[From: *Human translation, machine translation*. Papers from 10th annual conference on computational Linguistics, Odense, Denmark, 22-23 November 1979, ed. Suzanne Hanon and Viggo Hjørnager Pedersen.]

Contemporary Perspectives in Machine Translation

by

Rod Johnson,
Centre for Computational Linguistics,
UMIST, Manchester, UK.

THE FIRST GENERATION (1949-1966)

The history of Machine Translation (henceforth MT) is generally acknowledged to have begun in 1949 with a memorandum circulated by Warren Weaver to around 200 of his colleagues, many of whom had been engaged during the Second World War on cryptological work. Weaver said:

> "I have a text in front of me which is written in Russian but I am going to pretend that it is really written in English and that it has been coded into some strange symbols. All I need to do is to strip off the code in order to retrieve the information contained in the text."

Weaver's proposals, based on the fundamental premise that the process of translation is the same as the decoding of military and diplomatic messages, and hence similarly amenable to mechanical treatment, sparked off an enormous amount of activity in MT research. As described by Anthony Oettinger, one of the prominent workers in early MT:

> "The notion of.... fully automatic high quality mechanical translation, planted by overzealous propagandists for automatic translation on both sides of the Iron Curtain and nurtured by the wishful thinking of potential users, blossomed like a vigorous weed." (Oettinger, 1963, cited by Dreyfus, 1972, p. 3).

From Oettinger's tone here, we see already in 1963 an indication of the general disillusionment at the way quite promising early results seemed to lead consistently to a series of costly dead ends. This disillusionment culminated in 1966, after some $20 million had been spent in the previous decade by various United States government agencies on MT research, with the notorious ALPAC report (ALPAC, 1966), commissioned two years earlier by the United States National Academy of Sciences. The principal ALPAC conclusion, perhaps the most-cited reference in the whole of the MT literature, stated baldly:

> "....we do not have useful machine translation. Furthermore, there is no immediate or predictable prospect of useful machine translation." (ALPAC, 1966, p. 32).

The influence of the report was widely felt, both within the United States and beyond, particularly in the English-speaking countries. And its main recommendation, that research into full-scale MT systems should cease, effectively signalled the end of the MT boom and branded MT research with a stigma which was to remain for the next decade.

AFTER ALPAC

It is not difficult to see why ALPAC reached its conclusions. The systems which were used by the committee in their evaluations were the only ones which were sufficiently developed to participate in large-scale testing. These were all systems which had been conceived and begun in the 1950s, when the only formal models of language available were the information-theoretic models used so successfully by the cryptanalysts and implicitly proposed in Weaver's suggestion. In essence, the computer realisations of these models were firmly organised around very large dictionaries, with some degree of linear manipulation of input and output text mainly for morphological analysis and minor reorganisation of word-order based on purely local context. It is not surprising, in retrospect, that these early attempts, largely in translation from Russian to English, should have met with dramatic initial success, only to become bogged down subsequently by the inadequacy of the underlying linguistic theory. After the initial progress, the only improvements possible were restricted to *ad hoc* tinkering with local linguistic context and massive enlargement of dictionaries to cover one new special case after another. Major modifications to the programs to incorporate the new ideas rapidly being developed in formal linguistics would have meant more or less redesigning the systems from scratch, and that would have been too costly in terms both of money and of the reputations invested in the original design. It was inevitable that some time someone would call a halt to the growth of Oettinger's "vigorous weed", and the halt was called by ALPAC.

There was also, however, a positive side to ALPAC - in particular, the committee recommended that funds be transferred from research in MT to computational linguistics and to general linguistics, to allow linguistic theory to catch up with the pretensions of proposed applications. As a result many scholars engaged in MT moved across into the neighbouring fields of artificial intelligence, computational linguistics, information retrieval and theoretical linguistics, following the research grants to more fertile terrain. Moreover, the United States government did allow a small amount of MT research to continue - one system which was allowed to continue is now functioning, after a series of substantial modifications, as an operational system within the EEC in Luxembourg under the name SYSTRAN. Finally, the impact of ALPAC was less severe in countries where the political importance of language traditionally does not permit the linguistic complacency of the English speaking world. Notably, in the West, in France a team has been working continuously on MT at the University of Grenoble since 1961, and MT research has been supported without a break since 1962 at the University of Montreal by the Canadian government.

