[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.]

THE AMBIGUOUS TASK OF MACHINE TRANSLATION

G. C. Keil
Centre for Computational Linguistics
University of Manchester Institute of Science and Technology

Current optimism about the prospects of Machine Translation (MT) owes as much to activities outside its own sphere of interest as it does to the efforts of its own specialists. Developments in large-scale computer hardware have increased the feasibility of highly complex and storage-hungry systems, while the widespread introduction of small computer and word-processing systems into everyday life has brought more people into direct contact with the computer, rather than simply under its detached influence. Increased social acceptance and awareness of the computer in general seems to be occurring at just the time when the computer seems at last technically up to the job of MT.

For all practical purposes, the ALPAC report[1] cut off the money supply for MT research in name only; in all those fields from which MT draws its support, progress has flourished: artificial language processing, library science, computer-aided instruction, information retrieval, database technology, terminology and lexicography, semantics, theoretical and mathematical linguistics, and - dare I say it - artificial intelligence. This suggestion of progress by proxy is not to deny the contributions and undoubted successes of those centres which carried on after the ALPAC report broke both the euphoria and the bank; it is meant rather to highlight a fundamental difference between the circumstances which held in the 50's and early 60's, and those which hold now. Then, the funds which visibly went into MT produced disappointing results in MT, but proffered immeasurable benefits to related fields. Now at last, the once-bankrupt field of MT can turn the tables and profit from its erstwhile beneficiaries.

The remaining task of MT is by no means inconsiderable. Of course it has the derivative task of knitting together

techniques from numerous sources into a coherent and relevant whole; but there remain a set of problems uniquely associated with natural language processing - a set which can be characterised with the single word, "ambiguity." Ambiguity is manifested at every level of natural language, and the relationships between ambiguities at different levels can be highly complex. Ask a linguistically 'naive' informant for examples of ambiguity and he will quite readily come up with examples like:

> "glass - that can be something to drink from, or the stuff a glass is made from, or something to see yourself in."
>
> "car - something you can drive, or part of a train."
>
> "run - what I do to catch the morning train, or something a cricketer scores."
>
> "bread - something to eat, or a slang word for money."
>
> "ear - a thing on the side of my head, but if you say, 'give me your ear', you mean, 'listen to me.'"

The informant will probably think of these and other examples as differences in the "meanings of words". In fact, many apparent ambiguities are only secondarily lexical. "glass" is far less ambiguous if it can first be assigned to one of the noun classes 'count'/'non-count':

> "These two glasses are broken." (Mirrors or tumblers)
> "Glass is breakable." (Material)

Only then are we possibly faced with the mirror-tumbler ambiguity, i.e., after we have used morpho-syntactic information (or at least non-lexical information) to identify the sub-class. Given

> "This glass is broken" (Material, or mirror/tumbler)

it is still non-lexical information which signals the primary ambiguity of noun class.

The "car" example more clearly represents a lexical ambiguity; nevertheless, a quick glance at a few examples in context will show that the ambiguity is much more complex than

simple lexical distinctions can either identify or resolve:

"This is the restaurant car."
"Here is my car / the car I bought recently."
"My briefcase is in his car."
"His car is in my briefcase."
"Don't say car, say automobile."
"Cars will please queue here."
"Will you be going by car?"

The informant's other examples likewise reveal ambiguity at other levels in addition to lexical. The "bread" example implies ambiguity of linguistic register (slang or non-slang), "run" of grammatical category (verb or noun), "ear" of idiomatic usage (or of synecdoche). Many types of ambiguity do not directly involve lexical considerations at all. Is -ly an adverbialiser (as in "strictly"), an adjectiviser (as in "sickly"), either (as in "poorly"), or non-segmentable (as in "silly")? In the sentence, "I have fried eggs," is "have" an auxiliary or a verb? Given "I like bananas better than you" does "you" contrast with "I" or with "bananas"? - the ambiguity here being between two possible forms of ellipsis. "Tom saw the girl with binoculars" leaves in doubt who had the binoculars - Tom or the girl.

The detection of ambiguities is a more acute problem for machine analysis than it is for humans. Machine analysis must proceed stepwise in one fashion or another, e.g., linear/predictive, top-down, bottom-up; at each step it must identify potential ambiguities which may be resolved at a subsequent step. For example, the two utterances "He is going" and "He is angry" will each produce ambiguous interpretations of "is" at one stage - either copula or auxiliary. This ambiguity will in each case be resolved at a later stage when the verb phrase is analysed as a whole, excluding respectively the copula and the auxiliary. An utterance is said to be <u>inherently ambiguous</u> when ambiguities detected at any stage fail to be disambiguated by the end of the process of analysis, as in the example, "This glass is broken." A necessary quality of an MT system of any worth is that, for any given utterance, inherently ambiguous features discoverable by humans will also

be discoverable by machine.

