# Approaches to the Reduction of Ambiguity in Machine Translation

By L. E. DOSTERT

**Four basic approaches to the resolution of transfer-ambiguity are reviewed: (1) preparation of microglossaries; (2) formulation of linguistic operations to effect structural parallelism between the languages involved; (3) establishment of categories based on semantic functioning; (4) establishment of idiom classes. Structural analysis involved in a French-English translation is examined in detail.**

In research in the field of machine translation we are brought by the nature of our objective to consider languages as codes. If we accept that translation is the transfer of meaning from one language to another, then ultimately our task is to establish specific and automatic correspondence between the signs of a given language-code from which we translate (the source language), and the signs of another language-code into which we translate (the target language).

Since the signs in a given source language do not have one-for-one equivalence with those of a given target language, and since likewise there is no complete structural parallelism between the two systems, the object of research is the reduction and resolution of ambiguity in the transfer process. The word "ambiguity" here is not as precise as one would wish. To point out the problem involved more clearly the term "transfer-ambiguity" will be used in the present context.

In the case of two languages reasonably close in their lexical inventory and reasonably parallel in their structural characteristics, the reduction of transfer-ambiguity is relatively more feasible than when the source and target languages show great difference in their lexical inventories and marked divergence in their structural operations.

French and English are used in illustrating this discussion because these languages are reasonably close to each other and will be familiar to many readers.

The present objective of machine translation research aims at the translation of scientific literature. For this we have to arrive at a set of linguistic formulations, expressed in forms intelligible to a monolingual programmer, susceptible of reformulation in programming terms, and adequate to effect machine translation of scientific and technological writings.

In Georgetown we have recognized four basic approaches to the resolution of transfer-ambiguity:

(1) preparation of microglossaries;

(2) formulation of linguistic operations to effect structural parallelism between the languages involved;

(3) establishment of lexical classes or categories based on semantics functioning;

(4) establishment of idiom classes.

### Microglossary Approach

The microglossary approach relies upon the assumption that a given lexical item in the source language may have different equivalents — i.e. will translate into several separate items — in the target language when used in different disciplines. To illustrate, the noun *deck* used in a maritime context will have a different equivalent in French than when used in a text dealing with computers. Or the noun *la cour* in French will translate one way into English if used in a legal document and in another if used in a treatise on architecture. Likewise, the French noun *le sujet* will be translated normally as *person* in a legal context and as *subject* in a context dealing with grammar. Conversely the English noun *object* used in a grammatical context will translate as *complément* and in general context as *objet*.

The microglossary approach seeks to resolve these types of ambiguity by proposing the compilation of separate glossaries specific to individual disciplines with each source item having the minimum possible number of equivalents.

It should be pointed out that useful as the microglossary approach may prove to be, there are cases where it will not be effective — for instance when a word like *range* used in a text having to do with artillery can be ambiguous in the source text. At present, partial microglossaries have been processed for Russian-English in physics (Michigan) and in organic chemistry (Georgetown); and in physics for French-English (Georgetown). Only more advanced and much more complete research and experience will enable us to determine the extent to which microglossarization will be helpful in the reduction of transfer-ambiguity.

### Structural Analysis

Structural analysis involves the establishment of correspondence between the grammatical operations of the source and the target languages. It operates on four levels: first, on the morphological level; second, on the syntagmatic or phrase group level; third, on the syntactic or major sentence component level. It can also be said to operate on what can be called the word-class level. To illustrate this last level, the item *object* used as a noun will have a different equivalent in French than if used as a verb. The morphological, syntagmatic and syntactic techniques in the structural approach are illustrated by the following French statements for French-English translation.

*List of French Statements*

| | |
|---|---|
| 1. Nous parlons. | We speak. |
| 2. Parlons-nous? | Do we speak? |
| 3. Parlons. | Let us speak. |
| 4. Nous nous parlons. | We speak to each other. |
| 5. Nous parlons-nous? | Do we speak to each other? |
| 6. Parlons-nous. | Let us speak to each other. |
| 7. Parler. | To speak. |
| 8. Le parler. | The dialect. |
| 9. Sans parler. | Without speaking. |

The item *parl-*, stem of a French regular verb in group I (-er) has been entered in the stored dictionary, with its equivalent *speak* in English. Upon receiving the input item *parlons*, the machine identifies it in terms of stem and suffix. The operations then occur for sentences 1 to 6 as indicated in Fig. 1.

The stem having been identified, the morphological analysis then begins. The identification of the suffix *-ons* is the result of a series of consecutive no-yes decisions, thus: is it *-er* (infinitive); if not, is it *-ant* (present participle); if not, is it *-é* (past participle); if not, is it *-e* (1st person sing., pres. indicative); if not, is it *-es* (ditto, 2d pers.); if not, is it *-e* (ditto, 3d pers.); if not, is it *-ons* (1st pers. plur. pres. ind.). Having concluded the analysis on the morphological level, the next steps will be on the syntagmatic or on the syntactic levels.

By the word *item* we designate in this case the word which is the fulcrum of the analysis. By *item* $-1$ we designate the immediately preceding word (in left-to-right reading); by *item* $+1$, we designate the word immediately following.

The diamond-shaped figures in Fig. 1 represent the consecutive yes-no questions. The rectangular boxes represent the decision arrived at by the following steps:

Presented on October 23, 1958, at the Society's Convention in Detroit by L. E. Dostert, Director, Institute of Languages and Linguistics, Georgetown University, Washington, D.C.
(This paper was received on October 23, 1958.)

