PRESENTATION OF THE EUROLANG PROJECT

B.SEITE, D.BACHUT, D.MARET, B.ROUDAUD
SITE
12 rue de Reims - 94700 Maisons-Alfort - FRANCE
tel : +33 16 1 43 96 72 00
e-mail : D_Bachut@site-maisons-alfort.fr

ABSTRACT

International trade in general and particularly the Single European Market will bring about a considerable increase in the already huge documentation market. NLP products will contribute to improving the competitiveness of industry in this strategic field.

The industrial objective of EUROLANG is thus to provide efficient NLP tools and to give the European business community a better opportunity to maintain a command of multilingual technical and commercial communication.

The technical objective of the EUROLANG project is to build an MT/NLP toolbox, offering a wide range of 'open' and powerful tools, which reflect the state-of-the-art in computing and linguistic techniques. These tools will then be validated via the production of a multilingual MT system based on second generation principles.

With respect to the computing aspects, the main technical choices are : portability, maintainability, openness, possibility of evolution and ergonomy. This implies the use of standard techniques and tools (UNIX, X11, MOTIF, C, SQL, SGML).

The linguistic developments, based on a syntactico-semantic analysis, will follow an industrial methodology, implying formal linguistic specifications. The use of specialised languages will ensure a better separation between software and lingware, and thus better modularity. The three-phase translation process guarantees the multilingual aspect of the MT system.

INTRODUCTION

In the light of the communication society, which is leading industrial companies to become involved in NLP technology, the EUROLANG partners have decided to pool their technical, human and commercial resources in order to define and develop a new 'high quality' / 'low cost' MT system. This system will be based on state-of-the-art computing and linguistic techniques, and is aimed at the market for quality translation with post-editing.

To ensure the reusability of the linguistic resources and the possibility of evolution of the system, a powerful toolbox, providing different NLP tools and useful 'plug and play' functionalities, will be developed. This toolbox will be a linguistic platform providing a user-friendly environment to design and develop different kinds of NLP applications.

In the following paper, we present more specifically the aims and motivations of the project and the main technical choices, with regard to both computing and linguistic issues.

GENERAL PRESENTATION

Aims and motivations of EUROLANG

The technical objective of the project is to build an MT/NLP toolbox, offering a wide range of 'open' and powerful tools, which reflect the state-of-the-art in computing and linguistic techniques. These tools will then be validated via the production of a multilingual MT system based on second generation principles. This system will process five European languages (English, French, German, Italian, Spanish). Only ten language pairs (chosen according to the market and partner needs) will be dealt with : the eight pairs involving English language, and the two French/German pairs. The dictionaries will contain 50 000 terms in each language, and their equivalents in the other languages.

The general objective is to yield products at the end of the project, which will be targeted on the 'low cost' / 'high quality' market. This goal is made possible by the 1991 state-of-the-art MT technology and will considerably increase the share of the market currently occupied by the commercial MT systems.

The market is estimated at 12 billion dollars (Gartner Group figures), and is rapidly increasing. The following trends suggest a boom in this market :

- continuous increase in international trade,
- enormous need within the Single European Market,
- shortage and ever increasing cost of qualified translators,

- strategic need for a company involved in a competitive export market to possess high quality translations of technical and commercial documents rapidly available.

The Japanese were the first to appreciate the strategic importance of substantial investment in this field. The major corporations invest considerable sums and participate in projects such as EDR, ATR... The state-of-the-art in computing and linguistic technologies associated with skills to deal with different European languages seems to be a good opportunity for Europe to become a leader in this sector.

The linguistic quality of a translation is of course one of the major criteria involved in the evaluating of an MT system, however there are others (cf. [Roudaud 91]). Some linguistic phenomena are very complex, and thus very costly, to deal with, whereas another technical solution, less resource-consuming, could lead to an equivalent (or even better) result, as far as efficiency is concerned. The industrial issue implies that a compromise must be reached, between cost, efficiency and linguistic quality : the right solution to the right problem. For example, anaphora are a very difficult linguistic problem to solve, whereas in technical documentation the French pronoun *il* can more often be translated in English by *it*. In such a case, a better solution would be to propose *it*, as the default translation, and *he* and *she* as alternatives, so that the revisor can choose one or the other.

This last point shows that we must adopt a global approach for the design of an MT system. Not only is the MT kernel of major importance, but also a user-friendly environment, providing many useful tools for the end-user (translator, writer,...), must be designed. This is the reason why special attention is paid in this project to the design of the pre- and post-editing tools. For users, an MT system is only a part of the documentary system and integration and connection with other tools must be foreseen.

