[*Terminologie et Traduction* 1.1998]

ACHIM BLATT

EURAMIS: Added Value by Integration

1 Introduction

The acronym EURAMIS stands for European Advanced Multilingual Information System. This acronym refers to two generations of e-mail based client-server applications which are used in the European Commission's Translation Service (SdT) and give access to a number of services in the domain of natural language processing. The first generation is currently installed on all PCs of the Translation Service, whereas the second generation is installed on the PCs of those 300 users who already use local translation memory technology; it will gradually replace its predecessor.

The project was conceived and initially managed by JEAN-MARIE LEICK of the Commission's Directorate-General XIII. Following a call for tenders on the "Development of multilingual tools and their integration into multilingual services" (launched in 1994), actual work started in 1995. In 1996, the SdT took over the financing and management of the project, to which it had contributed manpower from the beginning. EURAMIS is based on the following general principles[1]:

- Storage of linguistic resources of general interest in one central place (the Linguistic Resources Database, LRD) in order to make them available to all users: in a translation service as large as the SdT, this is the only possibility to make sure that everybody has access to the linguistic data he may find useful;
- one internal format (pivot format with conversions to and from existing satellite applications) so that results from any module can be reused by any other;

1 See also LEICK'S paper in this volume p 52.

- mass treatment of linguistic data on a server by modular programs which are controlled by a service dispatcher.

These three aspects make it possible to reach synergies in a number of ways:

Central availability of data is a prerequisite for data sharing between different applications: here, the benefits from synergy consist mainly in the fact that data does not have to be duplicated. This leads to greater coherence and better maintainability. In addition, it turns out frequently that data which has been introduced for one application can, on second thoughts, be used by other applications.

The existence of one internal format makes it possible to combine the results of different programs into one result file - which is then much more useful than the individual results in isolation.

The dispatcher's generic command structure facilitates the reuse of modular programs so that they can be combined to new services - services which would otherwise have been too expensive to develop.

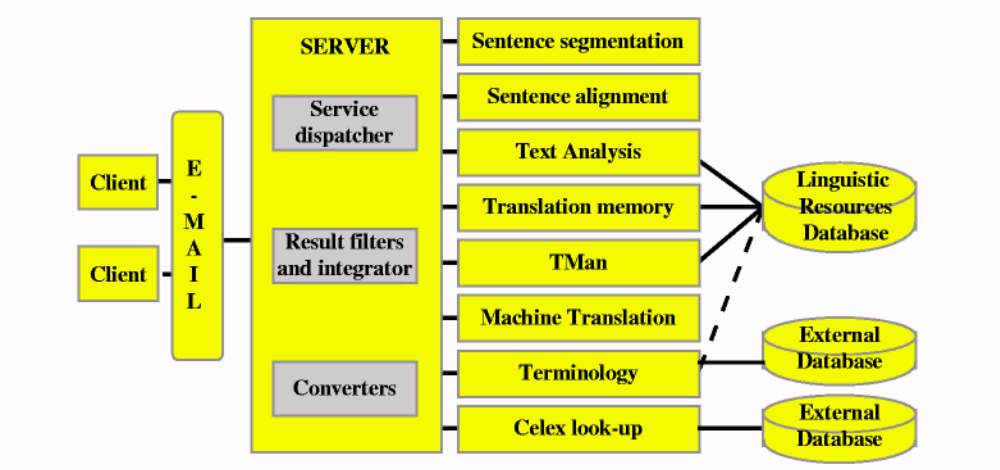
The following sections describe the technical background which makes such an integration possible, some practical examples which have already been implemented, and an outlook to future developments (see also LAVIGNE's paper in this volume, p 27).

2 Technical background

If the gentle reader is not so much interested in technical details, s/he can skip this section without losing the red thread.

2.1 *System architecture*

The following figure describes the general architecture of the EURAMIS system (terminology is not yet fully integrated in the Linguistic Resources Database - see also section 3.3):



The user creates a request by means of the EURAMIS client interface, a user-friendly Windows application. A request consists of

- the instructions by the user (the command file): which products are requested, which parameters have been set;
- the file(s) to be treated.

