DLT - AN INDUSTRIAL R & D PROJECT FOR MULTILINGUAL MT

Toon WITKAM
BSO/Research, Postbus 8348, NL-3503 RH Utrecht, The Netherlands

witkam@dlt1.uucp

Abstract

An overview of the DLT (Distributed Language Translation) project is given. This project is aimed at a new, multilingual MT system in the 1990s, which uses Esperanto as an internal interlingua. The system's architectural features, current progress and project organization are dealt with.

1. Introduction

DLT (Distributed Language Translation) is the name of a principle,
a design philosophy and a project. Within the area of MT, it represents another approach for steering between the hazards of low-quality output, endless prolongation of research and development time, restriction to narrowly-bounded subject fields, the geometric cost expansion when a new language is added, etc.

DLT is a concentrated high-tech effort to attain a product line of language translation modules in the 1990s. Together, these modules will constitute an interactive, knowledge-based, multilingual translation system, perfectly suited for operation on networked desk-top equipment.

DLT was conceived in 1979, in an environment with no historical ties to MT whatsoever. After patents had been applied for in 14 countries, the first publication followed at the conference on
"New Systems and Services in Telecommunications" in Liège [1980].

In 1982, the EEC granted a quarter of a million guilders for a DLT Feasibility Study, which was completed in 1983. A remarkable feature of the DLT design, highlighted in this study, was the use of Esperanto as intermediate language, with its own lexicon. This meant the adoption of an overall interlingual architecture, the most ambitious structure known for an MT system.

At the same time, the introduction of Esperanto into the MT scene of the 1980s aroused a lot of skepticism and prejudice. As it happens, this semi-artificial language (invented by an ophthalmo-logist towards the end of the nineteenth century) is not usually considered a respectable object of study among professional linguists.

2. Design philosophy

The research team at BSO considers Esperanto a valuable tool in language technology, and has motivated its use as the DLT pivot on rigorous systems engineering grounds:

- an overall interlingual architecture, i.e. an MT process of 2 main steps (instead of 3) fits extremely well into the outside operating environment, which consists of 'senders' and 'receivers' linked by a communications network; the interlingua (or Intermediate Language) is the 'semi-product' passed over the network, and should be independent of any source or target language in the system;
- the knowledge-based component of the translation process,

the world-knowledge inferencing system for resolving ambiguities is essentially language-independent and can therefore entirely be built in the interlingua; serving a multilingual system, this is an important economy-of-scale consideration;

- long-term development and maintenance of a complex translation and world knowledge system is a task that can only succeed with perfect man-machine interfaces for the system engineers; linguists, lexicographers, terminologists and other specialists must be offered quick and easy access to the heart of the translation machinery; this calls for an interlingua that is directly legible;

at the same time, the interlingua should be lexicologically autonomous and well-defined, the former eliminating the need

for re-paraphrasing in other languages, the latter being a prerequisite for distributed system development (language teams working to and from one common interlingua); Esperanto meets these requirements.

3. Prototype construction

In 1984, BSO set up a plan for a 6-year research and development project (75 person-years at the cost of 18 million guilders), aimed at a DLT prototype capable of translating at least one language pair (English-French). This plan received the support of the Ministry of Economic Affairs of the Netherlands, which granted an innovation subsidy of 8 million guilders. The first half of this 6-year schedule has now been completed.

A first prototype of DLT was shown to the press in December 1987. Though operating only slowly as yet, with a small vocabulary (2000 English words) and a restricted grammar, this laboratory model shows the various monolingual and bilingual processing steps of DLT in proper sequence [see also Fig. 1]:

1. Exhaustive parsing of the English source text. Two different parser implementations have been realized in the search for the fastest formalism: one is based on ATNs and BSO's graphic software environment (on SUN 3/50 workstations) developed for setting up, testing and optimizing ATNs, the other is based on APSG and the PARSPAT software system from the University of Amsterdam [Van der Steen, 1987].

The parsing process in DLT is breadth-first, syntax-only, and delivers dependency (not constituency) trees.

2. Surface translation (first half). Contrastive syntactic rules between English and Esperanto are applied here. This system of bilingual rules (250 at present) is based upon dependency grammar formalizations of both languages. The methodo-logical framework has been inspired by the work of the French linguist Tesniere and is comprehensively described in [Schubert, 1987]. Semantic considerations are disregarded systematically at this stage. The result is a (sometimes large) number of 'formally possible' parallel translations.

