[*Machine Translation Summit*, September 17-19, 1987, Hakone Prince Hotel, Japan]

THE ALVEY JAPANESE AND ENGLISH MACHINE TRANSLATION PROJECT

F. Knowles — G. Jelinek — M. McGee Wood

The so-called Alvey programme of Information Technology (IT) research and development, sponsored by the British Government, is now nearing the end of its five-year life-span. Approximately £300M has been spent on collaborative research between industry, government departments and academic institutions. The four main areas of activity have been: software engineering, man-machine interfaces, VLSI (very large-scale integration techniques), and IKBS's (intelligent knowledge-based systems). Within the IKBS area a special, corporate identity has evolved for natural-language processing (NLP), and ten projects have been funded in this specific domain.

One such project is the topic I and my colleagues wish to describe here today: it is the IKBS022 project entitled "Read and write Japanese without knowing it", initiated in 1984 at a cost of £0.25M of public sector funds (plus a matching component from the industrial "partner") and scheduled for a three-year track. The logistic basis of this project is - as might to be expected - collaborative: in this case the university-based researchers are working in as integrated a fashion as possible with ICL which has invested complementary — and considerable — resources into the project. These resources are chiefly represented by personnel on secondment to the R&D centres of the universities involved, but are also invested in the form of equipment such as ICL PERQ's and Sun Workstations etc.

The universities involved in this project are UMIST (University of Manchester Institute of Science and Technology) and Sheffield. At Manchester the unit conducting the work is the Centre for Computational Linguistics (CCL), headed by Prof. JC Sager and staffed by experienced NLP researchers working individually or collectively — on a number of different NLP projects, notably EUROTRA. The UMIST team has been concerned with developing an English-Japanese M(A)T system having a specific, "fine-tuned" capability - plus overall robustness - in the domain of computer documentation. The day-to-day managerial control of this R&D effort was performed by Mr. RL Johnson until his move to Switzerland a year ago. The R&D work proper was led until recently by Mr. PJ Whitelock who has now moved to Edinburgh University. Mrs. M McGee Wood was then appointed to take responsibility for the final stages of the Manchester-based work. The other researchers participating in the UMIST effort on Japanese are Mrs. H Horsfall, Mrs. N Holden (a native speaker of Japanese), Ms. E Pollard and Mr. B Chandler, on secondment from ICL to UMIST for the duration of the project.

The other university partner in this conjoint research project is Sheffield, via its Centre for Japanese Studies.

Although organisationally dependent on UMIST, the researchers at Sheffield — some 40 miles distant from Manchester — have necessarily enjoyed a large measure of discretion and freedom over how to pursue their task. The chief researcher at Sheffield is Dr. G Jelinek, a Japanologist with very long and extensive experience of using computers to process language material. His aim at Sheffield has been to develop his expertly structured and algorithmically configured teaching materials (a system known as the "Japanese-English Grammar Dictionary" and designed for teaching English-speakers how to read Japanese technical literature) into an M(A)T system for the translation of computer documentation from Japanese into English. It can hence be seen that the Sheffield and UMIST efforts complement each other totally, although the computational methods used to achieve their separate and disparate aims differ radically. Dr. Jelinek is assisted by Ms. M Gillender, Mr. M James, and - during this project — by Mr. G Wilcock of ICL who is on secondment to Sheffield.

I ought to mention my own role in this large project. I work at Aston University in Birmingham and was appointed Monitoring Officer by the Alvey Directorate: my principal duties are, firstly, to see that work is performed in accordance with the parameters and timetable set, and secondly, to attempt to alleviate any other problems which arise and conspire to frustrate the smooth progress of work, be those problems logistic or even the result of conflicts of view about how work should proceed. I am happy to report that no serious problems of either kind have arisen!

The university-based research and development on the conjoint two-way Japanese and English M(A)T system is scheduled to cease at the end of October 1987. At that stage ICL will take delivery of the two prototype systems: ICL's firm intentions are to press ahead — directly, in the case of the English-Japanese system and potentially via a consortium arrangement with a third party in the case of the Japanese-English system - with the transformation of both systems from prototypes to marketable products. The hope is that some sort of commercialisation can be achieved within the succeeding twelve to twenty-four months. It is of course clear that ICL will be able to call on the university-based expertise during this period — it may well be the case that extra resources can be found to permit the further aggregation of dictionary entries and, potentially, enhancement of the linguistic strategies embedded in the software. It is even possible that consideration may be given to the task of generating spin-off systems for different text types and certain different languages, such as Korean, for instance.