The client interface has in its knowledge base a number of inter-dependencies and constraints which prevent the user from launching requests which cannot be fulfilled, eg language combinations for machine translation which are not yet catered for.

The request is sent by e-mail to the server where a service dispatcher reads the command file and launches the programs needed in the appropriate sequence. Where necessary, the applications have access to the LRD, which is the storage area for linguistic data. The applications send their results back to the dispatcher, which passes intermediate results on to the subsequent application; final results are wrapped together and sent back to the user.

Program modules can be used in different sequences and for different purposes. For example, the conversion modules between text formats and pivot format and the sentence segmentation program are used for almost all services.

2.2 *LRD*

The LRD is the storage area for linguistic data of all EURAMIS applications. The aim is not only to have one common access mechanism to all linguistic data, but also to share common linguistic resources and thereby to exploit the synergies between the various data.

These objectives are met by one common database with application-specific data definitions: it is clear that translation memory information differs from, say, terminological information, but there are also intersections, eg main entry, administration information on users, change date and the like, possibly also on domain, references etc.

Data sharing applies not only to applications, but also to users: there is a logical separation between databases at user, workgroup and general level. By definition, every user has read access to all databases. Write access from group level onwards is granted by the person responsible for the database in question. Procedures have been designed for "harvesting" database contents to higher levels.

2.3 *Pivot format*

The EURAMIS pivot format constitutes the communication platform between the different EURAMIS applications; it uses an SGML definition which is as near as possible to HTML (the format used on the World Wide Web). A number of deviations from HTML and some extensions to it were necessary mainly for two reasons:

- Not all the formatting information which is encountered in documents created by word processing is supported by HTML (eg hidden text);
- it was necessary to introduce elements which are specific to the results of the linguistic programs.

In addition, EURAMIS pivot format uses Unicode (UCS-2): by using two bytes to represent each character, Unicode enables almost all of the written languages of the world to be represented using a single character set. This means that problems with Greek, diacritic characters and the combination of these are avoided.

A pivot document consists of a header whose most important element is the history, listing the applications that have worked on the file, and a body, containing the source document sentence by sentence. Each application that works on a pivot file leaves a trace in the history and adds its result to the sentence in question. In the following example, a document has been converted from Word6 to pivot format and segmented into individual sentences; from this point onwards, all subsequent applications add their results to each source sentence (explanations are put in bold):

history:
<appl appname="WINWORD6CONVERTER1.18" inst="#1" ...>
<appl appname="SEGMENTER" inst="#2" ...>
<appl appname="TM" inst="#3" ...>
<appl appname="MTF_CONVERTER" inst="#4" …>…

source sentence (inserted by sentence segmentation):
<s inst="#2" p="6" s="4">
Communication de la Commission sur les suites données aux avis et résolutions adoptés par le Parlement européen lors des sessions de septembre I et II 1996

translation memory - opening tags and information on database:
<TM inst="#3" NoS="1" SL="FR" TL="DE">
<Match LRD="blattac" FQ="88" ...>
<Admin ... DTY="Suites données" ... PYR="1997"> </Admin>

source sentence as found in translation memory:
<LRDS>Communication de la Commission sur les suites données aux avis et résolutions adoptés par le Parlement européen lors des sessions de mars I et mars II 1996</LRDS>

translation as found in translation memory:
<LRDT>Mitteilung der Kommission über die Folgemaßnahmen zu den Stellungnahmen und Entschließungen, die das Europäische Parlament auf den Märztagungen I und II 1996 verabschiedet hat</LRDT>

translation created (with markup of replacements):
<T>Mitteilung der Kommission über die Folgemaßnahmen zu den Stellungnahmen und Entschließungen, die das Europäische Parlament auf den Märztagungen I und <RC Ty="ROMNUM" no="0">II</RC> <RC TY="YEAR" no="1">1996</RC> verabschiedet hat</T>

