Chapter 10: Interlingual systems, 1965-1975

10.1: Centre d'Etudes pour la Traduction Automatique (CETA), University of Grenoble (1961-1971)

As we have seen (ch.5.5), the group at the University of Grenoble was set up in 1960 under the sponsorship of the Centre Nationale de la Recherche Scientifique, initially in conjunction with a MT group in Paris. In the following ten years it developed the interlingual approach in a system for Russian-French translation of mathematics and physics texts, which was tested from 1967 to 1971 on a corpus amounting to 400,000 words. In addition some trials were made with the system on SL texts in German and Japanese. (The definitive source for CETA is Vauquois 1975)[1]

The central feature of CETA was a 'pivot language' (Veillon 1968), Vauquois 1969, Vauquois et al. 1970, Vauquois 1971), an artificial language free of the morphological and syntactic constraints of natural languages. An early description of the formalism was given in 1962 by Vauquois (1966), but full elaboration did not commence until 1966. (The mathematical properties of the formalism were described at length by Veillon et al. 1967.) The formalism was designed primarily as an interlingua for syntactic structures, i.e. as the common 'deep syntactic' base of the languages in the system. (In fact, CETA's deep syntax went further than Chomsky's notion of 'deep structure' and represented semantic relationships, as we shall see.). Its lexicon, however, did not represent a common base; instead the pivot language conjoined the lexical units of whichever two languages were being processed (usually Russian and French). In other words, while the CETA pivot language was a true interlingua in syntax it was a bilingual 'transfer' mechanism in lexicon. Further, it was not intended that all sentences with the same meaning would be analysed as (or generated from) one unique pivot language representation. Nevertheless, although there were thus as many 'pivot languages' as there were SL-TL pairs analysed, all shared the same syntax and in this respect CETA considered their formalism as a first step in the direction of a 'universal language' (Vauquois 1975).

Analysis and synthesis in CETA proceeded in clearly separated stages. After pre-editing, Dictionary lookup (based on Lamb's approach, Ch.4.10) identified word-stems and affixes. The next stage, Morphological analysis, eliminated unacceptable stem-affix groupings, e.g. the segmentetation of *habilité* as HABILIT+É would be accepted (past participle of HABILITER) but not as HABIL+ITÉ, since the dictionary would record -ETÉ as the nominalizatioin suffix for HABILE. Morphological analysis was defined formally as a finite state grammar (cf.3.4), and in addition, analysis rules were coded directly into the program, i.e. not located in separate tables; consequently, as in all programs which embed grammar rules in algorithms, the morphological analysis programs of CETA were difficult to change; however, their merit was much faster processing (Vauquois 1975).

Syntactic analysis was in two stages. The first stage was a phrase-structure analysis (a context free grammar, using the Cocke parser). It produced the familiar structures of nominal and verbal phrases, and it included identification of discontinuous elements (e.g. *look... up, take... away*) triggered by information attached to the particles (*up, away*). For example, a partial phrase structure analysis of *Le remplacement des tramways par des autobus a permis un développement rapide de la circulation dans les rues de la ville* (from Vauquois 1971) is shown in Fig.16.:

[1] Vauquois' publications have been collected in *Bernard Vauquois et la TAO, vingt-cinq ans de traduction automatique: analectes. – Bernard Vauquois and machine translation, twenty-five years of MT: selected writings*, ed. C. Boitet (Grenoble: Association Champollion & GETA, 1988). See also: C. Boitet 'Bernard Vauquois' contribution to the theory and practice of building MT systems: a historical perspective', *Early years in machine translation: memoirs and biographies of pioneers*, ed. W.J. Hutchins (Amsterdam: John Benjamins, 2000), 331-348.

