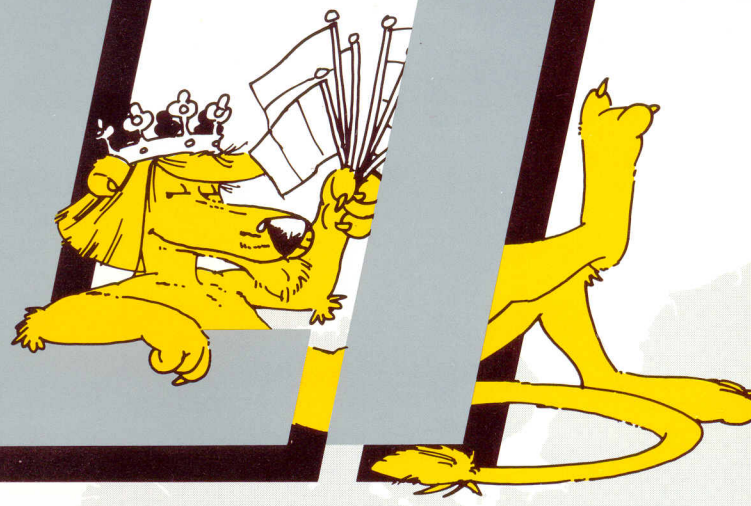
DISTRIBUTED LANGUAGE TRANSLATION

DLT



Distributed Language Translation – a multilingual system for computer networks.

Buro voor Systeemontwikkeling.
P.O. Box 8348, NL-3503 RH Utrecht, Netherlands.
Telephone (..31-/0-) 30 911 911.

DISTRIBUTED LANGUAGE TRANSLATION

Feasibility Study

of a Multilingual Facility for Videotex Information Networks.

A.P.M. Witkam,
Senior Consultant at BSO

BSO, Utrecht 1983

Published by BSO,
P.O. Box 8348,
3503 RH Utrecht,
The Netherlands.

CIP-gegevens

Witkam, A.P.M.

Distributed language translation : feasibility study of a multilingual facility for videotex information networks / A.P.M. Witkam. - Utrecht : BSO. - Ill.
Met lit. opg. - Met samenvatting in het Frans, Duits, Nederlands en Deens.
ISBN 90-70730-01-4
SISO 365.5 UDC 651.926:681.3
Trefw. : vertalen ; computertoepassing.

PREFACE

This report presents and defends a worked-out concept for one specific system: a system in the area of computerized language translation, traditionally known as MT (Machine Translation). The contours of this particular system, DLT (Distributed Language Translation), originated within the inspiring surroundings of the Dutch software-house BSO, back in the period 1979-1982, with a strong personal engagement of the author.
A grant by the Commission of the European Communities permitted a thorough feasibility study (September 1982 - October 1983) of this system. The present report has been based on this study.

What the reader will find here is not a systematic and purely objective evaluation of the MT field, or a complete strength/weakness investigation of various MT systems (such studies already exist). However, comparisons with other MT systems on various points seemed necessary to present DLT, and to explain its features. The result is the development and consistent defense of a new MT-system concept (DLT), in terms of reference of the current MT field. We at BSO are convinced that this will serve the interests of the European Commission and the professional community as least as much as another academic expertise report.

In line with the current EC policy of high-technology dissemination, this report will be widely distributed. It is aimed at MT specialists, computer-minded translators and computational linguists, at the heads of translation departments in industry and government. By exposing this design to the professional community, BSO hopes to receive useful response and support for continuation of its efforts. The report is also intended to find its way to the desks of planners and decision-makers in the publishing world and the computer and consumer-electronics industries.

From the outset, it was clear that the critical issue to be investigated was a linguistic one: the unambiguity of the Esperanto-based intermediate language. This explains why more than 75% of this report is of linguistic nature: Chapters III, IV and V. Readers not familiar with computer linguistics or translation are advised to browse through the illustrations of these chapters to get a general impression.
Chapter I gives background and motivation, Chapter II a general overview of the system in computer-industry terms. Hardware and software implementation is dealt with in Chapter VI, and the report concludes with a technical schedule for the next development phase (Chapter VII).

