4. Machine Translation

Martin Kay, Chairperson

Xerox Corporation
Palo Alto, CA 94304

Panelists

Margaret King, ISSCO
Jack Lehrberger, Université de Montreal
Alan Melby, Brigham Young University
Jonathan Slocum, University of Texas

A Manhattan project could produce an atomic bomb, and the heroic efforts of the sixties could put a man on the moon, but even an all-out effort on the scale of these would probably not solve the translation problem. In one sense, this almost goes without saying. The first ninety percent of the work that was invested in reaching the moon did not get the astronauts nine-tenths of the way there and the heat generated by the bomb did not increase steadily as the date of the first explosion approached. The translation problem is one whose solution must be reached incrementally. There will be no dramatic event to signal the end of the search and there is no single breakthrough that would assure success.

The translation problem is real and will in fact rapidly reach crisis proportions unless some action is taken. The problem cannot be alleviated by better language teaching, greater incentives for translators, or improved administrative procedures, worthy though these goals undoubtedly are. The only hope for a thoroughgoing solution seems to lie with technology. But this is not to say that there is only one solution, namely machine translation, in the classical sense of a fully automatic procedure that carries a text from one language to another with human intervention only in the final revision. There is, in fact, a continuum of ways in which technology could be brought to bear, with fully automatic translation at one extreme, and word-processing equipment and dictating machines at the other.

The productivity of the professional translator could almost certainly be greatly increased by technological aids which, though straightforward, are not all obvious. Powerful machine aids to translators could quickly be available and may be the best way to alleviate the translation problem in the short run. They will not arise as a natural by-product of work on fully automatic translation because, for the most part, they address such issues as communication among translators, identification of relevant secondary material, and special editing devices, rather than issues of syntactic analysis, pronominal reference and quantifier scope. The most valuable resources that a translator has for solving difficult problems are the text he is working on, other texts like it in the target as well as the source language, and his colleagues. At present, access to these resources is haphazard at best. But, improving it immeasurably is well within the scope of existing technology.

Another easily identifiable point on the continuum is occupied by human-aided machine translation. This could be a very different kind of enterprise both from fully automatic and machine-aided translation. By human-aided machine translation, we mean to refer to systems in which the machine, while retaining the initiative, works with a human consultant, who need not be a translator. Once again, the subtleties in the design of the system would not reside so much in basic linguistic questions as in how to recognize reliably when a difficulty of a certain type had arisen and how to communicate the nature of the difficulty to the consultant in such a way as to elicit a quick and unambiguous response. Especially in the early stages, a human-aided machine-translation system intended to produce output of high quality might well require at least as much work on the part of the consultant as a trained translator would take to do the job in the traditional way. However, two facts can be set against this. First, the consultant would not have to be a translator and could quite possibly be drawn from a much larger segment of the labor pool. Secondly, while the labor involved in translating a text grows in direct proportion to the number of languages into which it must be rendered, the work required of the consultant in such a man-machine team would grow much more slowly. Indeed, if those languages were closely related, it could be expected to fall off sharply as soon as that number exceeded one.

A substantial proportion of what follows will be devoted to upholding the panel's view that it is important for work to proceed in parallel on a number of different fronts. While fully automatic translation is the most adventurous, it is from this that we stand to learn most about language in general, and translation in particular. If fully automatic systems can be built whose performance exceeds that of present systems by even a modest amount, we should profit greatly as well

from their practical utility as from the theoretical lessons enshrined in them. Machine-aided translation can enhance the translator's productivity, though we have yet to discover how much enhancement is possible in this way. It could also be a source of invaluable information on how translators work. Human-aided machine translation can be expected to give better results than could be achieved with the fully automatic method, since the human consultant can be called upon to resolve otherwise unresolvable problems, but at an unknown cost. However, there are important applications, notably where one text must be translated into several languages, where the gains may be substantial.