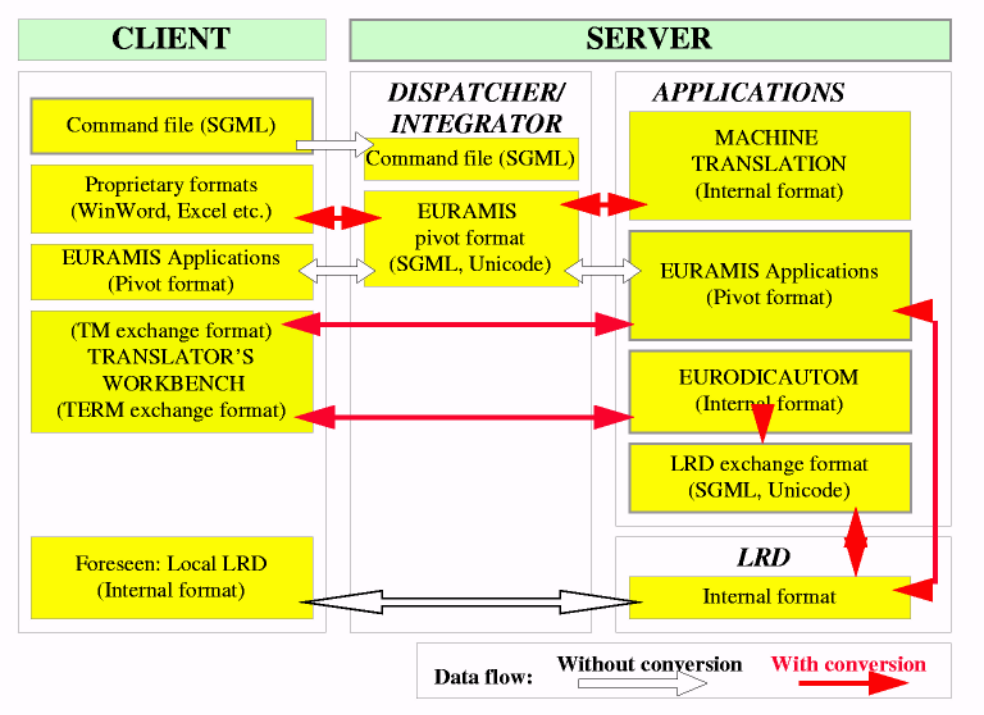
translation memory closing tags:
</Match>
</TM>

SYSTRAN result (inserted by MT filter) and closing tags of sentence:

<MT inst="#4">Mitteilung der Kommission uber die Weiterbehandlung der Ansichten und EntschlieBungen, die durch das Europaische Parlament auf den Sitzungen vom September I angenommen wurden, und II 1996
</MT>
</sxp>

Every application "knows" which parts of a document it has to see; it has to insert its results before the closing tag of the sentence treated (for SYSTRAN, which is an application external to EURAMIS, this task is carried out by a specific filter). This approach leads to one uniform treatment, ie all applications work independently of the order in which they are called.

The following graphic shows how the pivot format is used for exchanging information between applications. In order to make the pivot format work with existing applications which use their own external formats, a number of conversions have to be made:



3 Examples

Three examples of added value by integration are given in this section: the first example shows how synergies are created by sharing data between different applications (via the LRD). The second example demonstrates the integrative power of the EURAMIS pivot format. The third example shows how a number of existing programs can be combined in such a way that a new service is created without much additional effort.

3.1 *Reuse of TM data by other applications*

The Combined Nomenclature (CN) is the European Community's classification of goods. Depending on the level of detail, the length of its entries varies from one word to several lines, cf the following French and English examples (the numerical part has been left out):

(1a) Plomb sous forme brute
(1b) Unwrought lead
(2a) Poudre de cacao, avec addition de sucre ou d'autres édulcorants, d'une teneur en poids de saccharose y compris le sucre interverti calculé en saccharose ou d'isoglucose calculé également en saccharose, 5% mais < 65%
(2b) Cocoa powder, containing added sugar or other sweetening matter, containing >= 5% but < 65% by weight of sucrose, incl inverted sugar expressed as sucrose or isoglucose expressed as sucrose

Parts of the CN are used in a number of Community documents. It is therefore interesting to have all language versions of the different entries available in a central translation memory. Since they are also useful as reference material, these entries can also be obtained in MULTITERM format. Equally, since the entries are to a certain extent binding for other documents, they are suitable for use in TMan replacements (generally, these are partial replacements below the level of the sentence).