Is item $-1$ a p. p. 1 $p^1$ (pers. pronoun first person plural nominative); if not, is it p. p. 1 $p^2$ (ditto, reflexive); if yes, is item $-2$ p. p. 1 $p^1$? If yes, decision is "We speak to each other," or No. 4 in the list. If item $-2$ is not p. p. 1 $p^1$, is item $+1$ $-$*nous*, or hyphen + p. p. 1 $p^2$? If not, is item $+2$ a period? If yes, decision is *We speak*, or sentence No. 1. If item $+1$ is p. p. 1 $p^2$, and item $+2$ a ?, decision is "*Do we speak to each other*?" or No. 5 in the list. If item $-1$ is not p. p 1 p, and item $+1$ is p. p. 1 $p^1$, and item $+2$ is ?, decision is *Do we speak?* or sentence No. 2. If item $-1$ is not p. p. 1 p; and if item $+1$ is not p. p. 1 p, decision is *Let us speak*, or sentence No. 3 on the list. If item $-1$ is not p. p. 1 p, and item $+1$ is p. p. 1 $p^1$, and item $+2$ is *period*, the decision six occurs: *Let us speak to each other.*

A brief analysis of numbers 7, 8, and 9 in the list shows that: If suffix is identified as $-er$, is item $-1$ a prep. 1 (preposition, group I); if yes, translate the preposition and take decision No. 9, *without speaking*. If item $-1$ is zero, take decision No. 7, *to speak*. If item $-1$ is d. a. m. s. (definite article, masculine singular) decision No. 8 is taken, *the dialect.*

The formulation is susceptible of general application to nearly all French verbs and it handles the fundamental structure of declaration, interrogation, and command, and operates on the transitive and intransitive levels, as well as on the pronominal and noninflected levels.

### Semantic Categories

The establishment of semantic categories (sometimes referred to as lexical classes) is the least-charted area of machine translation research at this point. It involves the establishment of subclasses of major word classes (mainly nouns, verbs, adjectives and prepositions) based on the semantic function of the item. It calls for the resolution of ambiguity on a nonstructural basis, and without recourse to microglossarization or idiomatization. It involves the search for cues outside the morphological and syntactic levels.

The following example will illustrate the point:

The English statement, "Put your book on the table and put your sentence on the board," will call for a different translation of the preposition "on" in French. The first occurrence has to be translated by the French preposition "sur" and the second by the French prepositional article "au." The examination of the contextual environment will yield no clues in terms of structural data. Thus, the ambiguous item is preceded by an identical verb, by an identical possessive, and is followed by a common noun. Further scrutiny will reveal that both nouns can be classified as "nouns of surface" and that in the case of "table" the surface is *horizontal*, whereas the "blackboard" is normally thought of as being *vertical*. A similar context will yield a similar result. Thus, "Put your hat on the chair," will yield, "*sur* la chaise," and "Hang it *on* the wall," will yield "*au* mur."

A further illustration is represented in the sentence, "Pour a little acid in a little beaker." The item "little" functions here as an adverb in the first instance and as an adjective in the second instance. The determination of its grammatical function, however, can only be deduced from the fact that in the first instance we are dealing with a noun of *bulk* and in the second we have what can be called a noun of *unit*.

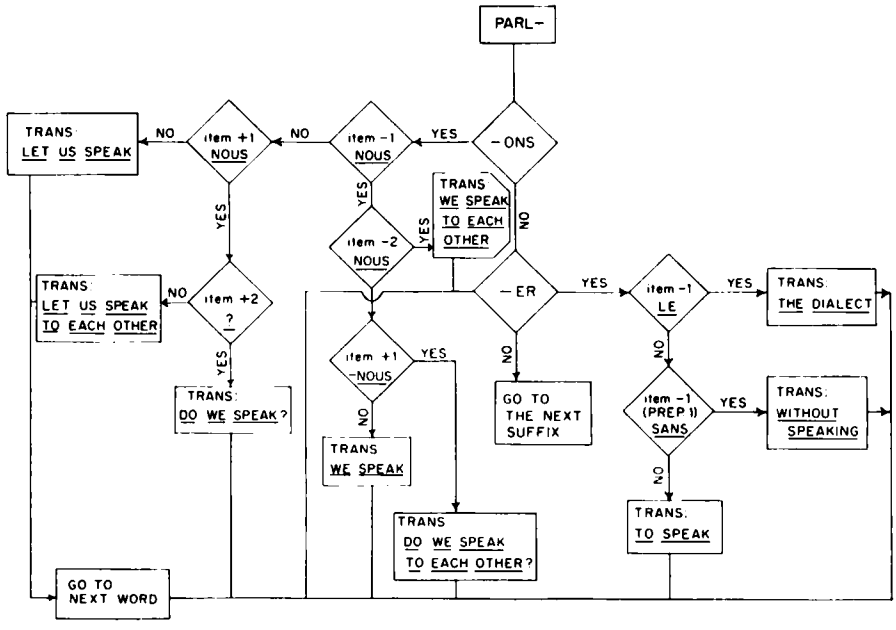


**Figure 1**

There is a class of ambiguity which cannot be resolved readily and that is when a single item in the source language has two distinct equivalents in the target language. Thus, for example, the English word *experiment* is rendered in French by *expérience*, but so is the English word *experience*. Thus, the sentence, "L'expérience a démontré . . ." could yield both, "The experiment has shown," and "Experience has shown." Since in French the article is present in both cases, no clue can be found within that context. The tentative solution to such a situation is to take the English equivalent *experience* alone and use it in all instances, even when *experiment* would be more accurate. The measure of ambiguity would be minimal. Even the use of *expériences* in the plural will not yield an absolute clue for the choice of either *experiences* or *experiments*.

### Idiomatization

The establishment of idiom classes is self-defined. An idiom is a cluster of items nontransferable as separate units. Thus, in isolation *right* will have a variety of equivalents which will not be acceptable in the cluster *right away*, which is treated in machine translation not as two items, but as a single lexical unit.