[From: Internat.conf. *Methodology & Techniques of Machine Translation*, British Computer Society, Cranfield, 13-15 February 1984]

DISTRIBUTED LANGUAGE TRANSLATION, ANOTHER MT SYSTEM

A.P.M. Witkam

BSO
Utrecht, Netherlands

MT systems on the drawing board today find themselves in a totally different environment than those conceived in the largely batch-oriented EDP world of the 1960s and early 1970s. In the era of 370s-on-a-chip, wide-spread local area and international packet-switching networks, the need for a new approach, commensurate with principles of distributed processing and personal computing, becomes evident.

DLT (Distributed Language Translation) is a proposed system for semi-automatic translation between written natural languages. It was conceived and first investigated within the software-house BSO in the Netherlands, during the period 1979-1982. After that, a grant by the Commission of the European Communities enabled a thorough feasibility study, the results of which were published at the end of 1983 [Witkam, 1983b].

The DLT project, a phased and long-term undertaking, aims at economic translations between European languages (starting with French, German, English, Italian) in the first place, but promises excellent extension possibilities for other languages (Japanese, Chinese, Arabic) as well.
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 original text, DLT translations can be made reliable and grammatically correct.

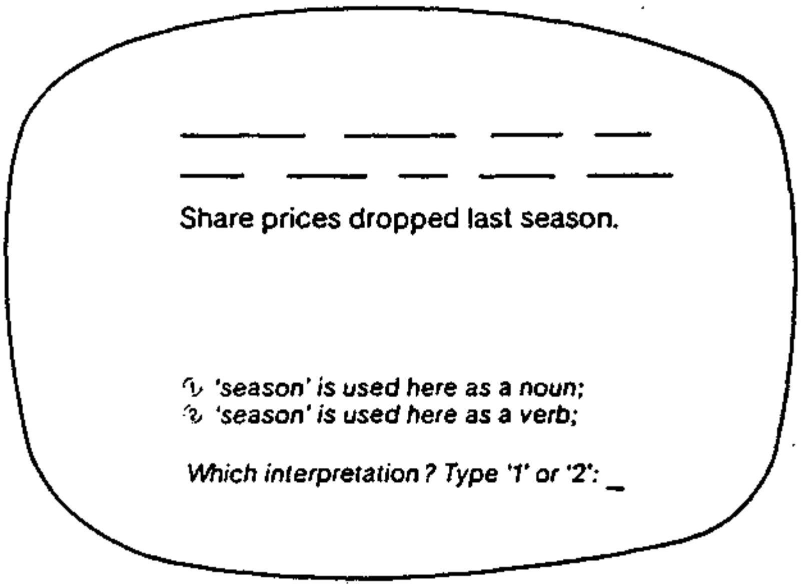
The operational environment of DLT.

DLT is a system to be embedded in computer networks and terminals. It consists of:

a. special equipment and human interaction at the sending terminal;

b. special equipment at the receiving terminal;

c. a special interface standard between these terminals.



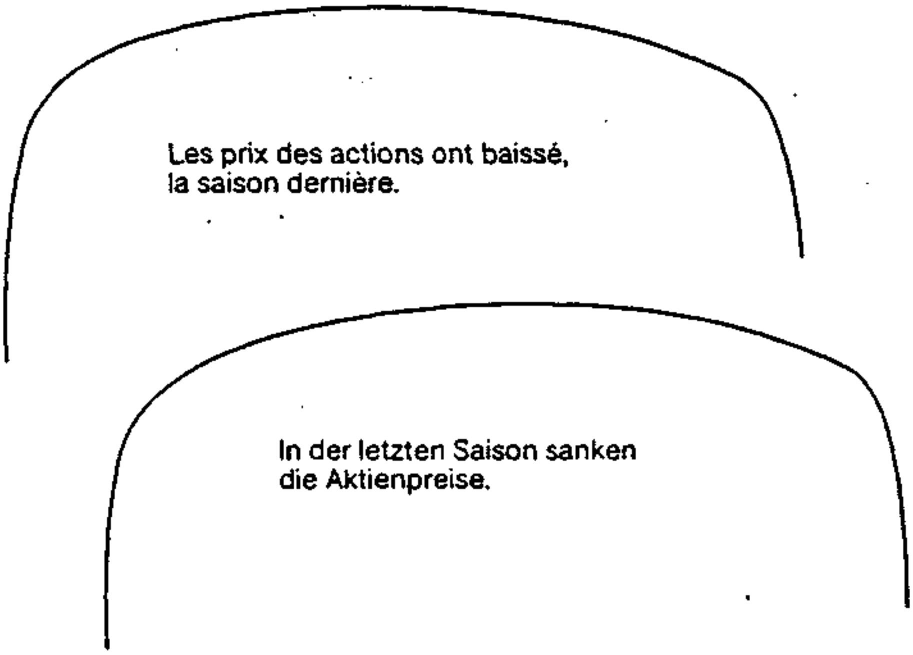
Share prices dropped last season.