3.2 *Integration of machine translation with translation memory*

As the pivot format example in the previous section shows, EURAMIS makes it possible to combine translation memory results with machine translation. This integration combines the advantages of translation

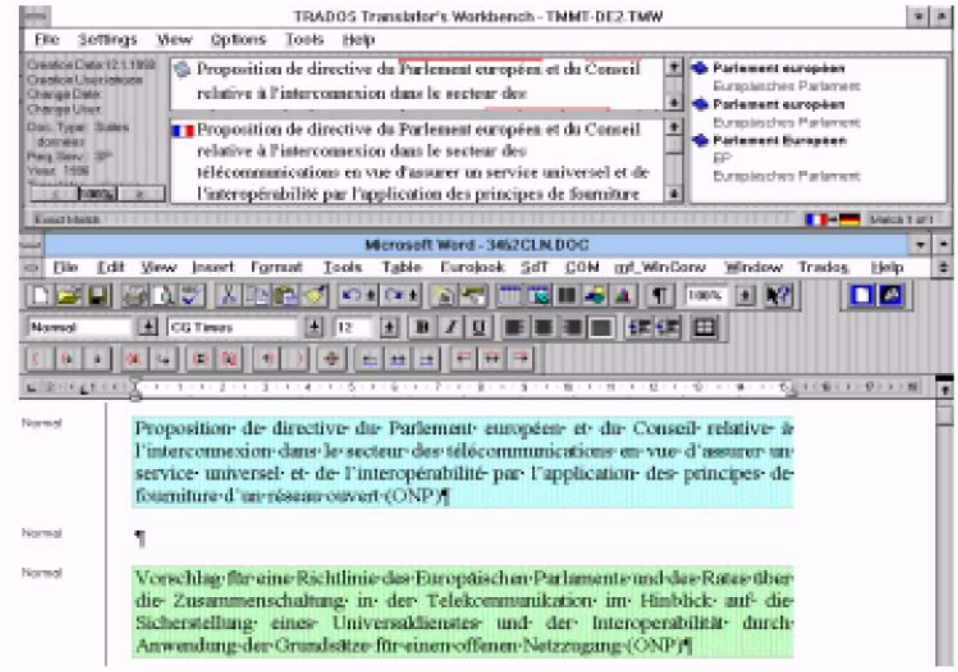
[2] It is foreseen to offer the same kind of integration with TMan (replacement below sentence level), eg for the language pairs for which no machine translation is available.

memories (better quality, based on human translation) with the advantages of machine translation (which always comes up with a suggestion).

This integration leads to a situation where it is no longer necessary to choose between TM and MT, but where there is an added value compared with either approach taken on its own: those who prefer working with a translation memory can still get suggestions from machine translation where translation memory could not provide satisfactory results; and those who prefer machine translation can use translation memories as an added value, eg in order to reuse binding translations or to be consistent with previous ones.

There are currently two ways to work with combined TM-MT results: depending on the users' own preferences and on the type of document to be translated, they can choose between Trados' Translator's Workbench (the SdT is acquiring licences of TWB to be used as a front-end for EURAMIS); or by directly editing a result file in native Word format.

