[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

Current Research in Machine Translation

Harold L. SOMERS

Centre for Computational Linguistics
UMIST, PO Box 88
Manchester, England

1. Introduction

The purpose of this paper is to give a view of current research in Machine Translation (MT). It is written on the assumption that readers are in fact more or less familiar with most of the well-known current MT projects, or else can find out more about them by following up the references given. My intention is to make some perhaps slightly opinionated remarks about certain of these projects, which, I will claim, have in common a direct line of descent from the classical 'second generation' design. I will then describe what I believe to be a significantly different set of current MT research projects - mostly rather less well-known - which form a heterogeneous group having in common only the feature that they in some sense reject the conventional orthodoxy that typifies the first group.

What this paper will *not* do, therefore, is to review the history of MT so far: see Hutchins (1986) or Nagao (1986/9) for the generally accepted version, or Wilks (1987) for an AI-oriented view. Nor will this paper attempt an exhaustive list of on-going research projects: Hutchins (1988) and JEIDA (1989) include such a list.

Two recent personal experiences have led me to the views on current MT research which I wish to elaborate here.

The first was two years ago when, at the previous conference in this series *(International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages)* at CMU in Pittsburgh, I participated in a panel session and addressed the question "Where will MT be in the next 20 years - honestly?" (Somers 1988). At that time I thought that the main developments would be investment in lexical development, and work on making the environment for MT more user-friendly, especially by having linguistically sophisticated MT-oriented word processing for post-editing. Looking back, I am struck by the fact that neither development really represented either a theoretical or methodological advance. The other thing that happened at CMU which made an impact was the reaction to the presentation of IBM's Peter Brown on the same panel (Brown *et al.* 1988a; cf. Brown *et al.* 1988b): it was hostile, to say the least, despite the fact that early results were not significantly worse than results of more orthodox systems. I joined the attack, the main thrust of which was to ask where was the linguistics, without realising that precisely what the research was doing was to question some of the fundamental assumptions underlying MT research since 1966, and try to find out which of them were really valid.

With hindsight, I can see that what this research was doing was saying that in the twenty years since ALPAC, the second generation architecture had led to only slightly better results than the architecture it replaced; so it was timely to question *all* the assumptions that had been accepted unequivocally in that period, *just to see what would happen*. I will return to this view later.

The second personal experience, which strengthened my opinion that the received view of the future of MT research was at best too restricted, or at worst totally misguided, was my attendance at an MT conference in Tbilisi, organised by the USSR's Vsyesoyouznyi Centr Perevodov, in November/December 1989. The significance of the Tbilisi conference was that many of the papers presented by Soviet researchers revealed that, due to the state of computer technology in the USSR, MT research in that country was about fifteen years behind the West, and following faithfully in its footsteps. It

seemed obvious that the mainstream of Soviet research would continue, as much as possible, to emulate the research of the West, including reaching the same, not entirely positive, conclusions some fifteen years from now. It occurred to me that they could be saved a lot of wasted effort if someone could indicate succinctly what those conclusions would be, and allow them to jump to the position I believe we in the West already find ourselves in, and embark on research projects which try to address these shortcomings. Perhaps part of this paper will provide such an indication.

2. What's wrong with the classical second generation architecture?

Let us start by considering the classical second generation architecture. Examples would be GETA's ARIANE system (Boitet & Nedobejkine 1981, Vauquois 1985, Vauquois & Boitet 1985), TAUM's METEO (Chevalier *et al.* 1981, Lehrberger & Bourbeau 1988), and the European Commission's Eurotra (Raw *et al.* 1988, 1989; Allegranza *et al.* forthcoming), and there are plenty of other systems which incorporate most of the typical design features. These include the well-known notions of linguistic rule-writing formalisms with software implemented independently of the linguistic procedures, stratificational analysis and generation, and an intermediate linguistically motivated representation which may or may not involve the direct application of contrastive linguistic knowledge. The key unifying feature is *modularity*, both 'horizontal' and 'vertical': the linguistic formalisms are supposed to be declarative, so that linguistic and computational issues are separated; and the whole process is divided up into computationally and/or linguistically convenient modules.

