The Errant Avocado

Peter Wheeler, EEC, Luxemburg

Approaches to ambiguity in Systran translation

Like the chairman in the panel game 'My Word', I am going to give you a sentence to be played with later, but whereas Frank Muir and Dennis Norden clearly have a reasonable amount of time in which to prepare their imaginative version of the original of some proverb or saying, you will have only a few minutes.

What I should like you to do is to work out the original French, or alternatively the correct English, for the following nonsensical sentence: 'In the group of about eight, the advice will approve asks for it some the sub-committee which the yard of justice creates a fifth general avocado station'.

Now the subtle thing about giving you this puzzle at this moment, is that those of you who are old machine translation hands do not need to listen to the next couple of paragraphs of my offering this evening, which will be the standard historical introduction, which you have probably heard a dozen times before.

The first large-scale machine translation development project took place in the United States at Georgetown University in the late fifties and early sixties when the U.S. authorities spent some 20 million dollars on developing a Russian-English MT system. The funds dried up, however, in 1966 with the publication of the ALPAC report which concluded that the results obtained were not sufficient to warrant further development, and that there was no likelihood of a significant advance in the field within the forseeable future. (We can all make mistakes!) The Georgetown system was nevertheless used both in the United States and in Europe for translation from Russian and many users found the output adequate for purposes of information-gathering.

Not everyone agreed with the outcome of the ALPAC report. Several of those involved in the Georgetown project decided to go it alone and continue development on a commercial basis. The most successful was Dr. Peter Toma, the developer of Systran, which represented a considerable improvement over Georgetown owing to its dictionary structure on the one hand and its more sophisticated parsing capability on the other. The Russian-English Systran system became operational in 1970 and has since been used extensively by the U.S. Air Force and other American government agencies.

By the time the English-French system was being developed in 1973, further sophistications had been introduced and in 1975, when the European Commission undertook a survey of free-syntax MT systems in existence at the time - 'Handbook of machine translation and machine-aided translation', Herbert Bruderer, North Holland 1977 - Systran came out on top. The Commission thus decided to purchase and develop further the English-French system and later decided to purchase and develop further the English-French system and later extended coverage to French-English and English-Italian.

Five years of development work followed, carried out either externally by contractors, or in-house by two renegade translators, detached for the purpose from the Commission's own translation service but administratively still part of it. This development work covered major improvements to the linguistic programs at the heart of the system, purchase of a whole package of utility programs for greater user-friendliness, and of course a massive increase in the dictionary component, from around 6500 entries when we bought the system to some 100 000 per language pair today.

After these five years, it was decided that the system had reached a point where it could produce usable translations for practical purposes, in other words, could supply the translation service of the Commission with output of a sufficiently high standard to be corrected and tidied up by translators and then returned to the requesting departments. About a year ago, therefore, the decision was taken to put Systran into experimental operation within the Commission's translation service in Luxemburg.

From these small beginnings a year ago we are now running at something like 400 pages a month in the couples English-French, English-Italian and French-English, and we hope to reach 1 000 pages a month by the end of the year.

Another major development for this year is that we have ordered English-German and French-German systems which will be operational by the end of the year, or the beginning of 1983, with a medium sized Stem dictionary and a fairly small LS dictionary, two concepts which I will come to later.

Briefly, a word on the practical procedures for using Systran within the Commission: when the head of the English, French or Italian translation section selects a document as being suitable for translation by Systran, the entire document is copied by a typist on to a word processor screen. (We use a Wang OIS 130 with a 10 megabyte disc, three workstations and two printers, and have just ordered a second 10 megabyte disc, seven more workstations and two more printers.)

While there are no particular input conventions which have to be followed, so that this stage represents copy-typing pure and simple, it remains one of the major time-constraints on the whole process, and consideration will have to be given at some time in the future to some form of optical character recognition. One of the factors holding us back from this is that the work the Commission translates comes from a great many different sources, and is thus in a wide range of typefaces, and we are doubtful whether any OCR can cope with this variety.

Finishing the typing, the typist enters a special document-handling program written for the Commission by a software sub-contractor, types in the document's number and the target language or languages into which it is to be translated, and the translation is thereby requested.

