MT at Grenoble in 1990
General presentation

Christian BOITET
GETA, Institut IMAG
(UJF & CNRS)
BP 53X, 38041 Grenoble Cedex, France

Introduction

This presentation has been prepared for ROCLing-III, as a support for a tutorial and a lecture. The tutorial is concerned by the current state of the art in *classical MT and Text Encoding*, as exemplified by the results of a long term research effort started in Grenoble almost 30 years ago.

The lecture is dedicated to the presentation of a new framework for MT, *personal MT*, and of the new potentialities and problems it offers. In particular, the *text encoding* problem presents interesting new aspects in personal MT.

The written support consists in four documents, three for the tutorial and one for the lecture, which are briefly summarized below.

I. The evolution of ideas in classical MT : B. Vauquois' contribution to the theory and practice of MT (1960—1985)

CNRS (Centre National de la Recherche Scientifique) launched research on MT in 1959-60. From 1960 until 1970, CETA (Centre d'Etudes sur la Traduction Automatique), under the direction of Pr. B. Vauquois, worked on *MT for the watcher* and succeeded in building an impressing Russian-French system of such quality that a French-speaking specialist could understand scientific and technical russian texts in various domains (nuclear physics, chemistry, linguistics, space sciences). This system was prepared on a corpus of 1 million words (4000 pages) and systematically tested on 40% of it.

In 1970, CETA was divided into 3 teams, one of which began research on AI at Grenoble, while another, named GETA (Groupe d'Etude pour la Traduction Automatique) and still led by B. Vauquois, switched to research in *MT for the revisor*. Based on new principles borrowed from AI and from modern linguistics, a generator of MT systems (Ariane-78) and a preoperational Russian-French system were developed, as well as numerous mockups and feasibility stodies on other languages. The aim of such systems is to automatically produce "first drafts" of translations, good enough for professional revisors to accept to revise them and for the revised translation to be produced more quickly and less expensively.

II. Software and lingware engineering in recent (1980-90) classical MT : Ariane-G5 and BV/aero/F-E

During the last decade, GETA made important efforts in technological transfer towards industry, in particular in the framework of the Machine-Aided-Translation National Project (CAT-NP, or PN-TAO in French, 1983-87). As a result, the techniques developed by GETA are beginning to be introduced in some industrial setting for operational use. Maybe the example of Japan and the perspective of the integrated european market are finally convincing industry people of the growing need to automatize the translation of large volumes of technical translation.

The perspective of industrial use has led to the emergence of a set of MT engineering techniques, which rely on a generator of MT systems, Ariane-G5, on a array of associated software tools, and on lexical and gramatical knowledge bases. MT systems of that type are analogous to expert systems : they are really usable only if they are specialized to particular text types, and, contrary to systems "for the watcher", if they are maintained and operated by teams working in close contact, in the context of the same organism.

In the framework of various cooperations, GETA has also participated in research on *MT for the translator,* that is, on computerized translation aids, which may be tools usable in various office automation contexts or integrated "translator's workstations". As a matter of fact, MT for the translator should better be called "Machine Aided Human Translation & Revision". It is the only way to help in translating small volumes of heterogeneous texts, for which real MT would be far too costly, and impractical in any case. It is also necessary as a complement to MT for the revisor, for the human revision of machine translations.

MT for the watcher, the revisor and the translator constitute nowadays "classical" MT. Of course, not all problems have been solved, and some are not likely to be in the near future. Nevertheless, these techniques can be used in appropriate settings, and are used today cost-effectively in various operational contexts.

III. Classical MT and the Text Encoding Problem : a Phonetic and Structural Encoding of Chinese Characters in Chinese Texts

MT is very often mentioned in the context of multilingual translation. Although it is quite difficult to find real situations in which texts of the same type have to be translated from all languages of a collection into all others of the same collection (but the example of the European Community and of other international organizations shows that the case may arise), it is rather common to find situations in which the same text has to be translated into many languages. But no widely available computer system today (OS) is able to handle at the same time even all languages written using the Roman alphabet, not to speak of Chinese, Japanese, Arabic, etc.

There exist of course computer codes for many, if not all, character sets used by today's languages, and, for storing purposes, it is enough to introduce special codes (tags) to indicate a change of coding. But what of the poor linguist trying to prepare a French-Chinese-Arabic lexical data-base? The Chinese and Arabic internal codes will give ununderstandable jumble on his screen or printer.

Hence, for the purpose of multilingual natural language processing (not storing), we have developed a generic technique of encoding (transcribing) language-specific character sets into a universally available character set., in a portable and readable way. The exercise is quite easy for languages using the Roman character set, or based on alphabets of comparable size. It is interesting to see how it can be performed on Chinese, on about 15000 characters, without the need of using a particular dictionary.

IV. Towards Personal MT : general design, dialogue structure, portential role of speech, text encoding

Since 1989, GETA has decided to explore a new approach, that of personal MT, ot *MT for the writer*. The problems of classical MT don't disappear, rather, they are posed in a new light. The idea is to offer MT to the authors of technical documentation, scientific papers, or even of books, letters, notes or (electronic) messages, the price to be paid being to accept to interact with the system in order to write better (terminological, grammatical, stylistic standardization) and to reduce ambiguities of all types (lexical, grammatical, semantic clarification). Hence, the aim is high quality MT, for the general public, and in general without revision, because it is impossible to put a revisor at the side of each writer.

Research is only beginning. The aim of the LIDIA project is to validate this concept and to attack the new problems it poses by constructing a small prototype, where the writing station will be a MacIntosh under HyperCard, the heavy linguistic functions being realized asynchronously on a MT server reached through a network. Among the most interesting points of this new concept, we may mention the integration of speech synthesis (clarification dialogue, output of translations), the use of reverse translations (for indirect verification of the translations), and the possibility of computer assisted language learning (by accessing the translations and the "natural" dictionaries supported by the writing station).

The use of national languages in teaching, culture, science and technology is growing around the world today. For personal communication, teaching more languages is certainly a good idea. But , in a multilingual environment such as Europe, that will never help authors who can write satisfactorily only in their own language and want to be translated in many other languages. The concept of personal MT is a first concrete proposal. For it to lead to usable solutions, it will no doubt necessitate the efforts of many research groups, and later heavy investments. As a matter of fact, while a demonstration prototype can be small, a system for the general public would have to rely on a very large vocabulary and on grammars covering all constructs of the languages of the system.

-o-o-o-o-o-o-o-o-o-o-o-