We now proceed to a succinct description of the systems themselves: we are grateful for permission to quote from reports submitted to the Alvey Directorate.

The Sheffield AIDTRANS Japanese-English M(A)T System

The AIDTRANS system is derived from a highly sophisticated and detailed analysis of Japanese grammar - developed by Dr. Jelinek — carried out for the purposes of teaching people without extensive linguistic training or talents how to "decode" Japanese text. This manual teaching system's core is the so-called Integrated Dictionary System (IDS) which has shown its worth over many years in teaching situations. The philosophy of IDS is to pack as much grammar and heuristics into the dictionary as possible. The emphasis of this approach, which is firmly rooted in classical contrastive linguistics, is hence on the elaboration of major grammatical and lexical resources for transferring from a given source language to a given target language. Perhaps it could be said that the comprehensiveness of the elaboration compensates for the monodirectionality of the resultant materials. More generally speaking, the IDS approach to translation is that each translation transaction is a unique event, the purpose of which is to retain the sense, rather than the sum of individual concatenated meanings, of an utterance as it is reformulated in a different language. From the point of view of M(A)T the primary implication of this philosophy is that - as a necessary preliminary - all potential translations into the target may need to be generated from the single source input structure; subsequently all except one of these potential translations need to be discarded - the one which is retained is the one which commends itself most by reason of having scored best during a battery of criterial tests. The M(A)T strategy best suited for achieving such an objective therefore reduces to an optimisation problem for the ordering and selection/deselection of alternatives.

This approach was perhaps best represented in the "classical days" of MT research by linear predictive analysis, relying on a single left-to-right scan of input. The chief remaining task is then to carry out multiple juxtapositional analyses of the input segments: the arbitrage of these analyses is usually directly driven by known, that is, stored statistical or tactical data appertaining to the particular constituent segments. The "finer" and more reliable this data, the better: in many cases, of course, the statistical or preference data may refer to actual terminals, or lexemes, rather than to non-terminals. This technique is clearly recognisable as an environmental analysis of valency, sensu largo, with the added feature that the valency values consulted or computed are prioritised: this is what drives the whole system in the fashion of a tournament until the "best" translation equivalence is elicited. At present well over 200 different types of juxtapositional linkage are recognised by the AIDTRANS system.

It is obvious that much effort is required to reconfigure a system of the above type for a different text typology; nonetheless it may well be cost-effective to do that in particular circumstances. The feasibility of the AIDTRANS approach appears to be already assured in terms of the present project — its future depends to a great extent on the results of the minute analyses of both Japanese text and lexicon which are currently under way.

The UMIST English-Japanese NTRAN System

The NTRAN system is designed to enable technical authors of computer software manuals who have no knowledge of Japanese to read and write such texts in Japanese. The prototype is hence designed to work within a particular text-type and accordingly incorporates only an appropriately configured restricted grammar and dictionary. The translation system is interactive in design, aiming to exploit the author's human expertise for the resolution of grammatical and lexical ambiguity. The system consists of the following modules: 1) English analysis; 2) Transfer; 3) Japanese generation; 4) monolingual and bilingual dictionaries.

The English analysis module is loosely based on Lexical-Functional Grammar (LFG). The F-structure serves as an intermediate representation which abstracts away from the surface constituent structure. This is a highly desirable feature for a translation system which involves two such very disparate languages as English and Japanese, the former configurational, the latter non-configurational. Japanese generation is based on categorial grammar.

A bottom-up parser allows compiled grammar rules to be interpreted by the PROLOG top-down interpreter in bottom-up fashion. This provides a significant improvement in parsing complexity whilst retaining the proportionality constant at the same value as that of directly interpreted DCG's. The F-structures are built directly during parsing and the well-formedness condition on the F-strucures is directly realised by the PROLOG interpreter's own unification algorithm. It will have emerged that the NTRAN system is implemented in PROLOG, principally for reasons of rapid prototyping.