For the input typist, the work is now finished, and she can go on to something else. At various times during the day, one of the small team running the Systran operation checks on his or her own word-processor screen to see whether any translations have in fact been requested. Unless a translation is urgent, the request will remain pending until several have been requested for the same language couple, as it is obviously more efficient to batch them together and process them together.

We then go into a different part of the document-handling programs; the translations selected are automatically concatenated together into a single document, and this document is sent down the telephone line by the 3780 telecommunications protocol to the IBM 370 158 computer which is some 15 miles away from the Commission's buildings.

Transmission is at 1 200 baud and takes about a minute per page. Depending on what else is waiting to access the computer, which is not a dedicated machine, there is a waiting time, which does not, however, normally exceed five or in the worst cases ten minutes and during which is is not necessary for anybody to be sitting watching the terminal. Once the text has been admitted to the computer, it is then translated.

The translation rate varies between 2.8 and 3.5 CPU seconds per 100 input words, the variation depending not only on the length of the text - longer texts being inherently more efficient - but also on which of the not strictly linguistic peripheral programs are running at the same time. After translation, the text is then transmitted back again, again at 1 200 baud.

From this point on the translation can take one of two routes.

Either, and more commonly, the translation is printed out on our 200 characters-per-second line printer and the printout sent to the translator, who makes his corrections by hand on the paper. These corrections are then, or can be, entered on to the word processor by one of the secretaries and the corrected version is then reprinted, usually on a 40 characters-per-second daisy-wheel printer, for sending to the reviser and then out to the requester.

Alternatively there are some translators - a small but growing number - who prefer to work directly on the word-processor screen themselves. The choice of which method to use, or indeed whether to use the Systran translation at all, is entirely at the translator's discretion. If the translator genuinely thinks that to correct the printout will take him longer than to dictate the whole thing from scratch, then he is quite at liberty to throw the printout in the waste paper bin. Or preferably, to send it back to us with a word or two on why he or she considers it to be unusable. We will then often do some work on the text, and resubmit it to the translator at the same time as he is doing the translation in the conventional way, and he is frequently surprised at the improvement which can be achieved with just a small amount of work.

We are finding, however, that most translators are prepared at least to give this new tool a fair trial, and that the corrections they make to the output prove very helpful in improving the system for the future. We have observed a significant acceleration in the rate of improvement of the quality of the output, now that we are receiving comments, suggestions, and occasionally howls of outrage, from some thirty different working translators instead of just from the pair of us.

The reason that the output from Systran has to go to the translator at all, of course, is that it is not always, to put it tactfully, 100% correct and not always limpidly clear.

However, to be fair, it nowadays seldom produces such comprehensive nonsense as the sentence I gave you at the beginning of the programme. Having thus neatly got us back to that sentence, I should like to use it to illustrate some of the linguistic problems arising when we try to make a computer carry out such a complex operation as translation, and some of the possible solutions we adopt to these problems. The sentence, you remember, was 'In the group of about eight, the advice will approve asks for it of the sub-committee which the yard of justice creates a fifth general avocado station'. Some of you will already have worked out some of what it really means, some of you will have worked it all out, and probably all of you have realized that this nonsensical sentence has been created by consistently taking the wrong possibility whenever a word has two potential interpretations.

Let us first of all define our terms, because in everyday parlance the word 'ambiguity' is itself ambiguous!

On the one hand there is ambiguity of part of speech. When a word can have two or more parts of speech, confusion may arise in the reader, although this rarely lasts beyond the first glance. One of the best known, and neatest, examples of such confusion or dual interpretation, is 'British push bottles up enemy'.

This type of ambiguity is what we call homography, and the resolution of homographies is fundamental to the success of a machine translation system such as Systran. The Systran French-English system has almost 70 different homograph routines, the English-source systems over 80, since around 50% of all words in English are homographic. These routines vary in length and complexity, and in the range of their applicability, - the longest is of the order of 150 lines of programming, the shortest being delightfully concise 'Bump to the end of the sentence and if the last word in the sentence is a question mark, conclude that the homograph under investigation is an interrogative adverb and not a subordinate conjunction' - but their purpose is always the same: to ensure that the various translation sub-programs within Systran at least start their work on the basis of the right parts of speech.