1 'season' is used here as a noun;
2 'season' is used here as a verb;

Which interpretation? Type '1' or '2': _

Fig. 1a. Text entry under DLT. After a sentence has been typed, the system (italics) interrupts the typist with a request for clarification.

Fig. 1b. Automatic display of translated text at receiver terminals.



Les prix des actions ont baissé,
la saison dernière.

In der letzten Saison sanken
die Aktienpreise.

The system permits text to be entered in (for example) English at one terminal, and subsequently to be displayed in French at another (possibly remote). A third terminal might present the same text in German, a fourth one in Italian, etc.

The translation process is in fact distributed over the network: one part takes place at the sending terminal, where the person who enters the source text also has to add some text clarifications, in a computer-initiated dialogue [see fig. 1a]. The other part of the translation takes place upon reception in the receiving terminal, completely automatically and unnoticed: only the translated text appears at the display screen there [fig. 1b].

Text entry (including editing), transmission and display will be handled by the usual word processing and data communications facilities. The language translation must be regarded here as an optional extra service, compatible with general terminal and communication interfaces.

Originally, DLT has been conceived for international videotex information retrieval and information distribution systems [Witkam, 1981]. Especially in Europe, but also in other regions of the world, a future rise of public videotex information systems together with satellite TV may create new language barriers that have to be resolved. This includes subtitling of news reports, interviews, documentary films etc.

In the future videotex mass consumer market, but also in the more near and partly already existing domain of professional on-line information retrieval, the emphasis is on the receiving of information. Though the user interacts by sending an information request, the main stream of data (abstracts or full text) is towards him. On the other side, the IP (Information Provider) generates text for a multitude of customers. This situation permits relatively low-cost text receiving equipment at one side, as opposed to relatively high-cost text generating equipment at the other side. The DLT design capitalizes on this balance.

Two other key-words characterize the environment in which DLT will operate: OOF (Office-of-the-Future) and PC (personal computing). In the OOF, desktop terminals will more and more replace paper trays. Electronic storage and transmission of information over LAN's (Local Area Networks) will be commonplace. For an international or multilingual staff, the provision of such a network with DLT is an ideal addition: within the supported set of languages, anybody can enter as well as read documents in his or her own language.

The entering of text will take place on WP (Word Processor), type of equipment. Text entry on WP's has become a normal practice in today's office. In an increasingly automated world, it is a process in which human activity is required, and this will probably remain so for a few decades. Even when speech input will catch on, human guidance and correction will be an indispensable part in the total text-entering process. DLT takes advantage of the presence of a WP operator, to restrict the cost of human assistance in the translation process. This process, or more exactly the part of it within the text-generating terminal is semi-automatic. The idea now is to use the same person both for usual WP tasks (typing, editing etc.) and for the addition of text clarifications at the computer's request: the so-called 'disambiguation dialogue [fig. 1a].

Text entering under DLT does NOT require the presence of a translator at the WP, and DLT is certainly not a tool for human translators. The latter is covered by so-called CAT (Computer Aided Translation) systems, of which the Weidner system has become the best known in recent years.

