Chapter 16: Other interlingual projects since 1975

In recent years the interlingual approach is most often associated with semantics-based projects, and these have usually been inspired by AI methods of analysis, particularly semantic parsing. Interlingual approaches based on older traditions have, however, continued. The concept of intermediary representations formulated in terms of logical formulae (present for example in the CETA and LRC concepts, Ch.10 above) is to be found in the SALAT project at the University of Heidelberg and in the Philips project at Eindhoven in the Netherlands. The argument for Esperanto as a MT interlingua has been revived in the ambitious multilingual project in Utrecht under the direction of Toon Witkam. And there are other conceptions of 'intermediary' representations to be found in a number of interactive systems (Ch.17 below)

16. 1: University of Heidelberg (1973-

The SALAT (System for Automatic Language Analysis and Translation) project began in 1973 in the Institut für Angewandte Sprachwissenschaft of the University of Heidelberg. The project is supported by the Deutsche Forschungsgemeinschaft (Sonderforschungsbereich 99 'Linguistik'). The SALAT design is an interlingual MT system with recourse to a knowledge database and inference rules. The system does not, however, have an interlingual lexicon. As in CETA (Ch.10.1), its interlingual features are restricted to syntactic and structural relations. Lexical conversion is through a SL-TL dictionary of equivalences. SALAT is similar to CETA also in that its interlingual syntax is based on logico-semantic foundations.

The system has the following stages (Hauenschild et al. 1979; Maier 1978): a two-step surface structure analysis of SL input (i.e. morphological analysis and syntactic analysis); transformational analysis into SL deep structure representations; disambiguation of SL structures (by reference to the knowledge data base); creation of TL deep structures (by reference to SL-TL dictionary and to 'deduction rules' for SL-TL structural transfer); and transformational synthesis into TL surface structures. The transformation of structures is implemented by tree transduction rules (Ch.9.14).

Though having logico-semantic representations, SALAT has also AI features. These are located in the application of 'knowledge-base' disambiguation and in the use of 'deduction rules' for SL-TL transfer. Interlingual representations are formulated as logical formulae, e.g. for *Das alpine Gebiet ist klein* (The Alpine region is small) the SL (German) representation is as shown in Fig. 36.

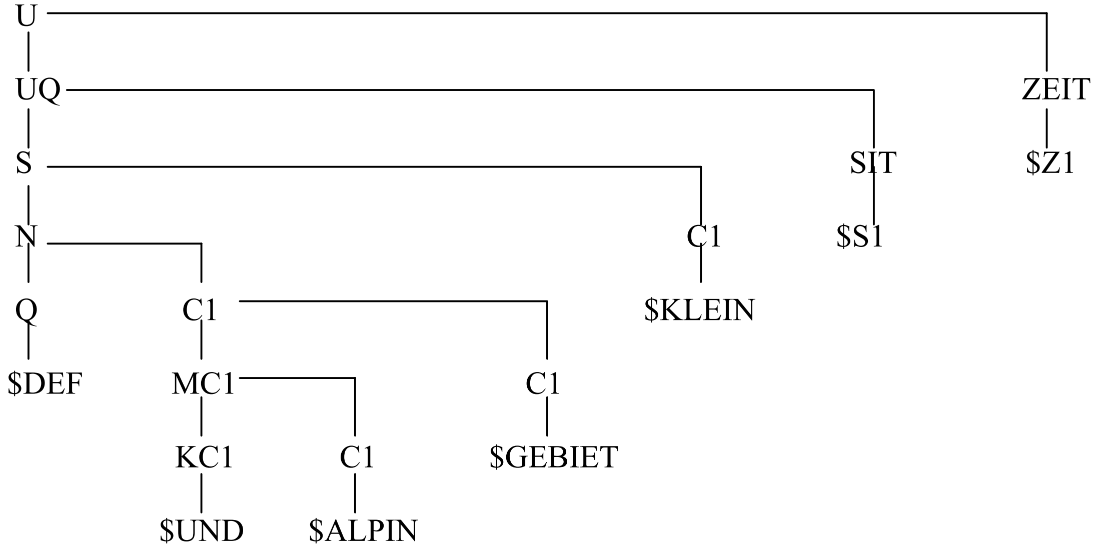

Fig.36: SALAT interlingual representation

For the translation of this structure into French, where *Gebiet* has to be rendered by *région* (geographical area) rather than *domaine* (scientific field), the transfer program refers to a definition of $ALPIN. This (logical) formula states that ‘Everything Alpine is something geographic’, from which it deduces that $GEBIET is also geographic and the TL substitute must be $REGION.

