' experience is typical (Cremers 1998). Beginning with the decision to introduce an MT system (Logos), the company found that it had to improve its terminology control. This in turn encouraged the control of the original English documents (manuals) to ensure consistency of terminology. Then a translation memory systems was added (first XL8, later Trados), and these were integrated into the translation process. Fortunately, the company had a centralized documentation department, which made it much easier for the managers to influence the writing process itself, to control document structure and formatting, and to adapt existing methods to automated translation tools and systems. As a result, schedules were devised to facilitate authoring, terminology work, review of source texts, the translation process (using MT or TM as appropriate for the document type or subject), and the revision of output. Software aids have been introduced to ease the flow: shell scripts, macros in Word, program integration, format conversion, text passage protection, etc.

Much of the integration required by translation services is, of course, now incorporated within the translation workbenches available commercially: the option of using term databases, translation memories and MT systems is now normal and expected. Development of new and more sophisticated systems continues at a number of companies: Xerox at Grenoble is a prime example, with particular emphasis on terminology management, translation memory systems, and project management (Chanod 1998). Further optimization of the documentation process is the goal of developments by STAR (Kaeser 1998). It was reported that their product will be incorporating a 'workflow system' that monitors and supports the whole process from initial receipt of the text to the final output.

The process may be illustrated with a translation agency sending jobs out to independent translators. First the workflow system receives the text from the author, converts formats as necessary, segments texts, and performs 'automatic pretranslation' using data identified by the system itself as the most suitable reference material. Then, it identifies the translator best suited to the particular job (type and volume of documentation, language combinations, current availability), and sends the archive file via modem or network to the translator. The translator will have a client version of Transit installed (TRANSIT Lite), together with a client version of the workflow system. This reads in the data, installs the translation job automatically and alerts the translator. The translator can now load TRANSIT, with the specified parameters for the job, and begin the work of translation – using the now familiar combination of terminology databases, translation memory (if appropriate) and MT resource. The workflow system tracks each job, sets and monitors deadlines, sends reminders, etc., and all transactions are logged and recorded.

Controlled language

The advantages of controlling the vocabulary, grammar and style of texts for translation has long been recognized in the case of full MT systems. The LANT service uses a combination of the METAL translation software and the Eurolang Optimizer, which LANT now markets (Caeyers 1997). It offers a translation service to large companies, in which an important component is the incorporation of facilities for writing documents in Simplified English (also based on the software developed for the METAL system.) The benefits are promoted as improved readability and understandability of documentation (whether translated or not), decrease of liability risks (by improved terminology consistency and clarity), and improved translatability.

As an example of improved understandability, the system (LANTMaster) would propose the uncontrolled:

When fasteners are removed, always reinstall them at the same location from which they were removed

by the controlled:

Always reinstall fasteners in the same location.

As an example of avoidance of liability risks, the uncontrolled:

Unscrew the plug-in unit from the wall and paint it.

is replaced by the controlled:

Unscrew the plug-in unit from the wall and paint the plug-in unit.

The advantages of control are: consist use of terminology, simpler sentence structures (often shorter sentences), lower levels of verbosity, lower semantic complexity, absence of ellipsis and indirect rhetorical constructions, lower levels of ambiguity, absence of pronouns and other anaphoric expressions. All contribute to clarity, effectiveness and reliability of the original document and at the same time increase translatability whether performed by an MT system or by a human translator at a workbench.

The practicalities of implementing a controlled language are being tackled by Cap Gemini (Van der Eijk and Van Wees 1998). Technical writers often find it hard to create documents in a controlled language, especially if they can no longer use constructions previously acceptable in the particular 'sublanguage' they have been writing in. Clearly, good training is essential. But also what is needed is supporting software for the authoring process. These can be tools which generate informative messages of guidance for authors, and tools which perform correction automatically.

Organizational impact

These developments have clear impacts on the personnel employed in organizations. There are not just the issues of careful planning for the introduction of new tools and systems, and of the adequate training of translators, technical writers and others involved. At the simplest level, the translators must become not just efficient and versatile users of word processors, but also familiar with terminology databases, translation memories and even MT systems. Above all, there are implications for the actual management of the service itself (King 1998). The introduction of a translation memory system (whether as part of a translation workstation or not) means that either translators themselves or one of their team must learn how to create the appropriate resource database – which may well mean learning how the alignment software works (and what its weaknesses may be) – that someone must be responsible for deciding which texts to incorporate and for keeping it up to date. The management and maintenance of terminology is an equally important issue: either each translator has responsibility for their own database, or a centralized resource is created under the direction and responsibility for a member of the team. If any of these tasks are contracted out, e.g. to an agency, then there is still the need for efficient liaison with translators.