It is well-known that second generation MT systems are very time-consuming. Predictions for the coming years show that workstations will very rapidly reach 50 or 100 MIPS, and that the cost of MIPS will continue to decrease. All these factors contribute to the reduction of exploitation costs of MT systems and to the improvement of delivery time/volume ratio.

Organisation and main technical choices

EUROLANG is a three-year EUREKA project which began in December '91. The first year will mainly consist of specifying both computer and linguistic developments. The two following years will be devoted to development, integration, tests and evaluation.

The global cost of the project is 684 MFF (about 140 M US$). It involves five European countries : France, Germany, Italy, Spain, United Kingdom. The partners of the project are :

- in France :
  - SITE group (prime contractor)
  - CAP GEMINI INNOVATION
  - CNET
  - GETA
  - LADL
  - MATRA SPACE MARCONI
- in Germany :
  - SIEMENS NIXDORF
  - KRUPP Industries
  - IAI Saarbrücken
- in United Kingdom :
  - RANK XEROX Ltd.
  - UMIST (Manchester)
  - University of Essex
- in Spain :
  - BDE
  - Universidad de Barcelona
  - Universidad autonoma de Barcelona
- in Italy :
  - LEXICON
  - THAMUS
  - Università di Salerno
  - Università di Pisa
  - Università di Torino

Most of the partners participate in projects in the field of MT/NLP (EUROTRA, ESPRIT...) and have practical experience. The other partners are industrialists with needs in this area, and they will be very active in the definition of the end-user stations (pre/post-editing...) and in the evaluation of the products.

The project is co-managed by SITE and SIEMENS.

SITE, prime contractor of the project, has the twofold competence : MT/NLP, in particular via its subsidiary B'VITAL, and industrial documentation management, with respect to writing as well as translation. SITE's translators have validated the quality of the translations produced using the ARIANE MT system and SITE is thus convinced that this technology can be used profitably in an industrial environment (cf. [Bachut 91a]). Unfortunately, the cost of such machine translation is currently so high, that the gain obtained by reducing the translators' work is lost when the cost of the CPU used is added to the human cost. Furthermore, the linguistic quality of the product is not a sufficient condition for improving the translators' efficiency : the translator workstation should be designed to fit the necessary ergonomic requirements.

SIEMENS NIXDORF, which is developing, maintaining, using (SIEMENS' translators use METAL) and commercialising METAL MT system (cf. [Slocum 83], [Schneider 91]), is particularly interested by the

definition of a common european NLP platform and wishes to improve METAL technology. This is the reason why SIEMENS NIXDORF decided to play an active role in the EUROLANG project. Its commercial experience is one major advantage in the commercial perspective of EUROLANG.

On these bases, SITE and SIEMENS NIXDORF 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 the industrial needs : portability, maintainability, openness, possibility of evolution and ergonomy.

TECHNICAL ASPECTS

Computer 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 with 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 independance of the lingware and the 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 is necessary to allow modification of the linguistic model, and thus modification of the linguistic information needed in the dictionnaries.

Representation of texts and characters in a multilingual environment is a crucial issue. Although works have 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 forseen : pre-editing and post-editing functionalities. Pre-editing functionalities will comprise conventional tools (e.g. spelling checker) 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 forseen 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, and other 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.

Linguistic aspects

The first application of the toolbox will be a multilingual MT system, based on second generation technology :

- the use by linguists of specialised languages, ensuring a better separation between the lingware and the software,
- a three-phase translation : analysis, transfer and generation, ensuring a better multilingual approach.

The underlying linguistic theory is based on a syntactico-semantic analysis, giving a deep representation of the text in an annotated tree structure (in which each node is 'annotated' by a set of linguistic features). The main tools used in performing such an analysis will be a slightly contextual parser, based on METAL parser, and a ROBRA-like tree transducer

(ROBRA is the tree transducer designed by GETA in the ARIANE MT system, cf. [Boitet 82], [Boitet 86]). The transfer phase makes it possible to translate words in context and the generation phase allows the linguist to specify the surface structure of the text, depending on the deep structure calculated.

Linguistic development methodology, already used by B'VITAL and SITE linguistic teams, implies formal linguistic specifications to describe the desired deep structure. These specifications will be perfomed using a specialised 4GL inspired by the GETA's static grammar formalism (cf. [Vauquois 85]).

Common linguistic interface structures are being defined (in the first project phase) to facilitate the plug and play mecanism between different linguistic components. These linguistic interfaces will consist of the definition of the minimal requirements which should be followed by the linguistic speficications of all the involved languages. This will also ensure a better multilingualism and make it possible to reduce the transfer phase between two languages.