We have been at pains to make it clear that the three methods of involving machines in language translation are not essentially different in kind but lie on a continuum. Except in a few special applications, some of which we will mention shortly, we do not foresee a time when translations of any quality will be produced without any human intervention whatsoever. In the so-called fully automatic method, the human plays the role of an editor, or revisor. His involvement begins only after an initial draft in the target language exists and it is for this reason that we remain content with the term "fully automatic". The other two methods involve him earlier so that he influences even that first draft.

The methods therefore differ as to how the person is involved. They also differ in the extent of his involvement. If the human partner can influence all the decisions that are made, he may be in a position to forestall sequences of errors, each resulting from the one before, thus reducing the total amount of his contribution. On the other hand, if a conservative system insists on having him confirm even those choices for which its own decision methods are substantially adequate, then the overall extent of his involvement may be increased. In any case, the utility of a given system in a particular situation cannot be assessed by a simple equation. The appropriate utility function involves at least the human cost, the machine cost, the quality of the result, and the nature of the consumer$_9$s requirements.

The consumer's requirements are, of course, crucial. The various types and degrees of automation in translation are, as we have seen, positioned along one dimension in a space of possible approaches to the overall problem. The different types of text and their consumers are another dimension and, not surprisingly, the two dimensions are far from independent. The type of technology appropriate to a problem, and the benefits to be expected from it, differ greatly with the type of the text to be translated and the use to which the result will be put. In the intelligence services, a great deal of translation is done for purposes of current awareness. The first priority is to know the subject matter of the document. It is also helpful to be able to discern the gist of the argument so as to discover whether it touches on certain key questions. A rough and ready translation, especially if it can be done quickly and cheaply, may give an excellent basis on which to decide which parts of a document, if any, need to be translated more carefully. Fully automatic translation, even of quite inferior quality, has already proven very valuable in this role.

Fully automatic translation, or some close relative of it, has also proved useful in recent years in situations where a sublanguage has come to be used, or where one can be readily imposed. Canadian weather reports are routinely translated by such a system. The system itself determines whether each translation unit – approximately a sentence – is within its capabilities. If it is, then it produces a translation, which is the one that will be used without human revision. If not, it presents the translation unit to a human collaborator, who makes the translation. We prefer to classify this with fully automatic translation because, though the machine does not translate everything that is translated, the translation it does is done entirely without human involvement even at a post-editing stage. The machine in fact translates eighty per cent of all translation units and readers of the reports prove unable to discern which parts were translated by machine and which by a human.

The success of this METEO system comes from the fact that meteorologists naturally write in a highly constrained subset of English. Fully automatic translation has also been successfully applied to the task of translating maintenance manuals for machines. The success of this does not rest on the existence of a naturally occurring sublanguage. In this case, the technical writers who prepare the manuals learn to follow a set of rules intended to ensure that their products will automatically translatable by simple means. The rules are straightforward and can be learnt in a two-week course. The machine translates the whole text without outside assistance and preliminary results encourage the belief that little or no editing will be required.

The features of a sublanguage that make it suitable for fully automatic machine translation are (1) restricted vocabulary, with consequent reduction in the number of words with more than one grammatical category, (2) small number of senses for each word in a given category due to the restricted semantic domain, and (3) restricted syntax resulting from the purpose of the text, e.g., instruction manuals may contain only imperative sentences and weather reports only declaratives. It should not be thought, however, that a sublanguage is simply a subset of the sentences of the standard language. The syntax of a sublanguage may differ radically from that of the standard language so that a

grammar of the latter would not cover the constructions of the former. Thus "Fair tommorrow" and "Winds from the northeast" are "sentences" in a weather bulletin. There are also closely related domains in which texts have a common syntax, and differ only in vocabulary. An extensive study of sublanguages, their restrictions and interrelations, will be important for determining the range of applications of fully automatic translation. The question of is a complex subject to which the report of another panel is devoted.

The following table summarizes our view of the three most interesting points on the continuum.

