SITE buys B'Vital
Relaunch of French national MT project

by Geoffrey Kingscott

Machine translation in the 1990s got off to a sensational start, when in January a series of announcements revealed a wholesale revitalisation of the MT scene in France.

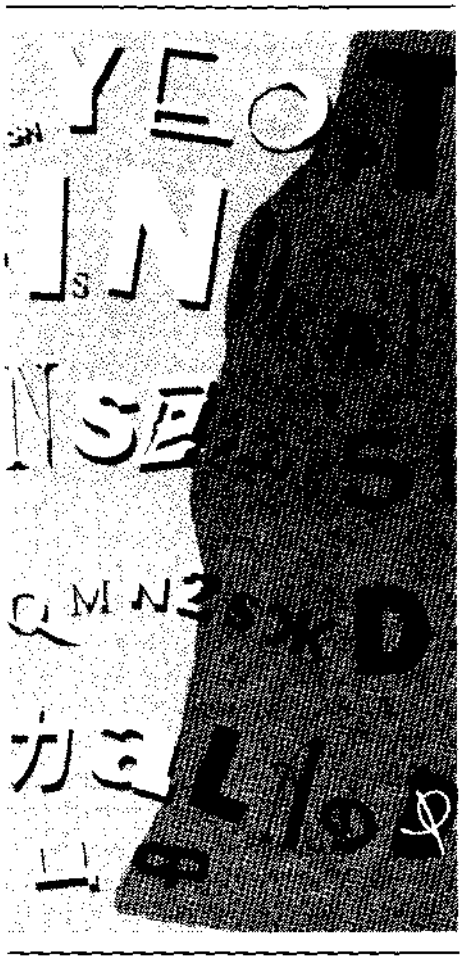
The main developments are that the French Projet National *to find and fund industrial applications of MT is to be relaunched, with the backing of CIGREF; the Ariane MT system is to be incorporated into an actual industrial application; and SITE, the documentation company chosen to spearhead all this activity, has acquired a majority shareholding in B'Vital, the Grenoble company formed to market the achievements of the main French MT research laboratory GETA, which is based at the Joseph Fourier University in Grenoble.*

This terse summary, however, is comprehensible only with background information on the dramatis personae, *and* Language International *editor Geoffrey Kingscott went to Velizy, on the outskirts of Paris, in January to talk to* **Erik Lebreton** *of SITE, and then on to Grenoble to talk to* **Professor Christian Boitet**, *head of the GETA project, and to the B'Vital researchers* **D. Bachut** *and* **R. Gerber**.

First, then, who are CIGREF?

The acronym stands for *Club Informatique des Grandes Entreprises Françaises*, which brings together the data processing managers of many of the big name companies in French industry and commerce, such as Aérospatiale, the BNP bank, Michelin, Dassault and Hachette. For the last two years the Club has been organising round-table meetings on TAO (*traduction assistée par ordinateur*, i.e. machine translation including machine-aided translation) and on the problem of translating large volumes of documentation.

This coincided with a decision by the French Ministry for Industry in July 1989 to move to a support programme for automatic processing of natural language (TALN - *Traitement Automatisé du Langage Naturel*) in French industry. A call for proposals was made. The most interesting proposal came from the company SITE, responding jointly with GETA and B'VITAL.



Who are SITE?

This acronym stands for *Sonovision ITEP Technologies* and the company resulted from a merger of France's two biggest technical documentation firms, Sonovision and ITEP. The company itself belongs to the Cora-Revillon group.

Sonovision started up in 1947, and ITEP in 1961, the latter company having been established by a group who had left the former to set up on their own. The new company employs some 1700 staff in 17 centres in France and counts two European sister companies (in Spain and Italy). Commercial agreements have also been concluded with a UK-based company.

SITE has its own in-house translation and interpreting department, with a team of over 60 translators, interpreters and terminologists, some 40 of which work at the company's head office at Vélizy-Villacoublay. Within the translation department, because of the large amount of aeronautics work handled by SITE (clients include Dassault, SNECMA, Saab, Aérospatiale and Air France), there is a specialist aeronautics section, head by former

French air force officer *Jacques Margelin*. Other specialist groupings bring together translators working on industrial translations and data processing translations. Although 85 per cent of the work is between French and English (in both directions), the in-house team also handle translation work between French and Spanish and French and German.

Other departments at SITE include a production nomenclature and coding department, a scanning centre, drawing offices, a technical writing department and a projects department including teams of data processing engineers developing electronic document management systems.

The 18-month first phase of the *Projet National* programme (starting in January 1990) at SITE — no name has yet been chosen for the project — aims at integrating Ariane into the production chain for documentation, a chain which starts with technical authorship in the source language and goes all the way to publication in the target language. This fits in well with the requirements already spelled out by CIGREF members into how an automatic translation system should perform. After the first phase, of testing in industrial conditions, a second phase would be industrialisation of a French to English and English to French product, with any revision of the system which might be necessary in the light of experience and new equipment. Phase three would see the extension of the system to other European languages, while in the longer term future SITE would look for other potential industrial users and European partners.