Thus, despite the negative publicity and financial discouragement which resulted from ALPAC, we find at the beginning of the 1970s new claims being made for the feasibility of a second generation of MT - claims based on lessons learned from the mistakes of the fifties and early sixties, on new ideas gleaned from parallel research in artificial intelligence, computational linguistics, information retrieval and programming language design, and a much more solid theoretical basis.

THE SECOND GENERATION

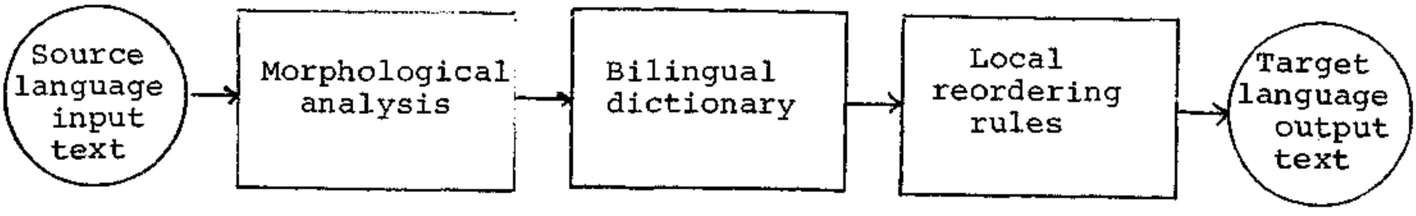
Ironically, while the ALPAC committee was formulating its drastic criticisms of first generation MT systems, many of the criticisms had already been foreseen and answers to them were already incorporated into early second generation prototypes, principally, and characteristically, in the French-oriented centres of Grenoble and Montreal. The essential insights which characterise second as opposed to first-generation systems are three - separation of the processes of analysis of the input text and synthesis of a (hopefully) synonymous text in the target language; recognition of the inadequacy of isolated surface forms in the source language as a basis for translation, and hence of the need for MT programs to incorporate a coherent linguistic theory; and separation of the expression of that theory as a system of linguistic rules from the algorithm or sequence of computer instructions which determines the way in which the rules build structures inside the machine. These insights merit consideration in rather more detail.

Separation of analysis from synthesis

The basic structure of a first-generation system can be seen in Figure 1.

Figure 1

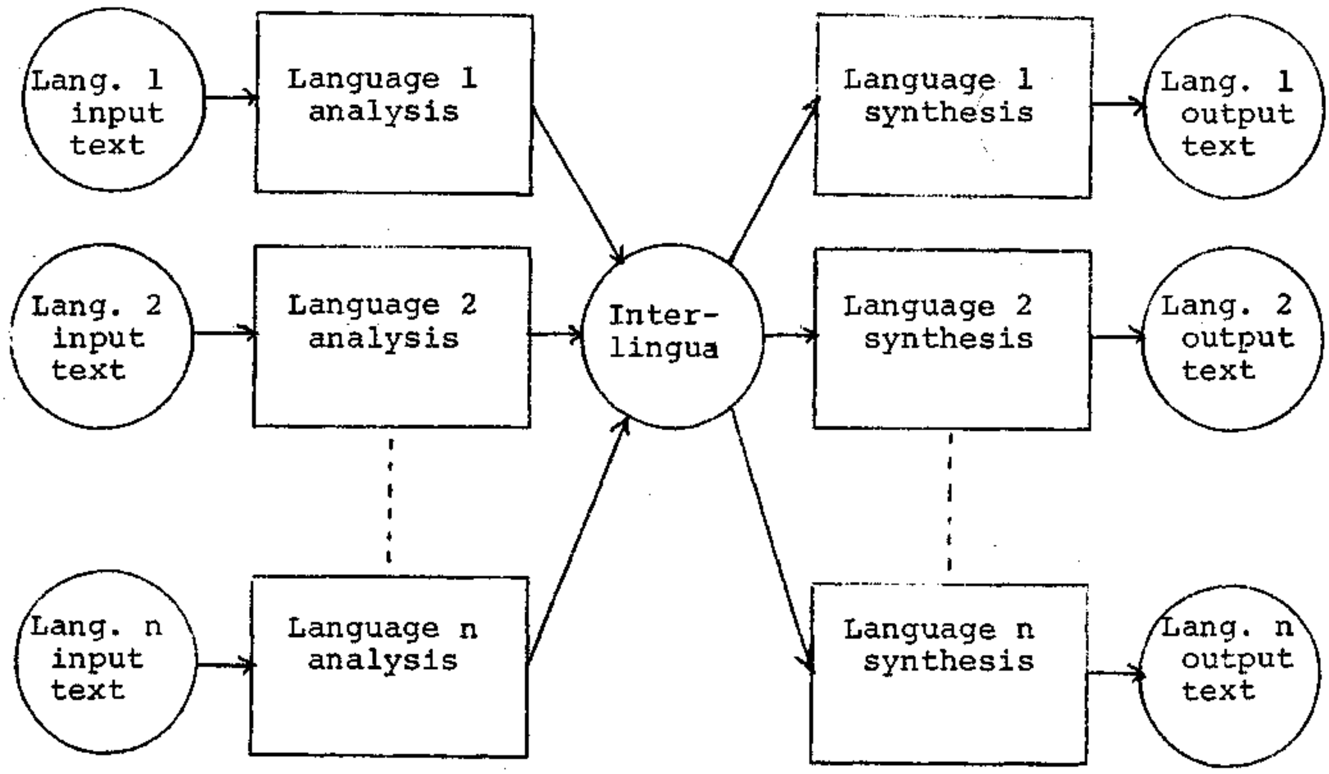
Schematic representation of first generation strategy