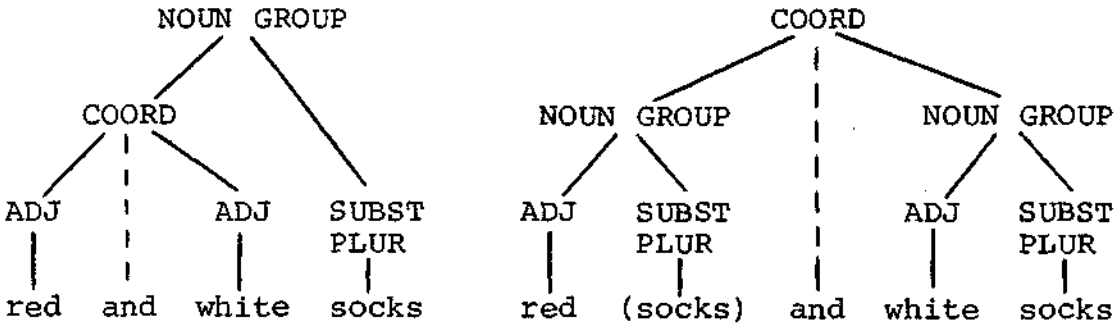
Many of the problems of ambiguity - and from now on the term will imply "inherent ambiguity" - become apparent only when two or more languages are contrasted. Does "know" in English translate into German/French as "wissen/savoir" or as "kennen/connaître"? Does "his" translate into Danish as "hans" or "sin"? Do these examples imply that "know" and "his" are inherently ambiguous in English? Care is needed here. If the answer is "yes, ambiguous in principle", then we are making some serious teleological implications: a word is ambiguous in one language because it has alternative equivalents in another. (What if no speaker of English knew any Danish, and vice versa? Would the word "his" then be unambiguous in English - or perhaps only "existentially unambiguous"?) If on the other hand we mean "tactically ambiguous", then to some extent we are predetermining the way we will handle ambiguity in an MT system.

Bilingual and multilingual contrast reveals another important dimension to the question of ambiguity - one which has telling implications for the methodology of MT. Take the following example:

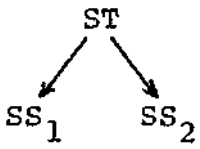
"Here are my red and white socks."
("Voici mes chaussettes rouges et blanches.")

Both versions are inherently ambiguous - but as it happens, identically so. The fact that the ambiguity cannot be resolved easily, if at all, ceases in cases such as this to be a problem, so far as translation is concerned. But somehow the MT system has to "know" that there is indeed no resolution problem, precisely to know that it needn't seek a solution - and that, as the following discourse will reveal, is an expensive undertaking. Let us examine the process which this facility implies, using the above example.

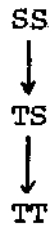
Analysis establishes that there are two possible interpretations of "red and white socks": the first assumes that "red and white" are coordinated adjectives modifying "socks", while the second assumes "red" modifies one group of "socks" (removed by ellipsis) and "white" modifies *a* second.



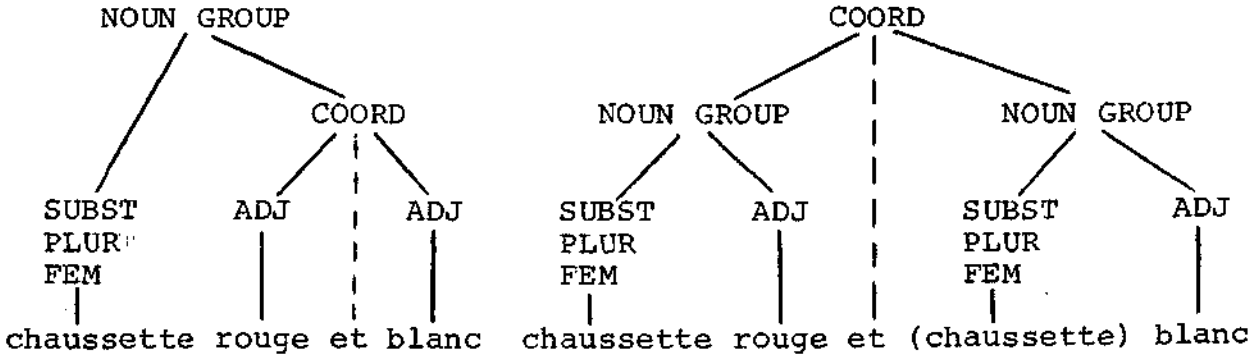
In other words, the source text (ST) results in the production of two alternate source analyses (SS):