While these are admirable design features, at least insofar as they seem to address the perceived problems of MT system-design pre-ALPAC, they also lead to several general or specific deficiencies in design.

In general, they reflect the preferred computational and linguistic techniques of the late 1960s and early 1970s, which have to a great extent been superseded. There are now several viable alternatives to the procedural algorithmic strictly-typed programming style; while in linguistics the transformational-generative paradigm and its associated stratificational view of linguistic processing (morphology - surface syntax - deep syntax) has become somewhat démodé.

The stratificational approach engenders two other problems which cast a shadow over second generation-style MT. First, there seems to be a tendency, once the general design of the MT system has been fixed, to go about the finer details, and the implementation, in a bottom-up manner: it is as if there is an attitude of doing what is known to be feasible (morphology, context-free parsing, incorporating simple semantic constraints, some tree transductions), seeing how far that gets you, and taking it from there. When the ideas run out, call whatever you've got an 'intermediate representation', do some 'transfer', and then a more or less deterministic generation of target text (this approach to generation in particular being criticised as long ago as 1985, at the first conference in this series, as being out of date (McDonald 1987:200ff)). A more appealing way to design an MT system would of course be to *start* by considering what sort of intermediate representation (or, more generally, what sort of contrastive processes) underlie the system, and then to consider how to analyse source texts into that representation, and how to generate target texts from it.

A second, perhaps more serious problem with the stratificational approach is the extent to which it encourages an approach to translation which I have called "structure preserving translation as first choice" (Somers *et al.* 1988:5). This stems from the commitment to compositionality in translation, i.e. that the translation of the whole is some not too complex function over the translations of the parts. This leads to a strategy which embodies the motto "Let's produce translations that are as literal as we can get away with" (cf. Somers 1986:84). Notice that this is in direct contrast with the human translator's view, which is roughly "structure preserving translation as last resort". This attitude can be seen again in discussions of the need to limit 'structural transfer' and to build systems which are essentially interlingual systems with lexical transfer. But we know very well the difficulties of designing an interlingua, even if we remove the burden of a 'conceptual lexicon'.

I must admit that I do not have a ready solution here. But it seems to me important to

recognise the limitations and pitfalls of the now traditional stratified linguistic approach to both processing and representation, so that even the apparently well established technique should not necessarily be assumed as a 'given' in MT system design.

I will end this section by making two other observations. The first is that all MT systems so far have been designed with the assumption that the source text contains enough information to permit translation. This is obviously true of non-interactive systems; but it is also true even of the few systems which interact with a user *during processing* in order to disambiguate the source text or to make decisions (usually regarding lexical choice) about the target text. Notice, by the way, that I want to distinguish here between truly interactive systems, and those which merely incorporate some sort of interactive post-editing. In fact, very few research systems are truly interactive in this sense (e.g. ENtran (Whitelock *et al.* 1986, Wood & Chandler 1988), and see below). However, the point I want to make concerns how MT researchers view this problem: it is seen as a deficiency of the system - that is to say, either the linguistic theory used, or its implementation - rather than of the text. Consequently, the solutions offered almost inevitably involve trying to enhance the performance of the part of the system seen to be at fault incorporating a better semantic theory, dealing with translation units bigger than single sentences, trying to take account of contextual or real-world knowledge. Of course these are all worthy research aims, but I think the extent to which they will address the problems they are supposed to solve is generally exaggerated.

The second point is the observation - whisper it - that despite nearly 25 years since the ALPAC report, results are not much better than those of the first generation systems which have over the same period continued to be developed (though probably with less invested effort overall): obvious examples are SYSTRAN (World Systran Conference 1986, Trabulsi 1988) and SPANAM (Vasconcellos & Leon 1985). As Wilks says of the former, "its real techniques owe a great deal to good software engineering, good software support..." (Wilks 1989:59). No one would deny that the second generation systems are more elegant, or even that they can be extended in a more principled way. But for all the investment, and the bold talk in, say, the mid to late 1970s, perhaps one could have expected better results. With the exception of METAL in the West, and two or three systems in Japan, notice that *all* commercial systems are first generation in design. Notice too that people buy them.