As the figure shows, the translation strategy is inextricably confounded with the properties of the two languages involved. The advantages of the second generation approach, separating source language analysis from target language synthesis, are clear; only one analysis module need be written for, say, Russian, whether the target language is to be German, English, French or any combination of these; and the design of the analysis of Russian can be entrusted to Russian speakers who need not have knowledge of all or any of the intended target languages. Similar considerations apply, *mutatis mutandis,* to the synthesis of any proposed target language. Logically, of course, such arguments taken to the extreme lead to the situation represented schematically in Figure 2, where all text is mapped internally onto an abstract universal language, or *interlingua.*

Figure 2

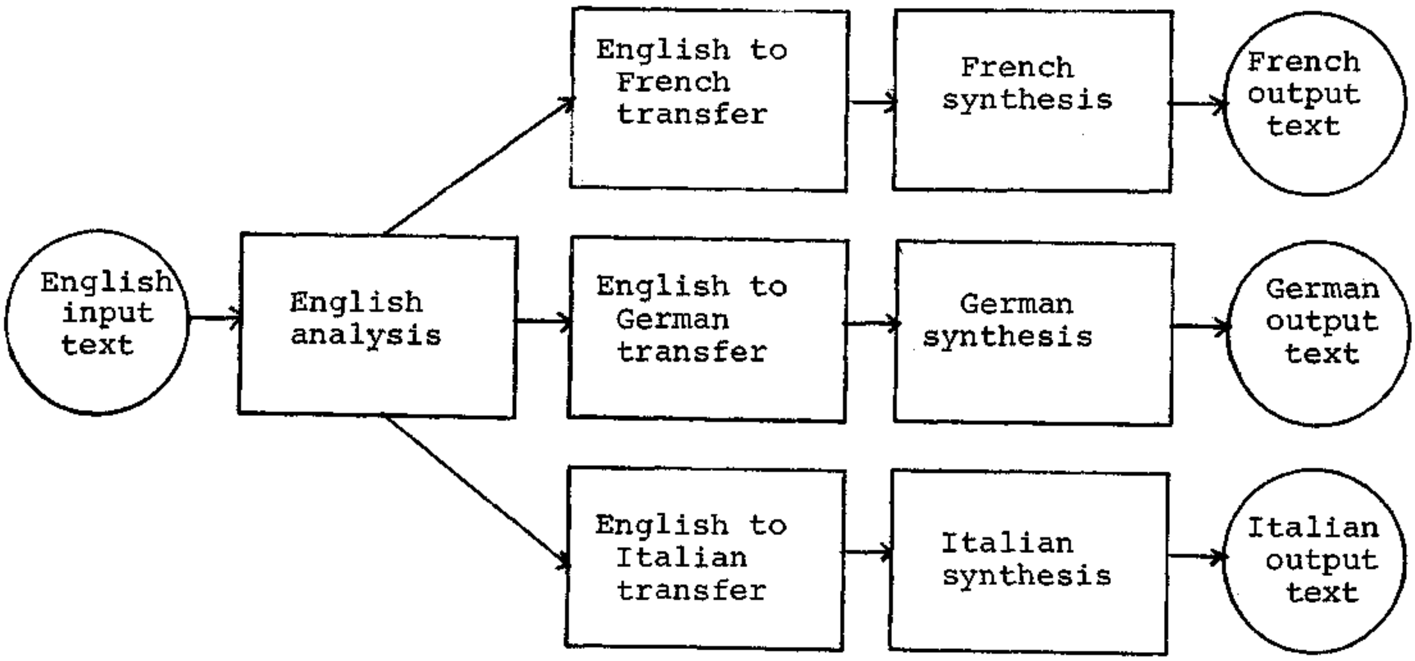
Translation scheme via interlingua



Very few workers in MT would claim, however, to have discovered sufficient universals of language and thought to propose a true interlingua, the claims of Esperanto and other eccentricities notwithstanding. The general consensus is rather towards a three-phase process of analysis, transfer and synthesis. Figure 3 shows a partial representation of a fragment of a transfer-oriented system involving translation from English to French, German and Italian.

Figure 3

Schematic of partial transfer-based system



Here, as much analysis as possible of English is done independently of the target language. Control is then passed to a transfer module which handles all structural differences peculiar to the particular language pair. Finally, the process is completed by a synthesis stage, which is again neutral with respect to the source language. Transfer-based systems are likely to continue to offer the most acceptable compromise for the foreseeable future.

- Incorporation of deeper linguistic theories

Consider an Italian-English first generation system faced with the text fragment

Il parlamento approvò il trattato

Morphological analysis followed by dictionary look-up would yield

il	*the* [ART, DEF, MASC]
parlamento	*parliament* [N, SING, MASC]
approvò	*approve* [V, 3RD PERSON, SING, PAST]
il	*the* [ART, DEF, MASC]
trattato	*treaty* [N, SING] */treat* [V, PAST, PART, SING, MASC]

Local rules can be formulated to exclude the sequence [ART] + [PAST PART] leaving only the interpretation

