Generalised Language Resources: EURODICAUTOM, SYSTRAN and EURAMIS - a Case Study

Jean-Marie Leick

Abstract

One of the challenges of the EURAMIS project was to find data representation methods capable of covering the language resources of terminology databases, machine translation and translation memories. The importation of EURODICAUTOM terminology data into the SYSTRAN dictionaries gave an indication of the obstacles to be overcome, but it was the exportation of SYSTRAN dictionaries into EURAMIS which revealed the major difficulties.

The main problems arose, in the case of EURODICAUTOM, from the lack of strict encoding and, in the case of SYSTRAN, from the algorithmic representation of morphology. Translation memory entries are less problematic provided that logical representation and organic implementation are kept separate.

The advantage of common representation is that the various applications access data from different sources using the same mechanisms, thus creating unexpected synergies. The SGML-based approach for logical data description facilitates exchange in TIF-alike format.

Jean-Marie Leick

Graduated in electronic engineering, but moved onto informatics engineering in the seventies. Currently in charge of the EC-SYSTRAN machine translation project within Directorate General XIII of the European Commission. Since 1992, head of the former MLAP (Multilingual Action Plan) of the European Commission which ended in 1995. While EC-SYSTRAN development was the main project of the MLAP, the need to combine different tools (translation memories, term banks, machine translation and others) led to the design of the EURAMIS project, which has now been taken over by the Commission's Translation Service. Work is currently in a transitional phase as a new European Union Action Programme, MLIS (Multilingual Information Society) is put before the European Council.

DG XIII of the European Commission

DG XIII deals with: "Telecommunications, Information Market and Exploitation of Research". Directorate E "Information Market, Information Industry and Language Processing" is based in Luxembourg. Unit E5 (Language Processing and Applications) is mainly running the Language Engineering Programme within Telematics in the 4th Framework Programme for Research and Technological Development of the European Union.

Jean-Marie Leick
Principal Administrator, European Commission - DGXIII/E5
EUFO 0/168, L-2920 Luxembourg

Tel: +352 430134525, Fax: +352 430132354
E-mail: jean-marie.leick@lux.dg13.cec.be

Background

At the end of 1992 I took charge of the Multilingual Action Plan (MLAP) at the European Commission, which aimed at lowering language barriers by means of information technology. As the proposals came flooding in for NLP-tools or translation aids, it struck me right from the beginning that every tool, whether it be spell-checker, grammar-checker, term bank or machine translation system, had its own dictionary. While the Commission already had powerful dictionaries, both for its own version of the SYSTRAN machine translation system and for the EURODICAUTOM term bank, these resources could not be used for other tools. Another invaluable, but not easily accessible, language resource was the very large volume of parallel texts in the official languages stored in the archives of the Commission's Translation Service and in the Union's databases (e.g. CELEX). Translators need to have all this data at their fingertips. To this end, a comprehensible toolset working on a common linguistic resources database was designed in the framework of the EURAMIS project (EURopean Advanced Multilingual Information System).

This paper discusses the first steps in this direction, which were the integration of the EURODICAUTOM and SYSTRAN dictionaries and their representation within the EURAMIS database.

Achievements

The main achievements, in chronological order, are:

- Importation of EURODICAUTOM into SYSTRAN
- Design of the EURAMIS Linguistic Resources Database (LRD)
- Exportation of SYSTRAN into the LRD.

Importation of EURODICAUTOM into SYSTRAN.

The question of importing EURODICAUTOM terminology data into the SYSTRAN machine translation system had been raised at intervals since machine translation was introduced at the Commission in 1977 and the conclusion had always been - mission impossible!

The problem was the fundamental difference in approach between an MT dictionary and a terminology database. MT entries necessarily reflect the most general translation of a term in a large variety of contexts, while terminology entries offer the most precise translation possible in a very specific context. Nevertheless, the wealth of terminology in the EURODICAUTOM base was tempting, and a feasibility study was conducted with Spanish as a source language. The results were so encouraging that they sparked off a revolution.

While the authors of the feasibility study were optimistic, there were plenty of pessimists to produce counter arguments and examples to show that the results of the feasibility study were biased and at any rate would not apply to other languages.