3. Current research directly descended from that architecture

I want to look now at current research projects which I take to be directly descended from the second generation architecture, and which therefore, in a sense, can be said to be subject to the same criticisms. The research projects in this group can be divided into subgroups according to which specific part of the problem of MT, as traditionally viewed, they try to address. So we have projects which address the problem of insufficient contextual and real-world knowledge; projects which seek a more elegant linguistic or computational linguistic framework; and projects where translation quality is enhanced by constraining the input.

For a while now it has been the conventional wisdom that the next advance in MT design - the 'third generation' - would involve the incorporation of techniques from AI. In his instant classic, Hutchins (1986) is typical in this respect: "the difficulties and past 'failures' of linguistics-oriented MT point to the need for AI semantics-based approaches: semantic parsers, preference semantics, knowledge databases, inference routines, expert systems, and the rest of the AI techniques" (p.327). He goes on to say "There is no denying the basic AI argument that at some stage translation involves the 'understanding' of a [source language] text in order to convey its 'meaning' in a [target language] text" *(idem).* In fact this assertion has been questioned by several commentators, e.g. Johnson (1983:37), Slocum (1985:16) etc., as Hutchins himself notes.

3.1. Incorporating AI techniques

Returning to the question of AI-oriented 'third generation' MT systems, it is probably fair to say that the most notable example of this approach is at the Center for Machine Translation at Carnegie-Mellon University, where a significantly sized research team was explicitly set up to pursue the question of 'knowledge-based MT' (KBMT) (Carbonell & Tomita 1987). What then are the 'AI techniques' which the CMU team have incorporated into their MT system, and how

do we judge them?

In the Nirenburg & Carbonell (1987) description of KBMT, the emphasis seems to be on the need to integrate discourse pragmatics in order to get pronouns and anaphora right. This requires texts to be mapped onto a corresponding knowledge representation in form of frame-based conceptual interlingua. More recent descriptions of the project (Nirenburg 1989, Nirenburg & Levin 1989) stress the use of domain knowledge. These are well respected techniques in the general field of AI, and we cannot gainsay their application to MT. But as 20 years of AI research has shown, the step up from a prototype 'toy' implementation to a more fully practical implementation is a huge one. And there still remain doubts as to whether the improvement of quality achieved by these AI techniques is commensurate with the additional computation they involve.

3.2. Better linguistic theories

It is normally said that a major design advance from the first to the second generation of MT systems was the incorporation of better linguistic theories, and there is certainly a group of current research projects which can be said to be focussing on this aspect. This is especially true if we extend the term 'linguistic' to include 'computational linguistic' theories. The scientific significance of the biggest of all the MT research projects - Eurotra - can be seen as primarily in its development of existing linguistic models, and notable innovations include the work on the representation of tense (van Eynde 1988), work on homogeneous representation of heterogeneous linguistic phenomena (especially through the idea of 'featurisation' of purely surface syntactic elements, and a coherent theory of 'canonical form') (Durand *et al.* forthcoming), as well as, in some cases, the first ever wide-coverage formal (i.e. computational) descriptions of several European languages. As much as anything else, Eurotra has shown the possibilities of an openly eclectic approach to computational linguistic engineering. Nevertheless, 'Eurotrians' will be the first to admit that the list of remaining problems is longer than the list of problems solved or even half-solved. 'Lexical gaps', usually illustrated by the well-worn example of *like/gern*, modality, determination, are just a few more or less purely linguistic problems that remain, before we even think of anaphora resolution, use of contextual and real-world knowledge and so on, already discussed.