The parliament approved the treaty

after grafting of the appropriate suffix onto the verb stem *approve.*

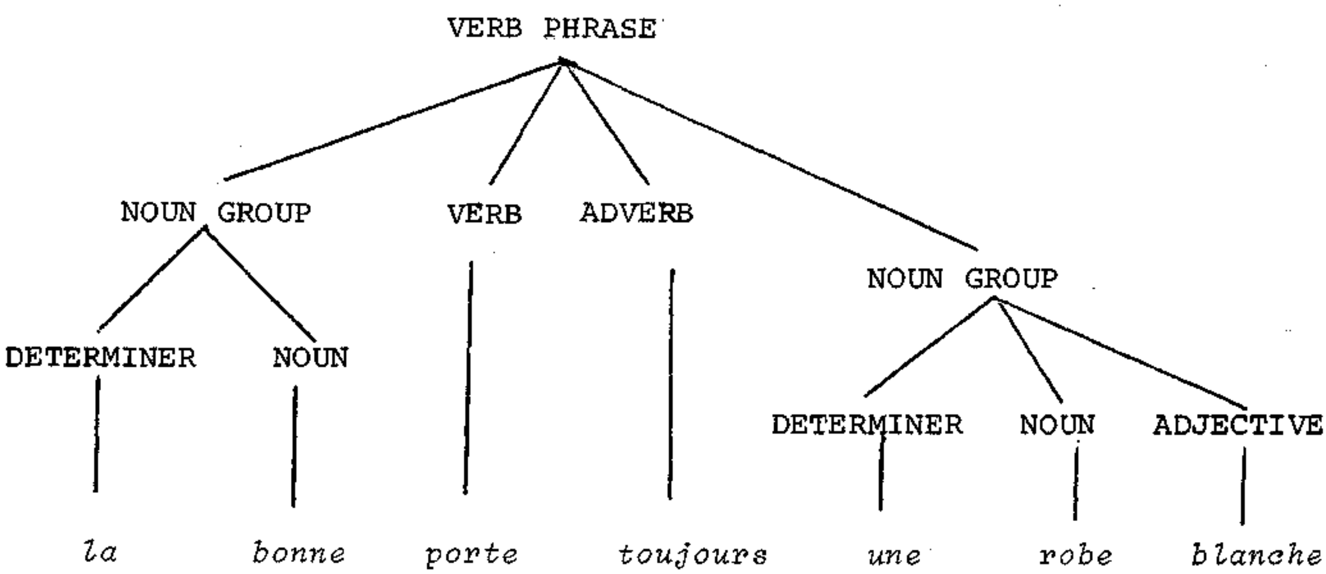
But consider now the French fragment

La bonne porte toujours une robe blanche

to be translated into English. The problem for the first generation approach is to discern that *porte* is a verb and not a feminine singular noun qualified by *bonne.* This is a problem which no local rule can in principle solve. The trick is to discover that there is no other local candidate verb, and that, given the context *....une robe blanche,* surface syntax dictates that a verb should precede, although it may be separated from the following nominal by an indeterminate number of qualifying adverbs. The answer is a generalised procedure which is able to discover the underlying structure, as, for example, in Figure 4, and to operate primarily on the major constituents *la bonne, porte, toujours* and *une robe blanche.* But the necessary formal linguistic theories and associated algorithms only become available when first generation systems were already in an advanced stage of development. Indeed, it is almost certainly no accident that the impressive advances in the formal theories of language in the years around 1960 should have coincided with the development of the first prototype second generation systems.

Figure 4

Constituent structure of *la bonne porte toujours une robe blanche*



- <u>Separation of algorithms from linguistic data</u>

The third serious shortcoming of first generation systems was the way in which any underlying linguistic theory which might have been present was inextricably bound up with the program itself. The disadvantage of this approach is clear: any modification to the linguistic model requires intervention by a skilled programmer; even then, as the linguistic model grows in complexity with development of the system, there comes a point where details of the model itself are lost in the intricacies of the program, and the system becomes so unstable as to make the consequences of even minor linguistic modifications potentially catastrophic. It is therefore a fundamental principle of second generation philosophy that it should be possible to entrust the expression of linguistic models to linguists, writing in a kind of very high level metalanguage designed explicitly for the representation of linguistic rules and needing only a minimum of knowledge of the design of the driving algorithms.