In the testing phase, some 5000 pages of documentation, in three types of text, will be processed, and the results compared with human translation of the same text in terms of quantity, time taken, deadlines met, and cost.

At the same time it is working on the exciting new Ariane project; SITE will also be involved in the production of a sophisticated lexicographical working station for the French telecommunications authority's research centre (CNET - *Centre National d'Etudes des Télécommunications*) In November 1989, SITE was informed that its proposals, made following a call for tenders in July 1989, had been accepted.

CNET's problem is that natural language processing applications are occasioning the use of voluminous lexical data. Because of the repetitivity of tasks, the difficulty of maintaining coherence with the volumes involved, and the complexity of relationships between the data, it was decided that what was needed was something like a lexicographical work station. The operator of such a station, who would need to have a sound knowledge of lexicographical principles, would develop and keep up to date a monolingual electronic dictionary.

The idea immediately occurred to SITE that such a station could be opened up to multilingual data, and it would be of interest not only to major dictionary publishers, but also to other countries' telecommunications authorities.

And if this were not enough SITE have also chosen this moment to launch a major distribution campaign for an electronic dictionary of its own production, AQUILA, which is available in English, German, French, Spanish and Italian. AQUILA is a system designed to operate and manage a multilingual terminology database. The system can accept entries in up to 14 languages and work with a combination of any two simultaneously.

What is claimed to be unique about AQUILA, differentiating it from other terminology systems for translators to use with their word processing programmes, is the specificity of each term. The context information contains three pieces of information: the "field of knowledge" in which the term is used (computer science, hydraulics, music etc.), the "sector of activity" of the client company or organisation (publishing, car manufacture, etc.), and the name of the client.

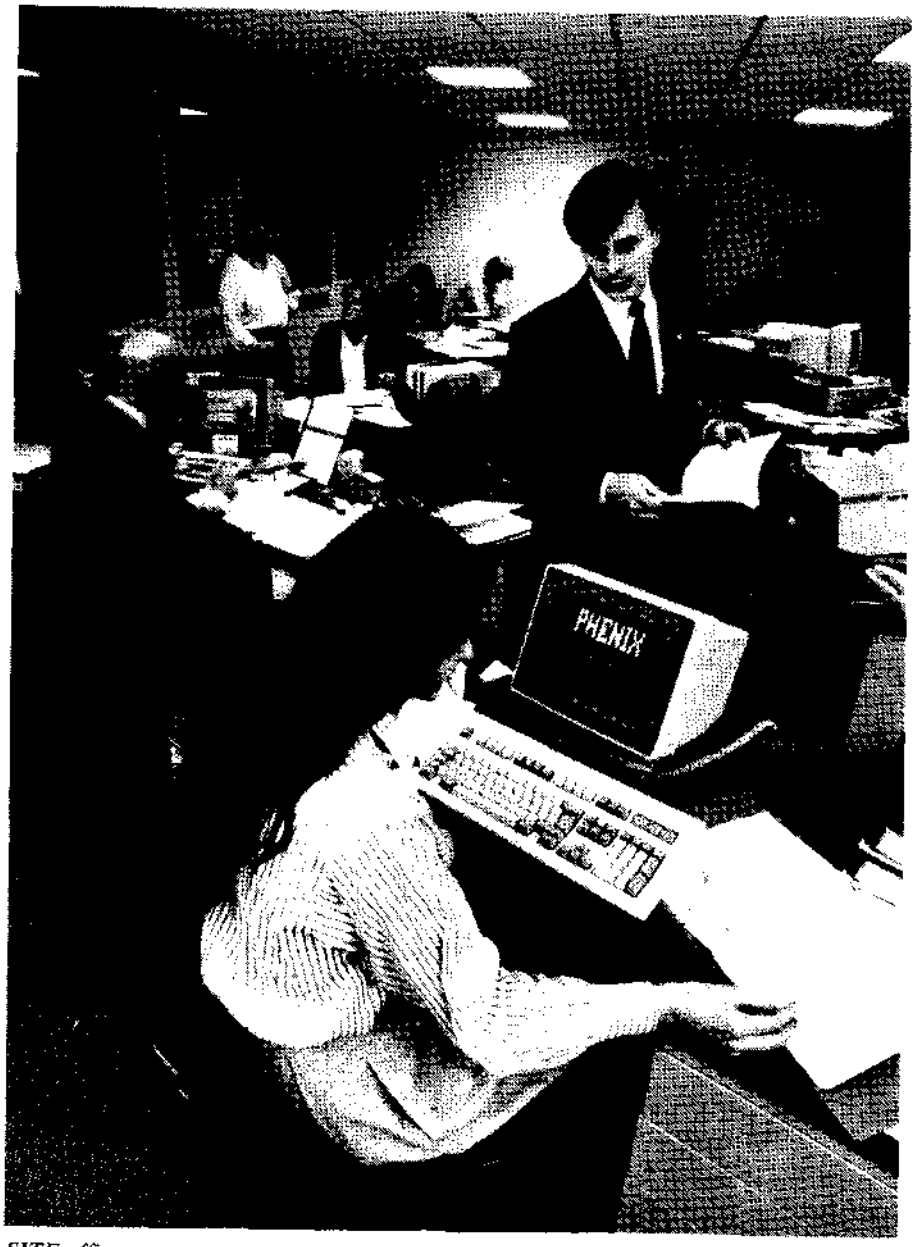
AQUILA is in fact a scaled down, packaged vesion of its big brother PHENIX, the first terminology management system to have been developed and used in-house at SITE. PHENIX is LAN-based (*Novell- Ethernet - 3 Con - Token Ring - Starlan*) and its ambition is to bring heavy-duty terminology management capacity into the microcomputing environment. SITE sees key benefits here both in terms of the PCs "*convivialité*" and in avoiding the higher running costs involved with mainframe applications.

