[*Terminologie et Traduction* 1.1998, pp.12-26]

JEAN-MARIE LEICK

EURAMIS - the ultimate multilingual blackbox?

1 Background

One of the fundamentals of the European unification is the mutual respect of cultural diversity between Member States, of which language is a most prominent representative. At the same time, with the Babel-tower story in mind, it is also a matter of concern about its feasibility and price. The number of combinations between languages raises with the square of the number of languages and the institutional translation services employ already more than 2 000 translators. So in view of the foreseeable increase of the number of official working languages in a near future it is worthwhile to investigate how information technology may alleviate the burden of multilingualism. This was already the rationale of the first "Action plan for the enhancement of information transfer between languages" of the Commission (DG XIII and SdT) back in 1977, the most visible result of which is a quarter of a million machine translated pages in 1997 by Commission officials, mostly non-translators.

When I took over responsibility in this action plan (commonly called MLAP, for Multilingual Action Plan) in 1992 I was struck by the richness of available multilingual resources on the one hand and the lacking information technology aids to access them on the other. Why IT was not doing what it can best: quickly find specific items in boundless masses of information items. Machine translation dictionaries had grown to 700 000 entries - available only to machine translation programs. EURODICAUTOM, with its 3 000 000 terminology entries, was only available through its consultation interface. The human translations archive with tens of thousands of documents in numerous languages was only available to IT freaks knowing how to scan-search the archives. CELEX, the most prominent repository of legal community texts in all official languages, was (and still is) organised into separate monolingual databases. And what about the translation solutions that every translator finds day after

day and which could save a lot of work for his peers, if only there was a means of sharing experience?

On the tools development and procurement side the situation was equally unsatisfactory, every development team concentrating on its problems, consultation of other's solutions being left to personal initiative. Also every new tool on the market came with its own dictionary, mostly not adapted to our language sub-universe - or with an empty memory to be filled by the user (eg translation memories).

So the idea of dumping all tools and resources into a big melting pot and to design a magic multilingual wizard coping for what is needed in whatever multilingual task, be it for translators, interpreters, terminologists or end-users, quickly became a LEICK-motif.

2 The first steps towards EURAMIS

A first move was the semi-automatic import of EURODICAUTOM terminology data into SYSTRAN machine translation dictionaries. These were increased fivefold, which had a dramatic influence (only) on very technical texts, made possible the use of SYSTRAN as lemmatiser for an EURODICAUTOM search engine and opened other opportunities still not fully exploited today. But it showed also, that a case by case approach with the complications of every tool's proprietary formats was cumbersome and inefficient.

In discussions with the colleagues in the Translation Service a general consensus on the desirability of a generalised "(inter-)institutional multilingual memory" with an integrated toolset for all multilingual needs was reached. But wasn't it a little bit lunatic? How could such a chimera become a real live project?

Being a public administration, and, even more in the framework of a DG XIII action plan, we had to proceed by call for tenders for the implementation. So I called in a working group with SdT's wise men to formulate technical specifications for a 3-year project. A contest was organised in order to find a name, the most important thing you have to do when launching an otherwise impossible project. H. PAESMANS got a lot of proposals, of which EURAMIS was retained, standing for EURopean Advanced Multilingual Information System. K VAN DER HORST came up with the idea that the treatment of a document should be like a train, with every tool attaching a specific wagon to it, the destinator then being able to look into every wagon in order to compile the right output for his purpose.

Together with the outcome of Esprit projects like Genelex and Multilex in mind, the technical specifications were finalised in August 1994. They were precise enough to describe the desired result, and generic enough not to pre-empt the details of implementation. It was just in time to proceed for a call for tender on the 1994 budget, and the proposal passed the CCAM with an unsurpassed 18 Mecu budget envelop approval for three years. This did not at all match DG Xlll's policy in the matter, so that at the end of the day the allocated budget reached only half of this amount, mainly coping for the machine translation part sunk into the overall project. After an initial financing of 1,5 Mecu by DG XIII B-part budget the specific Euramis project was taken over by SdT and financing continued on a much lower basis.

3 Design fundamentals

The first year of the project was essentially devoted to design. SGML (Standard Generalised Markup Language) was elected the formal description language, be it for data-structure description or for exchange (pivot) format description. The system was globally seen as a learning system, where existing resources in form of translated texts, terminology and dictionaries were the assets to start from and where language professionals could easily add new findings, first for their own needs, later for a larger audience, after validation. The injection of every new translated text would automatically keep turning the learning mechanism, the evaluation algorithm for optimised retrieval today still remaining in the pragmatics area.

Main design issues were the following:

- scaleable resources description,
- pivot format for document representation,
- inter tool communication scheme,
- scope management
- user profiling,
- multilingual service provision infrastructure.

3.1 *Scaleable resources description*

The first problem was how to describe entities like a pronoun with its translations in different languages and use the same formalism for a whole

document in different languages. T CARRASCO'S "Dragoman" MAT, standing for "Multilingual Aligned Text", was retained as being scaleable to the desired extend. The hierarchy of attribute-value pairs underneath this main concept allowed to further specify properties of the linguistic objects to be represented, the whole being described by a SGML formalism defined in an appropriate Document Type Definition (DTD). The description was convenient, but the transposal of the concepts into relational database schemata with efficient retrieval behaviour proved more difficult.