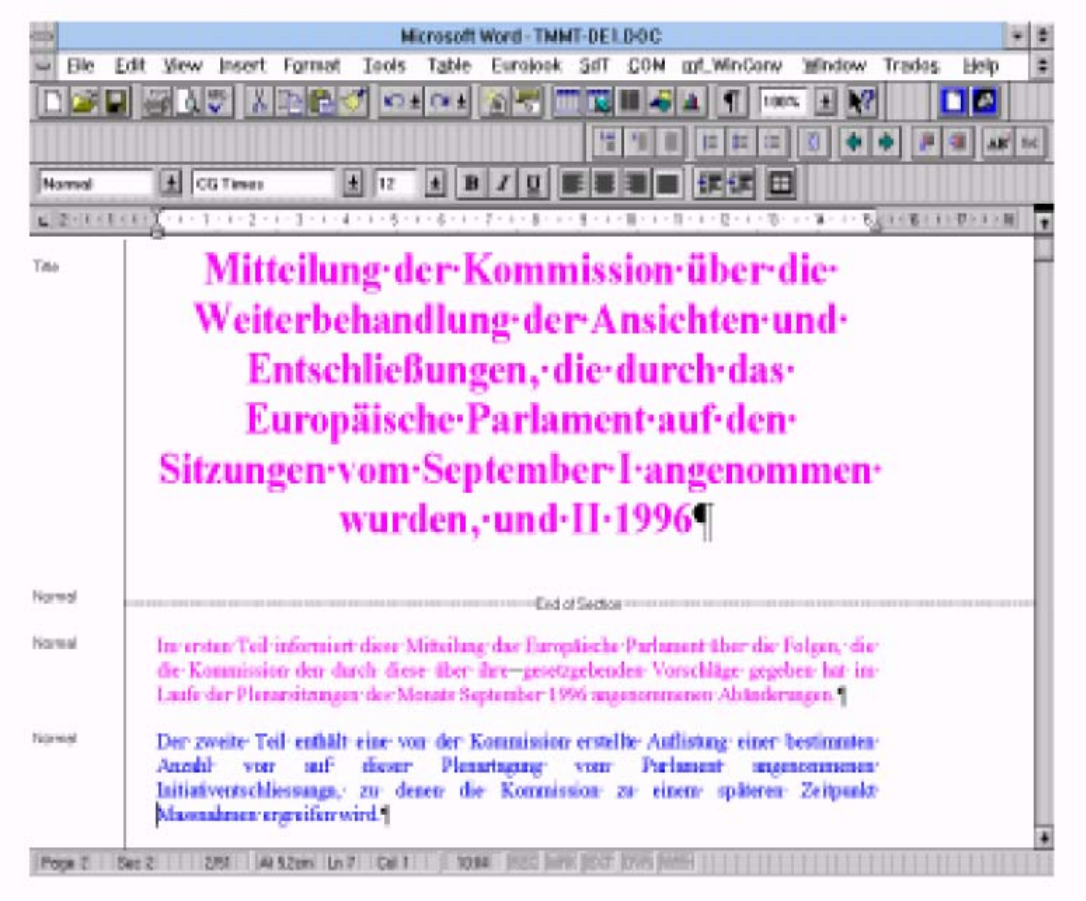
With TWB, results from TM and MT can be used in parallel, ie retrieval from central TM is imported together with MT output for every sentence; MT output receives a special attribute in order to warn users. During the translation process, different background colours are used: green for full matches, yellow for fuzzy matches, and grey for MT results:



Results can also be requested in native Word format, with the formatting of the source text being preserved to a very large extent: for MT results, even character formatting (eg bold, italic etc) at word level can be taken over (MT "knows" what is translated by what); for TM results, this can be done only from sentence level onwards. Since colours are rare in Commission texts, they can be 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.

As opposed to TWB, only one solution can be offered for a given sentence. This means that users have to set a fixed threshold for the degree of fuzziness they are prepared to accept for sentences from the translation memory. Below this threshold, MT results are taken.

The following picture shows the integration of MT and TM results in a Word document:



From the users' point of view, a TM retrieval is carried out with the filters they indicate, and the gaps left by TM are filled by MT. But this is not exactly what happens: in reality, the whole text is submitted to MT, ie even those sentences for which a translation has been found in TM. This means

that in principle, there is no difference to the creation of TWB output. The only difference is that after MT has worked on the whole text, a small filter fetches the TM results, or if need be, the MT results. This is not only technically simpler, but also offers linguistic advantages:

- Since MT sees the whole text, it has a better chance to find the referents of pronouns (important for the decision on how a pronoun has to be translated)[3];

  MT does not miss information which is valuable for resolving semantic ambiguities if the sentence where the information would come from happens to have been translated by TM; eg French "centrale" may have many translations in English, but it should most likely be translated as "power station", if in the preceding text, there has already been an occurrence of "centrale nucléaire".

