Chapter 15: Artificial intelligence systems (1970-

Researchers in MT have taken an increasing interest in the possibilities of applying artificial intelligence techniques, particularly during the last 5 years or so. In 1960 Bar-Hillel believed he had demonstrated the impossibility of high-quality MT when he argued that many semantic problems could be resolved only if computers have access to large encyclopaedias of general knowledge (Ch.8.3 above) However, it is precisely problems of text understanding involving knowledge structures which have been the subject of much research in artificial intelligence (Boden 1977). The relevance of AI research to MT systems has been obvious to many researchers, and AI techniques are applied increasingly in translation systems. Nevertheless there have been to date relatively few explicitly AI projects on MT as such, although in the future many more may be expected.

Language analysis in AI research has not been directly concerned with translation problems but rather with question-answering, summarizing, and general problems of language understanding. Characteristic of the AI approach is the abandonment of syntax-based models in favour of predominantly semantic models of structure. Semantic analysis is not seen as just the next stage after syntactic analysis, in effect to tackle problems left over from syntactic analysis, but as the central component of the system. Problems of syntactic structure are left, if necessary, to subsidiary operations.

The main features of the AI approach are the application of semantic parsing (based on semantic categories, e.g. 'human', 'liquid', etc.), the building of semantic (or conceptual) representations of the meanings of texts, and the use of knowledge databases to assist in the interpretation of texts. Typically included in the latter are representations of conventional event-schemata (e.g. what happens when going to a restaurant), normal inference patterns, and common-sense expectations.

For AI researchers it is "obvious" that no good MT can be possible without 'understanding'. They point to such sentences as:

Israel seized large quantities of new weapons from the U.S.S.R.

and:

Israel seized territory from Syria.

and comment that in order to produce accurate Russian translations, "a translator needs to understand that:

(1) The weapons were seized *not* from the Soviet Union, but presumably from a neighboring Arab country.

(2) The weapons were manufactured in the Soviet Union.

(3) The territory was indeed seized from (*not* manufactured by) Syria." (Carbonell et al. 1981)

The distinctions would be reflected in Russian by the translation of *from* in the first sentence as *iz*, and in the second as *u*.

More pertinently, AI researchers emphasise the complexities of disambiguation and selection of TL equivalents. In syntax-based MT systems, disambiguation operates through the 'embellishment' of syntactic representations by attaching semantic features as 'selection restrictions'. The problem is that such refinements are static (Ch.3.6). Such systems "do not have the capability to distinguish between all contexts in which an ambiguous word can take on its different meanings". For representations to be unambiguous the system needs to understand SL texts (Lytinen & Schank 1982)

Despite the evident interest of AI researchers in MT problems, there have in fact been very few projects directly concerned with MT within the AI community. Yet it can be argued that MT provides a useful "test of the rightness or wrongness of a proposed system for representing meaning, since the output in a second language can be assessed by people unfamiliar with the

internal formalism and methods employed." (Wilks 1975a). Only in recent years have there been any substantial MT projects applying AI methods. These have been primarily Japanese investigations, stimulated in large part by the 'Fifth Generation' project (Ch.18.2ff).

15. 1: Stanford University (1970-74)

One of the first to experiment with an AI semantics-based approach was Yorick A. Wilks in his prototype English-French MT system in the early 1970s (Wilks 1973a, 1973b, 1975a). Wilks had gone to Stanford University in 1970 after leaving the Cambridge Language Research Unit (Ch.5.2). It was at Stanford between 1970 and 1974 that Wilks worked on his MT experimental system. Wilks saw his work on MT as essentially a 'testbed' for AI rather than as research on MT as such, and AI has been Wilks' principal interest in subsequent years, first at the University of Edinburgh (1975-76), then the University of Essex (1977-84) and now the University of New Mexico (Wilks 1985). Nevertheless, although Wilks did no MT research himself at Edinburgh and Essex he was heavily involved in the Leibniz group and in the overall design and semantic aspects of Eurotra (Ch.14.2)