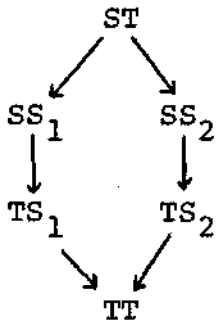
Upon transferring from the source-language structure to the target-language structure (TS), the utterance has been prepared for synthesis into a target-language text (TT). In the simple case, only a single TS occurs:



This example results in two target structures, TS_1 and TS_2, from SS_1 and SS_2, respectively:



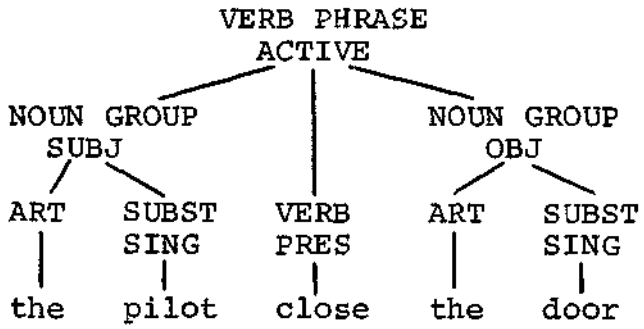
Each of these structures results in the identical surface text: "chaussettes rouges et blanches." Diagrammatically:



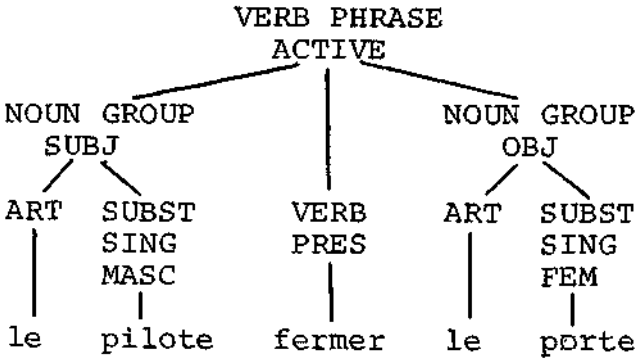
This procedure as described so far, however, still does not guarantee absolutely that the source and target texts are equivalently ambiguous. Consider the following example:

"The pilot closes the door."
("Le pilote ferme la porte.")

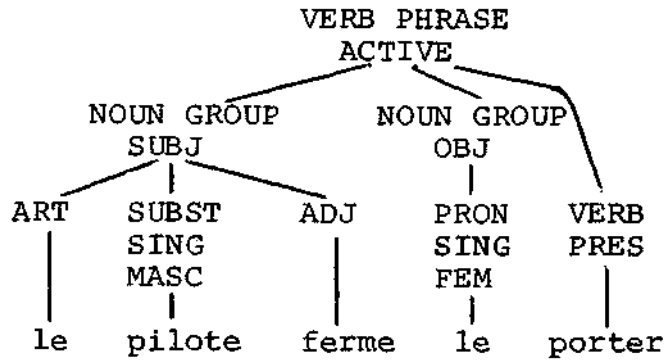
Analysis of the English source text produces a single SS,



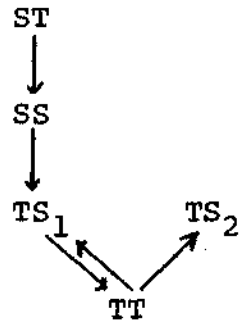
...which in turn results in a single TS from transfer,



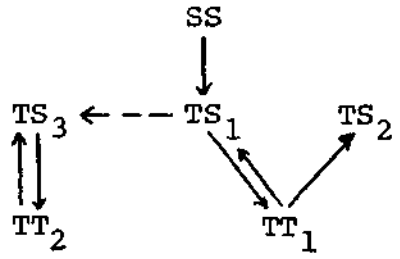
resulting in the single surface realisation, "Le pilote ferme la porte. But if we re-analyse the target text we derive two analyses - the SS given above, plus the following:



Diagrammatically, this phenomenon can be represented as follows:



It would of course be possible (barring objections on stylistic grounds) to backtrack at this stage to the original TS and derive an alternate TS_3, such as one which would produce the target text "La porte est fermée par le pilote." This solution has at least the advantage that analysis of the new TT produces only the TS from which it had been derived:



and - always presuming that reverse transfer from TS_3 does not result in alternate SS's - it may be concluded that translation has not resulted in the creation of new ambiguities in the target text.

