[From: *Translating and the Computer 14.* Papers presented at a conference… 10-11 November 1992 (London: Aslib 1992)]

EUROLANG PROJECT (EUROPEAN LANGUAGES)

J-J. Pérot
SITE - EUROLANG
BP 35, 94701 Maisons Alfort CEDEX, France

Abstract

The main aim of the EUROLANG project is to develop second-generation machine-aided translation products in three years covering the following language pairs:

French ⇔ English, German ⇔ English, French ⇔ German, Spanish ⇔ English, Italian ⇔ English.

In order to achieve this aim, some of the most renowned European industrialists and academics have decided to pool their technical, financial and human resources in order to define and develop a new "high quality, low cost" system based on state-of-the-art computing and linguistic techniques.

Background information

The EUROLANG project will cost approximately five hundred million French francs (90 million dollars).

The project started in November 1991, and will be completed at the end of 1994, at which time the EUROLANG products will be marketed.

By "machine-aided translation product", we mean a complete translation environment on a PC or workstation, in other words, a translation engine plus a set of tools upstream and downstream of the engine for pre-editing and post-editing, interactive terminology management, document text handling, and so on ... everything the translator or multilingual engineer needs to increase his productivity.

The system will process five European languages (English, French, German, Italian and Spanish). Only ten language pairs - chosen according to the market and partners' need - will be dealt with: the eight pairs involving the English language and the two French/ German pairs. The dictionaries will contain 50,000 general terms in each language and their equivalents in the other languages. Specialized terminological dictionaries will be developed by the partners and members of the EUROLANG Users Club for specific domains (aeronautics, data processing, etc.).

By the year 1995, these products should provide serious competition for Japanese products which, since 1982, have received considerable financial backing from the Japanese government and industrialists, as part of a medium- and long-term strategy aimed at the control of information technology.

The EUROLANG project will build upon the substantial investments made by the European authorities and industry over the last twenty years.

Translation market

The world translation market is a huge market estimated at 12 billion dollars.

As regards the traditional translation department, there

*. R&D Department, SITE - EUROLANG

is a great divergence between the real demand, the financial capacity of industry and the production capacity of the translation centers for a number of well-known reasons:

- The ever increasing cost of the translated page (currently between 45 and 90 dollars), which makes it impossible for any small company to undertake the translation of a large amount of documentation.
- The shortage of qualified translators, linked to the natural limitations on translator training opportunities.
- The low productivity of the traditional translation service (about one page per hour per translator).
- The strategic need for a company involved in a competitive export market to rapidly obtain high quality translations of technical and commercial documents.

In view of this situation, analysts estimate that only *5* to 10% of the demand is currently met.

There is no doubt that this problem can only be solved by the use of specialized software tools, designed to increase the productivity of translators.

First generation systems, based on the crude technologies developed in the 1960s, were unable to overcome this problem, judging from the turnover (between two and four million dollars), which serves to illustrate the disappointment of both industry and translators.

EUROLANG partnership

The starting point was the assessment by our translators of translations produced by the French ARIANE second-generation MT system on texts taken from aeronautical documents we had translated. Those texts were maintenance manuals and job cards. They were chosen because they were available in machine readable form and also because the bilingual terminology bank specific to the subject had been built up (over 4000 terms in French and English). The results over more than 500 pages were good enough to reach a post-editing time of only twenty minutes per page to get a quality equivalent to that of human translation.

Unfortunately, this system is run on mainframe and the cost of such machine translation is currently so high that the gain obtained by reducing the translators' work is lost when the costs of the CPU used are added to the human cost.

At the same time SIEMENS NIXDORF which is developing, maintaining and using the METAL MT system, was seeking to improve its system and interested in the definition of a common European NLP platform.

On these bases the SITE and SIEMENS NIXDORF Groups decided that it was necessary to develop a new MT system based on a considerably improved ARIANE and METAL technology, considering the advanced state of current computer technology and the evolution of linguistics. Technical choices are thus being made bearing in mind industrial needs, i.e. portability, maintainability, openness, possibility of evolution and ergonomy.

A number of major industrialists (KRUPP INDUSTRIES, MATRA MARCONI SPACE, CAP GEMINI, RANK XEROX, THAMUS-LEXICON, etc.) have also agreed to share their knowhow and resources.

