DESIGN CHARACTERISTICS OP A MACHINE TRANSLATION SYSTEM

M. King

ISSCO, University de Geneve

ABSTRACT

This paper distinguishes a set of criteria to be met by a machine translation system (EUROTRA) currently being planned under the sponsorship of the Commission of the European Communities and attempts to show the effect of meeting those criteria on the overall system design.

1. INTRODUCTION

EUROTRA is a machine translation system which so far exists as a detailed set of technical specifications. Work leading to the drawing up of the specifications started in February of 1978, and has been carried out on a collaborative basis by a group drawn from the Member Countries of the European Community, under the aegis of the Commission of the European Communities. The present author is responsible for the co-ordination of the technical work, and therefore makes no claim to be more than the synthesizer of ideas whose original sources are multifarious.

The system to be designed had to meet a number of very specific criteria, some of them coming from the particular needs of the European Community, some from an initial decision to carry out both planning and implementation of the system collaboratively. This latter meant that the system must be designed so that groups in the different Member Countries could work more or less independently during implementation.

In what follows, each section identifies a particular criterion, explores its consequences on the overall system design and tries to put these consequences into perspective by comparison with other systems. No distinction is made, for these purposes, between systems which are operational in the sense of producing translations routinely, for their bread and butter, and pilot systems which were developed primarily as experimental test beds for interesting theoretical ideas. Nor will any attempt be made to describe systems referred to in detail, or to put them into relationship with one another. Hutchins [7]fulfills these latter two tasks admirably.

2.MULTILINGUALITY

The most obvious special demand made of the EUROTRA system design was that it should be multi-lingual. The European Community had, at the time of starting work, six official languages. This has recently been increased to seven, and a further two languages are in prospect. n languages implies n(n-l) language pairs, so 9 languages gives 72 language pairs. All early translation systems were bi-lingual (e.g. the Georgetown system [5]), the analysis of the source language being done within the perspective of a particular target language. If this tradition were followed, 72 separate translation systems would have to be written, one for each language pair; a proposal which is clearly uneconomic. Some systems, although initially developed as bilingual systems, have made some attempt to move towards multi-linguality by allowing analysis or generation of a new language to take over techniques developed for some other language. This is the case, for example, with SYSTRAN [9] . Such an approach has obvious disadvantages: quite apart from the practical difficulty of modifying the treatment of linguistic facts in one language to deal with (optimistically) similar linguistic facts in another language and the consequent proliferation of historical remains carried over from one version of the system to another, it is by no means self-evident that the treatment appropriate for one language is appropriate for another.

It seems preferable then to separate out different stages of translation in such a way that as much as possible is done within the context of a single language. Thus analysis and generation within EUROTRA depend only on the grammar (in the widest sense) of the particular language being treated, and contain no reference to a particular target or source language respectively.

However, since translation clearly involves a mapping between two languages, there has to be a bi-lingual link between the analysis of the source and the generation of the target. This is the transfer stage of the translation process. There must clearly be as many transfer modules as there are language pairs, so in the interests of economy the transfer part should be kept as small as possible.

Other systems, (e.g. GETA [1,10,11]), have already adopted this approach to some extent. They differ from EUROTRA, however, in the level of representation aimed at.

Mu11i-lingual ity has obvious repercussions on the level of representation. Since it is not possible to take advantage of similarities between source and target the representation must go some way beyond an analysis of superficial syntactic structure. The choice of how far to go beyond it depends on a compromise between the ideal of producing a complete and explicit semantic/pragmatic representation of the text and the feasibility of reaching such an ideal in a system which has to work for its living within a relatively short time. GETA aims at establishing what is essentially a deep syntactic representation, based on the valency patterns of predicates. EUROTRA tries, at its 'deepest' level of representation, to characterize the semantic relations between constituents in the text via a set of relations based on an expanded form of case grammar, similar to the relations used by Wilks [12,13]. However, since the set of relations are defined as those useful for translation and are only 'universal' within the project, there is no attempt to reach a ideal, genuinely universal semantic representation.