By the early 1970's, those second generation systems which had been begun a decade earlier, and had succeeded in evading the ALPAC debacle, were beginning to produce impressive results, particularly GETA in Grenoble (Vauquois 1975) and TAUM in Montreal (TAUM 1971). The renewed claims for MT prompted by the success of these systems, together with a number of important external factors, were to lead by the mid-1970s to a resurgence of interest in MT far earlier than anyone could possibly have envisaged when the ALPAC report was published less than a decade earlier. It is the reasons for the re-emergence of MT as a serious prospect that we shall examine in the next section.

THE MT RENAISSANCE

We have seen that by the 1970s MT systems were being developed which had succeeded in avoiding many of the defects of first generation MT. This fact alone, however, would not be sufficient to explain the dramatic revival of interest in the field toward the latter end of the decade - the very fact that Machine Translation is once more a primary topic in symposia like this one bears witness to its renewed importance. We also have to take into account the coincidence of two other important contributory factors.

First of all, as a result of the ALPAC recommendations, a number of scholars had been working since the mid-sixties in the related fields of artificial intelligence, computational linguistics and information retrieval on the problem of "understanding" natural languages, particularly in English. Out of this work, the first mature results of which were beginning to arrive in the early 1970s, came a great number of new insights into ways of handling semantic and pragmatic information inside the computer.

Secondly, the increasing political importance of language, combined with the increasing cost of employing skilled manpower, was beginning to lead to serious backlogs of texts requiring translation. This was particularly true in countries and organisations like Canada and the EEC, where the principle of corporate bi- or multi-lingualism was constitutionally established. It is not necessary to enter into detail here on this issue which is discussed elsewhere in this collection (McNaught, this volume) and amply illustrated in, for example, CEC (1977) and Snell (1979). Suffice it to say that a variant of the TAUM system, TAUM METEO, has been earning its living translating weather reports since 1976 (Chandioux 1976), and that in 1976 the European Commission purchased their first SYSTRAN version for English-French translation. The most significant development of all, for Europe at least, began in February 1978, when the European Commission assembled together a committee of experts from all the EEC member countries to discuss the design of a common European MT system, which would represent a synthesis of all the advances of the previous three decades and be a true reflection of the current state of the art. The planning of EUROTRA is now in an advanced stage, and it is hoped that implementation can begin during 1980.