-

The following table compares the pros and cons of the two approaches:

Translator's Workbench	Integration in word processing
Pro	**Con**
choice between TM and MT results on a case-by-case basis, eg between a relatively weak fuzzy match and the machine translation result of the sentence to be translated	only one solution per sentence (to be predefined by means of a threshold)
interactive, ie document-internal repetition can be exploited immediately	more like post-processing (no immediate learning effect by the application)
immediate update of the database (in interactive mode)	sentence alignment always necessary for update
fully integrated with MULTITERM so that terminology can be consulted on the fly	stand-alone: terminology lookup by cut and paste or similar techniques
additional features such as concordance and coverage analysis	no additional features
Con	**Pro**
specific cleanup operation needed at the end of the translation process (to remove bookmarks and hidden text inserted by TWB)	simple reset of colours (supported by macro)
danger of accidental deletion of such hidden items by inexperienced users which leads to problems with cleanup	not applicable
space on the screen shared by word processing and TWB	full screen for word processing
TWB too complicated for many people	simple word processing
specific training needed for correct use	no specific training needed
licence needed for free-lancers and tele-workers	no licene needed for free-lancers and tele-workers

3 Apart from syntactic criteria (eg number and gender), pronoun resolution is based on heuristics where syntactic status of the candidate referent and its distance to the pronoun in question are the major factors. Since candidate referents are inspected across sentence boundaries, extra-sentential context must be available for optimal treatment

This comparison shows that a number of factors will influence the choice of the interface: personal preferences, background (TM-oriented or MT-oriented), specific properties of the document to be translated (repetitive or not), availability (or not) of manpower for subsequent alignment etc.

3.3 *Terminology from text: reuse of existing programs*

Automatic lookup of pertinent terminology for a given document is one example of reusing existing programs in order to obtain new services: SYSTRAN, the Commission's machine translation system, contains a very powerful morphological analysis which can be used as a lemmatiser. The result of this analysis is used for batch lookup in the central terminology database EURODICAUTOM. This retrieval is then upgraded by means of a converter which had originally been implemented for a completely different task.

3.3.a Background

A number of years ago, an effort was made to create routines for importing EURODICAUTOM data into SYSTRAN. It turned out that terminological entries are generally useful for machine translation only from a certain length onwards (there is always the risk of too specific translations for single words or smaller expressions). Depending on a number of factors, usage of EURODICAUTOM entries had to be restricted to entries from a certain length onwards. One factor is the general quality level of the language pair in question: for well developed language pairs, short terminology entries (which can still be very specific) might spoil more than they improve; if the user has indicated a specific domain for the text in question, shorter expressions can be taken from that domain.

Although expectations for the original goal had therefore not been entirely fulfilled, text-related terminology lookup emerged as an interesting by-product from this. In order to achieve this, a mock-up SYSTRAN translation is run with priority given to EURODICAUTOM entries; after the various stages of dictionary lookup have been completed, a list of all words and expressions encountered is produced. This list is then used for EURODICAUTOM batch lookup. This means that queries can be made from any of the SYSTRAN source languages to any official language of the European Union.

This service originally produced text files with the information requested by the user (it was possible to restrict output by indicating the EURODICAUTOM fields desired); frequently, too much output was produced, which limited usefulness considerably.