3. Main semantic analysis, entirely carried out in the Intermediate Language, by searching through a knowledge base of some 75.000 (present status) semantically related Esperanto word pairs, and by applying text-grammatical principles of cohesion etc. to the intermediate stage of the translated text [Papegaaij, 1986 and 1988].

This automatic disambiguation system, written in Quintus PROLOG, now largely serves as a rating (pre-ordering) of parallel surface translations, prior to the disambiguation dialogue which follows it. The DLT design offers a long-term perspective for steady improvement of this probabilistic component, ultimately by machine learning.

4. Disambiguation dialogue. The user is prompted to make a choice out of the possible interpretations listed on the screen. Note that these are parallel surface translations, back-translated ('paraphrased') into the source language. For the user, the disambiguation dialogue is a strictly monolingual affair, and free of linguistic jargon. In the present realization of the DLT prototype, mainly lexical ambiguities can be displayed.

5. Surface translation (second half). As Step 2 above, but now between the Intermediate Language and French. Some 500 contrastive syntactic rules have been implemented so far. Though the proliferation of parallel translations is less at this side of the translation process (due to the syntactic unambiguity of Esperanto and its lack of homonyms), it is not absent. If the target language happens to have a more refined "cutting-up- of-reality" in some concept area (like the proverbial 10 words for 'snow' in Eskimo), parallel translations will result. All the results of this step are in the form of dependency trees.

6. Additional semantics. TL-specific selection criteria are applied to select the right word. But because these criteria are knowledge-based (we are not talking of idiomatic phenomena), they are restated in terms of the IL, and the selection process is carried out on the intermediate stage of the translated text, using the Esperanto knowledge bank again. If the context does not provide enough clues, a default choice (e.g. the least specific word for 'snow') will be made. In contrast to the source language half of the system, there is no possibility for human intervention here.

7. Synthesis of the target sentence. In this tree-to-string conversion, the TL-specific word order is determined (including the application of elision and contraction rules).

4. Project outlook

BSO is now in the process of preparing for the next phase of the project (budgeted at 12 million guilders), in which the emphasis will be on large-scale dictionary and knowledge-base expansion, and relaxation of grammar restrictions.

Work done on DLT in the past 5 years confirms the feasibility of its architecture and its instrumental use of Esperanto. Some of the modifications to Esperanto thought necessary [Witkam, 1983] in its pivotal MT function appeared to be unnecessary as work on DLT progressed, i.e. the Intermediate Language is closer to Esperanto now than it looked like in the beginning. Criticizers' predictions that the IL would keep changing and would drift further and further away from Esperanto, have not been borne out.

The essence of DLT is not so much an attempt at an unambiguous interlingua, but rather: a split of the overall translation sequence into a form part and a content part, in which the former is arranged
as a double-direct and the latter as an interlingual MT process.
The double-direct process is the surface translation referred to above (Steps 2 and 5), which could be loosely characterized as "dumb syntax". The interlingual process corresponds to Steps 3 and 6 above, and contains all the semantics (including knowledge-based inferencing).

Meanwhile, the share of IL-based or IL-directed work in the overall translation sequence - form part as well as content part - has increased considerably, compared to the initial design of 1982 [see Fig. 2]. In the form part, which is bilingual and purely syntactic, the Esperanto IL plays the role of "metataxis partner" for every source and target language ('metataxis' is the contrastive-syntactic transformation of dependency trees). The content part is monolingual, i.e. the semantics is a question of IL-only. It goes without saying that such a design can only succeed by virtue of the fact that Esperanto is a well defined language on its own, with a well defined syntax and lexicon, and with the help of project staff fluent in or at least conversant with that language.

The logistics of DLT development draw heavily upon the existence of Esperanto resources in the widest sense: linguists with Esperanto training, Esperantists with degrees in languages or with extensive translator's experience, corpora of modern Esperanto texts, etc. Timely availability of these resources in sufficient quantity and quality demands some special organizational and promotional activity from the DLT entrepreneurs.

This means crossing international borders, including the border between EEC and COMECON countries. East European countries have a relatively large base of Esperanto speakers, and much on-going activity. The Budapest Eotvos Lorand University has a chair in Esperanto. In Poland, a new Esperanto center connected with the University of Bjalistok has been founded in 1987. In Bulgaria an international training center exists, and in Prague an International Center for Esperanto Terminology has recently been established.