THE NEXT GENERATION

It is not proposed that EUROTRA will be simply an enlarged version of existing second-generation systems. For all existing second-generation models, despite their superiority over earlier attempts, lack one important feature - a powerful semantic component. As early as 1964 Yngve had pointed out that

> "Work in mechanical translation has come up against a semantic barrier.... We have come face to face with the realisation that we will only have adequate mechanical translation when the machine can 'understand' what it is translating and this will be a very difficult task indeed..." (Yngve 1964).

Fifteen years later Yngve's statement is essentially just as pertinent as it was in 1964.

The kinds of problem which arise if we do not have an adequate semantic model are, for example:

- *homonymy:* in translating the English text fragment *All mechanics keep a file with a list of their tools,* how can the system know that *file* translates into Italian as, say, *archivio,* and not *lima?*

- *anaphora:* if we have *The pebble hit the window and it smashed (it),* how do we resolve the ambiguity of the pronoun *it?*

- *case/preposition:* again, from English into Italian, in *The contract must be signed by the end of the year,* how is the system able to reject the syntactically plausible *...dalla fine dell'anno* (agent) in favour of the correct *...entro la fine dell'anno* (time limit)?

In the first case, the system has to "know" that *files* are places where records are kept, without being fooled by the proximity of the generic term *tool,* closely related to the contextually inappropriate sense of *file* = *smoothing implement.* In the second, it has to "know" that when *pebbles* hit *windows* it is the *windows* that are likely to break, not the *pebbles.* Only in this way can the correct gender agreement be assigned to the pronoun in, say, French, if we assume, in the dictionary, c*aillou* = *pebble* (masculine) and *fenêtre* = *window* (feminine). Finally, in the third example, it must again "know" that *ends of years* cannot sign anything while time limits on signature are quite normal.

For evidence of a breakthrough of semantic processing it is customary to look to artificial intelligence, where the greater part of work on language processing has centred on the problem of mapping text onto knowledge structures - that is of a kind of "understanding". Artificial Intelligence researchers like Schank, Woods, Winograd and Wilks have achieved significant results in simulating aspects of human linguistic understanding on the computer.

It has long been argued by the sceptics, however, that to understand language in any general context what is needed is an encyclopaedic store of knowledge, and that "the number of facts we human beings know is, in a certain very pregnant sense, infinite." (Bar-Hillel 1964). Any attempt to implement an encyclopaedia within an MT system, even if we knew how to represent it, would clearly be impractical. The same sceptics point to the successes in Artificial Intelligence natural language understanding and show that they have been successful precisely and only because they have chosen to limit their universe of discourse to microscopic domains incomparably narrower than the breadth of knowledge required to translate even the most specialised texts. Woods (Woods *et al* 1972), they argue, parses into a data-base restricted to a small set of facts about lunar geology, and his program is reduced to a moron when faced with any but the small subset of possible questions about that limited domain. Winograd's "robot" SHRDLU (Winograd 1972) only "knows" about a universe of toy blocks. Wilks' (Wilks 1973) program produces good translation, certainly, but it only has a vocabulary of about 500 words.

All this is very true. But many workers in MT would argue that the sceptics, while correct in what they say, have missed the point. The central problem in translating technical text by computer is not one of "understanding" the source text, it is a problem of *disambiguation.* If the source text is not ambiguous then a second-generation MT system will translate it, and the translation will be a correct if not always an elegant one. The appropriate kind of disambiguation has been done - for example by Wilks - for a surprisingly high number of instances. To do it consistently in MT systems we will need to incorporate ideas from a wide variety of sources - especially from Case Grammar (cf. Rosner and Somers, forthcoming), from the much-maligned syntactico-semantic theory first proposed by Katz and Fodor (Katz and Fodor 1964) and, inevitably, from Artificial Intelligence semantics.

