[*Ambassador Interntaional* March 1982]

Computing ahead of the linguists

by PETER WHEELER and VERONICA LAWSON

While the much-loved stories that a computer once translated the Russian for "hydraulic ram" into the English phrase "water sheep" or "the spirit is willing but the flesh is weak" into "the vodka is excellent but the steak is not to be recommended" are probably apocryphal, undoubtedly genuine is the following gem from the European Commission's Systran machine translation system: "la Cour de Justice considère la création d'un sixième poste d'avocat général," rendered into English as "the Court of Justice is considering the creation of a sixth general avocado station".

The dwindling band of opponents of machine translation clutch anxiously at such a howler, as conclusive evidence that MT cannot work. To those actually working in this rapidly-evolving field, it is firstly a piece of light relief, a bit of a giggle in the working day, and secondly a challenge: since "avocat" means not only "advocate" but also "avocado" how does one make the machine distinguish between them?

The obvious approach, and one which was tried out in the very early days of the Commission's experiments on Systran, is to use a special "topical glossary", in which ambiguous words can be given a different meaning for each specific field. Under topical glossary "C", for Court of Justice, for example, "avocat" would mean *only* "advocate", "huissier" would mean only "bailiff" and not "office messenger," and so on. It became evident, however, that within the context of the European Community translation workload, such an approach was not sophisticated enough — given the enormously wide range of subjects with which the European Community institutions are concerned, and on which their translation services have to provide translations into any or all of seven (and in some rare cases eight) official languages, a document concerning the Court of Justice was just as likely to be talking about import quotas for avocado pears as about a submission made by one of its Advocates-General.

Similarly, while the topical glossary approach might have specified that the French word "ventilation" in a mining context means "ventilation" but in a statistical context means "breakdown", what was the poor machine to do with a document produced for the Mines Safety and Health Commission on "la ventilation des statistiques sur les accidents dues à la mauvaise ventilation?"

In consequence, the use of Systran topical glossaries has been almost entirely abandoned, at least as far as the Commission is concerned. (This last proviso is an important one, since the US Air Force and some of the industrial companies who use Systran, and who translate in more circumscribed subject fields, do find that the topical glossary approach suits their needs.) The approach the European Commission has taken instead is one of extremely painstaking dictionary coding, often of a level of considerable complexity.

To understand this, let us look very briefly at the way Systran works. The text to be translated is loaded on to the computer via some form of terminal, and all the words in the text are looked up in the dictionary held in the computer's memory. This main, or stem, dictionary contains one, and only one, basic meaning for each source-language word. (In the case of "avocat," this is "avocado") The Commission's stem dictionaries contain about 60,000 entries in each of the pairs French-English, English-French, and English-Italian.

The second stage is to disambiguate the words which may have more than one part of speech. At first glance, the two words "la porte" clearly mean "the door", but depending on what else is present in the sentence they may equally well mean "carries it". The problem is worse when translating out of English — about 50% of all English words are homographic. (Hence "Army push bottles up enemy.") Having decided which part of speech each of these homographs actually is, and having saved the other options for later reference in case it has made a mistake, the system then analyses the text sentence by sentence, establishing the diverse syntactic relationship between all the words in it — adjectives to nouns, subjects to verbs to prepositions to nouns, adverbs to verbs, and so on.

After this phase, which is the heart of the whole process and takes place in five distinct passes, the system now refers to another dictionary (at the Commission, this more sophisticated dictionary contains 40,000 entries in each of the three language pairs) to refine the meanings chosen.

We remember that the stem meaning of "avocat" is "avocado". In the second dictionary, however, are listed semi-fixed expressions such as "discours d'avocat, plaidoyer par...avocat" etc., each coded to have the translation "advocate". This is where "AVOCAT (modified by) GENERAL = ADVOCATE" should have been, but alas wasn't. It is now.

More generally, the dictionary will ensure that "avocat" occurring in a text in enumeration with other professions — judges, solicitors, even wheeltappers' 'arkers — will be translated as "advocate". And further that "avocat", if detected as the subject of a verb which in turn is labelled as requiring a human subject (to feel, to be of the opinion) or as being able to start a subordinate clause (to state, to say, to consider), will, once again, be given the translation "advocate". Avocados, after all, are unlikely to feel, and they never state anything. And so on.

Having selected the appropriate meaning for the context, insofar as the work put into the dictionary has allowed it, the system then goes on to synthesise the correct morphological forms of the target language, and to rearrange words which have a different order from one language to the other.

The work of making the dictionary entries — in effect trying to anticipate every possible occurrence of a word in advance — is a slow and painstaking process, and one in which the insight and experience of the working translator are essential. Originally designed by Dr Peter Toma for translating Russian into English for the US Air Force, Systran was bought by the Commission in 1976 to tackle English-French. Two more pairs (French-English and English-Italian) followed, and five years of development work ensued, under the overall linguistic control of two of the Commission's translators, detached for the purpose from the translation service, but administratively and hierarchically still part of it. These five somewhat ivory tower years were necessary to develop the system to a point where it might perhaps be useful to the translation service, a point which was judged to have been reached in the spring of 1981.

From this time on, some 3% — a small proportion but one which is expected to grow — of the Luxembourg translation service's French, English and Italian output has been initially produced by Systran. The machine draft (still containing a number of avocados which should be advocates) is tidied up on printout paper or on a word-processor screen by "traditional" translators, and their comments, suggestions and howls of rage are used as the bases for further work.

In the decade since Systran came out, however, there have been many developments in machine translation and the related fields of linguistics, computers (cheap storage!) and artificial intelligence. The Commission decided to exploit European expertise in these areas by initiating an advanced machine translation project, Eurotra.

Whereas Systran was originally designed as a bilingual system, Eurotra is to be multilingual from the start. It is also to be extensible, so that new languages, subjects, even research can be incorporated. A collaborative effort, its various language modules will be created by independent groups, to encourage research in all the EC member states. The system will in fact be highly modular, made up of distinct but compatible parts; and it is intended to be portable from one make of computer to another.

As yet, admittedly, Eurotra exists only as a set of detailed specifications. Still, the group, having worked part-time on a low budget, are in process of obtaining a £9,000,000 grant from EC institutions and governments, and they hope to have a pilot system translating by 1984. This is to translate Commission texts of 10,000 words in one subject area between a few languages. By 1987, a full-scale prototype may be translating a wide range of material between all available EC language pairs. That means up to 42 now, and 72 if Spain and Portugal join.