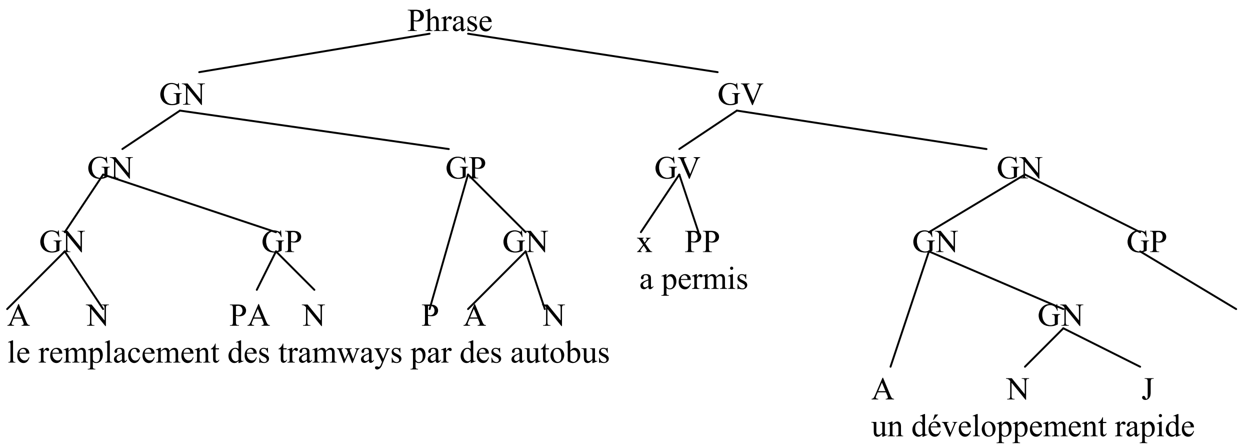

Fig.16: CETA surface syntactic analysis

In the second stage, the 'surface syntactic' structure was augmented by dependency relations, e.g. in a clause the verb was marked as 'governor' and the noun phrase as 'dependent'. Where a sentence could have two or more syntactic interpretations, the system produced a 'surface syntactic' analysis for each. In this dependency tree representation, analyses entered the Transfer stage where they were converted into 'pivot language' representations. These representations were in propositional logical form, consisting of predicates and their arguments or (in Tesnière's terminology 'actants'). To this end, lexical units were classed as either predicatives or non-predicatives: predicatives included adjectives and adverbs as well as verbs and non-predicatives were nouns and articles. In the 'pivot language' representations the arguments of predicatives could be either non-predicatives or other predicatives. Transformation into pivot language syntax involved therefore the 'semantic' (propositional logical) analysis of dependency relations and the removal of word-classes (GN (=noun phrase), GV (=verb phrase), N, J (=adjective), etc.), with the result producing an abstract tree representation such as Fig.17 (Vauquois 1971) for the sentence above:

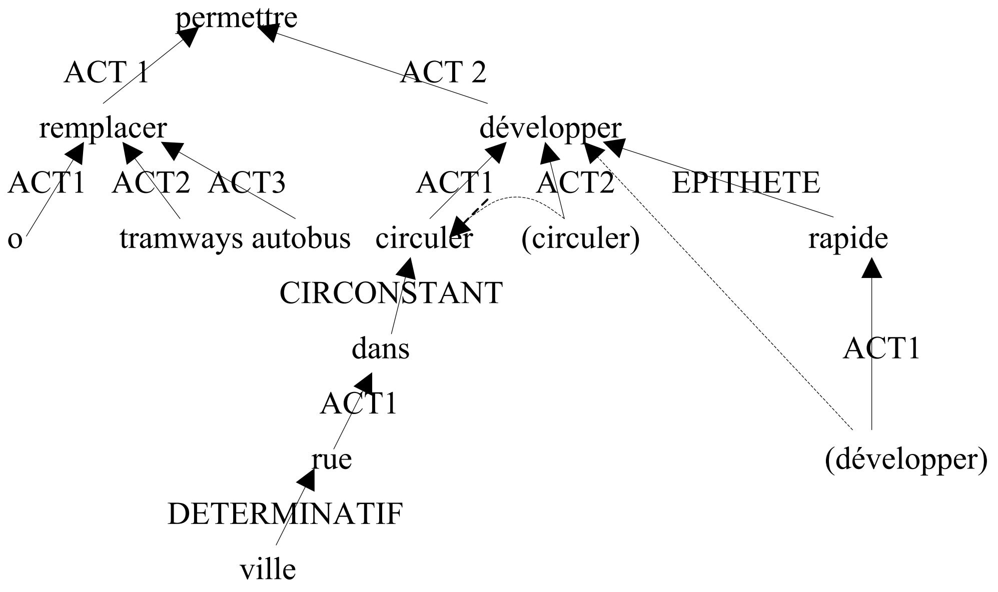

Fig.17: CETA 'pivot language' representation

(where *circulation* is regarded as both what is developing (agent, ACT1) and what is being developed (object, ACT2), and *rapide* is both modifier and predicative of *développer*). At the same time the semantically anomalous analyses were 'filtered out' by checking the compatibilities of the consituent SL lexical components on the basis of information in the SL dictionary.

Such a tree was the source for TL synthesis. It began with the substitution of SL lexical units in the pivot language representation by their equivalent TL units. In Syntactic synthesis, units were examined for their potential word-classes and for dependency relations with other word-classes. First, a predicative was located and its arguments checked as possible NP dependents. If one argument was itself a predicative, e.g. DEVELOPPER in the tree above, the posssibility of a clause structure was also investigated (i.e. *...que la circulation se développe rapidement...* as well as NP *développement*). Then the argument nodes (Act 1, Act 2, etc.) were replaced by appropriate categories (V, NP, Adj, etc.), elements were reordered to conform to TL surface syntax and the synthesis of TL words was begun (e.g. Act 2 (DEVELOPPER) became either V(DEVELOPPE) or NP(DEVELOPPEMENT)). Morphological synthesis completed the process by producing the correct surface forms (including the editing of variants, e.g. *le* → *l'* before *a, e, i, o, u*.)

An appraisal of the system in 1971 revealed that only 42% of sentences were being correctly translated and that readers found only 61% comprehensible. Although over half the incorrect sentences were caused by input errors, insufficient computer capacity or defects in programs which could be easily amended, it was found that the remainder lay beyond the system's capabilities. The main trouble was the rigidity of the levels of analysis; if morphological analysis failed because the dictionary had no entry for a word or did not record all homographic variants, then this affected all subsequent processes; if syntactic analysis failed to parse any part (however small) of a sentence, it was rejected. In addition, the parser was inefficient: it attempted too many partial analyses which came to nothing, and it produced too many analyses which had to be 'filtered out' later. What was needed was a parser which did not use its full armoury of analysis for every simple phrase structure but reserved the more complex parts for only complicated sentence structures. Finally, it was concluded that better synthesis would be possible if information about the 'surface' forms of SL sentences was also transferred; in the existing system information on choice of subject noun, use of passive, subordination of clauses, etc. was largely lost during conversion to pivot language representations, but such information could help considerably the selection of appropriate TL expressions.

A change in computer facilities in 1971 encouraged the Grenoble team to rethink the design of their MT system. From this date, the group, now called Groupe d'Etudes pour la Traduction Automatique (GETA), has worked on a system based on the 'transfer' approach (Ch. 13.3 below).

10. 2: Mel'chuk's 'meaning-text' model (1965-1976)

The syntactic representations of CETA were clearly influenced by the dependency grammar of Tesnière (1959), as the use of the term 'actant' reflects. However, the most direct influence on the CETA approach to MT system design was the 'meaning-text' model of the Russian linguist Mel'chuk (Mel'chuk & Zholkovskii 1970, Mel'chuk 1973), as the CETA researchers have acknowledged (Vauquois 1975, Vauquois et al.1970). Mel'chuk's model is stratificational in conception, like the analogous but nevertheless distinct and independent theory of Lamb (1966). Mel'chuk's model was developed for an English-Russian system in collaboration with Olga Kulagina and others (1967, 1971), although it was not implemented for internal political reasons (Ch. 11.5 below). In the course of years, since its original formulation in 1965 by Zholkovsky and Mel'chuk it has grown into a fully-fledged linguistic theory no longer specifically designed for MT application (Mel'chuk 1981)[2]. Nevertheless, it has remained more firmly rooted to the practicalities of MT analysis than Lamb's more theoretical speculations.