The lexical entries in the monolingual dictionaries contain information pertinent to each individual word's number and specificity, information which is incorporated into the F-structure during look-up. A series of feature co-occurrence restrictions (FCR's) is also applied at this stage. These FCR's serve to reduce the size of the dictionary entries by acting as redundancy rules across large numbers of lexical entries. The parsing process itself is invoked by the interaction of the FCR's and the instantiated dictionary entries. User-driven disambiguation occurs as necessary. The F-structures then have subcategorisation information appended to them which specifies the semanto-syntactic environments in which given lexical items may occur. Mappings from functional roles to argument positions are effected on the basis of this information and the S-structure which emerges is passed to the transfer module where interaction with the bilingual dictionary results in the substitution of the English S-structure's values by their Japanese semantic equivalents. The second stage of transfer effects the translation of attribute value-pairs which do not have semantic values. Any other contrastive information — which may be necessary for determining the topic in Japanese, for instance - will be appended at this stage to the resulting S-structure for the target string. Deep-to-surface mappings then take place, producing a Japanese F-structure from the prior S-structure. Japanese subcategorisation information is then applied and as a consequence the appropriate Japanese grammatical particle is associated with the relevant Japanese lexical units. The resulting string is finally reordered according to linear precedence rules.

It should be understood that in the final stages of the project most work in hand is directed towards enhancing the robustness, disambiguation, and coverage aspects of the system. The main thrust has of course been to get the linguistics right!

"The Englishman's Keyboard"

A useful adjunct of the Alvey MT project involving Japanese has been the development - not yet quite complete - of the so-called "Englishman's keyboard". This is a system which enables the non-native speaker of Japanese, for whom the AIDTRANS SYSTEM is designed, to bypass the problem of inputting Japanese-language data "indirectly", i.e. by means of Roman-script equivalences. The translation system already possesses, of course, a Japanese-script input module but users need to have a knowledge of how to read kana and kanji or — in the latter case - at least an ability to find out the readings of kanji characters from a standard dictionary. "The Englishman's Keyboard" is a system which allows users to input kanji by using visual information rather than readings and it consequently much more suited to users with a minimal knowledge of Japanese script. The editor simply displays the most common radicals on the screen and the user selects, via his mouse, the radical required; the complete set of characters associated with this radical is then displayed for a new selection by mouse.

Let us conclude with a recapitulation of the context and profile of the British Alvey project in MT. Fully automatic high-quality machine translation of unrestricted text is now generally recognised to be an impracticable proposition for the immediate future. The commercial, political, and scientific demand for cross-language communication has nevertheless meant that machine translation systems and — more particularly — sophisticated "translator's aids" which fall well short of ideal or quasi-ideal performance continue to be developed. Advances in computer programming languages and techniques, as well as in linguistic theory and the formal encoding of complex knowledge are allowing the achieved performance of state-of-the-art translation systems - that is to say, necessarily R&D prototypes - to improve rapidly, while more realistic perceptions of the precise capacities of such systems are informing the development of various strategies for optimizing these systems by minimising their "quality shortfall"; they also minimise but particularise in a most interesting way their human users' contribution to the overall task of translation. The Alvey programme seeks to accelerate the process of getting this expertise on to the market, in fact of reducing the gap between state-of-the-art know-how and the functionality of commercial products.

The most commonly used strategy for human/machine cooperation, i.e. human post-editing of more or less crude raw machine output, can be carried out only by a bilingual, that is to say, a fully competent translator, and the usefulness of this strategy is hence restricted merely to bringing speed and some terminological consistency to the overall translation process. Conventional interactive machine translation suffers from the same limitation, as queries may be presented in either source or target language. In pre-edited machine translation human intervention is restricted to the input, source language stage and is thus made available at the cost of unacceptable restrictions on input, and is therefore not appropriate for the translation of existing, independently composed texts.

Both the Sheffield AIDTRANS system and the UMIST NTRAN system attempt to combine - in very different ways and to different extents, not least by reason of the differences of task - the advantages of these strategies. Quoting the NTRAN system by way of exemplification, we can state that this is a monolingually interactive system, guaranteeing an English monolingual technical writer accurate Japanese output without post-editing. Any ambiguities encountered by the system at any point in the translation chain are referred to the user via an English-language query; even selection between alternative Japanese translation equivalents for an English word is done by means of English glosses held in the Japanese dictionaries. Thus the restrictions on input text are simply those of grammaticality and freedom from ambiguity - desiderata, of course, for any technical text in any language. Hence the NTRAN translation facility is available to any technical author of the source language who writes computer documentation, whilst accuracy of output is ensured by intelligent interaction with that self-same author.

I would like, on behalf of the Alvey Directorate and on my own and my colleagues' behalf as well, to thank the organisers of the "MT Summit" for the opportunity they extended to us to present our paper and to gain from participating in an event which we believe will subsequently be seen by researchers and funding agencies alike as a very important and influential watershed in the development of MT/MAT worldwide.