Several research projects have taken a more doctrinaire view of linguistics in that they have explicitly set out to use MT as a testing ground for some computational linguistic theory. Most notable of these is Rosetta (Landsbergen 1987a,b) based on Montague grammar, but we could also mention again ENtran, which uses a combination of LFG and GPSG in analysis, and Categorial Grammar for generation. There are several other research projects based on specific linguistic theories including LFG (Rohrer 1986, Alam 1986, Kudo & Nomura 1986, Kaplan *et al.* 1989), GPSG (Hauenschild 1986), Categorial Grammar (Beaven & Whitelock 1988), Functional Grammar (van der Korst 1989), Situation Semantics (Rupp 1989), and, though it may be regarded as more of a programming technique than a linguistic 'theory' as such, Logic Grammar (Huang 1988, McCord 1989, Isabelle *et al.* 1988). In all these cases, I think it is fair to say that under the stress of use in a real practical application, the linguistic models, whose original developers were more interested in a general approach than in working out all the fine details, inevitably crack.

A good example of this is suggested by Carroll (1989). Looking at Rosetta, he shows (pp.37f) how the all-important isomorphy principle found in and adhered to in the prototype Rosetta2 system is effectively abandoned in the expanded Rosetta3 project (Appelo *et al.* 1987:122): since some syntactic rules in Dutch do not correspond in an obvious way with English syntax rules (the example given is the Dutch 'verb second' rule), the isomorphy principle requires a dummy English rule to be added to the English syntax. Since this is not very elegant, a distinction between 'transformations' and 'meaningful rules' is introduced. As Carroll states: "This makes a complete mockery of the claim that the grammars are isomorphic. It would surely have been better to admit that their experience on Rosetta2 has shown that their various principles were no more than working hypotheses, which happened neither to work particularly well nor to be true in any legitimate sense" (p.38).

Other observers of the MT scene have made similar observations concerning the shaky relationship between linguistic theory and MT, none more outspoken than Wilks' (1989) observation that "the history of MT shows, to me at least, the truth of two (barely compatible) principles that could be put crudely as *Virtually*

any theory, no matter how silly, can be the basis of some effective MT and *Suc[c]essful MT systems rarely work with the theory they claim to"* (p.59; emphasis original).

3.3. Sublanguage

Obviously the most successful MT story of all is that of TAUM's METEO: a translation task too boring for any human doing it to last more than a few months, yet sufficiently constrained to allow an MT system to be devised which only makes mistakes when the input is ill-formed. Some research groups have looked for similarly constrained domains. Alternatively, the idea of *imposing* constraints on authors has a long history of association with MT. At the 1978 Aslib conference, Elliston (1979) showed how at Rank Xerox acceptable output could be got out of Systran by forcing technical writers to write in a style that would not catch the system out. I was bemused to see much the same experience reported again ten years later, at the same forum, but this time using Weidner's MicroCat (Pym 1990). This rather haphazard activity has fortunately been 'legitimised' by its association with research in the field of LSP, and the word 'sublanguage' is starting to be widely used in MT circles (e.g. Kosaka *et al.* 1988). In fact, I see this as a positive move, as long as 'sublanguage' is not just used as a convenient term to camouflage the same old MT design, but with simplified grammar and a reduced lexicon.

Studies of sublanguage (e.g. Kittredge & Lehrberger 1982) remind us that the topic is much more complex than that: should a sublanguage be defined prescriptively (or even proscriptively) as in the Elliston and Pym examples, or descriptively, on the basis of some corpus judged to be a homogeneous example of the sublanguage in question? And note that even the term 'sublanguage' itself can be misleading: in most of the literature on the subject, the term is taken to mean 'special language of a particular domain' as in 'the sublanguage (of) meteorology'. Yet a more intuitive interpretation of the term, especially from the point of view of MT system designers, would be something like 'the grammar, lexicon, etc. of a particular *text-type* in a particular domain', as in 'the sublanguage of meteorological reports as given on the radio', which might share some of the lexis of, say, 'the sublanguage of scientific papers on meteorology', though clearly not (all) the grammar. By the same token, scientific papers on various subjects might share a common grammar, while differing in lexicon. Furthermore, there is the question of whether the notion of a 'core' grammar or lexicon is useful or even practical. Some of these questions are being addressed as part of one of the MT projects recently started at UMIST, in which we are trying to design an architecture for a system which interacts with various types of experts to 'generate' a sublanguage MT system: I will begin my final section with a brief description of this research.