Apart from translators, what kind of personnel is required, particularly in a large organization? The final discussion session of the Geneva workshop identified the following. First there is the *project manager*, who controls and monitors a particular translation and/or documentation job through its lifetime, defines parameters (domain, customer, languages), partitions documents according to translator resources, ensures that deadlines are met, that resources are available as

required, and of course who manages the human resources – as in any other standard management activity. Secondly, there is the *language engineer* who acts as a consultant for the installation of technology required for particular services, proposes how tools are technically integrated most effectively, how computer and database resources are accessed and made available, etc. Ideally, this should be someone with qualifications and/or expertise in computer technology, and with an understanding of what translation involves. For the day to day technical tasks there could be a *translation technician* who would set up systems, load software, manage computer networks and databases, and performs general assistance. Fourthly, workshop participants saw the emergence of a new kind of post within services for translation and multilingual documentation. There are tasks such as text alignment for translation memory databases, preparation and management of documents pre- and post-translation, preparation of linguistic and terminological resources, etc. that do not necessarily have to be performed by a highly paid translator. These tasks could be done by *linguistic assistants* who might be secretarial staff sensitive to linguistic and translation operations.

At centre stage continue to be, of course, the *translators* themselves, responsible for effective use of the resources and for the quality of output, but ideally they should be involved only after the technical and resource preparations have been done by the support staff. Of course, in smaller organizations one or more translators might also be project managers, technical assistants, and 'linguistic assistants', but it is nevertheless useful (from an overall management perspective) to identify the different roles of those involved in a translation project.

The introduction of computer-based translation tools is not purely a technical and management issue. This understates the challenge. Most organizations tend to underestimate the effort required for re-engineering processes of multilingual document production. There is a tendency to continue with traditional working practices, but new technologies demand new thinking. The papers given at these recent EAMT workshops are a reflection of the ways in which organizations are seeking to make most effective and efficient use of the new tools for translators.

References

Blatt, Achim (1998): Workflow using linguistic technology at the Translation Service of the European Commission. In: EAMT (1998), 7-18.

Caeyers, Herman (1997): Machine translation and controlled English. In: EAMT (1997), 91-103.

Chanod, Jean-Pierre (1998): Multilingual tools at the Xerox Research Centre. In: EAMT (1998), 73-84.

Cremers, Lou (1998): Using automated translation in a corporate setting. In: EAMT (1998), 27-28.

EAMT (1997): *Language technology in your organization?* Proceedings of the 1997 EAMT Workshop, Copenhagen, 21-22 May 1997, Center for Sprogteknologi, University of Copenhagen, Denmark.

EAMT (1998): *Translation technology: integration in the workflow environment.* Proceedings of the 1998 EAMT Workshop, 2-3 April 1998, WHO, Geneva, Switzerland.

Kaeser, Paul (1998): Workflow automation of translation projects. In: EAMT (1998), 41-44.

King, Margaret (1998): Workflow, computer aids and organisational issues. In: EAMT (1998), 29-40

Lewis, Terence (1997): When not to use MT and other translation tools. In: EAMT (1997), 34-41.

Luecke, Dirk (1997): I18N/(MT + TM)2. In: EAMT (1997), 65-76.

Marty-Albisser, Doris (1997): Economics of translation tools – experience gained at Swiss Bank Corporation. In: EAMT (1997), 8-16.

McCormick, Susan (1998): A centralized approach to managing multiple lexical resources. In: EAMT (1998), 19-26.

O'Donoghue, Tim(1997): Localizing Canon's user documentation in Europe. In: EAMT (1997), 56-64

Theologitis, Dimitri (1997): EURAMIS, the platform of the EC translator. In: EAMT (1997), 17-32.

Van der Eijk, Pim and Van Wees, Jacqueline (1998): Supporting controlled language authoring. In: EAMT (1998), 65-70.