At the text generating terminal, DLT only requires knowledge of the source language and understanding of the context or

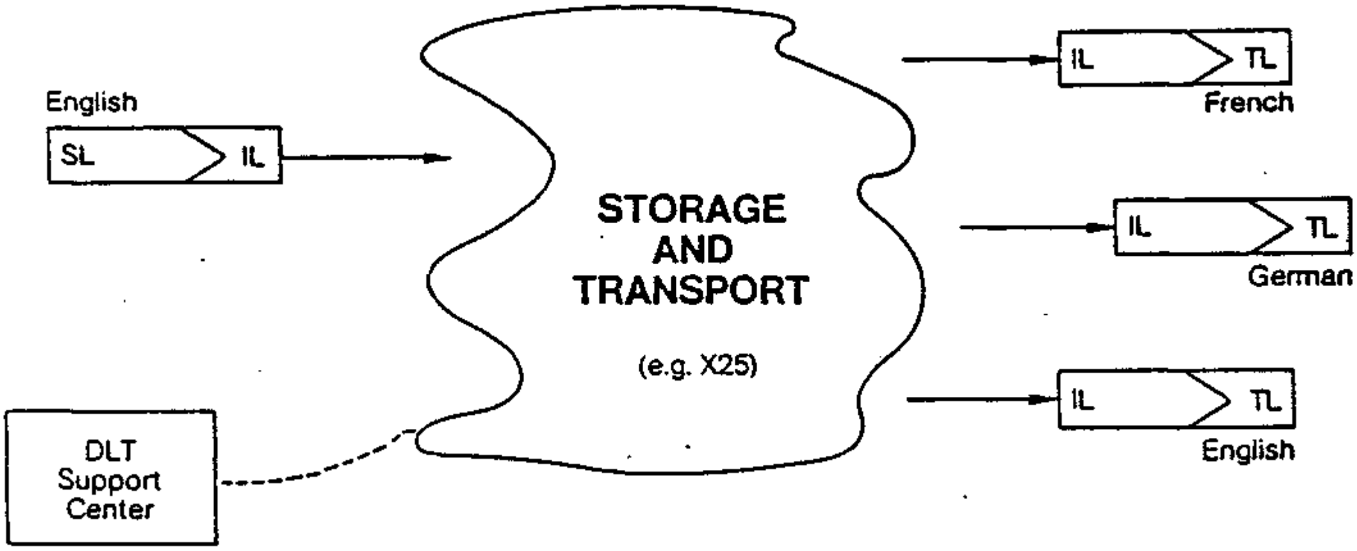


Fig. 2. DLT configuration for information distribution. The information provider is at the left side of the network. He enters his texts in English. The IL is used for all storage and transport, including long-term storage in a databank. The databank can be accessed by the information consumers at the right, where TL terminals convert the information to their home languages.

subject. When for instance the word 'bank' appears, DLT may ask the human operator for help, and he or she should be able to decide which sense ('side of a river', 'financial institution') applies. Sometimes, also basic grammatical concepts ('verb', 'noun') will occur in the man-computer dialogue [fig. 1a]. By and large, the 'disambiguation' work will be within reach of the well-educated secretary, who may experience it as a task enrichment over conventional typing work. For highly technical texts, the author of it will be the most appropriate person.

In contrast to MT systems that run on central mainframe or shared minicomputer configurations, DLT is entirely directed to the PC and communications environment, with all the required translation power distributed over the network and built-in into desktop equipment.
Fig. 2 illustrates the principal philosophy of DLT: terminals are separated by a storage and transport network, which can be thought of in abstract terms as a separation in space and time. This separation is bridged by the DLT intermediate language (IL, the interface standard between text-generating and text-receiving terminals. Storage and transmission of textual information in a multilingual environment take place in IL, a 'semi-product' of translation. The network simply passes this semi-product (no translation activity at all takes place within the network). In IL-form, text may be stored and filed temporarily or permanently, inside or outside the network, just like any other kind of computer data.