In SALAT the clear objective is to devise logical formulae both for ‘deep structure’ interlingual representations and for ‘knowledge base’ representations, on which can operate logical deduction rules for disambiguation and for SL-TL transfer. The logical foundations are the work of Suppes, Montague and Cresswell (Maier 1978).

16. 2: Philips Research Laboratories, Eindhoven (1980-

The semantic theory of the philosopher Richard Montague (1974) forms the foundation of the experimental MT system ‘Rosetta’ under development by Landsbergen (1981, 1982) and his colleagues at the Philips Research Laboratory. A ‘Montague grammar’ defines a language by specifying a set of expressions and their grammatical categories, and a set of syntactic rules prescribing how these expressions may combine to form new expressions and what the grammatical category of the new expression will be. Thus, an expression x of category N may combine with an expression y of category ADJ to form xy of category NP. The ways in which expressions are derived are represented by derivation trees (D-trees). For each syntactic rule a Montague grammar provides a corresponding logical composition rule which shows how the meaning of the expression constructed is derived from the meanings of its constituent expressions.

The assumption may be made that equivalencies can be established between the derivation trees (and the corresponding logical composition trees) of different languages, e.g. between the simple English derivation above and one in French for combining nouns and adjectives (in this sequence) into NPs. These equivalencies may in turn be represented by a ‘logical derivation trees’ which may serve as ‘intermediary’ representations. Fig.37 gives an example from Dutch and English (Appelo & Schenk 1985).

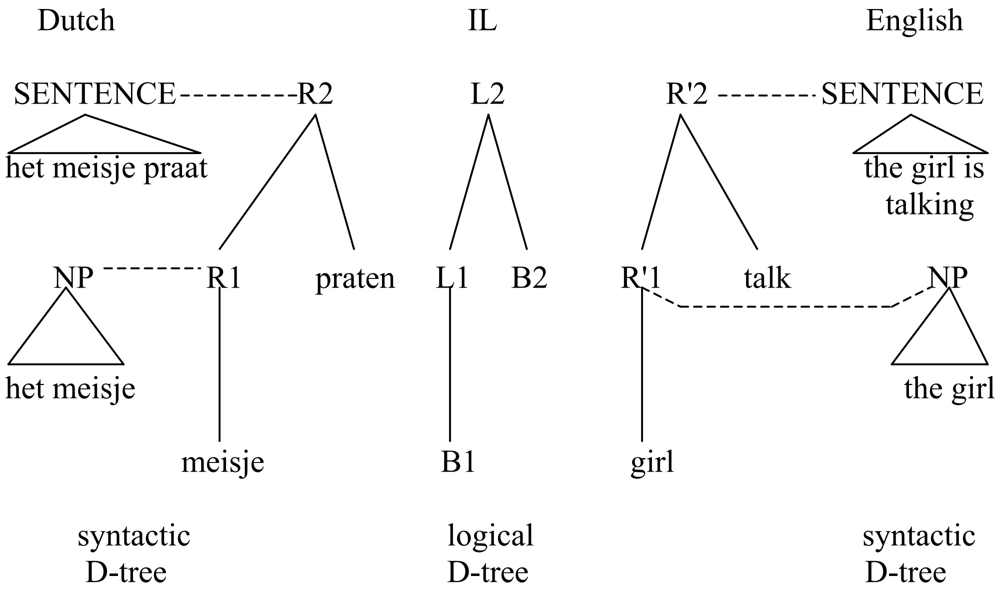

Fig.37: Rosetta representations

The aim of the Rosetta MT project is to explore this possibility in the construction of a multilingual system. Whereas in SALAT the logical formulae themselves are interlingual, in Rosetta it is their derivation trees which mediate between languages. The reversibility of the Montague grammar formalism means that grammars can be used both for analysis and for synthesis. As in SALAT, the scheme is highly speculative and as yet far from even experimental implementation. Initially the system will be designed for English and Dutch, with Spanish as the next language. Appelo and Schenk (1985) have begun by looking at the problems of idioms and tenses.[1]

16. 3: Buro voor Systeemontwikkeling, Utrecht (1979-

The idea of using an international artificial language as an intermediary language in an 'interlingual' MT system has been put forward on a number of occasions during the history of MT research. The most common choice has been Esperanto, not surprisingly in view of its widespread acceptance (Large 1985). Its supporters claim that Esperanto combines the regularity, consistency and 'universality' required of a MT interlingua with the flexibility towards new technical and scientific vocabulary which is the attribute of a 'living' language.