4. Some alternative avenues of research

In this final section, I would like to mention some research projects which have come to my attention which, I think, have in common that they reject, at least partially, the orthodoxy of the 'second generation and derivative' design, or in some other way incorporate some ideas which I think significantly broaden the scope of MT research. I make only a small apology for the fact that a number of these projects are being undertaken in our own research centre!

4.1. Sublanguage plus

One of the projects currently under way at UMIST is a sublanguage MT system, the research being funded by Matsushita Electrical Industrial Co. Ltd. Our design is for a system with which individual sublanguage MT systems can be created, on the basis of a bilingual corpus of 'typical' texts. The system therefore has two components: a core MT engine, which is to a certain extent not unlike a typical second generation MT system, with explicitly separate linguistic and computational components; and a set of expert systems which interact with humans in order to extract from the corpus of texts the grammar and lexicon that the linguistic pan of the MT system will use. The expertise of the expert systems and the human users is divided between domain expertise, and linguistic expertise, corresponding to the separate domain knowledge and linguistic knowledge (i.e. of grammars, lexicons, and contrastive knowledge). Using various statistical methods (see below), the linguistic expert system will attempt to infer the grammar and lexicon of the sublanguage, on the assumption that the corpus is fully representative (and approaches closure). From our observation of other statistics-based approaches to MT, we conclude that the statistical methods need to be 'primed' with linguistic knowledge, for example

concerning the nature of linguistic categories, morphological processes and so on. We are currently investigating the extent to which this can be done without going so far as to posit a core grammar, since we are uneasy about the idea that a sublanguage be defined in terms of deviation from some standard. The system will make hypotheses about the grammar and lexicon, to be confirmed by a human user, who must clearly be a linguist rather than, say, the end-user. In the same way, the contrastive linguistic knowledge is extracted from the corpus, to be confirmed by interaction with a (probably different) human. Again, some 'priming' will almost certainly be necessary.

4.2. Automatic grammar up-dating

A research project which I find particularly appealing concerns an MT system which revises its own grammars in response to having its output postedited (Nishida et a*l.* 1988, Nishida & Takamatsu forthcoming). A common complaint from posteditors is that postediting MT output is frustrating not least because the same errors are repeated time and time again (e.g. Green 1982). The idea that such errors can somehow be corrected by feedback from posteditors is obviously one worth pursuing vigorously.

The idea, as I understand it, is roughly as follows: there is a fairly traditional second-generation type English-Japanese MT system (MAPTRAN) whose output is postedited interactively. The postediting system PECOF (PostEditor's Correction Feedback) asks posteditors to identify which of the basic postediting operations (replacement, insertion, deletion, movement and exchange) each correction involves, with optionally a reason expressed in terms of other words in the text, or some primitive linguistic features, e.g. "replace *nl* by *"new word sequence"* where *nl* conflicts with *n2* in terms of *feature"*. What PECOF does with such a correction is try to locate the linguistic rule in the MT system responsible for the error, and then to propose a revision of it (typically an extension to the general rule to cover the particular instance identified), which must be confirmed by the posteditor.

The way it locates the error is also of interest. Translation errors are assumed, given the architecture of MAPTRAN, to come from errors in analysis, lexical transfer, or structural transfer. In order to locate which of these modules is responsible, the corrected text is subjected to reverse translation back into English using an MT system which is an exact mirror image of MAPTRAN, i.e. structural transfer precedes lexical transfer. The intermediate representations at each stage are compared with the corresponding original representations, and in this way the discrepancy is highlighted. Then, the appropriate rule can be located, and amended.

I believe that research on this system is still at an early stage, and it is obvious that only a certain category of translation errors can be dealt with in this way. But it seems to be a useful way of extending the grammar and lexicon of the system to account for 'special cases' (it is arguable that actually all MT consists of is special cases!) on the basis of experience, rather than relying on linguists to somehow predict things. If PECOF can also interact with the posteditor to see how generalisable a given correction is, then this is clearly an excellent way of developing a large-scale MT system.