Wilks' experiment investigated an AI-based 'interlingual' MT system, using a semantic grammar and inference rules. The SL text is first partitioned at punctuation markers and 'function words' (prepositions and conjunctions) into fragments, e.g. *I advised him to go* becomes '(I advised him) (to go)'. Each fragment is then tested against an inventory of templates, semantic frames expressing the 'gists' of (parts of) sentences in the form of triples of semantic features. For example, the template MAN HAVE THING (paraphrased perhaps as "some human being possesses some object") would be matched on a sentence such as *John owns a car*. MAN, HAVE and THING are intended to be interlingual semantic primitives which would be found as the principal semantic features of the words *John*, *own* and *car* respectively. Semantic formulas or definitions of words are constructed from semantic primitives, e.g. the formula for *drink* is: ((*ANI SUBJ)(((FLOW STUFF)OBJE)((*ANI IN)(((THIS(*ANI(THRU PART)))TO)(BE CAUSE))))). This is to be read as "an action, preferably done by animate things (*ANI SUBJ) to liquids ((FLOW STUFF)OBJE), of causing the liquid to be in the animate thing (*ANI IN) and via (TO indicating the direction case) a particular aperture of the animate thing; the mouth of course" (Wilks 1973a). The semantic analysis of lexical items goes no further than necessary; in this context there is no need to distinguish *mouth* from other apertures. The notion of preference is a central feature of Wilks' method: SUBJ displays the preferred agents of actions and OBJE the preferred objects or patients, they do not stipulate obligatory features of agents and patients and thus allowance is made for abnormal usages, e.g. cars drinking petrol. In this way, Wilks' preference semantics can cope with many types of metaphorical expressions without adding to the complexities of dictionary entries (Wilks 1975b). The final stage of the analysis produces a dependency network of semantic relations on the basis of the case links specified. Thus, our example sentence *The watch was sold by the jeweller to a man with a red beard* might receive the analysis shown in Fig.33:

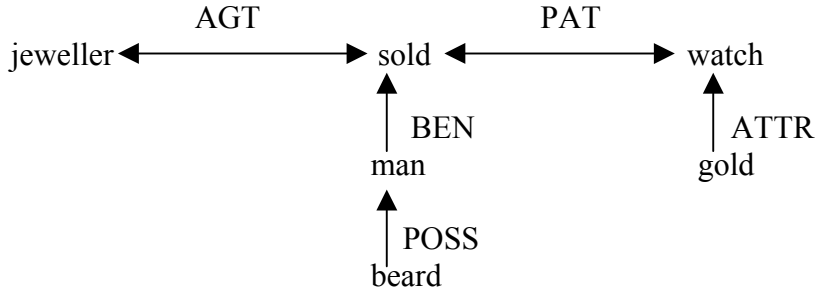

Fig.33: Wilks semantic representation

At this point relationships between the networks of fragments are established; thus a temporal phrase (*during the war*) might be tied to the 'action' element of an earlier or later fragment. It should be noted that ties are made not only within sentences but also across sentence boundaries, since the basic unit is not the sentence but the phrase (fragment). Some ties involving pronominal reference make use of 'common sense inference' rules. For example, in the sentence *The soldiers fired at the women and we saw several of them fall* the linking of the pronoun *them* to the noun *women* rather than to the other noun *soldiers* is made on the basis of a common sense rule stating that if an animate object is hit then it is likely to fall. In other words this rule establishes a causal relationship between the components of the templates of the fragments in question.

The distinctive feature of Wilks' analytical method are thus the use exclusively of semantic features in the 'parsing' of phrases, the use of preference semantics and common sense inference rules, and the analysis of discourse relationships. At no stage is there any reference to syntactic structures or indeed to the boundaries of sentences. Grammatical categories such as noun and verb have no role, not even in the resolution of homographs: to identify the verbal sense of *father* in a sentence such as *Small men sometimes father big sons* the program needs only to find that the semantic formula with CAUSE as its 'head' is the only one which will fit the other 'heads' in an acceptable template. In Wilks' system then semantic representations are reached without recourse to previous syntactic analysis.