During this feasibility study, I received valuable advice and constructive criticism from a diversity of competent professionals in the general linguistics and MT field: Louis des Tombe and Steven Krauwer from Utrecht, Heinz-Dieter Maas from Saarbrücken, Guy Bourquin and his team from CRAL (Université de Nancy II), and Bente Maegaard from Copenhagen. In addition, I have to thank Victor Sadler, John Wells and Ivo Lapenna of the Esperanto Academy.
I would also like to express my gratitude to all persons inside and outside BSO who encouraged and stimulated me by expressing their interest and who helped me with useful hints and suggestions.
Finally, I am indebted to Alex Olde Kalter, who devotedly assisted me during the whole study, kept me alert and contributed invaluably by literature investigation and many other supporting work, including the addition of section IV-4.
To Andy Simons I am indebted for his inspiring advice on hardware implementation.

Utrecht, October 1983 Toon Witkam

CONTENTS. Page No.

IV. PRINCIPLES AND DEFINITION OF DLT's INTERMEDIATE LANGUAGE (IL).

V. TEXT TYPE, TERMINOLOGY AND TRANSLATION QUALITY.

VI. DLT'S IMPLEMENTATION: HARDWARE AND SOFTWARE.

VII. PILOT PROJECT DEFINITION.

APPENDICES

DLT FEASIBILITY STUDY 1982-1983

SUMMARY

This report contains the results of a 14-month study for the European Commission, on the feasibility of DLT (Distributed Language Translation). DLT is a proposed system for semi-automatic translation between written natural languages. It was conceived and first investigated within the software-house BSO/Automation Technology b.v. of The Netherlands, during the period 1979-1982. The present feasibility study serves as a conclusion of the 'drawing board' phase.

The DLT project, a phased and long-term undertaking, aims at economic translations between European languages (starting with French, German, English) but promises excellent extension possibilities to include others (Japanese, Hebrew, Arabic). The type of text to be processed can be characterized as 'informative', ranging from technical instruction manuals to scientific literature abstracts and from business reports to nuclear waste disposal regulations. Stylistic effects, connotations and other subtleties ("reading between the lines") can generally not be preserved. Apart from that, and at the cost of more or less reflecting the structure and wording of the SL text, DLT translations will be reliable and grammatically correct.

Operational environment.

DLT can best be regarded as a coherent set of built-in facilities in text generation and text display hardware, i.e. it is used in conjunction with word-processors and videotex terminals. DLT is typically not a batch system, and therefore not particularly suited for translation of OCR-entered archival texts. It is however very attractive for information distribution networks and international on-line services (public as well as private), and compatible with modern synchronous data transmission and packet switching techniques.

Attempts to translate natural language texts by computers, a dream pursued since WW II, only produced partial and modest results. The specialist field of MT (Machine Translation) progressed slowly, and major breakthroughs failed to come. A revival of interest during the last six years, in Europe and the Far East as well as in North America, together with a diversification of 'multilingual facilities', forms the background against which DLT must be considered.

Multilingual computer facilities today range from word-processors with foreign language dictionaries to large mainframe systems producing raw translations in batch mode. The former can be typified as CAT (Computer Aided Translation), the latter as FAT (Fully Automatic Translation). In CAT, a human translator or editor does the job, seeking assistance from the computer if and when he or she likes it. There is however a third type of configuration, which we call SAT (Semi-Automatic Translation), where the computer does the job but consults a human attendant if and when the computer decides to. It is this type of system, in which an automatic process is supplemented by a computer-initiated interactive dialogue, where we must locate DLT.

In the SAT concept, the computer relies on human help for clarifying the meaning of words and constructions that are ambiguous to the computer (but not to the human, who is supposed to know the context of an SL sentence).
Apart from the difference in master-slave relationship between CAT and SAT, the latter is much nearer to FAT than the former. But whereas FAHQT (Fully Automatic High-Quality Translation) remains - by general concensus - an unattainable goal for the next five decades or so, SAHQT (Semi Automatic High Quality Translation) presents itself as a realistic goal-setting (especially for a software-house with a wide experience in man-machine dialogue design), and that is why this course is embarked on with DLT.

Translation Process Architecture.

