[From: *Information Processing 1965*: Proceedings of the IFIP Congress 65 (Washington, D.C.: Spartan Books, 1965)]

SEMANTIC PROBLEMS OF MACHINE TRANSLATION

R. M. NEEDHAM

University Mathematical Laboratory
Cambridge, England

INTRODUCTION

It is by now almost a cliché to say that semantic problems are the main obstacle to progress in Machine Translation. My purpose here is to sort out to some extent what the problems are and on what line solutions should be sought; the paper is therefore directed to MT rather than to semantic problems in academic linguistics. This emphasis determines my choice between two classes of problem for discussion: firstly, questions about the semantic relatedness of pairs of grammatically related utterances *(preservation of meaning under transformation)* and secondly, problems of discovering which of several possible interpretations of a piece of discourse is appropriate *(semantic ambiguity).* I am mostly interested in the second. This is not to say that the first would not figure in an MT process at all, but that it is in such a process logically subsequent to the second.

DEALING WITH AMBIGUITIES

This second type of problem, loosely called *dealing with ambiguities* or some such, itself appears in two guises. One occurs when a sentence may apparently be parsed in several significantly different ways, all renderings being equally consonant with the grammatical rules, so that we may have to reject some on a ground that may informally be stated as "on such and such a parsing, the sentence is saying something we don't think it is very likely to be saying" (it is *deviant).* The other occurs when a word in a text may be translated in a variety of ways within one parsing, and some of them are similarly distasteful. It is to be emphasised that in each case I said that *some* of them are to be rejected, and *not* "all but one of them." I doubt whether anyone believes that the rejection process is straightforward: however, a simple experiment will indicate just where some of the problems are. If we take a sentence in another language, preferably one that does not require too much grammatical rehash on translation, and write down a number of possible alternative renderings for the words, we may, as a purely human matter, write down some of the reasons why some of the sentences are rejected.

Consider the sentence (from Camus. *La Peste)*

> *L'un d'eux, le chapeau de paille en arrière, une chemise blanche ouverte sur une poitrine couleur de terre brûlée, se leva à l'entrée de Cottard.*

A rather hack translation is:

> One of them, with a straw hat on the back of his head, and a white shirt open over a terracotta-coloured chest, got up when Cottard came in.

Now,

- Why is *paille* not rendered as *strawcoloured?* For syntactic reasons—it is not an adjective here.
- Why is *paille* not rendered as *flaw* (in the sense of *fault*)? Because the sentence requires a material, and a flaw is not a material.
- Why is *paille* not rendered as *chaff*? A tricky one, this. Because *chaff hat* is not a thing we say, or because hats are not made of chaff?

Anyone interested may amuse himself by writing a careful blow-by-blow account of just why the erroneous parts of the following rendering must be rejected:

> One of them, a chaff cap in arrears, a blank coat-of-mail unfortified on a brisket

with the appearance of a parched estate,
got up when Cottard came in.

CORPUS OF STORED INFORMATION

The one example I gave, however, would be text for a long sermon, although it did not exemplify the whole gamut of possible reasons. What we now have to consider is the possibility of mechanising these decisions, i.e., of making them on the basis of a statable set of rules and a known corpus of stored information.

Case 1

Case 1 need not detain us. Anyone should be prepared to admit, at least provisionally, the possibility of this sort of thing.

Case 2

Case 2 is much harder: a) how do we know that a material is needed, and b) how do we know that straw is a material?

A. — How Do We Know that a Material Is Needed?

The elementary answer is that we do not know any such thing, and the reason given for rejecting this sense of *paille* is incomplete. There is no good reason to suppose that we must have a material here: consider *chapeau de marin*, a sailor's hat. It is much more the case that, although various kinds of word will do, *paille* in the sense of flaw is not one of them. Query, is this because 'flaw' is an unacceptable type of word, or is it that it is not one of the acceptable types of word? Since we are not dealing with a closed system, it is not trivial to ask which is the marked case. Again are we making some intrinsically bad assumption when we talk about 'types' of word?