[2] The most substantial descriptions are: I.A.Mel'čuk: *Opyt teorii lingvističeskix modelej 'Smysl↔tekst': semantika, sintaksis* (Moskva: Nauka, 1974; [new ed.] Moskva: Škola 'Jazyki russkoj kul'tury, 1999); I.A.Mel'chuk, *Cours de morphologie générale*, 5 vols (Paris: CNRS, 1993-1999).

In Mel'chuk's 'meaning-text' model there are five levels or 'strata' of linguistic representations are: phonetic, phonemic, morphemic, surface syntactic, deep syntactic, and semantic. (For MT purposes the first two are unimportant.) Surface syntactic representations include such grammatical dependency relations as 'subject-of', 'complement-of', 'auxiliary' and 'determinant', and structure of nominal groups. Its elements are the 'actual lexemes' (words) of the language. Deep syntactic representations are tree structures indicating valency relations among root lexical elements ('generalized lexemes'), such as 'agent', 'instrument' and 'location'. Semantic representations are abstract networks of semantemes (meanings of generalized lexemes) or of elementary semantic units (semes). For any given SL sentence there could be a number of different semantic representations each corresponding to a number of possible interpretations. As in CETA, every strata would 'filter out' any syntactically and semantically anomalous representations from lower levels.

One important feature of the 'meaning-text' model is the incorporation of discourse indicators at both syntactic and semantic levels. Representations include links between pronouns and their antecedents, and the include an indication of which elements are the topic (or 'theme') of the sentence ('what it is about') and which are the comment (or 'rheme') of the sentence ('what is said about the topic'). At the deep syntactic and semantic levels there appear in addition indications of which elements are 'new' in the text, i.e. have not been mentioned in previous sentences, and which are 'given' (known or inferrable from earlier text, or which readers may be presumed to be familiar with already, i.e. can be presupposed.)

A second important feature of the 'meaning-text' model was the extensive treatment of semantic relationships. The researchers established a set of some 50 'lexical functions' which linked generalized lexemes at the deep syntactic level. These lexical functions include, as one would expect, such relations as: synonymy, e.g. *shoot* and *fire*, antonymy, e.g. *victory* and *defeat*, and conversives, e.g. *fear* and *frighten*. They include also many other relations rarely (if ever) formalised in MT, and only recently considered in AI (and even in that field, rarely in the thorooughness of the 'meaning-text' model.). For example, a verb and an agentive noun, e.g *write* and *writer*, *prevent* and *obstacle*, and a verb and its causative form *lie* and *lay*. They include also phraseological and idiomatic constructions, such as indications of the typical or 'idiomatic' verb for expressing particular relations to a given noun, e.g. the inceptive verb for *conference* is *open* but for *war* it is *break out*. Likewise, the causative verb for *dictionary* is *compile*, for *foundations* it is *lay* and for a *camp* it is *set up* or *pitch*. Finally, as a last example of a lexical function: the realisational or implementative verb for *order* is *fulfill*, for *law* it is *observe*, for *promise* it is *keep* and for *obligations* it is *discharge*. (In many respects, it was a conceptual forerunner of AI notions of semantic representations, Ch.15 below.)

The lexicographic aspect of the model was explored in depth. It resulted in the elaboration of the concept of an 'explanatory-combinatorial' dictionary, designed for the automatic generation of texts from a given semantic representation (Apresyan et al. 1969). The basic principles were that it must be "fully sufficient for a smooth, idiomatic and flexible expression of a given meaning", i.e.it must provide explicitly all the information necessary for the correct choice and usage of words in a given context. It was combinatorial in the sense that it showed the combinability of lexical items in utterances; and it was explanatory in the sense that through the lexical functions it provided semantic interpretations of combinations. Fragments of dictionaries were elaborated for a number of semantic 'fields', e.g. words for emotions (Iordanskaya 1973).