After returning to the United Kingdom in 1975, Wilks suggested refinements to his preference semantics model. In *My car drinks gasoline* the system accepts *car* as filler in the template for *drink* because although expecting an animate entity there is no competitor. Rather than just accepting a preference-violating structure, the system might interpret the abnormal reading. The mechanism for this, Wilks (1978) suggested, is the incorporation of 'thesaurus information' as well as primitives in formulas (cf. the CLRU project, Ch.5.2) and 'pseudo-texts' expressing knowledge about lexical items. Thus, the formula for *car* might include the thesaurus item 'engine': (((((@engine PART) OBJE) (SELF USE)) WAY) ((SELF MOVE) GOAL) (MAN USE) (OBJE THING). The pseudo-text for car would include 'statements' (in formulas) to the effect that: (1) a human injects a liquid using a tube; (2) liquid is in the car; (3) the engine uses the liquid; (4) the car moves; 5) the human in the car moves; etc. Given *My car drinks gasoline*, this pseudo-text is used to make a projection on the basis of the similarity of *drink* and 'inject' (via the thesaurus entry for inject) that the sentence matches the fourth statement, i.e. 'engine uses liquid'. This projection supplies an interpretation of the sentence.

There are some similarities between Wilks' formula representations and Schank's conceptual dependency representations (Schank and Wilks were both at Stanford in the early 1970s.) But the main influence on Wilks was that of the Cambridge Language Research Unit and its director Margaret Masterman (Ch.5.2). Wilks' templates bear much similarity to the Cambridge group's semantic message structures, his primitives match closely the CLRU ones in Richens' interlingua, and of course there is the application of the thesaural concept. The CLRU influence was even more transparent in Wilks' research at the Systems Development Corporation, Santa Monica, during 1966-67. In his report of this program for the semantic analysis of philosophical texts (Wilks 1972), he acknowledges also the influence of Wittgenstein on his conception of semantic relations (i.e. 'family resemblances') and semantic message forms. It was this work which formed the basis of his Stanford MT experiment.

15. 2: Yale University (1978-82)

There have been two projects at the Department of Computer Science in Yale University which have experimented with 'knowledge-based' MT. Both adopt interlingual representations based on the 'conceptual dependency' representations developed by Roger Schank (1975).

The essence of Schank's approach is that modelling human understanding of language requires the representation of meaning in terms of 'primitive' semantic relationships ('conceptual

dependencies') which express not only what is explicit in the surface forms but what is also implied or can be inferred. Thus, a conceptual dependency representation of *John punched Mary* would be as shown in Fig. 34 (simplified from Schank 1973: 226):