Given that the MT product is designed for use in industry, a certain number of characteristics are essential to the final system :

- the system should always provide at least one translation,
- when several translations are possible, the presentation of the different solutions to the revisor should be user-friendly,
- unpredicted phenomena or new words should not block the whole translation process (robustness).

CONCLUSION

Considering the industrial objective of the project, EUROLANG will provide not only an efficient European MT system, to compete with Japanese and other MT systems abroad, but also an unequalled NLP platform for large scale NLP application developments.

This industrial goal will not prevent all linguistic and computing developments from being based on the current state-of-the-art technology. To ensure the durability of the toolbox, an R&D stratum will prepare for future versions of the product, in which new tools may be added and old ones may be improved.

The EUROLANG system will thus give the European business community a better opportunity to maintain the command of multilingual technical and commercial communication, which is crucial for developing international cooperation and for safeguarding all language specificities.

BIBLIOGRAPHY

[Alonso 88]
ALONSO, J., "A model for Transfer Control in the METAL MT-System", COLING 88, Budapest, 1988.

[Bachut 91a]
BACHUT, D., & al., "Industrialisation d'un système de TAO français-anglais pour la documentation technique", Génie Linguistique 91, Versailles, 1991.

[Bachut 91b]
BACHUT, D., & al., "Traduction et Terminologie : expérience et perspectives industrielles", 2èmes journées scientifiques du RLTT, Mons, 1991.

[Boitet 82]
BOITET, Ch., & al., "ARIANE-78 : an integrated environment for automated translation and human revision", COLING-82, Prague, 1982.

[Boitet 86a]
BOITET, Ch., "The French National MT-Project: Technical organization and translation results of CALLIOPE-AERO", IBM Conf. on Translation Mechanization, Copenhague, 1986.

[Boitet 86b]
BOITET, Ch., "Current Machine Translation systems developed with GETA's methodology and software tools", ASLIB, London, 1986.

[Boitet 87]
BOITET, Ch., "Current state and future outlook of the research at GETA", MT Summit, Hakone, 1987.

[Chandioux 76]
CHANDIOUX, J., "METEO : un système opérationnel pour la traduction des bulletins météorologiques destinés au grand public", Meta 21, 1976.

[Chappuy 83]
CHAPPUY, S., "Formalisation de la description des niveaux d'interprétation des langues naturelles", Thèse 3e cycle informatique, Grenoble, 1983.

[Gross 75]
GROSS, M., "Méthodes en syntaxe : Régime des complétives", Editions HERMANN, Paris, 1975.

[Gross 90]
GROSS, M., GUILLET, A., "Modèles Linguistiques", Traitement des Langues Naturelles, Ecoles d'été du CNET, Lannion, 1990.

[Hutchins 86]
HUTCHINS, W.J., "MACHINE TRANSLATION : past, present, future",

Chichester, Ellis Horwood series in Computer and their applications, 1986.

[Isabelle 78]
ISABELLE, P., & al., "TAUM-AVIATION : description d'un système de traduction automatisé des manuels d'entretien en aéronautique", COLING, 1978.

[Kugler 91]
KUGLER, M., & al., "The Translator's workbench : An Environment for Multi-Lingual Text Processing and Translation", MT Summit III, Washington, 1991.

[Perschke 89]
PERSCHKE, S., "EUROTRA", MT Summit II, Munich, 1989.

[Roudaud 91]
ROUDAUD, B., "A procedure for the evaluation and improvement of an MT system by the end-user", Workshop on Evaluation of MT Systems, Ste Croix, 1991.

[Schneider 91]
SCHNEIDER, T., "The METAL System", MT Summit III, Washington, 1991.

[Schütz 91]
SCHÜTZ, J., & al., "An Architecture Sketch of Eurotra-II", MT Summit III, Washington, 1991.

[Scott 89]
SCOTT, B.E., "The LOGOS System", MT Summit II, Munich, 1989.

[Séité 91]
SEITE, B., "Enjeux du TALN en gestion documentaire : Stratégie de SITE en ingénierie documentaire", Salon International des Industries de la Langue, OFIL, Paris, 1991.

[Slocum 83]
SLOCUM, J., "A Status Report on the LRC Machine Translation System", First Conference on Applied Natural Language Processing, ACL, Santa Monica, 1983.

[Uchida 89]
UCHIDA, H., "ATLAS", MT Summit II, Munich, 1989.

[Vauquois 85]
VAUQUOIS, B., & CHAPPUY, S., "Static grammars : a formalism for the description of linguistic models", International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, Colgate University, 1985.

Divers
Miscellaneous