Wilks, moreover, has shown that it is possible to perform very sophisticated disambiguation, including cases of the types suggested above, *without* recourse to encyclopaedic "knowledge", but with rather simple, even naive-looking, dictionary entries based on structured combinations of semantic primitives. The formalism for dictionary entries used by Wilks is a LISP representation, and the labels he uses for constructing his semantic "formulas" are rather idiosyncratic, so that his examples are difficult to read directly. But, for instance, his dictionary entry for *buy,* rendered roughly into English, would be

> "buying is what humans do when they acquire things by giving money in order to use them" (cf. Wilks 1976, p. 168).

It is also theoretically possible (cf. Boitet 1977) that the representation and strategies used by Wilks can be accommodated within the basic framework of an advanced second generation system, so that there is no theoretical obstacle to the incorporation of a powerful semantics *à la* Wilks into the next generation of MT systems.

CONCLUSIONS

The history of MT has not always been a happy one. Many mistakes were made and insistence on outworn conceptions and outdated theories has done little to enhance the reputation of the computer as an alternative translator. But we may have reached a point where external pressures make research into MT more than an esoteric academic pursuit. MT researchers believe that they now have the answers to many of the problems that have in the past proved so intractable. It is now clear that a high proportion of even the most serious problems of disambiguity can be solved by new and powerful semantic methods, and that the time is ripe for the third generation of MT systems.

It seems to me that the present climate is summed up most appropriately in the final remarks of the excellent review paper by Hutchins (1978):

> "There is now a mood of quiet optimism in MT research: it is a mood which should not be taken lightly."

REFERENCES

ALPAC (1966) *Language and Machines: Report of the Automated Language Processing Advisory Committee,* Washington DC, National Academy of Sciences National Research Council.

Bar-Hillel, Y. (1964) 'The present status of automatic translation of language' in *Advances in Computers* (F.L. Alt, ed.), New York, Academic Press.

Boitet, Ch. (1977) *Mechanical translation and the problem of understanding natural languages,* Grenoble, GETA.

C.E.C. (1977) *Overcoming the language barrier,* München, Verlag Dokumentation, 2 vols.

Chandioux, J. (1976) 'METEO: un système operationnel pour la traduction automatique des bulletins météorologiques destinés au grand public', *META,* 21, pp. 127-133.

Dreyfus, H.L. (1972) *What computers can't do,* New York, Harper and Row.

Hutchins, W.J. (1978) 'Machine translation and machine aided translation', *Journal of Documentation,* 34/2, pp. 119-159.

Katz, J.J. and Fodor, J.A. (1964) 'The structure of a semantic theory' in *The structure of language* (Fodor, J.A. and Katz, J.J., eds.), Englewood Cliffs, Prentice Hall.

Oettinger, A.G. (1963) 'The state of the art in automatic language translation: an appraisal' in *Beiträge zur Sprachkunde und Informations Verarbeitung* 1/2 (Marchl, H., ed), Munich, Oldenbourg Verlag.

Rosner, M.and Somers/ H. (forthcoming) *Case in linguistics and cognitive science, ISSCO* Working Paper no. 40, Geneva, Istituto per gli Studi Semantici e Cognitivi (to appear in 1980)

Snell, B.M. (ed.) (1979) *Translating and the computer,* Amsterdam, North-Holland.

TAUM (1971) *Projet de Traduction Automatique de l'Université de Montréal* (Colmerauer, A., ed.), Montréal, TAUM.

Vauquois, B. (1975) *La traduation automatique à Grenoble,* Paris, Dunod.

Wilks, Y.A. (1973) 'An Artificial Intelligence approach to Machine Translation', in Schank, R.C. and Colby, K.M. (eds.), *Computer Models of thought and language,* San Francisco, Freeman.

Wilks, Y.A. (1976) 'Parsing English II', Chapter 9 of Charniak, E. and Wilks, Y.A. (eds.), *Computational Semantics,* Amsterdam, North-Holland.

Winograd, T. (1972) *Understanding natural language.* New York, Academic Press.

Woods, W.A., Kaplan, R.M. and Nash-Webber, B. (1972) *The Lunar Sciences Natural Language Information System: Final Report,* BBN Report 2378, Cambridge, Mass., Bolt, Beranek and Newman Inc.

Yngve, V.H. (1964) 'Implications of mechanical translation research', in *Proc. Amer. Phil. Soc.,* 108/4.