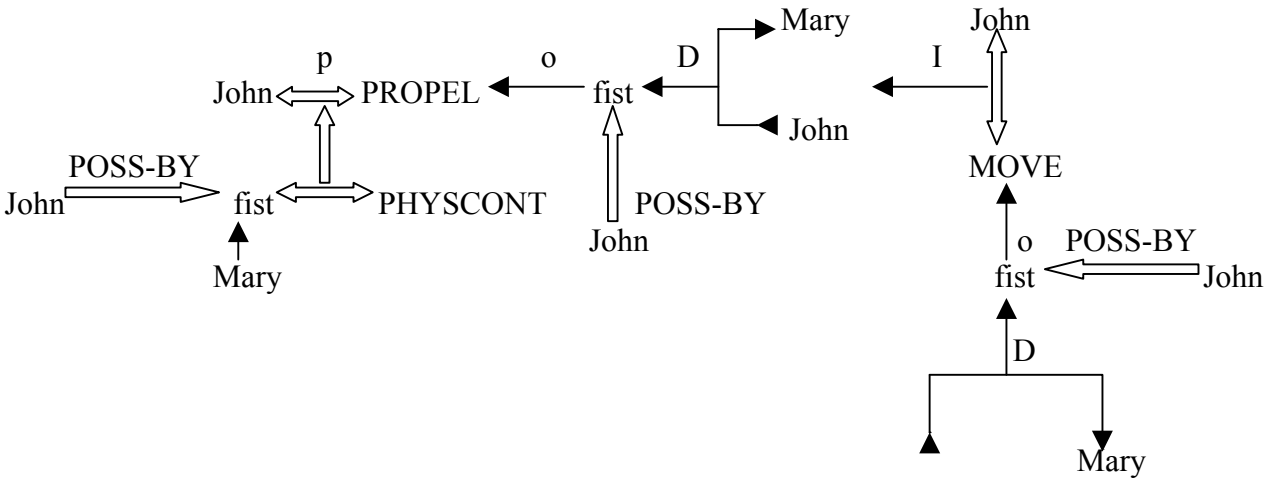

Fig.34: Schank conceptual dependency representation

This representation can be 'paraphrased' as: At some time past, John applied a force to the object John's fist, in the direction from-John-to-Mary; he did this by moving his fist in the direction from-John-to-Mary; his action of applying force to the fist caused John's fist to be in contact with Mary.

Conceptual dependency analysis requires therefore databases of 'contextual knowledge'. Applied to a MT system, the approach can be schematised as in Fig. 35 (Carbonell et al. 1981):

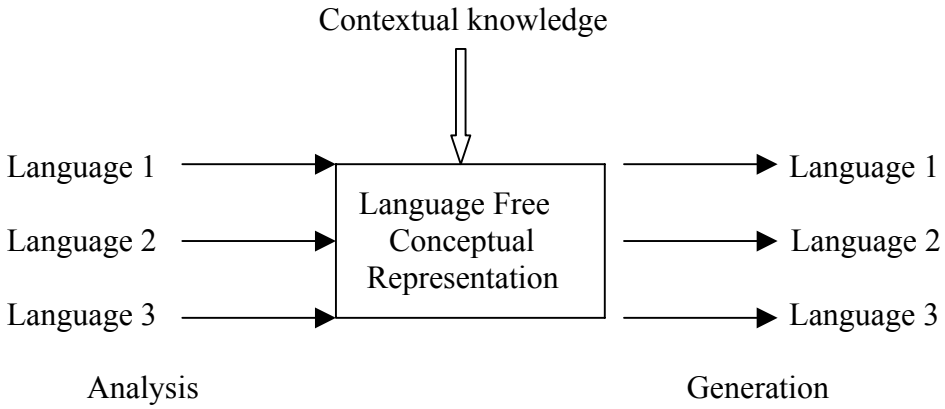

Fig.35: Yale 'machine translation paradigm'

In the rudimentary interlingual MT system developed in 1978 by Carbonell, Cullingford and others (1981) a simple English text, the report of an accident, is analysed into a language independent conceptual representation by referring to 'scripts' about what happens in car accidents, ambulances, hospitals, etc. The resulting representation is the basis for generating Russian and Spanish versions of the original report.

The analysis stages of the system were as follows: the first module was the English Language Interpreter which analysed the SL input text into a meaning representation; then a second module identified and tagged referential objects (people, places, static things); then the 'script-

applier' connects the text with an appropriate 'script' in its database of scripts (the knowledge database). The result at this point is the kind of representation shown below.

The basic task of text understanding is performed by the 'script applier': locating input to relevant parts of a script, linking new input to what has gone before, and making predictions about what is likely to follow. In each activity it uses 'world knowledge' to make explicit the connections which are only implicit in the text itself. Among the most important inferences it makes are those concerning the completion of causal chains of events, i.e. in this case the 'scripts' of accidents, ambulances and hospitals. For example, it makes the crucial connection (not stated in the SL text) that the person taken to hospital must have been injured in the crash. This is possible because it 'knows' what ambulances and hospitals are for. Other inferences establish pronominal references across sentences.

