[*Machine Translation Summit*, September 17-19, 1987, Hakone Prince Hotel, Japan]

INTERLINGUAL MT - AN INDUSTRIAL INITIATIVE

Toon Witkam
BSO/Research
NETHERLANDS

P.O. Box 8348
3503 RH UTRECHT

INTRODUCTION

This paper describes nature and progress of DLT. DLT (Distributed Language Translation) is a long-term research and development project undertaken by BSO/Research in the Netherlands. It is aimed at high-quality translation without post-editing, a not very narrowly limited range of informative texts, a diversity of languages, and networked personal computer equipment of the 1990s and beyond.

The project started in 1982 with a Feasibility Study funded by the Commission of the European Community (Hutchins, 1986, pp. 287 - 291). Since 1984, the government of the Netherlands and BSO each support 50 % of the project. Annual budget averaged around 1 million dollar for the last three years, and still 5 million has been reserved for 1988-1991. Commercialization is not expected to begin before 1992.

BSO (Buro voor Systeemontwikkeling) is a private software company based in the Netherlands. Its business is entirely service-oriented (customized software) and increasingly international, covering various high-tech areas. The DLT project was initiated as a non-commercial research activity.

GENERAL CHARACTERIZATION

The outstanding features of the DLT system design can be summarized as follows:

1. Interlingual Architecture.

DLT is one of the very few MT systems under development with a genuine interlingua: an intermediate language (IL) that comprises a lexicon as well as a syntax. DLT has adopted Esperanto for this purpose: a natural language according to Chomskyan criteria, but at the same time the natural language that - due to its structural regularity and transparency - is the least remote from formalized systems.

The fact that Esperanto is a language in its own right, with an autonomous grammar and lexicology, makes DLT appear as a double translation system. Indeed a contrastive-syntactic transfer and lexical substitution take place twice, once into and once out of Esperanto. But the extensive use of the intermediate Esperanto stage for semantic disambiguation (see item 3) and the distribution of the translation process over two main subprocesses (see item 5) justify the designation 'interlingual' for the overall system's architecture.

2. Worldwide Multilinguality.

DLT's interlingua offers excellent long-term prospects for connecting a diversity of language types. Though Esperanto word root material is from European origin (mainly Romance and Germanic language), the sentence structure is Slavonic and the word formation mechanism and invariance of morphemes rather rank the IL as an agglutinative language, with clear features of an isolating language as well.

Therefore, in addition to a main effort for the language pair English - French, preparatory studies have been undertaken in the framework of DLT for Finnish and Hungarian, and work on Chinese and Japanese is under negotiation now.

3. Artificial Intelligence.

To achieve proper, knowledge-based interpretation of natural language texts, techniques for problem solving, reasoning and understanding, belonging to the realm of Artificial Intelligence (AI), have been incorporated in the DLT system design, in such a way as to exploit its interlingual architecture, and with a tremendous growth potential for the future evolution of the system.
(Papegaaij, 1986a; Schubert 1986)

4. Interactivity during Text Entry.

Despite AI-supported disambiguation, fully automatic high-quality translation will remain Utopian for the next five decades or so. In the DLT translation mechanism, consultation of the author of the source document has been chosen as the best way to add the necessary human intelligence to the machine process. Therefore, DLT is closely interlinked with modern text entry facilities (increasingly referred to nowadays as "authoring environment", and featuring such additional services as spelling and style correction), prompting the author with questions on the interpretation of certain words or passages. This closed-question disambiguation dialogue is entirely in the language of the source text, and void of linguistic jargon, which makes DLT a system for monolingual end users.

5. Networking.

Compared to the traditional configuration of MT on a central mainframe computer, the DLT system represents a radical change (Fig. 1): the translation takes place in the sending and receiving terminals attached to the network, through which the IL is passed as a semi-product. Each terminal has its own copy (on CD-ROM or a similar device) of the translation software, complete with knowledge bank and dictionaries for a specific language. There is no question of the sharing of processors or storage capacity: the only thing shared between the DLT users is the IL, which will be pushed to become the industry standard for multilingual text flows in networks.