The 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 work done has been 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 inter—national development of EUROTRA, there is a remarkable difference

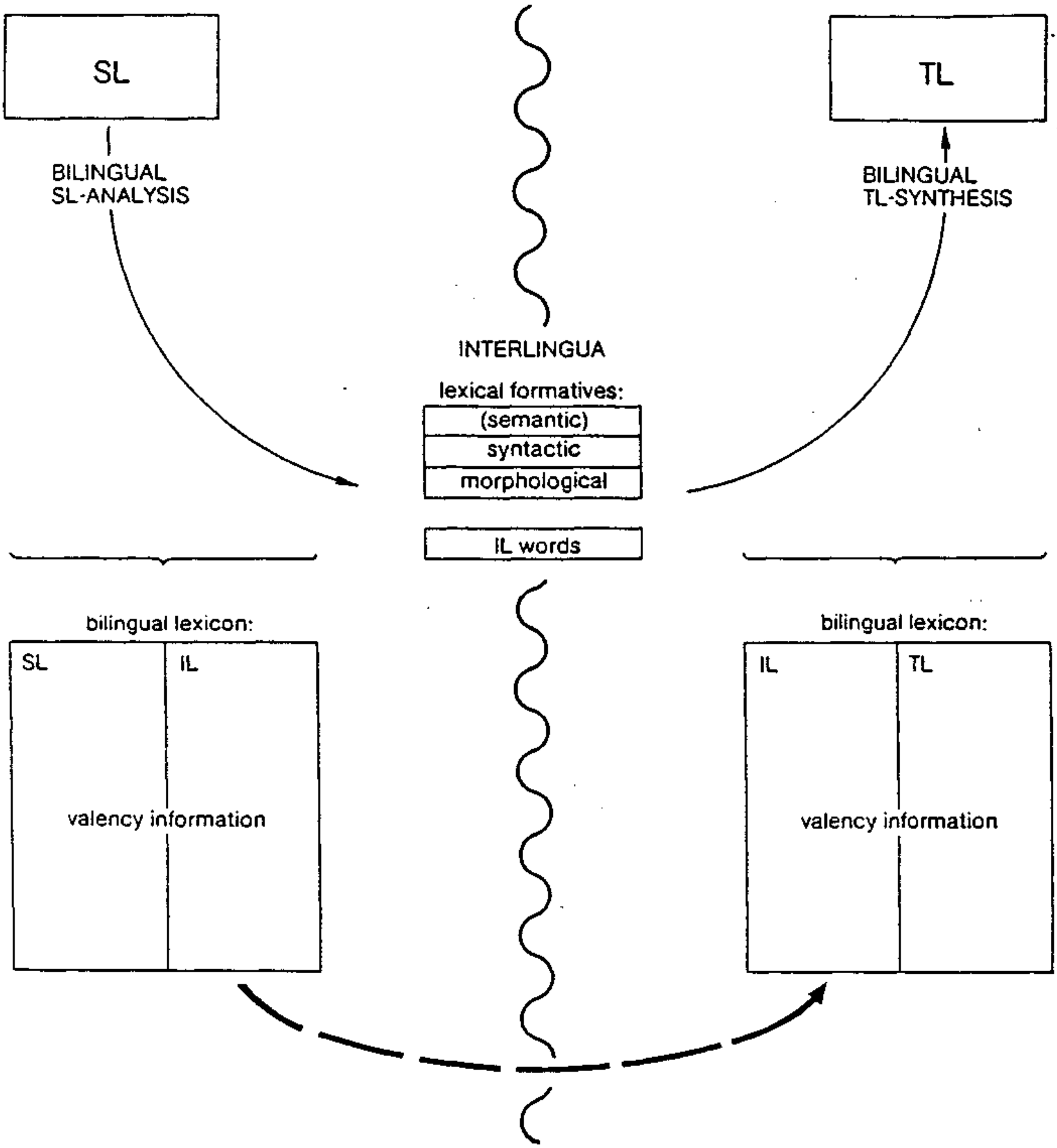


Fig. 3. Interlingual configuration, featuring an IL with lexical formatives only. The dashed arrow symbolizes the 'wide' lexical bridge that is formed by the presence of comprehensive IL dictionary columns at both sides of the interface. This characterizes DLT.

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 [fig. 3].