An illustrative extract of a semantic representation is given for the sentences:
Friday evening a car swerved off Route 69. The vehicle struck a tree. The passenger ... was killed....

```
SCLAB3:
        SCRIPTNAME $VEACCIDENT
        MAINCON EVNT4
        SCENECONS (EVNT4 EVNT17 EVNT33)
        INFERENCE ((EVNT20 EVNT26) (EVNT14))
        SCORECARD (EVNT33)
        EVNT4:
            VALUE ((ACTOR STRUCT0
                          <=> (*PROPEL*)
                          OBJECT PHYSO)
                   TIME (TIME))
            LASTEVENT (EVNT3)
            NEXTEVENT (EVNT20 EVNT14 EVNT7)
        STRUCT0
            CLASS (#STRUCTURE)
            TYPE (*CAR*)
            SUPERSET (*VEHICLE*)
            ELEX (AUTOMOBILE)
            SLEX (AUTO)
            SROLES
                   (($VEHACCIDENT . &VEHICLE)
                   ($DRIVE . &VEHICLE1))
        PHYS0
            CLASS (#PHYSOBJ)
            TYPE (*TREE*)
            ELEX (TREE)
            SLEX (ARBOL)
            SROLES
                   (($VEHACCIDENT . &OBSTRUCTION))
        EVNT14:
            VALUE ((ACTOR HUM0
                   FROM (*HEALTH* VAL (*NORM*)))
                   TOWARD (*HEALTH* VAL (-10)))
                   TIME (TIME17))
```

The main structure (SCLAB3) indicates the type of script ($VEHACCIDENT), the main story in EVNT4, episodes of the story (SCENECONS) and the inferred events or actions (INFERENCE). The representation EVNT4 expresses the crash itself: a 'structured physical object' (STRUCT0) hitting (*PROPEL*) an 'unstructured physical object' (PHYS0). The representations of STRUCT0 and PHYS0 indicate that these are respectively a *VEHICLE*, specifically an AUTOMOBILE, and a *TREE*. Connected to EVNT4 are other events, including the result (EVNT14) that someone (HUM0) died, i.e. went from state of normal health to state of non-health.

From such a representation the system generates TL sentences. These are not, therefore, translations of SL sentences, but expressions of the text as interpreted, i.e. paraphrases. The generator would, for example, take the part representation EVNT4 and 'expand' it as (*A car hit a tree Friday evening*):

```
((ACTOR (#STRUCTURE TYPE      (*CAR*)
                    SUPERSET  (*VEHICLE*)
                    TOKEN     (STRUCT0))
◄——► (*PROPEL*)
OBJECT (#PHYSOBJ TYPE (*TREE*)))
TIME ((WEEKDAY FRIDAY) (DAYPART EVENING)))
```

producing the Spanish version:

EL VIERNES AL ANOCHECER UN AUTO CHOCO CONTRA UN ARBOL.

There are three possible Spanish translations of *hit* (i.e. PROPEL): *pegar, golpear, chocar*. The representation is claimed to provide sufficient information for the correct choice to be made (via a 'discrimination-net', i.e. decision tree)

The Yale researchers admitted that what they produced were 'retellings' of the SL text in another language and not 'translations'. However, they insisted that they were "not trying to simulate a professional translator", but rather "to simulate an average nonprofessional person with a working knowledge of two languages who is asked to read a text and then to reproduce its content at a desirable level of detail." (Carbonell et al. 1981). It was considered more important to convey the sense unambiguously than to preserve the structure and style of the original.

A little later at Yale, Lytinen & Schank (1982) experimented with a MT model based on a more abstract type of knowledge base. Instead of scripts relating to specific stereotypical event-schemata, the Yale group now proposed MOPs (Memory Organization Packets). These are representations of knowledge common to many different situations. For example, whereas a 'script' for visiting a doctor would specify a sequence of having a medical problem, making an appointment, going to surgery, sitting in waiting room, having treatment, etc., a MOP would cover all 'professional visits' to lawyers, dentists, doctors, etc.

The Yale experiment, called MOPTRANS (MOP-based TRANSlator), has been designed to read newspaper stories about terrorism in Spanish and French, and to translate these stories into English. The MOPTRANS system uses two types of knowledge: packaging knowledge (the abstract event sequences referred to) and abstraction knowledge. The latter represents conceptual relations, e.g. that to arrest someone is to GET-CONTROL of them, and that police searches are types of FIND.