3.2 *Pivot format for document representation*

One of the major practical problems experienced in the past with machine translation was the proprietary text formats of the text processors. For the needs of EURAMIS a pivot format was defined that reduced the problem to the needed essentials: linguistic content had to be separated from presentation, except where presentation had some semantic meaning, like header, enumeration, emphasis etc. All proprietary presentation specialities are encapsulated allowing their reinsertion after treatment. If this approach reduced the impact of the format war between text processing producers, it did not prevent to cause considerable delays in trying to follow the irresponsibly undocumented format incompatibilities amongst successive versions of the same text processor imposed on the user community by the market leader in the domain. As pivot format HTML, the native Internet page format, was taken as a starting point. A few extensions to HTML had to be added. First the 2-byte Unicode was chosen as underlying character coding scheme in order to avoid from the beginning the calamities IT has experienced with European accented characters, not to speak about Greek. Then some encapsulating tags had to be added to vehiculate specific presentation characteristics.

3.3 *Inter tool communication scheme*

Tools have to communicate between themselves in a well defined format. The end-user interface has to interpret this format for profiling issues. Profiling is needed because a language professional would most probably have other needs than a secretary in an operational service. The train idea, with the tools attaching their specific results as "wagons" to the original document, was realised via application specific tags in the pivot format. In this way the original text could be completed by its machine translations

into different languages, the results of the translation memory retrievals, the text specific glossary processed by the EURODICAUTOM search engine and so on, depending on the request. The secretary having requested only an automatic translation will get the sentences having had a perfect match in translation memory merged with those coming from machine translation, all of those presented in a format as close as possible to the original by a simple filtering process in the user interface.

3.4 *Scope management*

The delimitation of scope of MATs is very practical in order to allow users to quickly introduce into their private domain translations directly useful for their current work, without hampering the overall learning process of the system. In a learning process any new information must be weighted against existing information in order not to learn rubbish. Especially in linguistics this is a painful exercise, as the human perception has a tendency to unconsciously limit the meaning of a word to the context it was found in. So general validation of new entries might be a longish verification process, whereas the scope limitation allows to directly benefit from non-generalised findings. In modulating the scope definitions to workgroups, units, DGs etc, the generalisation process may be staged, so that in the course of the migration of a concept to general use the scope definition may be more specifically expressed as a specialised field, text type, sublanguage, expression or idiom. This process is conceived as "harvesting", in which the organisation of the evaluation and decision body seems to be the more difficult part.

The EURAMIS project was from the beginning aimed at a large audience. Every official in the European Institutions is more or less often confronted with multilingual problems. His first reaction is to call for help of a colleague fluent in the requested language, the results often not being in line with what SdT's professionals would have proposed. So in divulging Euramis, be it only as enhanced machine translation, the SdT could influence positively the multilingual capacities of the operational services and guide them to confirmed translations. But the efforts of the Translation Service were more often directed to the immediate needs of translators, probably because being more rewarding and not so frightening as the perspective of 15 000 potential moaners. As member of DG XIII I always had difficulties to try to get these diverging perceptions into one approach. The user profiling issue in the EURAMIS project, established customisable standard profiles for translators, terminologists and developers, but also

end-users! The profiles provided for different presentations of appropriate functions in a unique parameterised user interface. The irony of fate is that this brilliant perception will most probably be obsolete by the advent of intranet techniques before EURAMIS will be made available to a wider user group.

3.5 *Multilingual service provision infrastructure*

A central multilingual service provider was originally designed as a value added e-mail server. This allowed to circumvent the difficulties of organising and supporting a large direct access application distributed amongst 35 DGs or assimilates, not to mention extra-Commission users. As the ins and outs of the services were most of the time pure documents and the response times in tens of minutes did not seem shocking, the successful scheme of the translating mailboxes for machine translation was extended. In order to provide more complex multi-function services a function dispatcher was established allowing to launch remote applications and combine the results into a common response. This enables future integration of tools available on the market, provided they are useable without human intervention (batch operateable). Today the e-mail concept might favourably be replaced by an intranet solution presenting the same advantages but not the longish response times of e-mail.

4 Status

The EURAMIS project was the last bounce of the 17 year old MLAP, of which I had the honour of being the pall-bearer in DG XIII. The transfer of responsibility from DG XIII to SdT for EURAMIS took place in 1996, for SYSTRAN as of 1st January 1998. The language relevant activities of Directorate E of DG XIII are now concentrating on shared cost projects with Member States within the MLIS (Multilingual Information Society) programme and within the 4th Framework Programme for RTD in the Language Engineering special task group.

The status of the project is presented in a contribution of ACHIM BLATT, the current project leader. A lot of progress was registered, with unsurpassed alignment and translation memory retrieval results. User profiling, format conversion, efficient retrieval were so many headtwisters, for which a solution had to be found. Yet the overall progress did not meet the original enthusiastic expectations due to a number of adverse

phenomena: personnel fluctuations with the contractor, withdrawal of DG XIII financing, serious performance lacks in the current implementation, among others, seem to prevent the project to get a critical mass. At the same time other ad-hoc, but well working, applications emerge outside the EURAMIS framework and cover some of the needs originally aimed at by EURAMIS. Outside the Translation Service, in the Commission's services, EURAMIS is still completely unknown, even in its simplest expression.

5 Conclusion

The original idea is still pending full operational materialisation, but a number of publications and presentations have certainly contributed to the fact, that CAT-products on the market increasingly combine different linguistic tools into one workbench. These products presently do not nearly master volumes and complexity of existing Commission content, but this might only be a question of time. So one day EURAMIS might strike back - under the name of Microsoft-Multilingua, the ultimate multilingual blackbox - enterprise version, in which case the welltried paradigm of know-how against product exchange between continents would get another confirmation.

JEAN-MARIE LEICK