4.3. Different designs for different users

Boitet (1989:1) has characterised the development of MT systems as follows: early systems were for the 'watcher', providing "informative rough translations of large amounts of unrestricted texts for the end user"; with the spread of the idea of postediting, systems could be said to be for the 'revisor'. When interactive systems became available, they could be described as being for the 'translator'. Most recently, systems for the 'writer' are emerging (see below). What I like about this classification is the recognition that there are different MT systems for different users, and I feel that this observation should inform future research projects in a basic way. Both systems for the revisor, and for the translator, in the above classification, still make the assumption that MT systems are essentially for bilingual users, an assumption that was almost unassailable throughout the 1970s and early 1980s. At UMIST we challenged that assumption (Johnson & Whitelock 1987) with the design of ENtran, a system for use by a monolingual writer of the source language, while a parallel project at Sheffield University investigated the possibility of an MT system for target-language monolinguals (see Knowles *et al.* 1989), adding 'MT for the reader' to Boitet's classification.

In general, proposals for new MT projects should *start* from a review of users' needs, with a wide variety of potential users defining a wide variety of types of MT system. This attitude is reflected in a recent UMIST proposal for a Malaysian National MT project (Somers & McNaught 1990): in Malaysia, the national language academy (Dewan Bahasa dan Pustaka, DBF) is undertaking the task of translating from English into Bahasa Malaysia (BM) huge quantities of scientific textbooks, in connection with their policy of replacing English with BM as the language of instruction. DBF seeks a solution using MT: for the given language pair, no suitable system exists (this is mainly because the heterogeneous nature of the texts rules out the use of systems designed with a specific domain in mind). An additional problem faces DBF: there is a shortage of trained translators; furthermore, because until recently the language of higher education was English, many writers find it difficult to write about technical subjects in BM without undue influence from English grammar and style (or other languages in some cases, notably Arabic in the field of theology). These factors suggest that an MT-based solution will require new system designs built up from existing BM computational linguistic tools. We have proposed what we call 'interactive *user-driven* MT', which is HAMT (human-assisted MT) where the user is a domain expert (perhaps a teacher) with reasonable but not native-speaker fluency in the source language.

The proposed system design is a development of the English-BM translator workstation SISKEP (Tong 1987), which combines source and target language dictionaries and thesauri, bilingual dictionaries, and morphological processing in a workstation environment. In our proposal, translation will take place as a sort of cut-and-paste activity, where the user selects with a mouse portions of the source text from one window, which s/he then 'pastes' into the target window in the appropriate place. But in the pasting process the text portion is input to a translation module, which will attempt a translation, with interactive disambiguation and/or choice of target lexical item in pop-up windows, if necessary. The idea is that it is the user who controls the translation process, rather than the system as in existing HAMT systems: with practise, the user will quickly be able to identify the size and type of text portion the system can translate easily (without excessive interaction) and correctly. The user determines the order in which the items are processed, and builds up the target text in this way. Note also that the underlying MT system is fairly restricted: since full-sentence MT is not necessarily envisaged, except in the case of simple sentences, the coverage of the grammar of the system need not be very broad or complicated. Correspondingly, the number of alternative (and often incorrect) translations offered will be greatly reduced. The problem of undue influence from English in the target text can be addressed by the incorporation of an 'interference style checker', i.e. style-checking software which is specially primed to look for and react to errors arising from foreign-language interference. This is an idea which has been pioneered by Lexpertise (a software company based in Neuchâtel, Switzerland) for whom we have been working on a version for Japanese users writing English (Holden & Somers 1989).

The overall design of the MT system proposed here builds on the strengths of state-of-the-art MT (i.e. good translation at the word and clause level, with quality deteriorating at sentence and paragraph level), and, we hope, divides the translation task between the user and the system in an appropriate manner.

4.4. Dialogue MT

Earlier I mentioned in Boitet's classification of MT systems the idea of 'MT for the writer'. A recent research direction to emerge is an MT system aimed at a user who is the original author of a text to be composed in a foreign language. We already mentioned the ENtran project, which embeds this idea in a fairly standard interactive MT environment, where interaction with the machine is aimed at disambiguating the input text. An alternative scenario is one where the interaction takes place *before* text is input, in the form of a dialogue between the system and the user, in which the text to be translated is worked out, taking into account the user's communicative goals and the system's translation ability.