The software specialists feared that performance would deteriorate with the addition of millions of new entries, and the resulting volume explosion on the disks. If panic reigned, this was understandable.

The Challenge and the Strategy adopted

The nature of Eurodicautom entries presented a real challenge:

- 600 000 multilingual entries
- No formalisation of entries — only guidelines
- No grammatical information (such as word class or gender)
- No morphological information.

In short, Eurodicautom entries are meant to be understood by a human reader rather than a machine.

On the positive side, the project offered:

- Corresponding entries in up to 9 languages
- Large SYSTRAN dictionaries for cross-checks and for producing language-dependent morphological tables for automatic recognition of word classes.

Hence, we had at our disposal considerable language resources which could be exploited automatically.

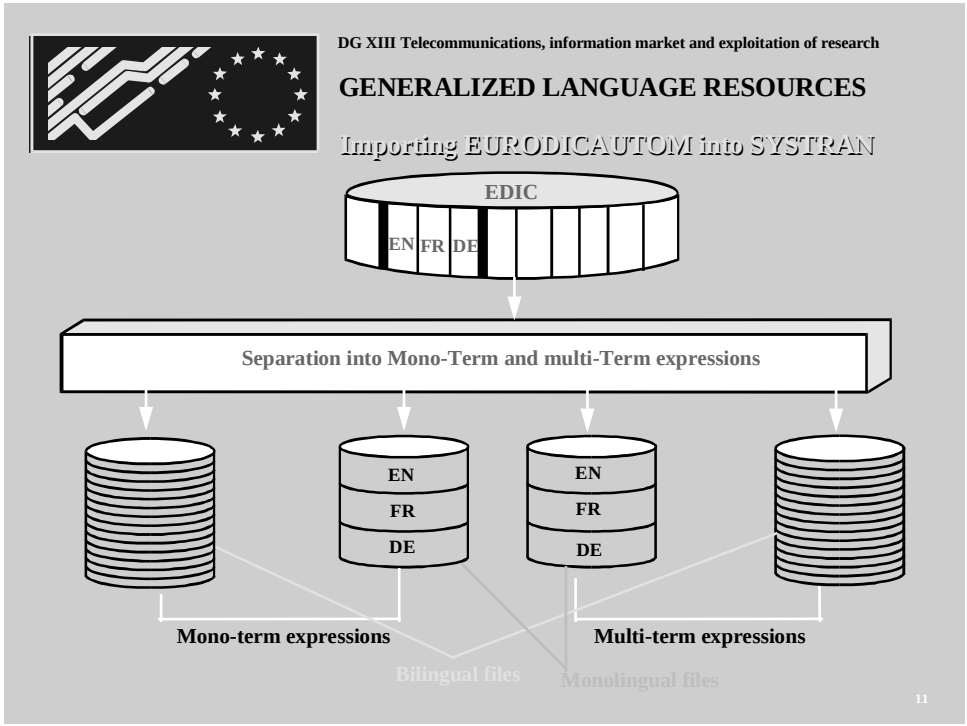
The following approach was adopted:

- Filtering of suitable entries
- Semi-automatic importation with plausibility checking at every stage
- Careful fine tuning of strategies.