In the description of the CETA system it was made clear that was not a fully interlingual system. It is true that CETA could deal with some syntactic equivalences, e.g. the structure dominated by *développer* in Fig. 17 above could represent either a subordinate clause with *développer* as finite verb or a noun phrase with *développemnet* as a verbal noun. In this respect CETA corresponded to the 'deep syntactic' level of Mel'chuk's model. However, CETA lacked much of the detailed paraphrasing operations present in Mel'chuk's model, which result from the indication of complex semantic relations.

The interlingual features went far beyond those of CETA. Not only were its representations genuinely interlingual, and not restricted to syntactic equivalences, but Mel'chuk also recognised the importance of retaining SL 'surface' information about theme and rheme, the choice of subject noun, the use of passive, subordinatiin of clauses, etc. which could help substantially in the selection of appropriate TL forms. It is only to be regretted that no actual implementation was possible, although whether it would have been in fact any more successful must be doubtful.

10.3: Linguistics Research Center (LRC), University of Texas (1970-75)

Research on MT was revived at Texas by a contract from the USAF Rome Air Development Center for research on a German-English system. The project was established in 1970 with Rolf A.Stachowitz as principal researcher, while Winfred Lehmann continued as overall director of the Linguistics Research Center (LRC). The main sources for information on this project are the reports by Lehmann and Stachowitz (1970, 1971-75, 1972a.)

Some work had continued intermittently on the earlier German-English model (Ch.4.11), although mainly it seems on expansion of the dictionary: by 1970 the German dictionary contained 40,000 items and the English 77,500 items (Lehmann and Stachowitz 1971), very little more than in 1965 (although additional information for a further 47,000 English items was available from the Russian Master Dictionary, which had been deposited at LRC, Ch.4.11.) The new system was, however, to be on a larger and more ambitious scale: a fully automatic 'interlingual' system.

At an early stage of the project, during 1970 and 1971, the Linguistics Research Center held a series of study conferences and individual consultations involving prominent linguists and MT researchers. The results were summarized in a 'feasibility study on fully automatic high quality translation', to which were appended a number of papers by participants and consultants (Lehmann and Stachowitz 1971). There was certainly no uniformity of opinion about the future prospects and direction of MT, but the LRC group felt able to conclude that their research was on the right lines. The major problems of MT were no longer computational but linguistic. Theoretical research in linguistics supported the 'universal base' hypothesis: "the surface structures of any language can be related to such a universal base. Since the universal base in turn can be used for deriving the surface structure ofany language, the universal base can serve as the intermediary language between any source language and any target language." In other words, the 'interlingua' approach to MT presented a feasible model. The earlier LRC model had been essentially a syntactic 'transfer' model (Ch.4.11), now structural analysis was to go to a universal 'deep' level common to any SL and TL. The project's aim was, therefore, to be the development of methods of analysis and formalisms of representation which could be applied to any pair of languages.

In fact LRC had further ambitions. MT was to be just one part, albeit important, of a general system, the Linguistics Research System (LRS), later to be renamed METAL (Mechanical Translation and Analysis of Languages). LRS was to be of sufficient generality for application to other aspects of natural language processing; there was particular interest in applications in information retrieval. The ultimate goal was a system which could recognise and produce synonymous sentences, by deriving 'canonical form' (i.e. semantic interlingual) representations from sentences and generating all 'surface' realizations of such representations (Stachowitz 1971)

Unfortunately, most linguistic theory was inadequate for MT purposes: linguistic research had dealt "primarily with syntactic analysis of individual sentences, and hardly at all with semantic problems and discourse analysis." Transformational grammar had proved to be inefficient as a model (Ch.9.11); more hopeful were dependency grammars and grammars based on the string analysis model which Harris had developed at the University of Pennsylvania (Ch.3.5 above). The latter formed the foundation of the LRC procedures. However, while syntactic problems seemed tractable, the neglect of semantic problems hindered advances in MT; in particular, there were the differences of ‘world views’ reflected in the vocabularies and semantic relationships of languages. Nevertheless, it could be assumed that “closely related languages, like English and German, are similar in expressing their semantic distinctions overtly and covertly, and even in their surface structures; accordingly, they are relatively easy to translate into each other.” The LRC group were encouraged to concentrate on these two languages: “for the development of the technology of machine translation, systems designed for related languages are accordingly recommended at this time as an immediate goal. Medium-range goals (Russian-English) and long-range goals (Chinese-English) should also be planned.” In fact, LRC did start up again research on Russian using as a foundation the Russian Master Dictionary. A Russian-English project was therefore established in parallel to the German-English one, based on the same methodology and approach to MT system design and using the same computational techniques (Roberts & Zarechnak 1974)