Regarding the major translation system architectures: Direct, Transfer and Interlingual, it must be emphasized that DLT has been conceived as an interlingual system, lexically as well as grammatically. To this purpose, we make use of a modified subset of Esperanto as IL (Intermediate Language), and a large portion of the report is devoted to the description and grammar definition of this interlingua.

The interlingual architecture implies a process consisting of 2 major steps (SL-analysis, resulting in IL, and TL-synthesis, departing from IL), which fits extremely well to the outside operating environment (distribution of the translation process over sender and receivers in an information network). The IL must be seen as a narrow bridge, a compact exchange of information between SL- and TL-modules, extending across (volume-tariffed) telecommunication networks.

Comparing DLT with a current competitive approach, the impressive development of EUROTRA, there is a remarkable

difference between the former's IL and the latter's interface structure: EUROTRA has adopted an intermediate tree representation with complex labels, covering semantic as well as surface syntactic and morpho-syntactic variables, i.e. abstract formatives. DLT's IL, on the other hand, basically consists of a linear string of lexical formatives.

In both approaches, the intermediate structure must have some 'added value' compared to the original SL-input: it must be void of the peculiarities and idiosyncracies of the SL, and further processable by TL-oriented modules. In particular, it should be free from ambiguities.

Where EUROTRA seems to tend towards storing more and more abstract information into the interface structure, DLT has sought to reach the above aim by careful design of its Esperanto-based IL, exploiting the experience and the linguistic characteristics (invariant and autonomous morphemes, Greenberg's agglutination index: 1.00, transparency of grammatical structure, a relatively precise system of prepositions) of an already existing, semi-artificial language. One could say that the modified Esperanto used for DLT incorporates a tree structure in itself, complete with morpho-syntactic labels (grammatical endings, particles and function words). Valency boundness information is preserved in IL dictionary entries.

Whereas a transfer system like EUROTRA attempts to limit (for obvious economic reasons) the size of the SL-TL transfer operation to a bare and straightforward substitution of lexemes (SL-words are replaced by TL-words), a fully interlingual system like DLT profits from the presence of full-blown IL dictionary columns at both sides of the SL-TL watershed. In DLT, translation can rely extensively on the level of valency boundness, which compensates the absence of abstract semantic relation labels. Still, the advantage of modular system development by separate SL- and TL-teams is retained, and familiarizing with the IL grammar and lexicon now takes the place of harmonizing on a common abstract labelling interface.

The limitation of DLT's intermediate structure to a linear string of lexical formatives has 2 practical advantages which much determine the overall shape of the system: quick inspectability (for developement and maintenance) and compactness (for low-cost transmission).

Unambiguity of the IL.

The main issue in the underlying study has been the unambiguity of DLT's Esperanto-based IL, a prerequisite for a fully automatic translation step from IL to TL. To this

purpose, 'unambiguity' has been more precisely defined here in terms of IL-parsability by a simple parser, not involving 'deep' semantics or knowledge-of-the-world, but relying on (morpho-)syntactic and (IL-dictionary based) valency information. The IL-grammar, which is described in the body of the report, has been built by adding 3 modification 'layers' on top of the basic layer of common Esperanto, each of which contributes to the IL's unambiguity. The modifications include: a strict prescription of word and word group order, introduction of a limited number of new function words and particles, a consistent use of punctuation, and insertion of a universal separator element. Special care has been taken to avoid space-consuming or obtrusive extralingual elements that could unfavorably affect the IL's compactness and inspectability.

Checking the resulting IL's unambiguity in a contrastive overview with syntactic ambiguities in SL's (English, German, etc.) has been an important part of the study. In the report, over 100 examples illustrate the IL's resolving power to distinctly represent the alternative readings in various areas of structural ambiguity, including notorious stumbling stones of MT: part-of-speech, function words, PP-ambiguity, verb nominalization, anaphorics, etc. Though it must be admitted that even such a comprehensive enumeration still does not prove the IL to be completely ambiguity-free, a proof is given for the extendibility of the IL's ambiguity-resisting power. An algorithm for automatic separator insertion, explained in this connection, guarantees a safe handling of accidental (and therefore difficult to predict) syntactic ambiguities.
The same algorithm protects the IL against the systematic ambiguity widely present in conjunction and modifier scope (following, in certain cases, an interactive disambiguation dialogue).