The system is illustrated by Lytinen & Schank (1982) on the Spanish sentence:

La policia realiza intensas diligencias para capturar a un presunto maniatico sexual...

to be translated (roughly) as:

Police are searching for a presumed sex maniac...

Given the multiple senses (i.e. vagueness) of *diligencias* ('search', 'determine', 'shop', 'go' according to context), it can initially be assigned only a very general structure *DO*. However, the goal or purpose of this action (indicated by the FOLLOWS-FROM relation expressed by *para*) is to GET-CONTROL of someone (the general unspecific sense of *capturar*). This instantiates a MOP (called M-GET) specifying the sequence KNOW + FIND + GET-CONTROL (i.e. that when one

wants to get control of an object, one first has to know where it is, and then one has to find it) Consequently it can be inferred that *diligencias* actually refers to the structure FIND. Since the police are the agents of this FIND, it is now possible to replace FIND by the more specific POLICE-SEARCH. As a consequence, a new more specific MOP can be set up, viz. M-POLICE-CAPTURE, with the sequence POLICE-INVESTIGATION + POLICE-SEARCH + ARREST. In this MOP, ARREST corresponds to the GET-CONTROL of the M-GET, permitting the final recognition of the sense of *capturar* as 'arrest'.

It is stressed by the Yale researchers that this system does not require rules which look for specific lexical items in order to disambiguate *diligencias* (as in most syntax-based MT systems). The recognition of its sense as FIND follows from the occurrence of GET-CONTROL, and this procedure could apply to any item represented as *DO*. The rules which use abstraction knowledge and packaging knowledge (MOPs) are independent of any particular language and its lexical relationships.

It should be noted that the MOPTRANS approach does not ignore syntactic information completely. As Lytinen (1985) illustrates with the simple sentence *John gave Mary a book*, the parser would build a conceptual dependency representation for *give* in which the 'person' (*John*) already identified could be either the ACTOR or the RECIPIENT of the transaction. Syntactic rules would identify *John* as a subject and thus confirm this 'person' as ACTOR. As a consequence MOPTRANS integrates semantic and syntactic analyses, with semantic parsing playing the dominant role.

15. 3: Other AI approaches.

In a sense, almost any AI project which involves the analysis of natural language text as 'conceptual' representations and the generation of surface text may be regarded as an embryonic interlingual MT system. Tucker and Nirenburg (1984), for example, describe the work of Wilensky and Arens in the Berkeley AI group as providing such a potential. The group has developed analysers and generators of English for natural language interfaces for the UNIX computer operating system (Wilensky et al. 1984). Both the parser PHRAN (phrasal analyser) and the generator PHRED (phrasal English diction) are claimed to be easily extendable to languages other than English, and such extensions of PHRAN have been implemented in Spanish and Chinese. Apparently a small English-Spanish translation system has been attempted.

With the exception of Wilks' system, all the groups mentioned above adopt basically Schank's conceptual dependency approach to 'interlingual' representations. Other AI projects have taken different lines.

15. 4: Massachusetts Institute of Technology (1971-74)

At roughly the same time as Wilks was experimenting at Stanford there was also a small-scale project at the Massachusetts Institute of Technology which attempted to integrate AI methods into a MT system. Sponsored by the Office of Naval Research, the research was done as part of the thesis of Gretchen Purkhiser Brown (1974) Brown's objective was a German-English system which implemented Winograd's PROGRAMMAR (Winograd 1972) in an interlingual MT design.

SL analysis involved initial morphological and syntactic analysis and then transformation into representations in terms of semantic primitives. The English TL forms were to be generated from semantic representations. For the disambiguation of SL analysis structures (i.e. those ambiguities which could not be resolved by syntactic means), the system was to include an 'understanding' component. Representations were to be checked for consistency and comprehensibility against a knowledge base, containing definitions of lexical items in terms of semantic primitives and information about the 'real world'. One of its tasks would be, for instance, the determination of the antecedents of pronouns on the basis of semantic information about relationships within the text and of conceptual hierarchical relationships, e.g. that octopuses are sea

creatures. The 'understanding' component was also to deal with thematic information, given and new information, problems of metaphoric usage. Unfortunately none of the 'understanding' was implemented. At a result, the actual system (which was based on a very small text corpus) comprised little more than a conventional analysis and synthesis program.