	Fully Automatic Machine Translation	**Human Assisted Machine Translation**	**Machine Assisted Human Translation**	
Information Acquisition	can be quite cheap (revision excluded)		increased human efficiency	*advantages*
	requires effort and experience to read	N/A	more expensive and slower then FAMT	*disadvantages*
	technical material; possibly other material		almost any material	*text types*
	applications exist, greatly improvable		technology exists; might use FAMT to select candidates	*status and prospects*
Denotative Translation	faster and cheaper than human 1st pass	very high-quality especially multi-lingual	increased human efficiency	*advantages*
	human revision required	possibly high cost	high minimum costs	*disadvantages*
	technical material	technical material	almost any material	*text types*
	technology coming of age; applications exist (METEO); very large intermediate and long-term payoff	few or no existing prototypes; FAMT spinoffs possible in near term with suitable funding	commercial systems exist (e.g., ALPS); FAMT spinoffs could reduce costs in near term	*status and prospects*
Connotative Translations			increased human efficiency	*advantages*
			necessarily costly	*disadvantages*
	N/A	N/A	legal and religious texts; literature?	*text types*
			technology exists; greatly improvable	*status and prospects*

Legend

Fully Automatic Machine Translation (FAMT) refers to translation wherein the programs run in "batch" mode (off-line) and produce translations without human intervention; afterwards, human revision (post-editing) may be performed with a text editing program or via other means, if desired.

Machine-Assisted Human Translation (MAHT) refers to translation wherein the program is a fancy editing and dictionary concordance tool which the human translator uses to increase his efficiency by automating his access to word definitions and terminology correspondences. All initiative resides with the human, unlike FAMT and HAMT.

Information Acquistion refers to a situation in which translation is being performed for "current awareness" or "screening" purposes where a quick-and-dirty approach may be sufficient, at least to determine if a more careful translation is justified. No human revision (post-editing) is assumed.

Denotative Translation refers to an information dissemination situation in which the everyday and technical definitions of the words are meant, and where subtle nuances of a word choice are unjustified or even undesirable. This is typical of technical texts.

Connotative Translation refers to an information dissemination situation in which subtle nuances of word choice are very important, if not critical, in conveying the intended meaning of the text. This is typical of, for example, religious, and literary texts.

At the opposite end of the spectrum from current-awareness services are such delicate enterprises as the translation of legal statutes and political speeches. A lawyer in Finland can base his arguments either on the Finnish or the Swedish version of the applicable law, according to which he considers will most favor his client's case. All statutes must be translated and the translators must be at great pains to ensure that there is no construction, however perverse, that can be put on one version but not on the other. We do not foresee a time when any part of this job could be usefully consigned to a fully automatic or even a human-aided, system. On the other hand, it would be a prime candidate for machine-aided methods. There is constant need to compare one part of the text with others, and with other legal texts, to ensure consistency, and this is where these methods come into their own.

If our optimism about the future of mechanical methods in translation has increased during the twenty years during which it has been seriously pursued, it must be largely because of important advances that we perceive in theoretical and computational linguistics as well as computer science. Advances in computer science are the least contentious of these. The construction of most large internal memories was not available and external memory could be accessed only in a serial manner. The consequent inefficiency in the programs that were written is less important than the undue amount of effort that was required to make them work at all. A machine-translation program was large, even by today's standards, and each one produced in the sixties was a programming tour de force. The achievements are even more impressive for the fact that they were made without the aid of the compilers, editors and other paraphernalia that programmers now take for granted, and before the great value of certain programming practices and disciplines had been recognized.

Many of the important advances made in computational linguistics during the same period also tend towards the easier construction of more robust systems that can be more readily maintained. The most obvious examples come from the domain of syntactic analysis which is now universally thought of as a job to be done by a fairly general parsing program, coupled with a grammar. The parser embodies the necessary strategies and techniques while all knowledge of the particular language resides in a static data structure, namely the grammar. Associated with the grammar is a formal language in which a linguist writes rules from which the data structure is obtained automatically. This formal language is specially designed to facilitate the statement of linguistic facts and is largely decoupled from the grammar itself and from the methods that will be used to process it. This greatly increases the power that the linguist can bring to the job and his ability to modify the system in the light of experience.