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 idiosyncrasies 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 of an already existing, semi-artificial language, such as:

- invariant and autonomous morphemes (Greenberg's agglutination index: 1.00),
- transparency and regularity of grammatical structure,
- a relatively precise system of prepositions.

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 evident 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 [fig. 3].

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 inspectabi1ity (for development and maintenance) and compactness (for low-cost transmission).

The unambiguity of the IL.

The main issue in the DLT feasibility study has been the unambiguity of the Esperanto-based IL, an obvious prerequisite for a fully automatic translation step from IL to TL. To this purpose, 'unambiguity' has been more precisely defined in terms of IL-parsabi1ity 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 feasibility study report [Witkam, 1983], 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,
- insertion of a universal separator element.

Special care has been taken to avoid space-consuming or obtrusive extralingual elements that could unfavourably affect the IL's compactness and inspectability.

Though it is hard to give a "proof" of the IL's unambiguity, at least at this stage of the project, its soundness has been checked on the basis of a contrastive-linguistics approach: in various areas of structural ambiguity, including notorious stumbling stones of MT (part-of-speech, function words, PP-ambiguity, verb nominalization, anaphorics, etc.), the IL's resolving power to distinctly represent the alternative readings of the SL (English, German, etc.) original.
Moreover, an algorithm for automatic separator insertion guarantees a safe handling of accidental (and therefore difficult to predict) syntactic ambiguities, and thereby secures the extendibility of the IL's ambiguity-resisting power.
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).

The 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 7 years, assuming a continuous