Faced with the need to translate 'British push bottles up enemy', Systran would call on homograph routine No 38 to decide whether 'British' is a plural noun or an adjective, on No 25 to decide whether 'push' is a singular noun or a plural verb, and then again on the same routine to resolve whether 'bottles' is a singular verb or a plural noun. The resolution of 'bottles', of course, will be helped by the decision already taken on the problem of 'push'.

The homograph resolution programs within Systran consist of thousands of questions about the syntactic context of the word to be resolved. These questions, and the conclusions the routine is to draw from the answers, are written in a macro language specific to Systran, and I am grateful to Margaret Masterman and to the Cambridge

Language Research Unit for the sterling work they have put in under contract to the Commission to develop a program which has automatically annotated the routines into plain English.

The aim of this program is that linguists should be able to look into the routines, and bring their linguistics expertise to bear on them - spotting inconsistencies in the approach to real 'working' syntax, adding in possibilities which have been overlooked - without the need to learn the Systran macros. The automatic annotation program even has the startling benefit of enabling a non-specialist such as myself, a mere Systran mechanic, to be whisked across the Channel at vast expense to explain these complicated routines to an august audience such as yourselves.

Let us take as an example the homograph routine 57, which serves to disambiguate a word coded as being either an adjective or an adverb - a word such as 'fast', or 'very', for example. This routine starts as follows :

```
Set the A-word pointer to the first word before the current word.
Set the B-word pointer one word beyond the current word.
Set the C-word pointer one word beyond the B-word.

If the B-word is not a left bracket, go to HM570B.
Otherwise:
Starting from the B-word, scan along the sentence to the right:
.........
When a word is found, set the B-word pointer on it.
If the B-word is a cardinal number, go to HM57A.
Otherwise:
If the B-word is 'ENOUGH' go to HM57V.
Otherwise:
If the current word is the first word of the sentence, go to HM57C.
Otherwise:
If the A-word is an auxiliary verb, conclude that the current word is an adverb and go to HMADV.
Otherwise:
If the A-word is a finite verb, a past participle, a finite form of the verb 'to be' 'BE', an infinitive, a verb ending in 'ING',
```

finite form of the verb 'to have', 'HAVE', 'AND' or 'OR', a coordinate conjunction, 'HAVING', a pronoun, a quotation mark, a right bracket (round or square), a system control word, 'BEING' or 'BEEN', go to HM57C .....

HM57C: If the current word is 'VERY', conclude that the current word is an adverb and go to HMADV.
Otherwise,
If the current word is 'NEXT', go to HM57COD.

HM57COD: If the B-word is a comparative adjective or adverb or a superlative adjective or adverb, conclude that the current word is an adverb and go to HMADV.

Otherwise:
If the B-word is not 'AND' or 'OR', go to HM57CC
Otherwise:
If the C-word is not a homograph type 57 (adverb, adjective), go to HM57COM.
Otherwise:
If the C-word was a homograph which was resolved by being an LS or IDIOM, conclude that the current word is an adjective and go to HMADJ.
Otherwise ....

And so on, for page after page!

I should like to stress here the essentially pragmatic nature of the Systran programs. They would no doubt make an academic computational linguist shudder, but they work. They aren't trees, they have no taxonomic structure, they don't rely on Artificial Intelligence, they simply run up and down the sentence like a kitten on the keys, asking themselves the childishly simple question 'What might we find here in an everyday French sentence in the real world?'. and emitting a whoop of joy whenever they find what they were looking for.

Looking once again at our puzzle sentence, we immediately spot three homograph errors in it. 'Ask for her' is a translation into English of the French words 'la demande', but rather unkindly I have assumed that Systran has mis-analysed their parts of speech. 'La demande' may be not only a finite verb with an object pronoun in front of it, but also a definite article followed by a noun. It is this latter interpretation which is the correct one, and it is this also which gives a clue to the correct resolution of the second homograph mistake, namely 'which' given for 'que'. Here, homograph routine 50 has failed to distinguish between the relative pronoun 'que' and the subordinate conjuction 'que'. Because, of course, of the incorrect resolution upstream. If 'demande' had been correctly analyzed as being a noun, which in turn is coded as being likely to introduce a subordinate clause, the 'que' would in all probability have come out right as well.