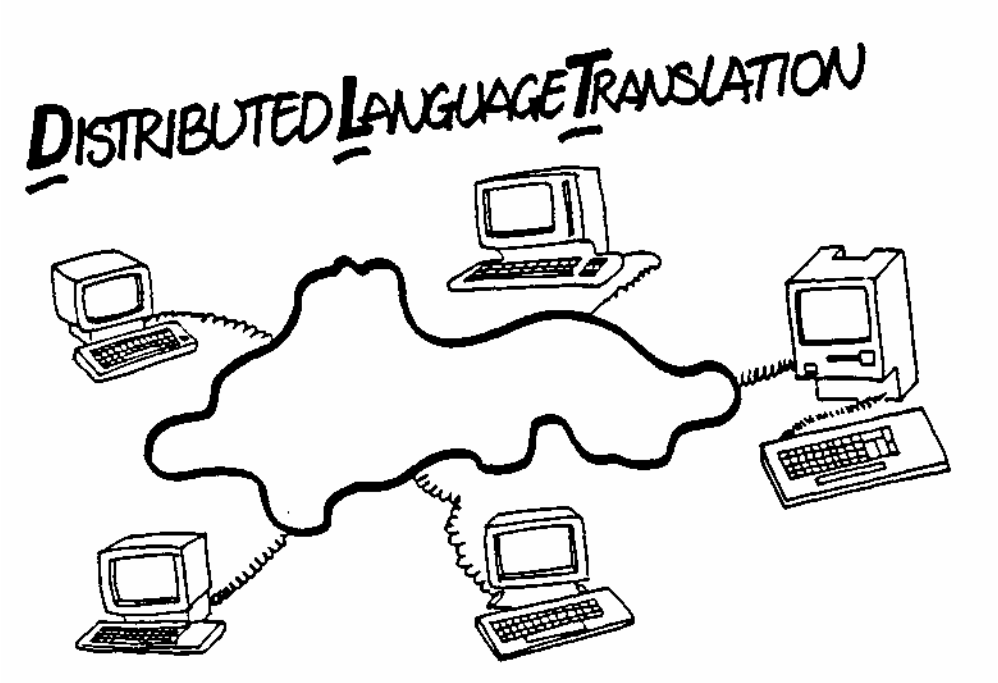

Fig. 1 In the DLT system, the translation process is split over the sending and receiving terminals. Through the network goes a semi-product (the IL).

RECENT PROGRESS

Prototyping work by a gradually increased project team (up to 16 researchers in 1987) lead to much refinement, new insights and a better understanding of the overall DLT design than was possible at the time of the 1982 - 1983 Feasibility Study. The most important progress and findings can be related to the following two areas:

- strengthening of the interlingual architecture;
- early and intensive use of AI techniques.

The strengthening of the interlingual architecture means several things: the share of the IL phase in the translation process has been increased, and the particular choice of Esperanto as IL has become more convincing.

In 1983, we envisaged a system with an IL kernel of moderate size, flanked by an SL-IL language pair system on the one side, and a IL-TL transfer system on the other. Both these subsystems have now evolved into heavily IL-based processes, leaving only a relatively small margin for SL (Source Language) or TL (Target Language) specific work. The SL-specific part includes exactly those elements that are not interesting from an AI point of view: morpho-syntactical details and SL or TL idioms and collocations.

At the same time, the importance of the IL as an intermediate stage in the translation process has become much more evident: all semantic disambiguation attempts, including discourse analysis, are being carried out in it. The crux of the DLT design in this respect is the presence of a world-knowledge base written in the very same IL which is practically identical to Esperanto.

Disambiguation of small sentence fragments was demonstrated at COLING '86 in Bonn (Papegaaij, 1986b). The latest progress (see the Melby test below) with context-dependent selection of the right word meaning for whole sentences largely relies on the early and intensive use of blackboard and other AI techniques applied at the IL stage, for which PROLOG has been used throughout.

The advantage of concentrating all AI-related (knowledge-based) processing at the IL stage, is evidently that its development needs not to be repeated for each SL an TL again. For this reason, most of the DLT development effort has gone into the IL-based system kernel.

THE MELBY TEST

The biggest hurdle in MT is that of ambiguity: how does the computer identify the right meaning of a given word or phrase? Most MT projects have come to grief on this point (Melby, 1986). The phrase "a simple model", for example, can conceivably refer to a mannequin who is either just naive or downright retarded; but of course it can also mean "a straight forward simulation", e.g. of an aircraft or of a process. The correct interpretation (and hence the right translation) can only be deduced from the context.

For a computer to draw the right conclusions from the context, it needs "knowledge of the world". Such knowledge is one of the essential features of expert systems, and a good many systems have already been developed which exhibit some of the characteristics of AI. However, the question whether true intelligence can be created artificially will - in the view of some researchers - only be answered with a convincing YES when a computer succeeds in translating general informative (not specifically technical) texts with a high degree of accuracy.

In the beginning of 1987, the DLT system completed its first critical test designed by an American team at the Brigham Young University in Utah under the leadership of linguistics professor Alan Melby. The crux of this test was to determine whether DLT is capable of solving the major problem of ambiguity and polysemy in translation.

The system was fed a selection of English texts which were hitherto unknown to the DLT researchers. These text samples had been drawn from a corpus of UN and EC documents amounting to some 500,000 running words. The samples themselves contained roughly 480 different content words, and the DLT team had been provided in advance with an alphabetical list of these words. However, in order to make it difficult, if not impossible, to guess the original context, a further 320 random words had been merged into the list. As an additional safeguard, the Melby team were given a copy of the DLT software and dictionaries, in advance of the test proper.