3.3.b Ongoing work

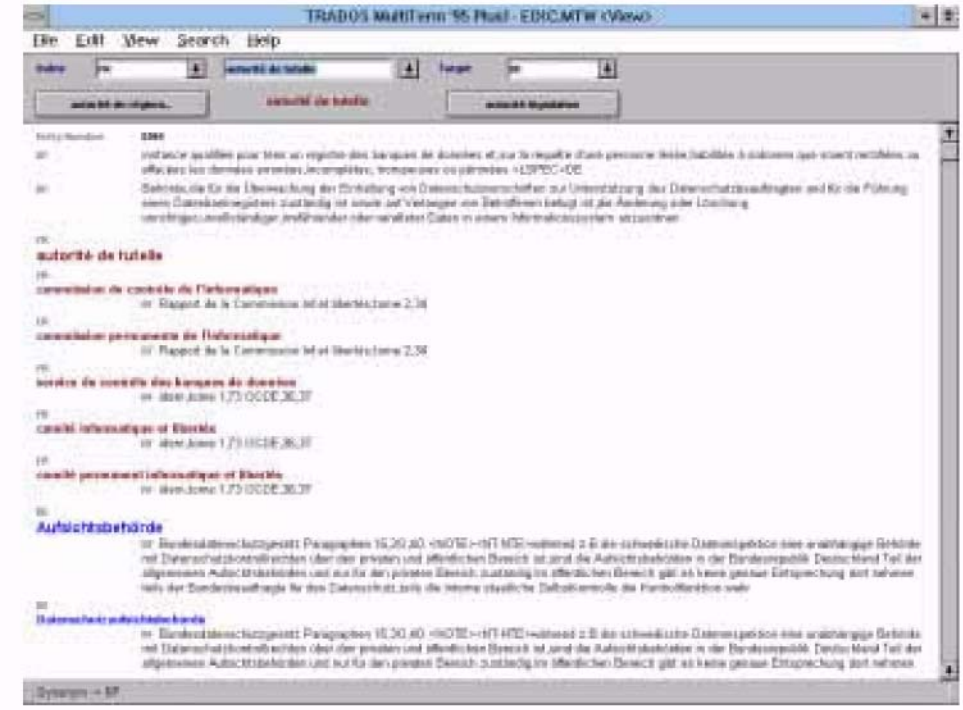
A new data model for terminological entries has been designed recently in order to ensure a smooth transition from the current EURODICAUTOM data structure and database technology to a more modern relational database management system (RDBMS) - see also the paper "EURODICAUTOM, gestion centrale de la terminologie", p.172. This approach will, among other things, offer the following advantages:

- It will be possible to define closed sets of allowed values; this means that certain typing errors will be excluded automatically, such as F1 (read F one), instead of Fl for finance;
- it will be possible to manage values: starting from an inventory of the values of a given field, it will be possible to unify the different variants with the same meaning (eg DE, dt, Ger etc for German), and thereby eventually arrive at closed sets;
- there will be a clear separation between the different synonyms which are attached to the same term: this will facilitate indexing and lookup;
- notes will be attached to individual synonyms rather than the whole entry;
- it will be relatively simple to derive from such a "development database" a very fast "distribution database" which is based on full text database technology (cf article on SdTvista).