Phased-approach benefits.

The feasibility of an unambiguous "Esperanto"-IL having been affirmed positively by the IL-design shown in this report, fully automatic IL-to-TL translation lies within reach. On itself, the construction of natural language synthesis modules (given an unambiguous interface) is commensurate with the current state-of-the-art and does not present a technological breakthrough. In this case however, the particular nature of the IL has an undeniable practical significance: derived from a human communication language, its accessibility and perspicuity make the IL not only an excellent development medium, but also an effective data input language for certain multilingual database applications (viz. the controlled environments of a videotex

information provider, a production department for technical manuals, the building of a consultative expert system, etc.).

The next phase proposed for the DLT project is therefore the development of a pilot system which includes at least one IL-to-TL module (with German as preferred choice for the first TL, and an estimated effort of 12 man-years over a 2-year period). As for the translation quality, the TL-output is expected to show imperfections in the use of articles and past tenses (though the study indicates how to overcome these defects eventually). Besides, a somewhat monotonous style must be reckoned with, because a natural variation of sentence patterns according to the topic (theme-rheme) is hard to realize. Proper translation of lexemes, which is determined by IL-TL lexical divergence and the capacity to formalize and add word choice procedures to the IL-TL dictionary, can be achieved by limiting the pilot system to the denotative register ('technical' texts) and one specific application field.

International business and economy, supplemented by international law, has been selected as pilot application area. The reason for this, apart from its wide commercial interest, is the availability of an already extensive and up-to-date terminology in Esperanto (i.e. in IL). Though further work is necessary, the existing amount of well-defined terms provides a sound basis to build an application-oriented translation system upon and again demonstrates the practical value of a genuine, comprehensive and easily accesible IL.

Long-term prospects.

The time-scale for bringing a complete, hardware-integrated multilingual DLT system (with at least 1 SL and 2 TL's) onto the market is approximately 5 years, subsequent to the limited pilot system proposed for the next 2 years.

The realization of a semi-automatic SL-to-IL analysis module, in which a careful use is made of questioning the typist, represents a crucial and characteristic part of DLT development, slightly different from existing and competitive efforts. Only global design features of the SL-analysis process can be revealed now: intervalwise, data-driven, single-pass LR parse, in step with the entering of words by the typist (integration of DLT within word-processor); stepwise quasi-parallel creation of IL-directed SL-trees along a (moderate) number of alternative parse trails. The accent is on fighting undeterminism by parallelism instead of backtracking. This approach is

favored by the relatively slow speed of manual typing (leaving gaps of 'free' processing time to dedicated processors) and the projected availability of high-capacity storage chips.