Despite its obvious attractions, Esperanto has rarely been studied in MT projects. The Saarbrücken team investigated the problems of translating into Esperanto as a TL (Ch.13.2), but only one MT project has so far studied the possibilities of using Esperanto as a MT interlingua. This is the recent project of A.P.M. (Toon) Witkam, a senior consultant at the Buro voor Systeemontwikkeling in Utrecht, the Netherlands. Preliminary investigations began in 1979. Witkam has presented his ideas at a number of conferences (Witkam 1981, 1984) and at greatest length in the report of a feasibility study (Witkam 1983) supported during 1982-1983 by a grant from the Commission of the European Communities. The favourable reception of this report influenced the Dutch Ministry of Economic Affairs in granting a six year contract (1984-1990) for the continuation of the project. A team of ten under the leadership of Toon Witkam has been set up, whose initial task will be to produce a pilot operational prototype for translating from English into French using Esperanto as the intermediary language (*LM* 19, Apr 1985, p.5). The long-term aim (Witkam 1984) is a multilingual system for translating between European languages of the Community (French, German, English, Italian), with eventual extensions to other languages (Japanese, Chinese, Arabic).

Witkam refers to his proposed MT system as Distributed Language Translation (DLT). The overall system would permit text to be entered in one language at a terminal, and displayed in other languages at remote terminals. The translation process would thus be distributed over a network: conversion of SL text into the interlingua (Esperanto) would take place in one location, and conversion into the TL text would take place in another part of the network. Like other current projects, DLT is intended as a practical economically viable system for translating 'informative' literature (abstracts, manuals, reports, regulations) with little attention to stylistic subtleties. It is also envisaged as a system necessarily involving on-line clarification of ambiguities, etc. through computer-initiated interaction at terminals with persons sending SL texts. DLT is not designed as a tool for translators; it requires only that those entering SL texts have sufficient knowledge to resolve ambiguities in the text; it does not expect them to suggest TL equivalents. The system is conceived within the framework of an international videotex information network. Witkam emphasises the importance of a good communication environment, word processing facilities and personal desk-top computers, in which DLT becomes a realistic proposition.

Most MT interlinguas and 'interface representations' are characteristically tree representations with complex labels and abstract formatives, e.g. the SUSY, GETA and Eurotra representations (Ch.13.2, 13.3, 14.2). By contrast, Witkam's DLT interlingua will be more like a natural language, "a linear string of lexical formatives". It will also be a true interlingua; whereas the 'pivot language' of CETA (Ch.10.1) and the intermediary representations of LRC (Ch.10.3) were interlingual only in syntactic structures and for lexical transfer the systems relied on bilingual

[1] A full account of the project was published as: M.T.Rosetta, *Compositional translation* (Dordrecht: Kluwer, 1994). See also chapter 16 of W.J.Hutchins and H.L.Somers, *An introduction to machine translation* (London: Academic Press, 1992).

dictionaries, the use of Esperanto makes the DLT system fully interlingual in both syntax and vocabulary. The advantages of adopting as MT interlingua an “existing universal auxiliary language” are claimed to be twofold: as a (semi) ‘natural’ language (used as an instrument of human communication since 1887), Esperanto has a richness and flexibility surpassing constructed logical interlinguas (cf. Andreev’s remarks on interlinguas in Ch.8.11 above), and it provides a ready-made standardised vocabulary based on common Indo-European roots. It is conceded, however, that Esperanto cannot function as a MT interlingua without some modification. Although an artificial (constructed) language, Esperanto exhibits such ‘natural’ features as homonymy, lexical ambiguity and structural imprecision. An important component of the feasibility study was therefore an examination of the modifications which were necessary to enhance Esperanto’s suitability as a MT interlingua.

Remedies for some of Esperanto’s known deficiencies are already available. For example the homonymy of *marko* (brand or stamp) can be avoided by using *fabrik'marko* (brand) and *post'marko* (stamp). The syntactic ambiguity arising from the multiple uses of *de*, which (like its French correspondent) may indicate commencement, possession, or agency, can be circumvented by the use of more precise prepositions: *disde*, *al*, *far*. These would also in part reduce the familiar problems of analysing prepositional phrases. With the aim of greater regularity and unambiguousness Witkam proposes further modifications. These include the strict prescription of word order (basically: subject verb object); the introduction of a limited number of new function words (e.g. for *whereas* and *besides*); the consistent use of punctuation; and the inclusion of special markers to indicate the antecedents of pronouns and the scope of coordinators.

A more serious defect of Esperanto, however, which cannot be so easily overcome is its lack of technical vocabulary. Esperanto permits national speakers to coin their own specialised terms; there is no standardisation. The DLT project is to take a “pragmatic” approach, adopting the form common to at least two of the languages English, French and German. In effect, the project will be building an interlingual dictionary for international technical terms from scratch, with all the dangers of pragmatic adhocness. The difficulties of terminological standardisation would appear to have been minimised in the interests of initiating the DLT prototype project.