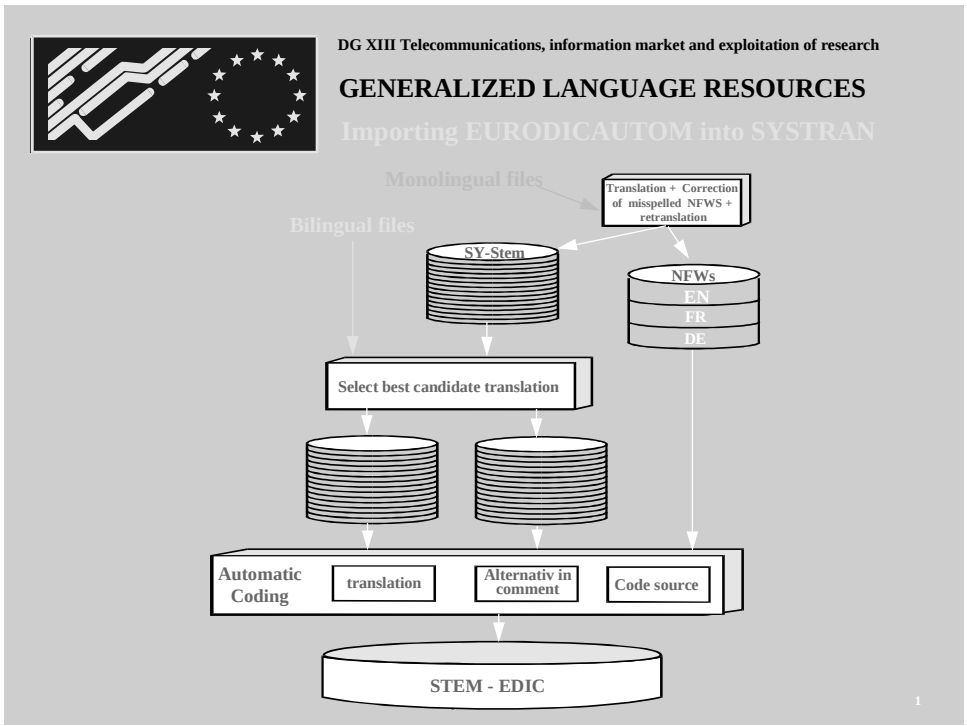
Filtering was done on noun groups. Verbs and adverbial groups were excluded.

Semi-automatic import was achieved by working out probabilities for corresponding entries in the various languages. Nouns could be identified in German (capitalised), verbs in English (preceded by "to"), other categories in Latin languages by sorting on endings, etc. The results were checked on sorted lists by the encoders.

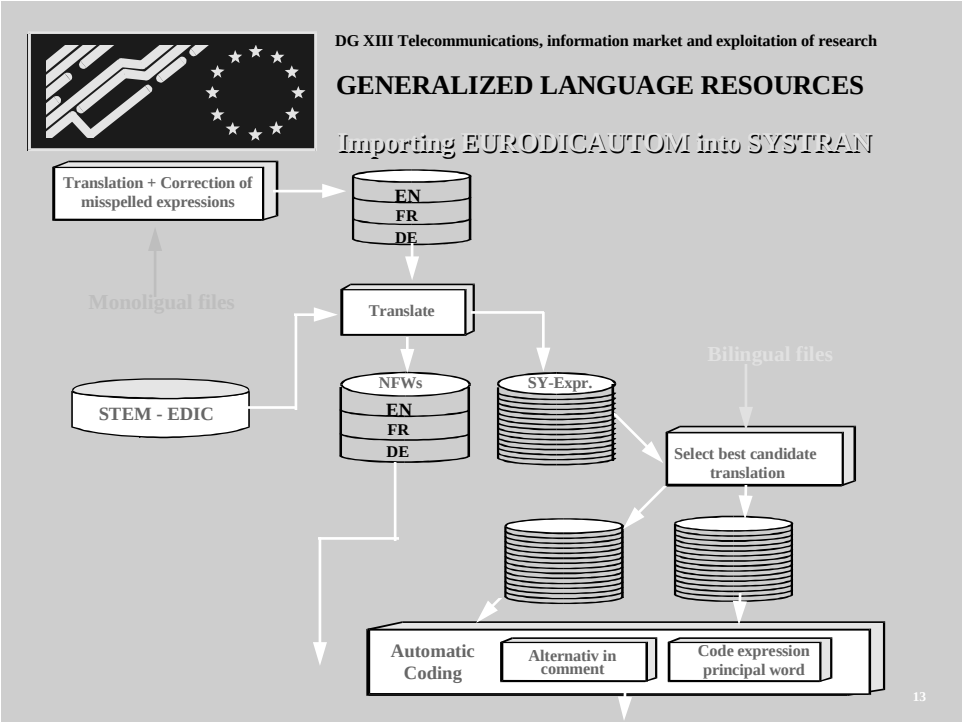
The following diagram shows the process in some detail.