There is one further implication which should be spelt out, which comes partly from the multi-linguality constraint and partly from the collaborative nature of the project. Since the same analysis module must provide input for a number of different transfer modules, and the same generation module accept output from a number of different transfer modules, the structure and content of that input and output must be very closely defined if the system is not to disintegrate into a number of mis-matched lumps. This consideration has led to the definition of an interface structure, to be used as a means of transferring results between the main modules, and indeed as a way of representing intermediate results within the main modules. More will be said about this interface structure in subsequent sections.

3. PRACTICALITY

The aim of the project is to produce a system which will be operational, at least within a limited domain, within five years from the start of implementation. The semantic relation level of text representation outlined in the last section is, in terms of a practical, working system, quite ambitious. Although experimental systems (e.g. Wilks, op. cit.) have proved quite successful, it would be rash to assume that an accurate semantic representation can always be established. But a working system cannot be allowed just to give up and produce no translation at all, especially if the informationis lacking only for some limited stretch of text. The system must therefore have fall-back mechanisms. Thus, the interface structure is also to include information on the valency boundedness of constituents, on their surface syntactic function, on their morpho-syntagmatic class, and on the morphology of terminal elements. Furthermore, the geometry of the interface structure defines the syntactic constituent structure of the text. All this information may be used during transfer in order to establish the correct lexical units in the target language. Where the semantic relations are unavailable, it may also be used by bi-lingual safety-net grammars in order to provide some translation rather than none at all. Such grammars can be imagined as producing translation in descending order of quality. In the worst possible case, the quality would be that of word-to-word systems.

Although the main justification for retaining as much information as possible about the source language text is its potential usefulness as fall-back information, it does fulfill another useful purpose, in that information about the surface form of the source text can often be very useful in selecting the appropriate surface form of the target language text.

The necessity to keep all these different kinds of information on a single data structure has led to the definition of a structure whose geometry is determined syntactically. The data structure is a general tree, where at each level of the tree one node is distinguished as being the node to which all other nodes are (syntactically) related. Thus, within a verbal phrase with a finite verb the distinguished node will be the finite verb node, within a noun group the noun and so on. (The analogy with dependency grammar [3]is clear.) This might be expected to lead to the semantic information, which is represented via labellings on the nodes of the tree, being represented somewhat unnaturally. In fact, although this does happen occasionally, it happens rather less often than one would expect, partly because semantic relations, once named, need not be ordered (cf. Charniak [2] for the inverse argument that if ordered they need not be named), partly because an extension of the data structure to allow for copies of constituents to be inserted in those cases where a single constituent plays two semantic roles (e.g. 'I told him to go') removes much of the difficulty. The awkward cases are those where intuitively, the dominant constituent semantically is defined to be the dependent constituent syntactically, e.g. 'the bottle of wine' ('bottle' is syntactically dominant). Even here, intuition tends to oscillate depending on surrounding context: 'He drank the bottle of wine' vs 'He broke the bottle of wine'.

4. COLLABORATION

One of the initial postulates of the system was that it could be designed and implemented collaboratively.

In practical terms, collaboration means individual groups working on the analysis and generation of their own language, joint teams constructing transfer modules and a separate group ensuring communication and co-ordination between the participating groups. Such an organisation is made possible by the strict division into analysis, transfer and generation modules already described.

It is re-inforced. however, by a further consideration. In planning such a project, it is preferable to draw as much benefit as possible from experience already existing amongst the participating groups: indeed, it is the experience of those who have co-operated in planning the system which has produced its overall design. But the fact that experience exists means that it is experience with particular techniques and strategies of language processing. Obviously, a group which has spent many years developing and improving a particular strategy will want to use the results of that work in working on EUROTRA. Therefore, the participating groups should be left considerable freedom to choose their own linguistic strategies.

This has immediate implications, if only because anarchy must be prevented from degenerating into chaos. The most obvious concerns the interface structure: its definition must be agreed by all parties, and all parties must agree to produce results conforming to that definition. For this reason, much of the last three years has gone on defining the interface structure.