The original list of content words formed the basis for the building up of "knowledge of the world". For each word in the list, the computer was supplied not only with the relevant translations but also the appropriate combinations of concepts. For the concept "dog" for example, an appropriate combination might be "to bark", but certainly not "to delegate". "To delegate", on the other hand, would be a suitable combination with "manager". In this way some 50.000 word combinations (basic units) derived from the original 800 words were stored in the knowledge bank.

The Melby Test was designed to evaluate both the SL-IL and the IL-TL translation stage. As to the former, results showed the system chose the right translation 32 % of the time, making its selection from an average of 18 alternatives. A further 12 % of choices had previously been marked (by professional translators) as "equally reasonable". The second stage scores were significantly higher (Melby, forthc.).

INCLUSION OF TERMINOLOGY

Present shortcomings are due to the still modest size of the knowledge bank and notably the absence of a mechanism for recognition of phrasal verbs and other collocations of the SL, a circumstance which considerably increases the load on the knowledge-based IL disambiguation.

Technical terms form a particular and frequent type of collocation. The timely creation of technical terminology in IL will therefore be a major factor in the future success of DLT. This matter is closely related to the introduction of semi-automatic techniques for lexicography and terminography, as well as to issues involving organization and standardization.

If the overwhelming yearly increase of new technical terms in common languages like English and Spanish can be coped with only by recourse to automation and AI techniques, the Esperanto-based IL with its regular word structure will certainly not be in an unfavorable position (Sadler, 1987; Schubert, forthc.).

AI-related processing with its intensive pattern matching will profit from the clear morpheme structure of Esperanto words. To exploit this phenomenon for DLT, a powerful intra-word grammar has been written in 1986. This word grammar will enhance not only semantic disambiguation, but also semi-automatic term creation.

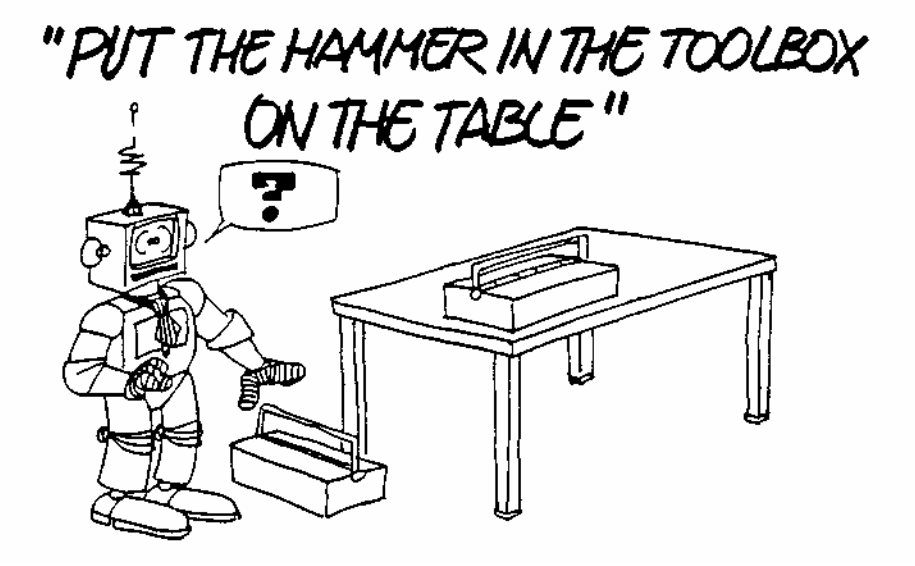

Fig. 2 Ambiguity is the biggest hurdle in MT, as it is in robot control.

REFERENCES

Hutchins, W.J. (1986): Machine Translation: Past, Present, Future. Chichester: Horwood.

Melby, A.K. (1986): Lexical transfer: A missing element in linguistics theories.
In: llth International Conference on Computational Linguistics. Proceedings of COLING '86. Bonn: Institut für angewandte Kommunikations- und Sprachforschung, pp. 104 - 106.

Melby, A.K. (forthc.): Report on the DLT test on lexical disambiguation. Paper to be presented at COLING '88, Budapest, August 1988.

Papegaaij, B.C.(1986a): Word expert semantics: an interlingual knowledge based approach. V. Sadler/A.P.M. Witkam (eds.) Dordrecht/Riverton:Foris

Papegaaij, B.C./V. Sadler/A.P.M. Witkam (1986b):
Experiments with an MT-directed lexical knowledge bank.
In: llth International Conference on Computational Linguistics. Proceedings of COLING '86. Bonn: Institut für angewandte Kommunikations- und Sprachforschung, pp. 432 - 434.

Sadler, V. (1987): AI-directed interlingual terminography in tomorrow's NLP systems.
In: Proceedings of the International Congress on Terminology and Knowledge Engineering. Trier, September 1987.

Schubert, K. (1986): Linguistic and extra-linguistic knowledge. Computers and Translation, Vol. 1, No. 3. Osprey (Florida): Paradigm Press.

Schubert, K.(forthc.): Interlingual terminologies and compounds in the DLT project.
In: Proceedings of the International Conference on Machine and Machine-Aided Translation. Birmingham, April 1986.