As in the contemporary CETA project, which had also adopted the ‘interlingual’ approach, the basic stages of the LRC system were: analysis of SL texts into an intermediary representation, and synthesis of TL texts from the intermediary representation. But, also like the CETA ‘pivot language’, the LRC interlingua was not a genuine interlingua. It was restricted to syntactic structures (‘universal’ deep structures); there was no attempt to decompose lexical items into (universal) semantic primitives, for example. Tosh’s earlier suggestions of ‘universal’ numerical semantic codes were not pursued; nor were his ideas of using Roget’s thesaurus to establish interlingual codes (Ch.4.11). The lack of semantic analysis meant that the system could not, for example, handle such semantic equivalences as *He talks a lot* and *He is loquacious*. Conversion of vocabulary from German into English was consequently made through a bilingual dictionary operating essentially at the lexical level.

Analysis was performed by three separate ‘grammars’ working in sequence. After morphological analysis and dictionary lookup, the ‘surface sentence’ was converted by a ‘surface grammar’ into one or more tentative ‘standard strings’. In this process certain elements discontinuous in the surface form (e.g. verbs such as *look...up*) would be brought together. In the second stage, the tentative standard strings were tested by a ‘standard grammar’ for syntactic well-formedness and each string accepted by the standard grammar was then provided with one or more phrase-structure representations, called ‘standard trees’. The result of such an analysis for the sentence *An old man in a green suit looked at Mary’s dog* is illustrated by the standard tree in Fig.18 (Lehmann and Stachowitz 1972a):