B — How Do We Know that Straw Is a Material?

We can only hope to mechanize the decision that straw is a material if we can discover this class-membership by reference to stored information. How reasonable this is will be discussed later.

Case 3

Case 3 is harder still. It brings up what is perhaps the thorniest problem of all, which is sometimes put in the form: for MT, do we need a dictionary or an encyclopedia? That hats are not made of chaff is a statement of fact and could (may well) be false. Hats are, after all, made from a wide selection of materials. It is not possible to rely on a collection of facts, to which a translation procedure could make reference, in order to find out what was and was not appropriate. Two different but strong reasons support this statement:

A.—Lack of Relevant Facts

No collection of all relevant facts could be made (it is not material to discuss whether this is logically or medically so).

B. — What Is Passage Discussing?

If the passage is talking about a chaff hat, then we must translate it as talking about a *chaff hat, whatever hats are made of.*

An important fact about the experiment and comments is that we have rejected alternatives because they do not satisfy some stated criterion. What right have we to expect that one or more of the possibilities *will* satisfy some previously stated criterion? The answer, I believe, is necessarily *none.* To require that everything in a text must satisfy some of a previously stated set of semantic criteria is to require that the criteria cater for everything that can be said. (It might be argued that they could cater for everything *meaningful* which can be said. To the extent that this remark is meaningful, it is false.) This is patent nonsense, and one of the great differences between syntactics and semantics: it is commonly supposed that a good grammar describes and delimits the whole corpus of grammatical sentences of a language.

From these remarks follows a basic fact about the rules and criteria to be used in making semantic decisions. They must be such that they can, when interrogated, simply give no answer, i.e., be inapplicable. This must be sharply distinguished from rules which always give an answer, even a noncommittal one. If rules do this, then they cover anything which can be said and will accordingly be trivial.

CLASSIFICATION

A customary answer to some of the problems I have raised is CLASSIFICATION. (An appeal to classification has already been made in the exam-

ple.) I now want to consider what classification does in this context, and what its scope and limitations are. Suppose we go back once again to the case of *chapeau de paille*, where a correct treatment of *paille* depends, among other things, on a knowledge of the kinds of material hats are made of or are said to be made of.

It might be possible to assemble a list of such substances, and simply remember the list and refer to it. It might be possible: but it is not, for a hat can at any time be made of a new substance, and can even more easily be described as being made from a new substance. This is exactly the same thing as saying that we cannot define the class of things hats can be made of, and it is clear that if classification is to help it must do something else. However, if we require that what comes after *chapeau de* belongs, when it is in the sense of material rather than use, to any other class than the inapplicable one just mentioned, then we are certainly not in a position to make the selections and rejections which we would like to make. What then can classification do? Apparently we must have either an *ad hoc* class of no generality, which cannot be set up anyway, or else a class guaranteed capable of giving some wrong answers.

Here we must return to the point of the previous passage: we must accept that our semantic rules will sometimes be unhelpful. If there are several alternative possibilities, and after weeding out what is in common, there is no discriminatory information left at all, this is simply unfortunate. How much useful discrimination can be achieved is at present a matter for experiment or opinion.

It is hardly possible to give an exact statement of one's opinions on this point, but I would commit myself to the following assertion:

> It is not reasonable to expect a mechanical procedure to discriminate between *straw* and *chaff* in the example above, other than on the dubious ground that, other things being equal, *paille* is more normally to be rendered *straw*.