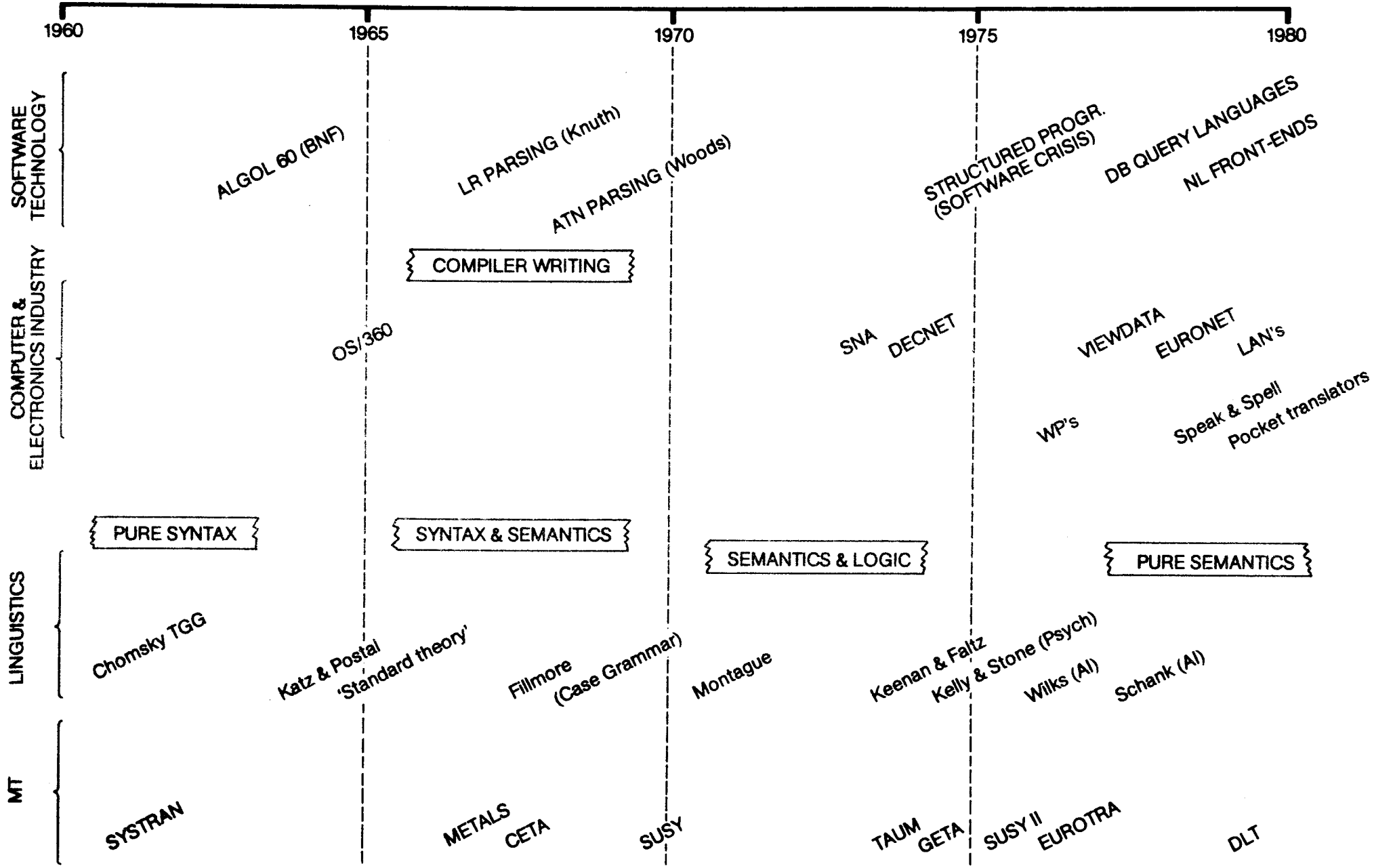
As for the interactive disambiguation dialogue, this will not be initiated before an entire input unit (a sentence) has been entered, and only after an automatic-disambiguation attempt has failed. The system must generate questions which expose the presence of alternative interpretations, without using linguistic jargon or a language other than the SL (the typist, who may be the author of the SL input text himself, is only supposed to know its context and the SL). Moreover, the analysis module will contain an algorithm to optimize the order of questions and thereby reduce their number and the load on the typist. These are non-trivial tasks for the SL-analysis module, which require a more detailed study within the time-frame of the proposed pilot project. A simulation of the interaction dialogue should be part of this study.

On a prolonged time-scale, DLT offers scope for a gradual quality improvement towards stylistically correct TL-output, by intra-IL syntactic mappings (the first system releases will produce TL-output which, though grammatically correct, still reflects the structure of the SL-input). Further, a very gradual relaxation of the interaction dialogue may be achieved by implementing macrocontext-oriented AI disambiguation algorithms, operating on the IL in connection with an IL-based knowledge-bank. This remote future development will profit from the richness (lexicon, terminologies) and compact morphematic structure of the IL (whose unorthodox internal coding accelerates all string-matching operations), and is therefore closely connected with the specific DLT design presented today.

Final conclusion.

The central issue of this study, the feasibility of creating an unambiguous Esperanto-based IL, has been answered positively. Even when taking into account the large investment and non-trivial problems to be mastered in order to develop the ultimate (SL-to-TL) DLT system, the practical significance of only a partial realization phase (IL-to-TL) already justifies continued support in the form of a pilot project.

Fig. I-1. **PARALLEL CHRONOLOGY OF SOFTWARE, HARDWARE, LINGUISTICS AND MT.**



I. INTRODUCTION.

1. Challenge and innovation.

Almost every week now we are faced with articles, reports and cover stories describing Europe's worsening position in the so-called high-technology race. In key areas such as semiconductor technology, the manufacturing and marketing of mainframe and personal computers as well as consumer electronics, Europe appears to be lagging behind the USA and Japan - and the gap still seems to get wider.