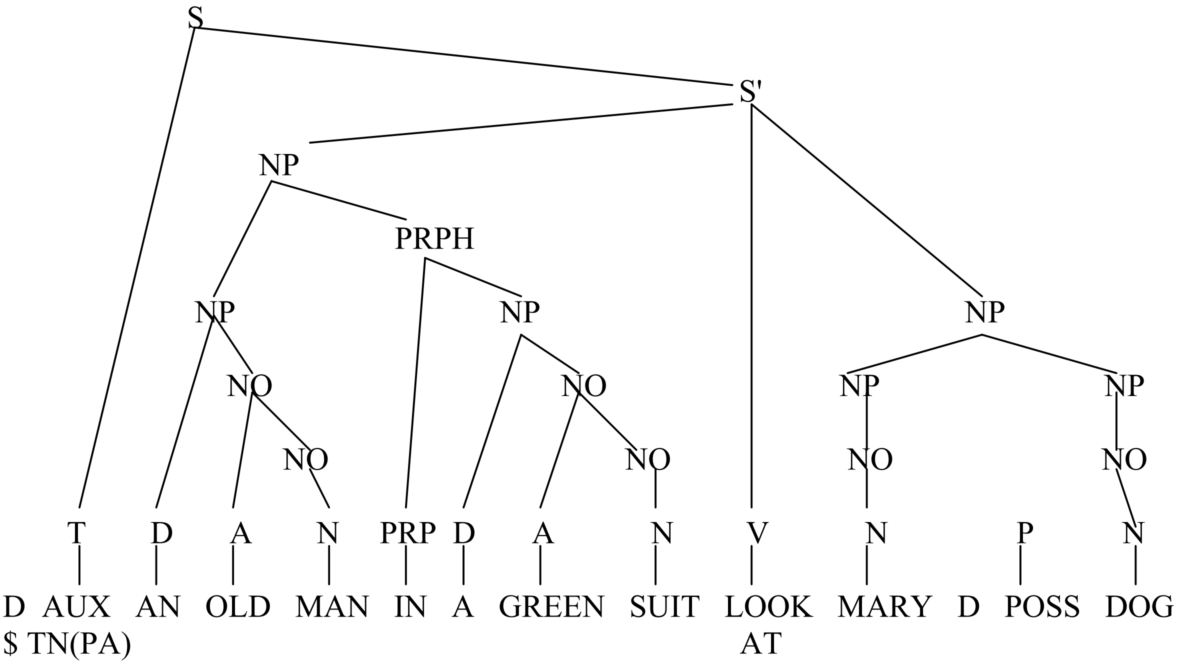

Fig.18: LRC standard tree representation

The third stage, 'normalization', filtered out semantically ill-formed standard trees by testing the semantic compatibility of syntactically related lexical items (referring to information provided in dictionary entries.) Each standard tree was then converted into a 'normal form' (or several 'normal forms' if the sentence was genuinely ambiguous). In this 'deep structure' representation relationships between items were expressed in terms of 'predicates' and 'arguments' or, alternatively, 'entities' and 'attributes' (much as in CETA representations, Ch.10.1). For the standard tree above would be derived the normal form in Fig.19:

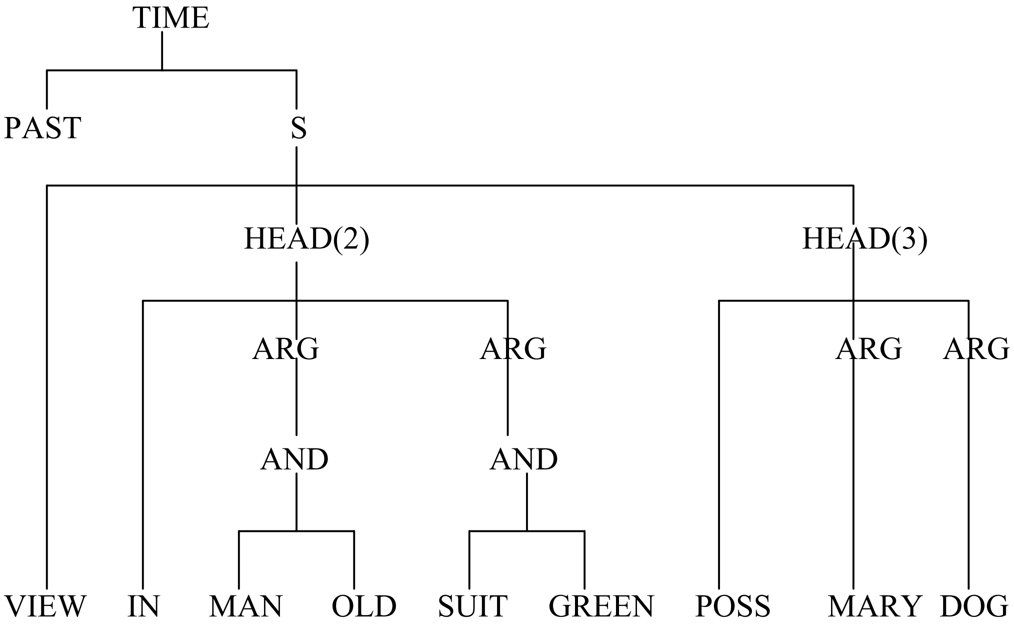

Fig. 9: LRC 'normal form' representation

The conversion from the standard tree into this normal form involved the identification of *in* as the predicate element of a tree with *the green suit* and *the old man* as argument elements; and the recognition of the adjectives in these noun phrases as arguments of their respective head nouns. The result is a dependency-style semantic representation intended to be independent of language-specific surface and phrase structure forms. It could not be completely language independent, however, because the system lacked semantic decomposition of lexical items. (The conversion of 'normal forms' into 'canonical forms' would have been the next stage of analysis.) Consequently, such semantic equivalences as *He ignored her* and *He took no notice of her*, could not be indicated since these sentences would have different 'deep syntactic' structures.