The word *reasonable* is to be emphasized. We cannot brand it as self-contradictory to suppose the discrimination possible, for there is nothing self-contradictory in the notion of an explicit class of materials from which hats are not made, which we look up and find *chaff* in. There are few pieces of practical knowledge which cannot be incorporated in a classification somehow, but that is not the point. Granted the impossibility of assembling all the facts, it is equally impossible to make a classification embodying all the facts. However, it is as erroneous to suppose that classification does nothing as to suppose that it does everything. It would be very hard not to grant the status of *material* to straw and chaff, and to deny it to a fault. And this, as we have seen, can be helpful. The boundary comes in a place determined by considerations of reasonableness rather than of logic or linguistics, and it is this very fact that has led to so much of the controversy about classification for semantic analysis. The question is much more like that of determining the correct level of income tax than it is like determining the present vocabulary of English (which could in principle be done by experiment) or the truth of Fermat's last theorem (which can in principle be done by thought).

Given this, it is clear that there is no absolutely *correct* classification; only more and less useful ones. And there is no way of avoiding the problem which occurs in all subjects where classification is seriously used: in a case of difficulty, was it that something was wrongly classified, was it that the scheme of classification was wrong, or were we trying to do something which it was unreasonable to expect our method to achieve? This problem came up in a very sharp way in pioneering experiments done at the Cambridge Language Research Unit in 1956-1957, in which *Roget's Thesaurus* was used as a semantic classification for the kind of purpose I have been talking about.[1,2] (It is from these experiments and much subsequent work by the Unit, that my present views are derived. They are, however, purely personal.) The procedure then adopted depended upon looking up each word in the index to the Thesaurus, finding there the list of headings or sections in which the word appeared, i.e., finding how it was semantically classified, and looking for repetitions among the lists thus found for the words in a sentence.* Often an expected repetition failed to occur. Were we wrong in expecting it? Had the word been inadequately treated in the Thesaurus? Was the very structure of the Thesaurus such that repetition could not be expected? If the Thesaurus were amended to deal with this case, would something go wrong somewhere else? No way could then be found of answering these questions. Something further can, however, now be said about some of them.

*The reasons why repetitions were sought are described in the reports of the period. What concerns us here is that the experiments depended upon certain words occurring in certain classes.

FORMS OF UTTERANCE

Progress has been made in attacking the first of them, i.e., where repetitions can be expected, or, more generally, where a word should be expected to be a member of a certain class. To return to our original example, this is the problem of knowing that, among other things, a word in the class *material* is acceptable after *chapeau de.* It can reasonably be claimed that we know this ourselves because *chapeau* is a THING, and one of the ways a THING can be qualified is by giving the MATERIAL of which it is made. While it is unthinkable to have said in advance that hats can be made of straw, it is much more possible to say in advance that THINGS are made of MATERIALS; and if we have a way of recording this assertion we have embarked on a new kind of classification—the classification of *forms of utterance.* This must be carefully distinguished from a grammatical description of forms of utterance; and we must see why the kind of classification in question is not merely an extension of grammar. As with the earlier kind of semantic classification, we must be prepared to fail to recognize what we have in terms of an existing corpus of semantic information, and not to mind, in particular, if this happens when there is no ambiguity (no choices to be made). This contingency is wholly alien to grammatical analysis, and should it occur it is upon the grammatical analysis that we must lean. Whether the very form of the semantic rules is also alien to grammar would be a matter for investigation.

The problems of classification of the present sort, where a part of a text is recognised as an instance of some more general semantic assertion, has been very extensively pursued by Margaret Masterman.[3] It remains unclear how many kinds of utterance must be recognized; the simplicity of the statement of a standard form in Miss Masterman's work is very striking, but there could still be a great many possibilities. The feasibility of such a classification must be determined by experiment, and the experimental design is not easy enough to expect rapid results.