The consortium has decided to integrate some of the most highly skilled European academics, notably the teams which were involved in the EUROTRA project, to benefit from their expertise and to prepare for the industrialization of new technologies for the next decade. Few people realize that in this highly sophisticated domain at least ten years is required for the development of a prototype fit for industrialization.

Building upon several thousand man-years of R&D and industrial development, we will be able - as early as the end of 1992 - to demonstrate a prototype on a workstation covering at least two languages, which will prove the feasibility of future developments as well as the ergonomics and the productivity of the products.

Technical aspects

As already mentioned, the main objective is to provide a powerful toolbox, containing tools dedicated to linguistic developments. One of the most important characteristics of such a toolbox is the implied reusability of its components. A plug-and-play strategy will thus enable the linguists to develop different kinds of applications using the existing "components" of the toolbox. A "lingware workbench" will provide all the facilities required to specify, implement, test and maintain these applications. To facilitate the communication between the tools and external systems and thus enable the plug-and-play strategy, an API (Application Programming Interface) will be defined.

The toolbox will be designed in such a way that new tools can easily be added, ensuring its durability. Any evolution of the "state of the art" in computational linguistics can thus be rapidly taken into account in the EUROLANG product.

Most of the initial tools will be specialized languages, allowing the developer (or linguist) to handle concepts he is used to. Such linguistic languages will consist of 4GLs (4th Generation Languages, i.e. specialized programming languages adapted to specific developments) and the associated compilers and interpreters. This architecture ensures a better independence of the lingware and software (for instance, the pattern matching mechanism is part of the software and should not be programmed by the linguist), and consequently a better linguistic modularity.

Lexical and textual data bases are also needed in the toolbox, to enable an easy management of the lexical and textual resources. The lexical data base will provide a user-friendly interface to add or modify terms. A flexible underlying model allows modification of the linguistic model, and thus modification of the linguistic information needed in the dictionaries.

Representation of texts and characters in a multilingual environment is a crucial issue. Although work has already been undertaken to solve this problem, no general standard exists as yet and an external and an internal representation should be designed, taking into account any standard or recommendation (e.g. Text Encoding Initiative recommendation).

A general exchange format, based on SGML (Standard Generalized Markup Language), will thus be defined for both lexical and textual data. It will guarantee the openness of the system by allowing the reusability of the lingware.

The final MT environment will provide a user-friendly translator's workstation. Two kinds of functionalities are foreseen : pre-editing and post-editing functionalities. Pre-editing functionalities will comprise conventional tools (e.g. spelling and grammar checkers) and enhanced functionalities (e.g. tools to handle new words and predict their linguistic behaviour). Post-editing functionalities will comprise functionalities needed by any translator (even to translate ab initio) and functionalities specialised for MT revision. Among all the foreseen functionalities, the following are worth underlining: direct access to dictionaries, management of successive annotations, intelligent search and replace manipulations, easy access to alternative translations offered by the MT system, request for information concerning the MT system, translation project management and specific word processing functions.

To ensure that the system is portable, developments will be made in C or C++ portable language (ANSI), under UNIX. Graphics will be produced under X-WINDOW/MOTIF. The standards currently in force will be respected (SQL, SGML, etc.). Although UNIX has been chosen for the developments during the project, the PC world (with WINDOWS) is one of our future objectives.

"High quality, low cost"

The computing power of PCs and workstations has drastically increased, allowing the implementation of sophisticated linguistic and computing technologies. This means that the products proposed by the consortium will target the "high quality, low cost" market, with an average return on investment of 18 dollars per translated page, that is, for industrialists, a return on investment in less than one year.

The system is aimed at the market for quality translation with post-editing of technical and commercial documentation, multilingual querying of information systems and personal MT.

Conclusion

In summary, the EUROLANG project can be viewed as a response to the following challenges:

- We have to convince industry and translators that quality products could increase productivity.
- EUROLANG should give the European business community the command of multilingual technical and commercial communication which is crucial for European integration and should surpass the Japanese competition in this strategic field.
- More generally, we will contribute to European efforts to control information technology. This is of strategic importance, given the great labour income which is related, directly or indirectly, to the management of technical information and that there is a high probability that this trend will continue with an ever increasing automation of the production process.