The challenge of the DLT project is therefore as much an organizational as a technological challenge. Cooperation with Hungarian, Finnish and Czech linguists has already begun, and preparatory work has been arranged with collaborators in East Asian countries. Diversity of language types is a deliberate aim
for the multilingual DLT system of the 1990s, and the designed architecture makes it technically feasible. It can be hoped that, against the background of 'glasnost' and 'perestrojka', fruitful and effective East-West cooperation may add to the success of a system for international use and of general interest to the growing community of computational linguistics students and researchers.

REFERENCES

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

Papegaaij, B.C., & Schubert, K. (1988): Translating Text Coherence. Dordrecht/Riverton: Foris.

Schubert, K. (1987): Metataxis. Contrastive dependency syntax for machine translation. Dordrecht/Riverton: Foris.

Van der Steen, G.J. (1987): A Program Generator for Recognition, Parsing and Transduction with Syntactic Patterns

[dissertation] University of Utrecht.

Witkam, A.P.M. & Hillan, J.J. (1980): Resolving Language Barriers in International Videotex Communication. In: *New Systems and Services in Tecommunications*, Cantraine, G. & Destine, J (eds.). Amsterdam: North-Holland, pp. 143--153.

Witkam, A.P.M. (1983): Distributed Language Translation: Feasibility study of a multilingual facility for viedotex information networks. BSO, Utrecht.