CONSTRICTING A THESAURUS

Another line of inquiry is into methods of constructing a thesaurus or similar semantic classification. If we inspect *Roget's Thesaurus,* it is obvious that it is the product of a good deal of high-powered thought, but it is not obvious what exactly were the principles employed. We seek almost in vain for any general statement about how a section should be constructed, or what criteria should be used for deciding whether to put a particular word in a particular section or not. It is also unclear on what principle it is to be decided whether to have more than one heading in a particular sub-field: for instance whether the agents of an activity should be in separate sections, e.g., *Deceiver* and *Deception* in Roget, or together, e.g., *Killing* in Roget. Attempts to sort this kind of thing out rapidly suggest that some kind of root-and-branch attack should be made, but how? The extreme case of a root-and-branch attack is to suggest that the whole could be set up on the basis of statistical examination of a mass of nondeviant text. This would lead to a collection of association coefficients to be followed by an automatic classification procedure. If this could be done, one would at any rate be assured that there was some uniformity about the product. However it is *prima facie* impracticable to do any such thing. Arguments about the practical impossibility of actually processing enough data are too familiar to need reiterating. It is nevertheless of interest to consider what would happen in the hypothetical case of our being able to process an indefinite amount of text and perform indefinitely complex calculations upon it.

Firstly, no matter what quantity of text we processed, we would not have examined instances of everything which could be said. An automatic classification would therefore have to make inferences from what had been said to what might be said. There is no evidence that this inference would avoid inserting just the same kind of potential errors into the classification as are introduced by a human classifier making what is, after all, the same kind of inference. The introduction of potential errors is not of itself a criticism: it was argued above that this is a necessary accompaniment of a usable semantic classification. It is not the case, then, that a statistical approach would produce the (delusory) "perfect semantic classification," and I have no evidence that it would produce a less erroneous, or less badly erroneous, version than other methods.

Secondly, although it would be possible to exclude, given an indefinite amount of material, any stated source of noise in the method of analysis, this definition of what was to be excluded could only be made on the basis of the results of the experiment. And it is not clear that this process would terminate.

LIMITING THE CLASSES

The bugbear, on the other hand, of humanly constructed classifications is the variability of the products of different people, who interpret the classes in different ways. It is possible to use various devices which have the effect of limiting the freedom of choice of the person making the classification, and thus perhaps ensuring greater reliability.

One of these is to limit drastically to 50 or 100, instead of Roget's 1,000, the number of classes used.[3] The classes are then quite sharply different, and while it may still be a subject of debate whether a word should go in them or not, it is less likely to be difficult to decide whether a word goes in one or other of two related classes. Several variants of this approach have been investigated at the C.L.R.U., and it certainly seems to be easier to apply than more multifarious classifications.

It is also possible to limit the scope for human judgement in another way, by getting people to write down small sets of words which are related in some way which they can be taught to recognize (for example, using a well-defined notion of "synonymy in some context"), and then using machines to sort out a classification from the assembled results, rather than having predetermined classes which the people must use. This approach has been studied by K. Sparck Jones;[4] it appears to be relatively easy to assemble the data, but quite difficult to process on an adequately large scale, though the obstacles do not seem insuperable as with the purely statistical method.

CONCLUSION

This paper has been a brief survey of a large subject. By keeping narrowly to the discussion of one or two particular cases, and some reflections not too far separated from them, there is no doubt that I have missed many of the possibilities in semantic studies. I hope, nevertheless, that some indication has been given of the kinds of problem which arise, and the approaches which can be made to studying them. Much remains to be done, and the time for experiment on the vast scale is not yet, but progress is certainly being made.

REFERENCES

1. P. M. Roget, *Thesaurus of English Words and Phrases*, London, 1953.
2. M. Masterman, *The Potentialities of a Mechanical Thesaurus*, read at the International Conference on Machine Translation, M.I.T., abstract in *M.T.3*, 1956.
3. M. Masterman, *Semantic Message-Detection for Machine Translation, Using an Interlingua*, Proceedings of the 1961 International Conference on Machine Translation of Languages and Applied Language Analysis, London, 1962.
4. K. Sparck Jones, *Synonymy and Semantic Classification*, Ph.D. Thesis, University of Cambridge, 1964.