The case of the odd word 'some', on the other hand, is a real trial. The French original, of course, is 'de', and I think it would probably be fair to say that 'de' is currently my biggest headache.

Quite apart from the very complexity of the homograph resolution itself, even if homograph routine 3 has sorted out that a given 'de' is a preposition rather then an infinitive particle or a partitive article, a subsequent part of the Systran programs, written in order to give special translations to prepositions, often leaps in and interferes with the preposition's meaning. For example, as the verb 'reduire,' is coded as governing 'de' and as this 'de' is in turn coded as 'to be translated as 'by'', a phrase such as 'reduire le temps de freinage', tends to come out as 'reduce the time <u>by</u> braking' instead of 'reduce the braking time', because the preposition has been considered as starting a prepositional phrase, the instrumental complement of the verb, rather than as indicating the adnominal relationship between two nouns.

The critics, in particular academic, of early systems such as Systran level the charge that there comes a point at which one part of the system inevitably interferes with an earlier part, and overturns what was a correct resolution. While a lot of their charges, particularly with regard to Systran, are exaggerated, and have in any event become rather muted over the past couple of years, in this specific case I fear they may be right. If anyone has any suggestions as to how this problem can be resolved, I should be most appreciative. Once we have put these three homograph errors right, our sentence now reads 'In the group of about eight, the advice will approve the request of the sub-committee that the yard of justice create a fifth general avocado station', and we have now to turn to our second type of ambiguity, otherwise known as polysemy.

Consider another sentence: 'Because the clock did not work properly, the conductor failed to make the connection and the bus had to operate with just a driver'. What image does this conjure up: a cheerful Cockney bus conductor, perhaps, sitting at the breakfast table at home and enjoying an extra cup of tea, unaware that the kitchen clock is running slow, and that in consequence his mate, cursing and swearing, is already trundling his bright red double-decker out of the depot without him?

Highly plausible, and quite wrong!

In fact, the sentence comes from (or might have come from: actually I wrote it myself) a report on the failure of a piece of electrical equipment. What has led us astray are the words 'clock', 'conductor', 'connection', 'bus' and 'driver'. Note, though, that there is no problem with their part of speech - most of them are unhomographic, and 'clock' has been correctly resolved in our minds as a noun. The problem is simply that these nouns each have more than one meaning, all of them can be items of electrical equipment.

Since normally we do not read a sentence out of context like this, usually the surrounding sentences would tell us whether we are reading about Cockney clippies or chunks of copper wire. The computer does not have this world knowledge - although the proponents of AI-type systems hope to give it this at some point in some undefined future - and so we have to spell all the options out and give the computer idiot-proof directions for choosing the right one.

Systran is a system based very firmly on its dictionary. This, too, is a reproach often levelled at it by those who think that computerised translation should be based on some transcendentally universal deep meaning. To them, my reply is along the lines of or rather, on a pyramid of dictionaries. Right at the bottom, and holding the whole structure up, is the Stem dictionary. This contains nothing but single-word entries: 'maison' - 'house', 'voiture' - 'car', 'chameau' -

'camel'. Where a collection of words has and can have only one meaning, then a Stem-type dictionary is sufficient. If we take a sentence such as 'the rabbit eats the egg' and we assume that our homography analysis is correct, that rabbit is a noun and not a verb meaning to 'talk at length', and that 'egg' is another noun and not a verb meaning 'to urge', then the sentence is unambiguous and can be translated from a simple word-for-word dictionary such as the Systran Stem: 'le lapin mange l'oeuf'. But sentences need highly complex and expensive computer systems - or highly qualified and expensive translators - to translate them.

Relying on the Stem dictionary alone is what has caused Systran's downfall in our sample sentence.