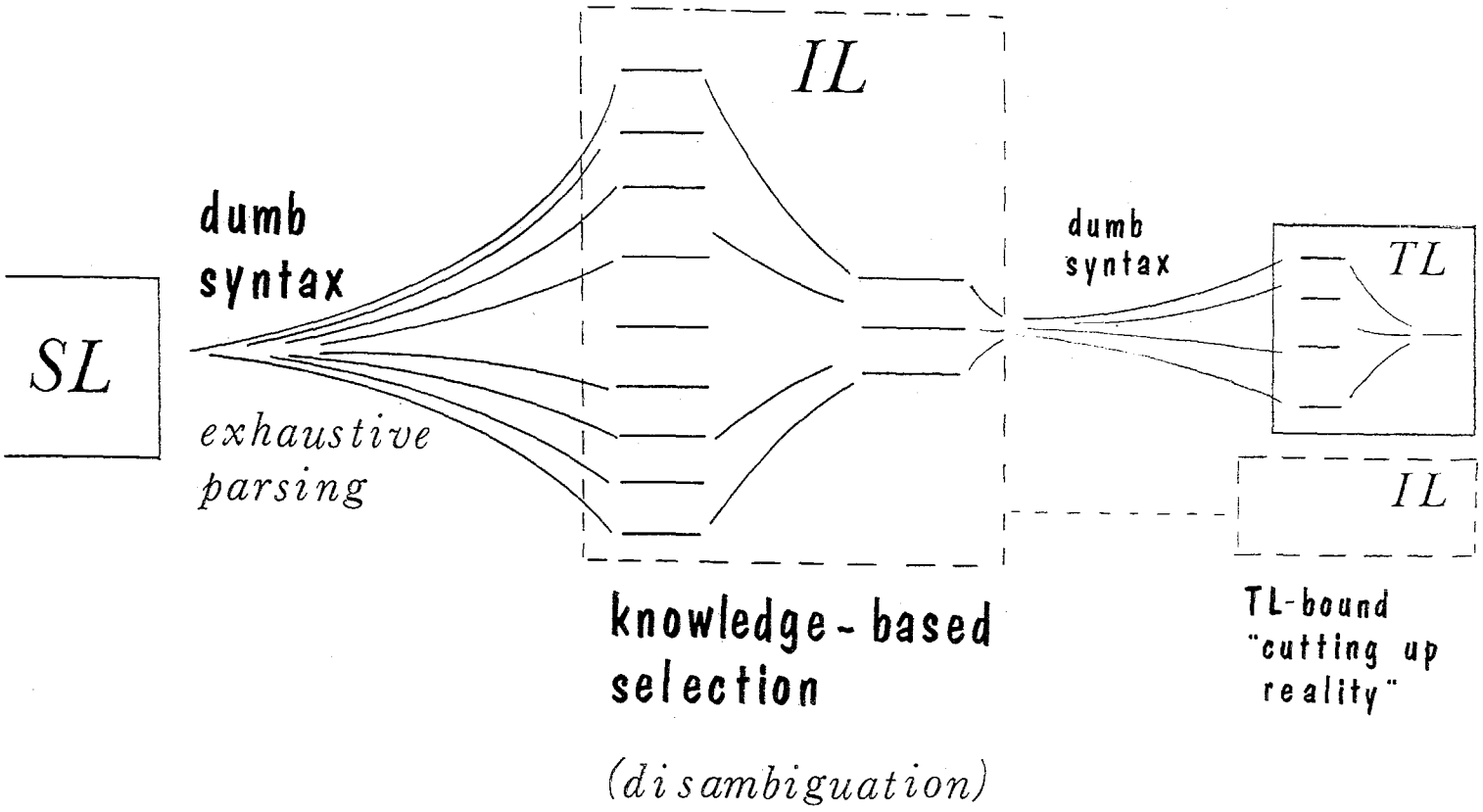

FIG. 1. Schematic visualization of the overall DLT translation process.

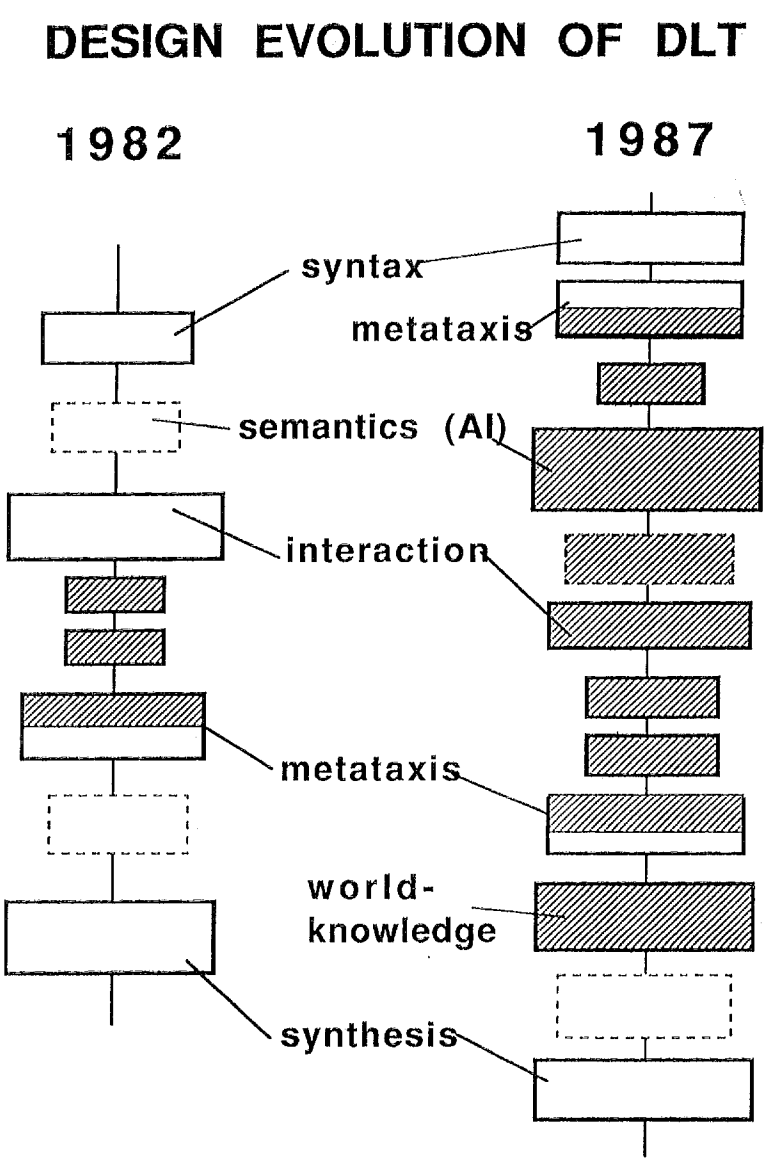

FIG. 2. An impression of the sequence of process steps for the 1982 and the 1987 designs of the DLT translation from source to target language (top-to-bottom). The shaded parts indicate where the Intermediate Language is involved.