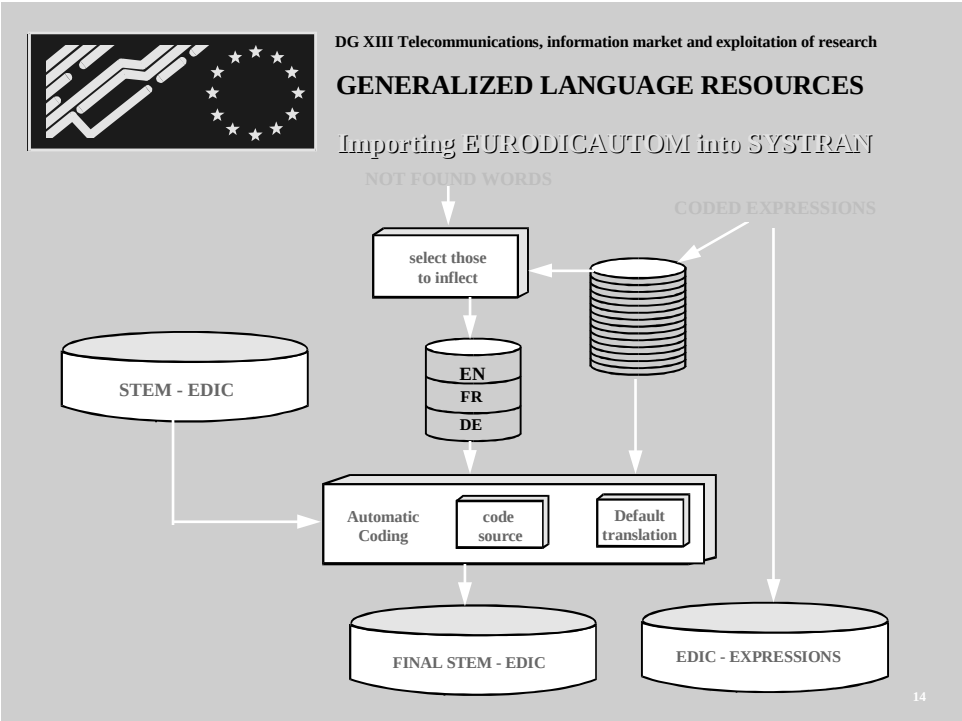
The general idea was to separate into monoterm and multiterm expressions and to deal first with monoterms.



The new monoterms were used to produce an MT dictionary, which in turn was used to translate the multiterm expressions. The selection of the best candidate translation via probabilistic methods then generated the new expressions dictionary.

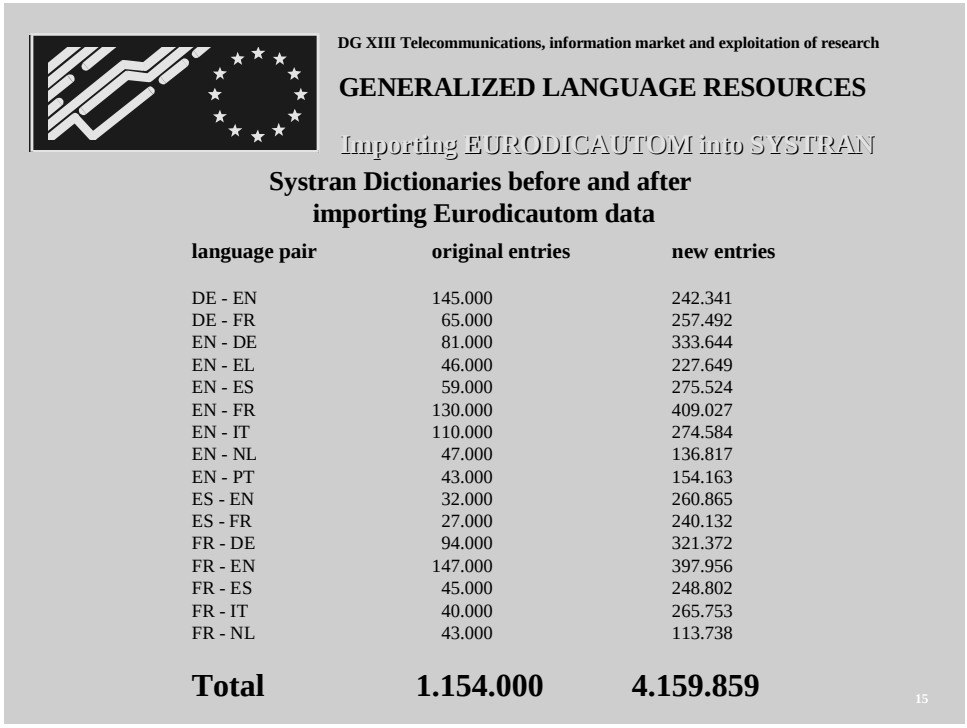


The final EDIC-STEM dictionary was generated from information relating to the principal word within the expressions.