To start at the beginning, always a wise point to start, 'group of about eight' is the Stem translation for the French noun 'huitaine', and in many contexts would be the correct one: 'une huitaine de traducteurs - a group of about eight translators'. When governed by 'dans', however, 'huitaine' no longer means a group of about eight of anything, but a week. 'Dans la huitaine' therefore means 'within the week'. To enable the system to select this correct translation, however, the fixed expression 'dans la huitaine' *cannot be* entered into the single-word Stem dictionary, since it isn't a single word, but has to be coded into one of the system's three or four more sophisticated dictionaries. Within the context of the more sophisticated entries, a fixed phrase such as 'dans la huitaine' or 'en fin de compute' or 'pour ainsi dire', are the simplest form, and are known for Systran purposes as simple *Idioms*.

Whenever this group of words is found together, in absolutely the form in which the group has been coded into the dictionary, it will receive the specific idiomatic translation given.

One step higher up the ladder of complexity are *Limited Semantics* expressions, which are always noun phrases, and which differ from Idioms in being allowed to inflect. If we jump now to the middle of the sentence, it is clear that 'yard' is the Stem translation of the French word 'cour' but we see instantaneously through our world knowledge that since it is linked to 'justice' what is required here is a different translation of 'cour', namely 'court'. Some diligent Commission coder, therefore, coming across the howler 'yard of justice' will code into the LS dictionary the expression 'cour de justice' and give it the translation 'Court of Justice'. The word 'cour' (and, incidentally, 'justice' as well) will now carry a flag in the Stem telling the system that if it encounters this word at the main Dictionary Look up stage it must branch into another dictionary and check whether there is a longer-than-one-word match.

Our sentence now reads 'within the week, the advice will approve the request of the sub-committee that the Court of Justice create a fifth general avocado station'.

I should perhaps make it clear that this laborious process of gradually getting closer to something that it recognisably English is not what happens when Systran is actually running - what we are doing in the hour allotted to me is encapsulating the sort of development which has taken place over the past five years. The nonsense sentence that we started with represents the sort of quality we were getting out of Systran in 1975, the version that I intend to have reached by the end of my talk will be approximately comparable to the standard we can now realistically expect from any random text in a field for which a reasonable amount of dictionary work has been done.

Let us now look at this strange word 'advice'. Some of you are no doubt ahead of me here, and have realized that it is the Stem translation of the French word 'conseil'. How are we going to make this word come out correctly as 'Council'? 'Conseil des Ministres' and 'Conseil Européen' are simple, of course, as would be 'Conseil d'administration', for example, we can code them as LS expressions like 'Cour de Justice', but what about 'conseil' on its own? Let us consider how the human reader knows that 'conseil' in a given text is a 'council' and not some 'advice'. Surely, what gives us the clue is the context, which is the smart academic word for 'what else is to be found in the sentence'. And at this point our very junior Systran coder realizes - an experienced one would have realized it long ago - that he or she has not completed the work tackled so far.

In addition to the various identifiers attached in the dictionary to a word to give it the correct morphological forms in both source and target, provision is also made for the attachment of additional codes, whether syntactic (eg. 'always transitive', 'Noun Clause Opener') or semantic (eg. 'profession', 'financial'). To our expression 'cour de justice', therefore, the coder should have added the curious code 'ENPRIS' which indicates that this is some form of enterprise or body. Other examples of 'ENPRIS' entries would be The British Computer Society or Kings College, London.

This in turn allows our coder to write an expression somewhat more complex than any we have seen so far, a so-called conditional limited semantics expression, or CLS.

Broadly, this is an expression which selects a particular translation only if certain conditions are fulfilled in the sentence. In the case of 'conseil', an expression can be written and entered to the dictionary to say: 'Scan all the way back to the beginning of the sentence, or forwards to the end of it, and see if you find a word coded 'ENPRIS'. If you do, translate 'conseil' as 'council''.

Once again, like the homograph routines, this approach to the ambiguities of words is highly pragmatic, with all the disadvantages that this entails. Coding 'conseil' as I have suggested, for example, would lead to a wrong translation in the case of a phrase such as 'le counseil offert par la Commission', in which 'conseil' does indeed mean 'advice'. But two related factors have to be borne in mind here - on the one hand, everything produced by Systran, at least for the Commission's internal purposes, will pass through the hands of a human translator before it reaches the customer. And secondly, Systran's well-meaning efforts are under the contant scrutiny of a team of eagle-eyed linguists. If we find, therefore, that we have guessed wrong, and that 'conseil' in the same sentence as an 'ENPRIS' means 'advice' more often that it means 'council', than we can just delete the dictionary entry. The aim is always to get more hits than misses, and to rely on the goodwill of the post-editing translator to correct the misses.