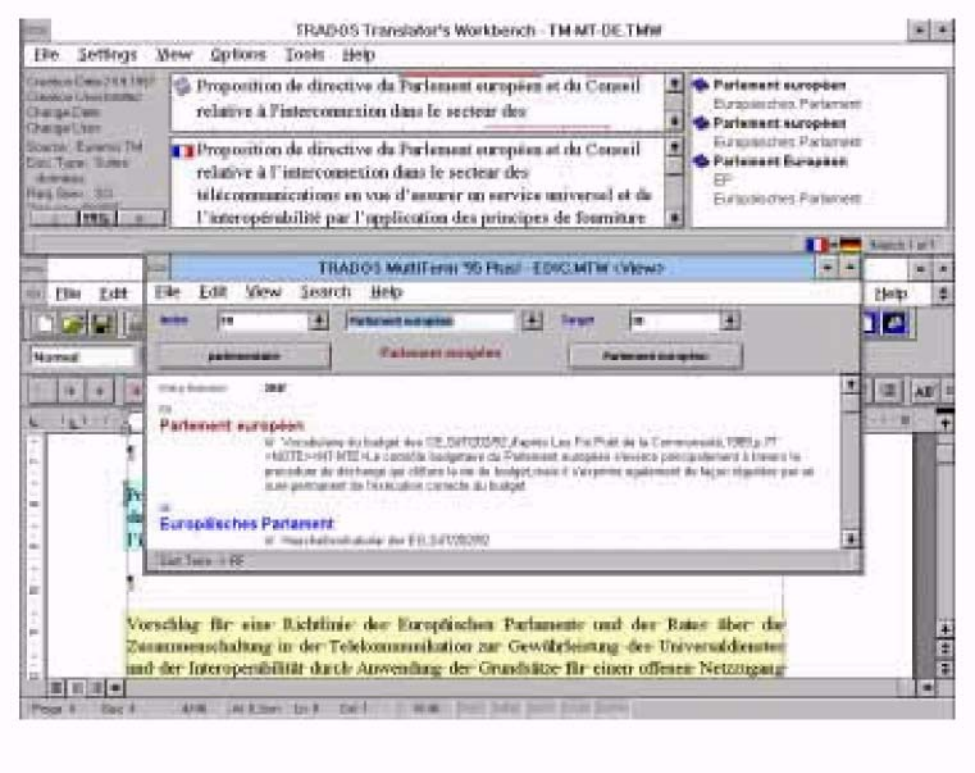
In order to be able to migrate data from the old model to the new one, a converter from native EURODICAUTOM to a neutral SGML structure has been implemented which already deals with most of the problems encountered with the old model: synonyms are separated and notes are attached to the correct synonym; a number of wrong values are mapped to their correct forms (eg F1 is replaced by FI) etc (for more details see FONTENELLE's paper, p 222).

3.3.C Combination of programs

The existing programs have been combined with the new EURODICAUTOM-to-SGML converter, a relatively simple SGML-to-MULTITERM converter has been added, and the result is a much more useful service which offers a text-related terminology retrieval in MULTITERM import format:



The advantage of this is that the terminology retrieved can now be used in the TWB environment, ie that translators who work with a translation memory are presented automatically with the terminology which has been found for the sentence they are translating at a given moment:



4 Outlook

The potential of EURAMIS as an integrator of applications has by far not yet been exhausted. Additional production chains based on existing applications as well as new developments will create new services. Example: when perfect sentence alignment is achieved for certain text types, and with some extensions to EURAMIS TM technology aiming at very fast access to alignment files, it will be possible to extract reference documents for a given document automatically, align these reference documents reliably, create *ad hoc* TMs from those alignments, run a retrieval on the *ad hoc* TMs and return the document-related results to the user. In such a scenario, a user can get a TM for a given document without even having to know where to look for it.

All the new products and services will make it more and more difficult for the average user to determine which product should actually be used in a given case. This problem can partly be solved by providing an expert system which suggests the most suitable treatment for a given text. It will

not come as a surprise to the reader if, for this purpose, existing modules are rearranged and a few adaptations are made. As far as TM treatment is concerned, a recommendation could be based on the following analyses:

- Calculate the degree of internal repetition of the source document (including fuzzy sentences): a high value favours treatment with Translator's Workbench;
- find the most pertinent server TMs and calculate the overall coverage of the source document: a high value favours TM treatment in general.

The next step would then be to integrate such an expert system upstream in the production management, so that the queries necessary for the preferred treatment can already be launched and the results can be saved into a working directory before the translation request even arrives on the translator's desk.

ACHIM BLATT
Translation Service
European Commission
Luxembourg

Bibliography

BLATT A (1996) "The EURAMIS Project" 131/134 in LAUER A / GERZYMISCH-ARBOGAST H / HALLER J / STEINER E *Übersetzungswissenschaft im Umbruch. Festschrift für Wolfram Wilss zum 70. Geburtstag* Tübingen

BLATT A / MARTINS P (1997) "EURAMIS, The European Advanced Multilingual Information System" 3/5 *The ELRA Newsletter 2,* 1

REINKE U (1997) "Integrierte Übersetzungssysteme. Betrachtungen zu Übersetzungsprozeß, Übersetzungsproduktivität, Übersetzungsqualität und Arbeitssituation" 97/106 *Lebende Sprachen* 37, 3