The Results

The final result? More than 3 million entries!



The development of the whole procedure and the initial import took 180 man/months.

This exceeded initial estimates fivefold! (The original workload estimate was 34 man/months), but still resulted in a 600 entries per man day rate. The first two-yearly update only took 20 man-months for 750.000 added entries, giving a rate of 1.700 entries per man-day. It included the development of comparison functions allowing to identify and classify changes to existing entries.

Improvement of translation quality was:

- for technical documents: up to 60% or more
- for other documents: 2-10% or more, depending on the language pair and the type of document

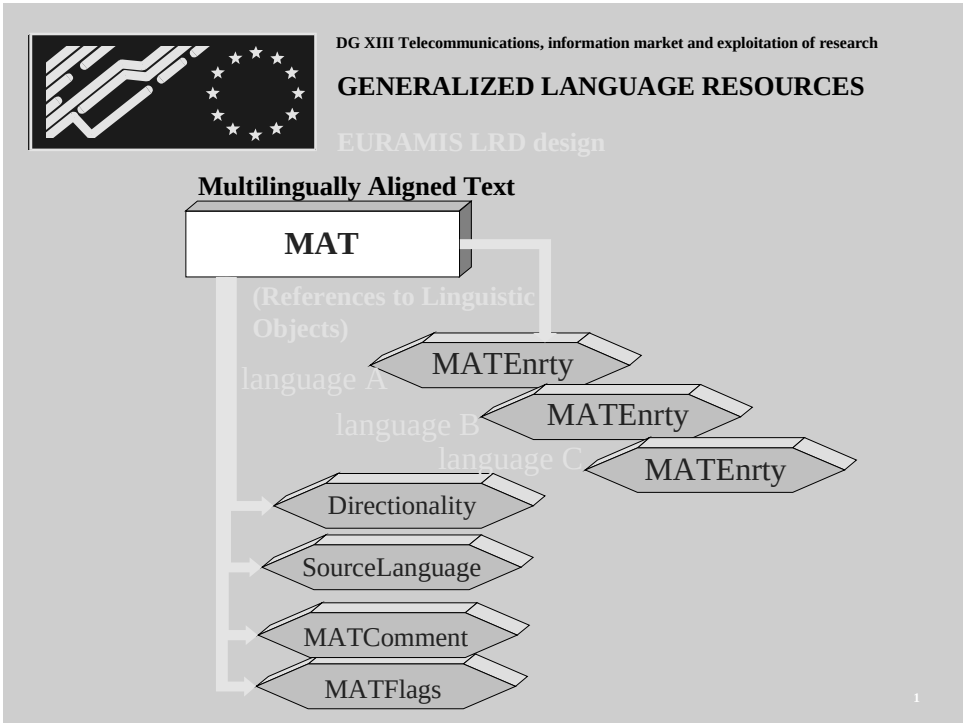
Quality enhancements are measured by translating a document twice, once with the EURODICAUTOM dictionaries and once without. An enhancement of 60% means that the number of sentences containing improvements minus the number of sentences containing deteriorations divided by the total number of sentences gives 0,6.