A similar approach can be adopted in order to make 'commission' come out as 'Commission' instead of 'committee' or 'sub-committee' and this then just leaves us with the delightfully surreal concept of the general avocado station.

I should like to take the case of 'avocat' to discuss one possible avenue to the resolution of ambiguities, an avenue which has now virtually been abandoned by the Commission of the European Communities, however. This is the so-called Topical Glossary route, in which ambiguous words are given alternative meanings depending on their subject field, and the texts entered for translation are themselves assigned a code to indicate the subject field they belong to. Using this approach, one could give 'avocat' the meaning 'avocado' with a topical glossary code F for fruit, and the meaning 'avocate' with the topical glossary code L for legal.

Fine in theory, because then when the head of the English translation section finds in his in-tray the document from which this sentence is supposedly extracted, and sees at a glance that it concerns the Court of Justice, he can order a translation by Systran specifying that in cases of ambiguity, the 'legal' meaning is to be selected.

Five years ago, we at the Commission had high hopes of this approach, but it proved to be unworkable, precisely because of the wide variety of subjects covered by the Community institutions and translated by their translation services. A Court of Justice document in which the word 'avocat' appears is just as likely to be about import quotas for avocado pears as about some learned opinion by one of the Advocates-General.

Similarly, while 'ventilation' means 'ventilation' in a mining context and 'breakdown' in a statistical context, how is the Topical Glossary system to cope with a text dealing with 'ventilation des statistiques sur les accidents survenus à cause de la mauvaise ventilation dans les mines'?

With a couple of very specific exceptions, therefore, the use of topical glossaries has been abandoned by the Commission. On the other hand, some of the commercial companies using Systran, and by definition translating in a more circumscribed field, do make use of topical glossaries, for example when two firms working in the same product field use different words to describe the same component.

In the case of the Commission, almost the only use of topical glossaries nowadays is in connection with the minutes of meetings, which as you no doubt know are written in the present tense in French and have to be rendered into the past in English, and vice versa. Asking for a text to be translated using topical glossary 'M' allows this transformation to be carried out automatically.

It also, incidentally, allows one to specify that the word 'president' is to be translated in the minutes of meetings by 'chairman' and not by 'president', although this in turn is not adequate for the minutes of meetings of bodies such as the ECSC Consultative Committee, which are sometimes chaired by the President of the Committee, sometimes by a chairman, and where even on occasion the president (president) of the Consultative Committee may be present at but not in charge of a sub-committee meeting being chaired by a chairman (president). Systran still has some way to go before working that one out!

Work is still going on at the Systran Institute, the Commission's sub-contractor, on a system of typology categories which will be attached to certain very specific words to enable the system itself to detect, at paragraph or even sentence level, what is the topic being covered. Something similar allegedly works reasonably well in the US Air Force's Russian-English system, but we still have some way to go.

And so, pending a success in this particular endeavour, instead of just specifying that 'avocat' in Court of Justice means 'advocate' we have had to do dozens of dictionary entries: 'avocat' modified by 'general' is an 'advocate', 'avocat' as the subject of a verb of speaking or thinking (two activities for which avocados are not reknowned) is an 'advocate', 'avocat' in apposition with a word semantic-coded 'PROFES' for 'profession' is an 'advocate', and so on.

Which only leaves us with 'station'. A typical Stem compromise for a word with as many meanings as 'poste'. Whether it is a radio set, or the place where a guardsman stands, or the cabin where a train driver or tractor driver works, 'station' is not quite right for any of them, but nor is it absurdly wrong. It is not even wrong, just rather biblical, to use 'station' for a person's job. But of course, after five years of development we will not settle for a system which contents itself with Stem meanings, we will ensure that someone codes 'poste' governing 'de' in turn governing a word coded 'PROFES' to give the meaning 'post' or 'position'.

'Within the week, the Council will consider the request from the Commission that the Court of Justice create a fifth post of Advocate-General'.