Who exactly are GETA?

In 1959-1960 the French main research body, CNRS (*Centre National de la Recherche Scientifique*) started a study of MT, and decided to take this further by setting up, in Grenoble, CETA (*Centre d'Etudes sur la Traduction Automatique*). From 1960 to 1970 this worked on a Russian to French translation-for-information-purposes system, *TAO du veilleur*. In 1970 CETA split up into three teams, one of which was GETA (*Groupe d'Etude pour la Traduction Automatique*). Based on new principles derived from artificial intelligence (the subject of study of one of the other teams, and from current work in linguistics, a new MT generator, Ariane-78, was developed, and a "pre-operational" Russian to French system, while models and feasibility studies were done on other language pairs. This was the *TAO du réviseur*, the idea being to produce, automatically, draft translations good enough for a professional translation reviser to use, in preference to translating the text himself *ab initio*.

The 1980s saw GETA making a major effort to put its systems into operational use, in particular within the framework of the French *projet national*. Interest was also shown in developing translator work stations, the *TAO du traducteur*.

Since last year, 1989, GETA has also been working in a new area, in attempting to make automatic translation available to technical writers and others who are not necessarily linguists, the *TAO du rédacteur*, which



SITE offices

implies quality MT not requiring revision. This research is still at an early stage, but a project, which has been given the title LIDIA-1 (*Large Internationalisation des Documents par Interaction avec l'Auteur*), aims at constructing a small prototype, with a Mac PC using HyperCard serving as an "authoring station" or "editing station" (*station de rédaction*), while major operations, carried out in asynchronous mode, are transferred to an MT server accessed over a network. Among some innovative ideas being incorporated into this new concept are voice synthesis (for clarification dialogues, and outputting translations), the use of back translation as an indirect way of checking the accuracy of translation, and the possibility of "assisted language training" (*apprentissage assisté des langues* by which access can be given to the translations and the "natural" dictionaries supported by the authoring station.

Professor Boitet believes that this new concept of "personal MT" is the first really concrete proposal to the problem of the author of an article or a memorandum who wants to write in his own language and yet be understood in the eight other languages of the European Community, or indeed in languages of other parts of the world.

"Le concept de TAO personnelle est une première proposition concrète. Pour qu'elle débouche sur des solutions utilisables, il faudra sans doute que de nombreux groupes de recherche s'engagent dans cette voie, et que les investissements suivent. En effet, si un prototype "démonstrateur" peut être de petite taille, un système grand public devra reposer sur un très gros vocabulaire et sur des grammaires couvrant toutes les constructions des langues traitées."

Professor Boitet will be describing the LIDIA project in more detail at the COLING-90 computational linguistics conference in August 1990 (see the Calendar).

Who are B'Vital?

Established in 1985, B'Vital specialises in language applications and the development of management tools, based on the Ariane system, i.e. in finding and developing practical uses for Ariane. The acronym stands for Bernard Vauquois Informatique et Traitement Automatisé des Langues. It has itself developed a system, BDTAO (*Bordereaux et Dictionnaires pour la TAO*) to make it easier to access Ariane dictionaries. Among its other products are *SCRIBERE*, a high-level formatting software, and *19/20*, a system for recognising the lemma or dictionary-entry form (lemmatisation) and generating word forms (conjugations, declensions) in the French language.

B'Vital had in its first period been associated with Sonovision, one of the two companies which were merged to make up SITE, and with SG2, another company which was masterminding in the 1980s the *Projet national de traduction assistée par ordinateur*, which enjoyed some French government funding. B'Vital's role at that time was in the fields of language specifications, developing grammars for French-English and English-French translation systems, training of terminology and indexing teams, and the specification of a new basic software for MT.

Bernard Vauquois

Professor Bernard Vauquois was the leading French figure in machine translation from 1960 until his death in September 1985. A scientist, astronomer,

mathematician, and computer specialist, he founded and directed CETA and GETA. A selection of his writings has recently been brought together, edited by Christian Boitet and published by GETA as *Analectes — Bernard Vauquois et la TAO, 25 ans de traduction automatique*.

B'Vital, 35 rue Joseph-Chanrion, F-38000 Grenoble.
SITE, 11 avenue Morane-Saulnier, F-78143 Vélizy-Villacoublay.