In the same period we have come to understand, not just how a general-purpose parser can be constructed, but how to make these parsers more effective by the application of some very general principles. In particular, we have come to appreciate the value of the notions of complete parsing and of nondeterminism. By complete parsing, we mean simply the requirement that nothing shall count as part of the final result except inasmuch as it is part of an analysis of an entire sentence. The practical value of this apparently obvious restriction would be difficult to overestimate. A parser that incorporates it largely releases the grammar writer from concern for when it would be incorrect to allow an analysis that would be correct in another environment, a concern which consumed much time on the part of the designers of early translation systems.

General methods for implementing nondeterminism go together with this and have an equally liberating effect. These methods are useful in situations where, at any given stage of the process, a number of conflicting possibilities are open, any number of which could lead to useful results. In particular, a complete parser, in the present sense, must pursue lines of attack that seem reasonable on the basis of local eviden-

ce, but which may or may not lead to a complete analysis. A general method for handling nondeterminacy releases the programmer from all concern for how the machine will contrive to follow up on all possibilities; how and when it will return to the choice point and restore the exact state as of that moment, how it will follow all possibilities once, but none more than once, and so forth. To the extent that early translation systems faced these problems at all, they did so on a case by case basis, and at great cost in human labor.

The panel was also impressed by the advances that have been made in general linguistics and our overall understanding of the workings of human language in recent years. Various classical problems--noun-noun compounds in English, ambiguities of prepositional attachment, conjunction, pronominal reference, and many others--have been made the object of intensive research with results that are direct relevance in the construction of translation systems.

The panel thought it had every reason to assume that progress on the relevant fronts would go forward at least as quickly as in the past. Most of the members confidently expect to see some major new fully automatic systems in use during that period. In particular, it is hoped that EUROTRA, a very large-scale collaborative European effort, will result in a working prototype. In addition, there is work in progress in a number of places on computer-based work stations for use by translators. It is not clear what these will incorporate, but it is likely that they will explore some new parts of the large space of possibilities that exist in machine-aided translation.

The panel made no prediction about just which areas of research were likely to fall before the inexorable advance of theoretical linguistics but felt that the future of more technological areas was easier to foresee. There will, in all probability be more flexible methods of syntactic analysis, capable of relaxing requirements in the face of constructions that would not meet with a pundit's approval. Either they will fall back on more permissive rules or they will modify the sentence to "correct" the apparent "error". It is also expected that wider use will be made of parsing devices which, while allowing for nondeterminism in a general way, will be able to make fairly accurate judgements about the paths that are most likely to lead to a successful solution, and so concentrate on these first.

The panel was in agreement on the achievements of the past and on the desirability of following a number of parallel paths in the future. The disagreements concerned the extent of the optimism in future successes that those past achievements warrant. Some members took the view that advances in computer science and computational linguistics, important though they are, do not go to the heart of the problem; they make easier what once was hard but they make nothing possible that once was impossible. Linguistics has made advances of which it can feel justly proud but which, while they may indeed go to the heart of the matter, barely scratch the surface of what needs to be done.

All agree that it would be unjustifiable to devote an excessive proportion of the available resources to fully automatic systems while neglecting cheaper and more modest approaches with more certain short-term payoff. Unless caution is exercised, both in promises made and policies followed, there is a high risk that taxpayers and administrators will call all too soon for a second ALPAC report whose effect on the entire field of computational linguistics will be altogether more devastating than the first.

Some members believe that sponsors have grown more realistic in their expectations, that so long as they are involved in a continual dialog about the progress of the work they can be made to understand the problems, and that they have the fortitude to withstand unreasonable pressure from their superiors and their electors. They no longer, for example, expect full translations of arbitrary texts, but are content with texts from suitably restricted domains.

It is claimed earlier in this report, and agreed upon by all the panel members, that fully automatic translation is the line of attack whose benefits, if realized, would be greatest. Futhermore, its success would contribute greatly to the successes of all other approaches. The subscribers to this view are impressed by the extent to which the designers of early systems were overcome by the sheer complexity of the design and programming task that they had undertaken so that the systems they built cannot be taken as a measure of the technology that linguistics, even the linguistics of that day, could support.