Since the DLT system is intended to be fully interlingual (in lexicon as well as syntax) and the interlingua, ‘modified Esperanto’, is not an abstract representation but a regularised language, SL analysis and TL synthesis represent in effect two ‘translation systems’: from the SL to Esperanto and from Esperanto to the TL. Although described as an ‘interlingual’ MT system, DLT is (as Witkam (1983) readily acknowledges) in fact a network of bilingual MT systems with ‘modified Esperanto’ at the centre. Only the economies of a full multilingual system can justify the added complexity.

Each bilingual MT system comprises a separate set of analysis and synthesis programs. Witkam envisages the systems for conversion of SL texts into Esperanto texts as versions of ‘direct’ MT systems, while the systems converting Esperanto texts into TL texts would be basically designed on the ‘transfer’ principle. Only the SL-Esperanto ‘direct’ systems have been described in any detail (Witkam 1983: III.50-97). This part of DLT is semi-automatic, operating via computer-initiated interactions. It is intended to implement a single-pass SL-parser; words will be sought in SL dictionaries during text input, syntactic analysis will operate word by word (an ATN parser, with a ‘moderate’ degree of parallel parsing), input errors (e.g. misspellings) will be corrected automatically as far as possible, semantic disambiguation will occur at ends of sentences first by reference to lexical and valency information and then by interactive consultation with the person entering the text (in a manner similar to that of the MIND system, Ch.17.8) and eventually, it is hoped, by reference also to a ‘world knowledge’ database. The ‘direct translation’ strategy is evident from the explicit orientation of SL analysis to the lexical and structural features of the Esperanto interlingua (e.g. an English *-ing* clause is analysed as a pronoun and finite verb; English words are treated as homonyms only if there is more than one possible Esperanto output, etc.) It is evident also in the preservation of SL sentence styles in the Esperanto interlingual string, and therefore subsequently in TL sentence output, e.g. French *Ils traversèrent la rivière à la nage* will appear in English as *They crossed the river swimming*. In the long term, it is hoped that DLT will incorporate conversion modules within the interlingual stages to reflect TL stylistic preferences, and produce the more idiomatic English *They swam across the river*.

The Esperanto-TL 'transfer' system (intended to be fully automatic) is sketched in less detail. Most attention is paid to the details of the more regular syntax and morphology of the 'modified Esperanto' interlingua and its ATN formalisation (Witkam 1983: IV.28-111). Three stages are envisaged: conversion of Esperanto strings into tree representations; bilingual transfer (tree transduction and lexical substitution); and TL synthesis from tree representations.

A striking feature of the DLT process is the amount of tree-string and string-tree conversion: SL strings to IL-oriented trees; IL trees to Esperanto strings; Esperanto strings to IL (Esperanto) trees; IL trees to TL trees; and TL trees to TL strings. The inclusion of an Esperanto string representation may appear to sceptics to introduce unnecessary complexity: why cannot the IL-oriented trees from SL analysis be directly converted into TL trees without going through the intermediary stages of IL strings and IL trees? In other words, why not a 'transfer' system with Esperanto-inspired 'interface' representations? The fact that the DLT interlingua Esperanto, being a language in its own right, is much more accessible (to system developers as well as users) than the abstract interlingual representations of other MT systems, would seem not to be in itself sufficient justification for the greater complexity of processing.

A key feature of DLT is the emphasis on the application of computational techniques which are well tested; hence the selection of ATN parsers and an orthodox approach to design features. Slightly more innovative is the choice of Prolog as the programming language (still relatively new in natural language processing). In general, DLT is not intended to be an experimental system as regards its MT techniques and strategies; it is experimental principally in the exploration of an interlingual system based on a refined Esperanto. Future developments will indicate whether Witkam's confidence in the flexibility and regularity of Esperanto is justified or whether the sceptical predictions of DLT's excessive complexity are fulfilled.[2] Whatever the ultimate success of DLT, it represents so far the only MT project incorporating a full-fledged interlingua; previous efforts have either been semi-interlingual, primarily syntactic (e.g. CETA and LRC) or primarily lexical (e.g. the CLRU thesaurus), or have not been implemented (e.g. the 'meaning-text' model). In this respect the DLT multilingual project is as ambitious and innovative as the Eurotra project.

[2] Reports on the DLT project were published from 1986 to 1989 by Foris Publications (Dordrecht) in a series entitled Distributed Language Translation. See also: K. Schubert 'Linguistic and extra-linguistic knowledge', *Computers and Translation* 1 (3), 1986, pp. 125-152; and chapter 17 in W.J.Hutchins and H.L.Somers, *An introduction to machine translation* (London: Academic Press, 1992).