The examples given suggest a process of forward and backward analysis, transfer and synthesis which is repeated until stability is attained. At each task step (a "task step" is represented by one arrow in the diagrams) ambiguities must be recognised - or remembered from previous task steps - and represented as alternate productions. In other words, the process produces a bidirectional series of potentially one-to-many mappings, which may converge: the control structure of the process is therefore a multigraph, and stability is represented by the existence of at least one TT from which the set of all possible traversals back to the original ST includes visits to all and only the SS's generated as a result of the detection of source-text ambiguities. Such a process is possible, but the following observations have to be made:-

- the process is expensive. Each task step alone consumes considerable computing resources, and the number of steps increases in geometric proportion to the number of alternative productions resulting from any single step;

- the examples were atypical in that they were simple: typical translation units (generally, whole sentences) contain multiple ambiguities, and the number of alternatives at any task step is the product of the number of ambiguities in the translation unit;

- the process is incapable in itself of deciding whether resolution of ambiguity is justified in any single case. An ambiguity discovered by machine may not even be immediately apparent to a human reader, who has recourse to world knowledge, wide textual context and - occasionally - common sense;

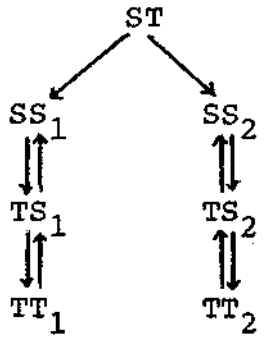
- this process is applicable only in the following circumstances:

  (a) there are no ambiguities, either in the source or the target text (in which case the process confirms the fact);

(b) there are parallel ambiguities in the source and target texts (in which case the process discovers the parallels);

(c) of the possible alternate target texts, there is at least one for which either (a) or (b) holds.

There remains a condition, however, in which the process can provide no solution: that is, where there are inherent ambiguities in the source text which do not map onto equivalent ambiguities in the target text. That would have been the case if French had been the source language in the above example, and an English translation were sought for "Le pilote ferme la porte." The diagram would have been:



The only possible way to obtain a correct translation in this case is to resolve the ambiguity literally at source. Whether or not such a solution can ever be practicable in an MT system for all possible instances of ambiguity is still an open question; what is certain is that any solution, if it exists, is going to be highly sophisticated and equally highly expensive.

The conclusion to the discourse above must gladden the heart of any translator (or indeed anyone else) who fears or resents the intrusion of the computer into traditional areas of human activity - but it is no more than people involved in MT have always maintained: that machines cannot replace humans, at least so far as ultra-high-quality translation is concerned, in any forseeable future. The obvious, perhaps the only, application of pure machine translation is in bulk production of hack translations, and in the real world the merits of MT

will be judged by purely pragmatic rather than esoteric criteria. There are many areas in which machine translation is adequate, and many others in which it will have to suffice simply because human resources are not available to do the work: examples are (a) translation of routine material such as sub-committee minutes, draft reports and memoranda; and (b) production of rough translation for content scanning, prior to selection of material for quality translation. Of ultimately far greater importance, however, is the part MT can play as one aspect of an integrated translation facility which includes not only "pure" machine and "pure" human translation, but also a variety of "hybrid" translation processes, in which the human can aid the machine (through pre- and post-editing of machine translations, or through interaction on a computer terminal with the translation process itself), or alternatively, the machine can aid the human (through text preparation and editing facilities, document and information storage, classification and retrieval, on-line dictionaries and term banks).

It is not easy to appreciate from "outside" the positive advantages which MT and its by-products can afford to humans. Resistance to change, especially to automation, is understandable; but resistance through ignorance of the alternatives is irresponsible and self-defeating. Automation at all levels of professional translation is inevitable, and largely already here; but the sensible design and application of that automation urgently requires the advice and expertise of the very individual who is most reluctant to participate positively in the dialogue - the professional translator himself, whose daily life will be most affected by the changes. The "ambiguous task of machine translation" is not so much technical as it is social - and that is one ambiguity which can only be resolved by humans.

References

1. Report by the Automatic Language Processing Advisory Committee (ALPAC), Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C. (1966): *Languages and Machines: Computers in Translation and Linguistics.*