

ACHIM BLATT

Translation Technology at the European Commission: Description of a Workflow

The following describes the different translation technologies which are currently available in the Translation Service (SdT) of the European Commission, in order to exemplify their usage, they are described in a (maximalistic) workflow.

For almost any of the technologies described below, it is essential to have the source document available in electronic form. In order to give requesters an incentive to use electronic mail, the SdT has developed a simple and user-friendly interface, known as POETRY (Processing of Electronic Translation Requests), which allows users to send a translation request together with the document to be translated and, if possible, reference material. This request is then passed on to individual translators via WinSuivi, a management tool which makes it possible for the work to be allocated according to the target languages and products required.

If one abstracts away from organisational aspects (which parts are done by secretaries, which parts are done by translators etc), a difference has to be made between the preparation of a translation, and the translation itself. In the worst case, a translation memory has to be created from zero.

1 Building-up a translation memory (TM)

In the SdT, translation memories are increasingly often used in order to improve productivity as well as quality and coherence. If users want to use TM technology, possibly combined with machine translation or replacement tools, and if translation memories do not yet contain enough suitable data, they have to look for reference material which can be imported.

1.1 *Reference documents*

The Commission's translators have access to a number of on-line full-text databases, which can be searched for reference material.

SdTvista contains almost all the translations from 1994 onwards (except for confidential documents) plus many source texts and a number of reference documents. This allows users to check whether or not a document, or part of it, has already been translated and to retrieve pertinent source documents as well as their translations.

Queries are typically made on the basis of search strings, but they can be refined by means of additional filters (eg requesting service, year, translation type etc).

If interesting material can be found, it can be viewed or downloaded for further treatment. During the translation process, translators might also consult SdTvista in order to solve terminological problems.

CELEX is the full text database of the Office for Official Publications of the European Communities. It offers multilingual coverage of a wide range of legal documents. CELEX offers a user-friendly Internet access with a large number of different query types.

The SdT provides an automatic batch retrieval where, on the basis of references found in a document, CELEX identifiers are calculated and the corresponding documents are returned to the user. For example, the German sentence:

In der Verordnung (EWG) Nr 210/69 der Kommission, zuletzt geändert durch die Verordnung (EG) Nr 1171/96, sind die Informationen zur Verwaltung des Marktes für Milcherzeugnisse festgelegt, die der Kommission regelmäßig mitzuteilen sind

yields the identifiers 369R0210 and 396R1171, which are then used for remote CELEX queries of titles or complete documents. The example above leads to the following German and English titles:

- Verordnung (EWG) Nr 210/69 der Kommission vom 31. Januar 1969 über die gegenseitigen Mitteilungen der Mitgliedstaaten und der Kommission im Sektor Milch und Milcherzeugnisse
- Regulation (EEC) No 210/69 of the Commission of 31. January 1969 on communications between Member States and the Commission with regard to milk and milk products
- Verordnung (EG) Nr 1171/96 der Kommission vom 27. Juni 1996 zur Änderung der Verordnung (EWG) Nr 210/69 über die

gegenseitigen Mitteilungen der Mitgliedstaaten und der Kommission im Sektor Milch und Milcherzeugnisse

- Commission Regulation (EC) No 1171/96 of 27. June 1996 amending Regulation (EEC) No 210/69 on communications between Member States and the Commission with regard to milk and milk products.

1.2 Alignment

Interesting reference documents and their translations can be downloaded to the user's PC, and a sentence alignment request can be launched. The SdT uses its own aligner which is highly customised and therefore produces markedly better results than commercially available applications. Alignment results can be corrected by a special editor and later imported into a translation memory.

2 Preparation of work

If users want to work exclusively with their own local translation memory, they can start translating at this point. They can however request additional information coming from other sources via the EURAMIS client interface which gives access to a number of batch services.

2.1 TM

The SdT has developed server translation memories whose design facilitates data sharing and the combination with other products. Central translation memories are provided for each of the seven thematic groups of the SdT. In addition, there are a number of consolidated topical translation memories (eg most important European legislation and Court rulings, the monthly Bulletin etc). For each of these translation memories, a coordinator grants write access, sees to a coherent use of attributes etc. In addition to common TMs, there are personal TMs for every translator: there are no restrictions concerning read access, everybody can retrieve from everybody's TM.