A further guarantee of final integrability comes from an agreement to use a common basic software, manipulating an agreed data structure. The data structure is a chain graph, very like the Q-system Q-graph [3], whose arcs are labelled with trees conforming to the definition of the interface structure. It is manipulated via production system type rules | 4 | , internally unordered except of course for implicit ordering, but controlled by external means as in MYCIN |3]• Both the rules themselves and the control mechanisms are written in a specially designed language, intended to allow convenient and transparent expression of linguistic facts. No conceptual distinction is made between grammar rules and dictionary rules, both of which are written in the high-level language. Dictionary information may be very complex, including valency information, semantic information, information on surface behaviour and contextual information in addition to conventional morphological information. The unity of the data structures and of the high-level language for describing and manipulating them makes elegant expression of a great deal of information possible.

Several systems have contributed ideas to the definition of this part of the system. Both GETA (op. cit.) and TAUM [3] , amongst machine translation systems, have used external rules, and production systems as a whole have received wide-spread attention in AI. EUROTRA pushes the general approach to an extreme, by explicitely separating control, rules and computational model in a way which allows the same basic tools to be used in a variety of different ways.

5. EXTENSIBILITY

Multi-linguality has a further dimension which was not discussed explicitely in the second section: it implies the ability to add new language pairs at any time without having to re-write the pre-existing system. This is made possible by the overall modularity of the system. To add a new source language going to existing target languages a, b, c, it is only necessary to write an analyser producing a valid interface structure from source language texts, and three transfer modules transferring the new source into languages a, b, c. The existing target generation modules will then take over.

Similarly extension of the linguistic modules to cover new structures or new domains of discourse is simplified by the use of internally unordered production rules. Addition of new rules does not perturb the existing set of rules.

But extensibility was also defined to include extension to include new research results. In the current state of the art, certain linguistic problems such as resolution of pronoun reference depending on extensive use of world knowledge seem to be intractable within the framework of a bread and butter system. However, given advances in linguistics and in AI in the recent past, it is possible that they may become tractable. The general framework described in the preceding sections should prove flexible enough to allow the incorporation of new research results permitting the treatment of problems which, for the moment, have quite deliberately been left aside. On a less ambitious scale, the rigid separation of rules, algorithms and control should make it easy to experiment with new linguistic models.

6, CONCLUSION

This paper has outlined some features of the design of a multi-lingual, extensible machine translation system to be developed by a number of groups working in collaboration. An attempt has been made to explain the considerations leading to those features, and to set the system within an overall framework.

REFERENCES

[1] Boitet, C. Où en est le GETA début 1977? T.A. Informations, 18, 1977, pp. 3-20.

[2] Charniak, E. A brief on case. ISSCO, WP No. 22, 1975.

[3] Colmerauer, A. Les systèmes-Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Projet TAUM, Université de Montréal, 1971.

[4] Davis, R and King, J.J. An overview of production systems. In Elcock, E. and Michie, D. (eds). MI8: Machine Representation of Knowledge, John Wiley, New York, 1977.

[5] Georgetown University. Machine Translation Research Project. General Report 1952-1963. Paper No. 30, June 1963. Prepared by R.M. MacDonald. Project director, L. Dorstert.

[6] Hays, D.G. Dependency theory: a formalism and some observations. Language, Vol. 40, No. 4, pp. 511-524.

[7] Hutchins, W.J. Machine translation and machine aided translation. Journal of Documentation 34, 1978, pp. 119-159.

[8] Shortliffe, E.H. Computer-based medical consultations: MYCIN. American-Elsevier, New York, 1976.

[9] Toma, P. SYSTRAN as a multi-lingual machine translation system. In Commission of the European Communities; Overcoming the Language Barrier. München, Vlg. Dokumentation, 1977, pp. 569-581.

[10] Vauquois, B. La traduction automatique à Grenoble. Dunod, Paris, 1975.

[11] Vauquois, B. L'évolution des logiciels et des modèles linguistiques pour la traduction automatisée. T.A. Informations, 19, 1978.

[12] Wilks, Y. An artificial intelligence approach to machine translation. In Schank, R.C. and Colby, K. (eds). Computer models of thought and language. Freeman, San Francisco, 1973, pp. 114-151.

[13] Wilks, Y. An intelligent analyser and understander of English. Communications of the ACM, 18, 1975, pp. 264-274.