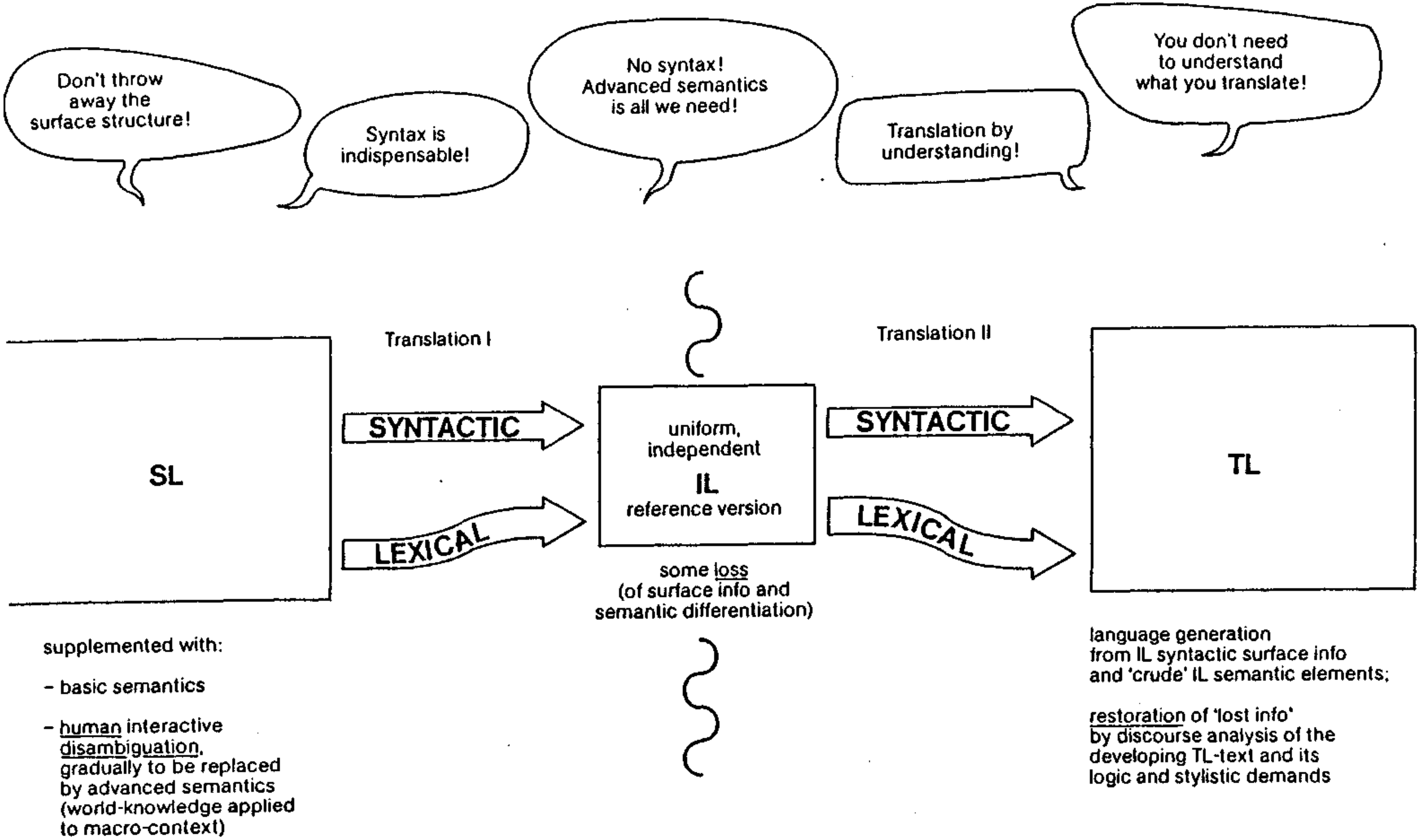

Don't throw away the surface structure!
Syntax is indispensable!
No syntax! Advanced semantics is all we need!
Translation by understanding!
You don't need to understand what you translate!
SL
Translation I
SYNTACTIC
LEXICAL
uniform, independent
IL
reference version
some loss (of surface info and semantic differentiation)
Translation II
SYNTACTIC
LEXICAL
TL
supplemented with:
- basic semantics
- human interactive disambiguation, gradually to be replaced by advanced semantics (world-knowledge applied to macro-context)
language generation from IL syntactic surface info and 'crude' IL semantic elements;
restoration of 'lost info' by discourse analysis of the developing TL-text and its logic and stylistic demands

Fig. 4. Profile and prospect of DLT's translation method against a background of conflicting views in linguistics and AI (Artificial Intelligence).

effort, phased over several pilot projects.

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 mentioned here:

- intervalwise, data-driven, single-pass LR parse, in step with the entering of words by the typist (integration of DLT into WP equipment),

- stepwise quasi-parallel creation of IL-directed syntactic SL-trees along a (moderate) number of alternative parse trails.

The accent is on fighting undeterminism by parallelism instead of backtracking. This approach is favoured 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 towards the end of the 1980s.

It should be noted that the interactive disambiguation dialogue [fig. la] will NOT be initiated before an entire input unit (a sentence) has been entered, and only after an automatic-disambiguation attempt has failed. If so, the system must generate questions which expose the presence of alternative inter—pretations, without using linguistic jargon. The analysis module will contain an algorithm to optimize the order of questions and thereby reduce their number and the load on the typist. Besides, a user-friendly dialogue will often (unlike the example of fig. 1a) necessitate automatic paraphrasing of the original clause or sentence.
These are non-trivial tasks for the SL-analysis module, which require a more detailed study within the range of a DLT pilot project. A simulation of the interaction dialogue will be part of such a 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 means of macrocontext-oriented artificial-intelligence techniques, operating on the IL in connection with an IL-based knowledge-bank [Witkam, 1983a]. 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.

Fig. 4 gives another characterization of DLT's translation mechanism, its design philosophy and the course of its future evolution.

References

Witkam, A.P.M. and Hillan,J.J. (1981). Resolving Language Barriers in International Videotex Communication. In: "New Systems and Services in Telecommunications", Cantraine G. and Destine J. (eds.), 143-153, North-Holland Publ. Co., Amsterdam.

Witkam, A.P.M. (1983a). An alternative strategy for steady growth towards high-quality translation networks. In: "Information Management Research in Europe (Proceedings of the EURIM 5 Conference)", Taylor P.J. and Cronin B. (eds.), 196-204, Aslib, London.

Witkam, A.P.M. (1983b). Distributed Language Translation: Feasibility Study of a Multilingual Facility for Videotex Information Networks. BSO, P.O. Box 8348, 3503 RH Utrecht, Netherlands.