As part of its coordinated efforts to alleviate the problem, the Commission of the European Communities (CEC) is trying hard to remove those factors that are believed to impede inter-European cooperation. One such factor is multilinguality, and a number of interesting language-related projects has been supported by the CEC during the past few years:

- SYSTRAN, a partly automated translation service, accessible via EURONET, and based on extended versions of the originally American system;
- the design of a multilingual keyboard for word processors (WP's);
- international terminology standardization in various fields, e.g. trade product names (in cooperation with the Chambers of Commerce);
- research and development for a completely new computerized translation system, EUROTRA, primarily aimed at the EC organization's own needs;
- the underlying feasibility study for another new computer translation system, DLT [see fig. I-2], differently structured and originally conceived for videotex networks.

At the same time, natural-language processing emerges as a spear-head technology. Computer and database access via natural language presents itself as the only and ultimate answer to the software crisis.
The launching of the Japanese 'Fifth Generation' project, with machine translation as one of its pilot applications, has had a Sputnik-like effect on the Western computer world.

Advanced software development such as for natural-language

processing and programming seems to be one of the few areas left, where the West still has a chance in the technological race with the Far East.
And especially Europe, one would think, is predestined to take the lead in the development of multilingual systems, in such a way that the initial drawback of internal language barriers can eventually be turned into a lever for software innovation.

Translation by computers or - as it has been traditionally called - MT (Machine Translation) has always been a difficult and challenging venture. Not only its persistent lack of success, also the continuing changes in the computer environment and the accelerated evolution of linguistics during the last couple of decades [see fig. I-1] contribute to that challenge.

Under these circumstances, the support in Europe of more than one MT innovation effort, appears to make sense. As in Gilb's dual programming technique, one expects the two development