15. 5: University of Essex and New Mexico University (1983-

The research by Xiaming Huang of the Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China, has been conducted since 1983 first at the University of Essex (Colchester, UK) and later at New Mexico State University (Las Cruces, USA). It is an implementation in Prolog of 'definite clause grammar' (a version of context-free phrase structure grammar), case grammar and Wilks' preference semantics. Huang (1984, 1985) describes a small-scale English-Chinese 'interlingual' system XTRA designed principally to test problems of ambiguity in coordinate structures, in relative clauses, in participle clauses, and in prepositional phrase combinations. Parsing incorporates both syntactic and semantic analysis, and interlingual representations are basically dependency trees with case frame slots and conjoined semantic features.

15. 6: Georgia Institute of Technology (1982-

Richard Cullingford, who had collaborated with Carbonell at Yale, has continued to investigate the AI approach at the Georgia Institute of Technology (Cullingford & Onyshkevych (1985). As before, the basic component is an interlingual representation based on the conceptual dependency approach. However, rather than drawing inferences from 'scripts' (or MOPs) in order to predict and build representations, the main sources of predictions are lexical entries. The entry of, for example, punch would be:

(propel actor (person)
obj (bpart btype (hand))
to (physcont val (bpart partof (person))))

The 'slots' for actor, obj(ect) and recipient (to) would be filled by appropriate items in the sentence analysed. The method has been tested on a small Ukrainian-English system. Ukrainian lexical entries contain information on case (nominative, accusative, etc.), gender and number in addition to semantic information. Cullingford's method of "lexicon-driven analysis" might be regarded as a refinement of the 'case frame' parsers (without the syntactic information), and as closely related to the 'word expert' parser of Small (1983), cf.19.3 below.

15. 7: Colgate University (1983-

The MT system being developed at Colgate University by Allen Tucker, Sergei Nirenburg and others (Tucker & Nirenburg 1984; Nirenburg et al. 1985) adopts an AI approach in that it has an interlingua and a knowledge base. There is, however, one important addition: the TRANSLATOR system is also intended to include representations of the expertise of human translators. In this sense, the project is an experiment in modelling human translation. The aim is a multilingual MT system for the four languages English, Japanese, Russian and Spanish.

In overall design TRANSLATOR is a 'conventional' interlingual system: morpho-syntactic and semantic analysis converts SL text into IL representations (producing normally sets of alternative interpretations), and a generation module converts IL representations into TL text. It differs in the insertion of an 'Inspector' module, the embodiment of an expert translator, which examines the alternative interlingual parsings of SL texts and decides which is the most plausible in the context by reference to a knowledge base. The Inspector is thus envisaged as a simulation of the human consultant in an interactive MT system, cf. below the early Brigham Young system (Ch.17.10) and the DLT system under development (Ch.16.3), or as the designers put it: "The

challenge... is to simulate the behavior of the post-editor and thus arrive at a high-quality translation *before* the target text is synthesized rather than after." (Tucker & Nirenburg 1984).

The knowledge base of TRANSLATOR has the following components (Nirenburg et al.1985): IL dictionary, SL-IL dictionary, IL-TL dictionary, SL grammar, SL-IL translator program (set of parsers), IL grammar, IL inspector, Inspector knowledge base, TL grammar, IL-TL translator program. The AI features of TRANSLATOR are the embodiment of 'language-independent' world knowledge in the IL dictionary, and the inclusion of an 'expert system' in the IL inspector.