EURAMIS TM queries can be fine-tuned by means of positive and negative filters ("only these" or "all but these"). The following filters are possible: requesting service (eg Directorate-General XXII), domain (eg agriculture), document type (eg letter) and individual translator (unique login).

2.2 Machine translation (MT)

MT is currently used in the Commission for 17 language pairs with a quality which varies considerably between the different languages: French-Spanish offers the best quality, followed by French-English, French-Italian and English-French; language pairs with German produce less acceptable results.

MT is available not only to translators, but to all Commission officials who are equipped with either a terminal or a PC connected to the internal network. The number of pages translated by the system has increased considerably in recent years: 170 000 in 1995, 231 000 in 1996, 260 000 estimated for 1997. MT is mainly used for

- the fast translation of short, repetitive texts with a standardised structure and terminology (mail, minutes of meetings, parliamentary questions, reports etc);
- the browsing of texts written in a language the user does not know;
- drafting purposes: users write a text in their mother tongue and request a machine translation in order to have a document drafted in something other than their native or main language.

MT is appreciated by translators because of its speed, its capacity to keep the original format, and as a terminology aid.

2.3 TMan

TMan replaces pre-defined strings (from words up to paragraphs) in the source document so that the resulting document is a mix of source and target language items. TMan replacements are based on a repetition analysis of the document type in question. This means that the use of TMan is only worthwhile if a document type is quite frequent and if it contains a large number of repetitive elements.

This approach is taken for a number of master documents and regular publications (eg the Bulletin) where many expressions can be found in every issue, and consequently the analysis of previous texts yields large expression databases to be used for subsequent issues of the same publications. For example, the following French text (from a prototype contract)

*La Communauté européenne, représentée par la Commission des Communautés européennes, ici représentée par Monsieur
Directeur général,
d'une part*

et la firme "...", dont le siège social est à ... ci-après dénommée "le fournisseur", représentée par Monsieur... en vertu de la délégation lui conférée par ladite société
would be translated as follows, where colours (put in bold here) show what remains to be translated:

La Communauté européenne, représentée par la Commission de las Comunidades Europeas, **ici représentée par Monsieur ...**, Director General, y, de otra, **la firme"..."**, con domicilio social en ... denominada en lo sucesivo "el proveedor", representada por el Sr ..., según el poder otorgado por la citada sociedad.

2.4 *Combination of products*

It is possible to combine TM results with MT (the same integration with TMan is foreseen): on the basis of the user requirements, a EURAMIS TM retrieval is carried out, and the remaining gaps are filled with MT results.

2.5 *Document-related terminology*

EURAMIS provides automatic terminology extraction of pertinent EURODICAUTOM entries, together with indirect access to TIS and EUTERPE, the terminology databases of the European Council of Ministers and the European Parliament. Since the Commission's MT system is used as a supporting application, this service is restricted to documents in English, French, German and Spanish. The user can choose between the following output formats: EURODICAUTOM import format (for data correction), a sequential RTF file, and MULTITERM input format.

Queries can be restricted in a number of ways (eg by indicating domain). Output can be restricted by selecting specific EURODICAUTOM fields.

As a first step towards migrating EURODICAUTOM to a relational database management system, a new database structure has been designed recently: among other things, formal restrictions are imposed on values and synonyms are put in separate fields. A conversion to this new data model (including some clean-up and normalisation) is already used for output in MULTITERM.

3 Translation proper

In principle, users can translate in two different working environments; depending on their own preferences and on the type of the document to be translated, they can choose between TRADOS' Translator's Workbench (TWB) and a simpler treatment completely inside word processing. With TWB, the products mentioned above can be used in parallel: retrieval from central TM is imported together with MT output; the latter receives a special attribute in order to warn users.

If the result is requested in native Word format, the formatting of the source text is preserved to a very large extent. Since colours are rare in Commission texts, they are used to convey information on result type and have to be reset after editing, eg blue for TM full matches, red for TM fuzzy matches, and magenta for MT.

ACHIM BLATT
Translation Service
European Commission
Luxembourg