Synthesis of TL (English) sentences proceeded first by the substitution of TL lexical elements in the 'normal form', then the conversion of 'normal forms' into 'standard strings' and finally the conversion of strings into 'surface sentences'.

The LRC system suffered, like CETA, from an inadequate method of syntactic analysis. There were obvious problems if parsers failed to produce a structure for a given input at some stage of the analysis. But there were equally problems with too many analyses. For example, the 'context-free' (bottom-up) parser in the 'standard grammar' component often produced too many phrase structure analyses, e.g. for nominal constructions with prepositional phrases (cf. Ch.3.6 above). The absence of intersentential and discourse semantics also allowed multiple 'normal forms' to be produced for a single 'standard tree'. Furthermore, since a single normal form could obviously be the source of many different (but semantically equivalent) surface forms the problems of synthesis were also multiplied. In the end, the complexities of the procedures became too great for a workable MT system.

However, from the beginning LRC projects were conceived as general-purpose systems designed also for other automated language processes. At the theoretical level, for example, the possibilities of 'language comprehension' systems were investigated. These would have included not only a semantic analysis component but also logical components and (AI-type) 'real world' knowledge and "awarenes" components (Stachowitz 1971) The research on information retrieval in the later years of the project were part of this activity. Most emphasis was consequently placed on the system's ability to produce single 'normal form' semantic representations from a great variety of surface forms, which had potentially much relevance for automatic indexing and abstracting (Lehmann and Stachowitz 1972a)

Funding for the project ended in 1975, and there was again a hiatus in MT research at LRC, until a new project was started in 1978 (Ch.13.4 below)

10.4: Forschungsgruppe LIMAS, Bonn (1964-76)

Founded in 1964 by Alfred Hoppe the LIMAS (Linguistik und maschinelle Sprachverarbeitung) research group in Bonn pursued lines of theoretical MT research similar in a number of respects to those of the Milan group (Ch.5.3). The basic premise was that computer language processing, including MT, must be based on a language-independent semantic syntax, a 'communicative grammar' expressing content elements and their relations (Hoppe 1966, 1967, Lehmann & Stachowitz 1972). A classification of content elements or 'semantic factors' (i.e. essentially semantic features such as 'place', 'location', 'interior', 'product', 'producer') was developed; every lexical item was coded positively or negatively for each of the finite set (about 80) of 'factors'; translation from English to German involved the comparison and matching of matrices of coded 'factors' both between languages and within languages. For example, German *nach* is either a preposition coded for 'time' or a preposition coded for 'place'; the selection of English *after* (coded for 'time') or English *to* (coded for 'place') depends on the coding of the associated nouns, i.e. *nach dem Mittagessen: after lunch; nach Köln: to Cologne* (Schweisthal 1967). Semantic constraints were also expressed in terms of 'case frames', although Hoppe did not use this terminology. For example, he referred to a 'Geschehensziel' appearing as a subject (*der Hund wird gefüttert*), as accusative object (... *füttert den Hund*), as the specifier in a compound (*Hundefütterung*), or a genitive attribute (*Fütterung des Hundes*) (Hoppe 1966). Clearly, Hoppe's 'communicative grammar' was a variant of semantic syntax. As in the case of Cecccato's similar approach, the LIMAS group was anticipating the development of 'interlingual' semantic grammars within the AI framework (Ch.15).

The LIMAS programme of research consisted therefore primarily in the laborious establishment of a lexicon of semantic factors, the construction of factor matrices for English and German vocabulary, the determination of analysis rules for deriving factor formulas, the development of matrix matching and conversion procedures, and the establishment of synthesis rules for deriving output text. The system was intended to be fully reversible, and potentially extendable to other languages (Zint 1967). From the late 1960s the emphasis turned much more to theoreticalexplorations of general text analysis (e.g. for information retrieval), and so not surprisingly the practical results were meagre: a 1975 demonstration of the LIMAS content analysis program was devoted to a short text of just 4000 words (24 sentences). The LIMAS group was funded by the Deutsche Forschungsgemeinschaft until 1976 (Hoppe 1984).