The interlingua (IL) is envisaged as a 'full' language with its own lexicon and grammar. As in the case of the majority of interlingual systems (cf.Ch.10), the designers of TRANSLATOR do not attempt componential analyses of meaning (e.g. decomposition into semantic primitives); the IL dictionary is thus essentially a set of SL and TL mappings. Entries in the IL dictionary include hierarchical relations ('is-a', 'consists-of', etc.), specification of 'sublanguage' context, and valency information, i.e. 'frames' specifying what types of items may be related to them in texts (or rather, in IL representations of texts). For the latter, the familiar 'case relations' are suggested: agent, patient, source, goal, instrument, etc. Finally, the IL grammar specifies the types of relationships and constituency structures in which IL dictionary items may occur in IL representations. Particular attention has been paid to the representation of time relations, space relations, and inter-clausal relations (Nirenburg et al. 1985)

Entries in SL and TL dictionaries are to be constructed on the same design. For example, the English word *we* might appear as:

(*speaker* ; 'we'
(isa human)
(subworlds everyday-world)
(agent-of process)
(object-of process)
(patient-of state)

and the entry for *computer* as:

(*computer* ; 'computer'
(isa device)
(subworlds computer-world business-world college-world)
(object-of use)
(instrument-of solve analyze)
(source-of information)
(consists-of CPU *memory* peripherals)

These examples illustrate the amount of 'world' knowledge incorporated in the IL dictionary.

As yet there are no details of the most innovative feature of TRANSLATOR, the Inspector expert systems. It is envisaged, however, that its knowledge base would include a 'short term memory' for information about the subject matter of the text being analysed (used for disambiguation), and a 'long term memory' of information relating to sets of texts on similar subject matters (e.g. sublanguage information).[1]

15. 8: University of Massachusetts at Amherst (1985-)

The project by McDonald (1985) aims to develop a SL parser which establishes the discourse decisions made by the writer in composing the original text, for use by the TL generator

[1] The project was continued at Carnegie-Mellon University (Pittsburgh, Pa.) by Nirenburg and Carbonell from 1987 onwards. Full descriptions of the project are given by K. Goodman & S. Nirenburg (eds.) *The KBMT project: a case study in knowledge-based machine translation* (San Mateo, Ca.: Morgan Kaufmann, 1991) and S. Nirenburg, J. Carbonell, M. Tomita, & K. Goodman *Machine translation: a knowledge-based approach* (San Mateo, Ca.: Morgan Kaufmann, 1992)

in synthesis. The parser is intended to identify the 'realisation class' for a given analysis, i.e. the set of rules which determined the particular surface form, such as selection of subject noun, selection of there-sentence form, selection of passive verb, etc. These realisation rules are then to be converted into equivalent generative rules of the TL. Except that the parser itself is AI-based, the basic idea is akin to the 'analysis by synthesis' approach at MIT in the 1960s (Ch.4.7).

15. 9: The future relevance of AI approaches in MT.

These examples of the AI approach to MT give some flavour of the complexity demanded in semantics-based methods. Their feasibility in full-scale MT must, however, remain doubtful. The problems of expanding AI techniques for a large-scale MT system are clear enough. The earlier research at Cambridge (Ch.5.2) showed the difficulties of defining semantic 'primitives' and applying them in the creation of large dictionaries. Even more problematic would appear to be the elaboration of all the scripts (or templates, or MOPs), the 'knowledge databases' and the inference rules to cope with a much wider range of possible texts than in the restricted domains of AI systems. In particular it should be noted that no AI project has yet tackled scientific and technical sublanguages, and that most research has concentrated on narrative texts (stories and newspaper reports) rather than expository or argument texts.

Naturally, AI workers are confident that their approaches are expandable. Thus, Carbonell et al. (1981): "Our experience ... suggests that the increase in... understanding abilities does not lead to any serious decrease in its speed. The combinatorial 'explosion' feared by Boitet has not come to pass". The MT experience has been frequently that the expansion of small-scale experimental prototypes into large-scale systems intended for operational implementation has often resulted in disheartening degradations in quality. Perhaps AI approaches will be more successful.

It needs to be stressed, however, that none of the AI workers are expecting their work to result in the near future in 'operational' MT systems. What they are engaged in is the fundamental research which is the necessary preliminary for more advanced MT systems. In a number of respects therefore, AI projects on MT are the successors of the 'perfectionist' projects of the 1960s. (cf. Ch. 3.10 and Ch. 8.2 above)