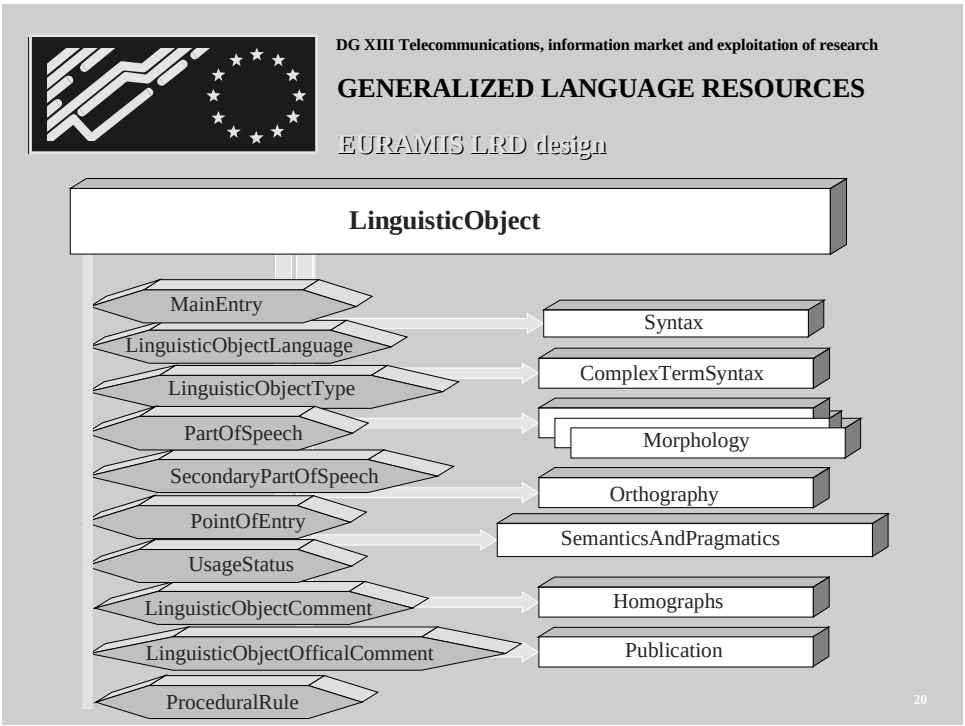
The fine tuning provided a better understanding of the Systran system and allowed to streamline coding strategies.

The EURAMIS Linguistic Resources Database (LRD)

The LRD has to cover the spectrum of multilingual entries from paragraph down to stem entries. In order to achieve this the multilingual aligned Text (MAT) - concept was defined as the generalized LRD-entry. Every MAT consists of corresponding

monolingual entries referring to linguistic objects with attribute-value pairs describing the details (see diagrams).





The structure is described by an SGML construct resulting in an SGML-database.

Exporting SYSTRAN dictionaries into EURAMIS-LRD.

After the design and initial implementation of the EURAMIS LRD stage, the natural progression was to consider the feasibility of exporting MT dictionaries into the LRD.

Human nature being as it is, there was again strong opposition. Of course the problems were very real:

- Contextual rules in the Systran translation engine
- Morphology, mainly for verbs (paradigms), is defined in Systran in order to optimise retrieval and is dependent on language
- Copyright considerations resulted in proprietary and public data sets
- The data volume explosion by the SGML tags was dramatic
- Character sets in Systran are quite antiquated.

The TIF-DTD was taken into consideration and the LRD-DTD adapted to contain SYSTRAN specific attributes and secrecy provisions. The export procedure from MT was kept separate from the import procedure into the LRD in order to take advantage of linguistic knowledge build-up specific to the LRD and to be able to represent morphology, for example, as an independent factor. Tag compression helped solve the problem of data volume explosion created by SGML-tagging, although the total volume is still estimated at some 20 GB of data for 5 million entries.

The main work was the definition of the trade-off between representing simple linguistic rules encrypted in the Systran specific coding by standard clauses understandable by linguists. This caused longish discussions and trials in order to isolate specific problems.

The import function for EURAMIS is still under construction. The fine tuning will have to be done after completion of the complete cycle.

Conclusions

The main benefits of this exercise have been the creation of synergies between developers of machine translation and terminology on one side and combined functions on the other:

- Scanning all concerned Eurodicautom entries allowed the correction of entries (spelling-errors, language clean-up, formal variants contradicting the guidelines a.s.o.)
- A combined function provides terminology look-up from text in combining the Systran retrieval with Eurodicautom lookup even for target languages not available in Systran.
- In using the domain identifiers in terminology entries a content identification function was developed allowing to prescan a document to identify by statistical methods the domain it belongs to.
- The use of Eurodicautom entries in Systran translations is of immediate use in machine translation especially for long expressions and lesser developed language pairs.

- In the future it is planned to use Euramis for entering terminology, so that the MT entries will be available to terminologists.

Some of the new products, such as terminology look-up from text, are available today. Others will be available in the EURAMIS system, with the first prototypes currently undergoing tests. We have already achieved a great deal, but there is still more to come!