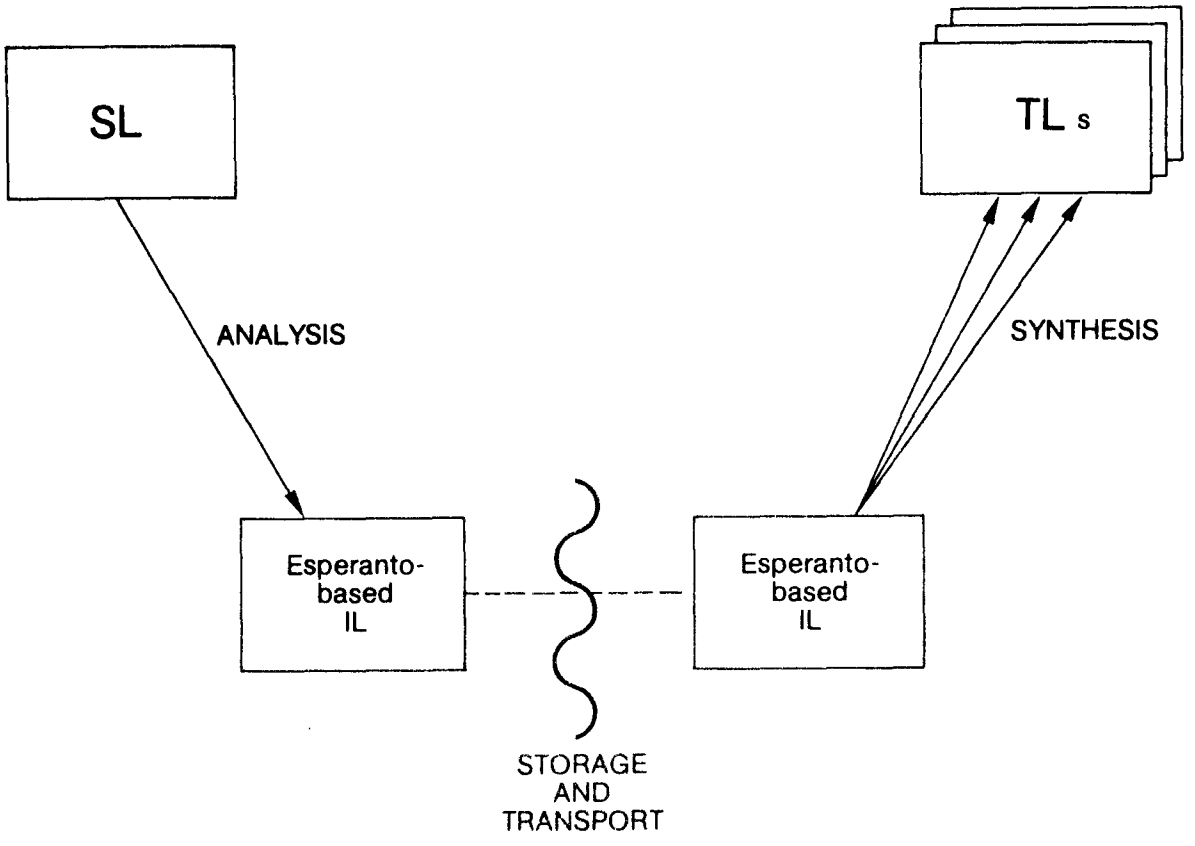


Fig. I-2. The DLT (Distributed Language Translation) system in-a-nutshell: an Esperanto-based IL (Intermediate Language) serves as a common semi-product (invisible to end users) in the translation of one SL (Source Language) to several TL s (Target Languages).

teams to reach their goal independently, by different paths. Whereas EUROTRA got off the ground in academic (largely linguistic) circles, DLT originates in the software industry. The subtle mixture of knowledge and ignorance on which the latter has been based at least guarantees an innovative approach.

2. The translation market.

In the interdependent world in which we live, translation has become a multi-billion dollar market with an annual growth rate of 9 to 10%. This was of one of the findings of a recent market investigation at the CEC's request [Van Slype, 1983; Van Dijk, 1981].

DLT was originally conceived as a translation system to be integrated within international videotex networks [Witkam, 1981], or - formulated more generally - information distribution networks. Though the present feasibility study indicates DLT's viability as a general-purpose translation system in the WP (word processor) and network environment, the system will probably first gain ground in situations where one IP (information provider) serves a large number of information consumers.

Such a situation also exists in exporting industries, where product documentation (operating instructions, maintenance manuals etc.) in various languages has become a major cost factor.
The attractiveness of computerized translation for this purpose will be boosted by new means of information distribution, such as on-line query systems and video discs. In this regard, we quote [Rolling, 1983: 173]: "...the most valuable assistance to machine translation will be provided by the new techniques of office automation and telecommunications...".

In the area of electronic publishing, including the dissemination of industry newsletters, state-of-the-art reports, scientific articles etc., the rise of new on-line full-text information systems depends on investments and moves (such as the French Hachette-Dassault fusion) in both the publishing and electronics industry, and defies market forecasts. In its wake however, transnational systems and an additional translation market will surely develop.

If TV manufacturers, eager to treat the consumer to new gadgets, succeed in selling TV s with teletex, paper output, games, stereoscopic depth-effects etc., they will also be able to create a demand for polyglot TV s.

The market investigation mentioned at the beginning of this section also revealed the existence of a latent market for low-cost 'quick and dirty' machine translation, where the output suffices to make the reader aware of what a foreign-language document is about. It must be underlined here, that DLT does NOT aim at that particular market. This should be clear already from the emphasis in this report on (public) videotex systems, international information distribution, publishing and consumer electronics. These objectives demand quality translations.

Next to information distribution, electronic mail is a potential market for computer translation. As facilities for world-wide telecommunications, personal computer networking, electronic mailbox services, videoconferencing etc. increase, the language-barrier would remain as a communications bottle-neck if there were no electronic translation.

Finally, national governments and international bodies such as the EC, the UNO etc. can have a big influence on the course and speed of electronic information distribution and automatic translation facilities. They can help to get public videotex out of its current stagnation [Maurer, 1983]. By installing advanced systems for their own needs, they can also guide third-world countries how to take advantage of advanced tele-communications and multilingual network technology, and liberate themselves from the predominance of English in the informatics age. In this respect, any recent translation market survey can only have disclosed the top of an iceberg.