The idea of automatic composition of foreign language texts was suggested by Saito & Tomita (1986), and is the basis of work done at UMIST for British Telecom (Jones & Tsujii 1990). In this system, the user collaborates with the machine to produce high-quality 'translations' of business correspondence on the basis of pretranslated fragments of stereotypical texts with slots in them

which are filled in by interaction. The advantage is that the system only translates what it 'knows' it can translate accurately, with the result that the system shows what MT *can* do, rather than what it cannot, as in traditional MT. Obviously though this strength is also a weakness in the sense of the severe limitation on what the system can be used for.

However, we can extend the idea to make it more flexible, and conceive of a system which has more scope concerning the range of things it can translate, with corresponding degrees of confidence about translation quality. This is the case in our recently started dialogue MT system (Somers *et al.* 1990) which we are working on in collaboration with the Japanese ATR research organisation: we are constructing a system which will act as a bilingual intermediary for the user in a dialogue with a conference office, where the user wants to get information about a forthcoming conference. It is thus a 'dialogue MT system' both in the sense that it enters into a dialogue with the user about the translation (cf. Boitet 1989), *and* in that the object of the translation is the user's contribution to a dialogue. Dialogue is a particularly good example of the problem, inherent in MT, that the translation of the text depends to a greater or lesser extent on the surrounding context (Tsujii & Nagao 1988). In other words, the source text alone does not carry sufficient information to ensure a good translation. We envisage a sort of MT 'expert system' which can play the role of an 'intelligent secretary with knowledge of the foreign language', gathering the information necessary to formulate the target text by asking the user questions, pushing the user towards a formulation of the 'source' text that the system can be confident of translating correctly, on the basis of some existing partial 'model translations' which have been supplied by a human expert beforehand.

The fact that the object of translation is also part of a dialogue (with another user) adds another dimension of complexity to the project described in Somers *et al.* (1990), but the idea of dialogue MT in general is an interesting development away from the current situation where the MT system makes the best of what it is given (and cannot really be sure whether or not its translation is good) towards a situation where quality can be assured by the fact that the system knows what it can do and will steer the user to the safe ground within those limitations.

4.5. Statistics-based MT

I would like to end approximately where I began, with the statistics-based approach to MT of Brown *et al.* (1988a,b). As I mentioned at the start, this work received a somewhat hostile reception at the previous conference in this series. Nevertheless, I think it deserves to be taken seriously as an avenue for MT research, and will make some comments here on the work.

As far as I understand it, the IBM researchers, encouraged by the success of statistics-based approaches to speech recognition and parsing, decided to apply similar methods to translation. Taking a huge corpus of bilingual text available in machine-readable form (3 million sentences selected from the Canadian *Hansard),* the probability that any one word in a sentence in one language corresponds to 0, 1 or 2 words in the translation is calculated. The glossary of word equivalences so established consists of lists of translation possibilities for every word, each with a corresponding probability. For example, *the* translates as *le* with a probability of .610, as *la* with probability .178, and so on. These probabilities can be combined in various ways, and the highest scoring combination will determine the words which will make up the target text. An algorithm to get the target words in the right order is now needed. This can be calculated using rather well-known statistical methods for measuring the probabilities of word-pairs, -triples, etc.

The results of this experiment are certainly interesting. Translations which were either the same as or preserved the meaning of the official translations were achieved in about 48% of the cases. Although at first glance this level of success would not seem to make this method viable as it stands, it is to be noted that not many commercial MT systems achieve a significantly better quality. More interesting is to consider the near-miss cases in the IBM experiment: incorrect translations were often the result of the fact that the system contains no linguistic 'knowledge' at all (and indeed this was one of the main criticisms at the 1988 conference). Brown *et al.* (1988a:11) admit that serious problems arise when the translation of one word depends on the translation of others, and suggest (p.12) that some simple morphological and/or syntactic analysis, also based on probabilistic methods, would

greatly improve the quality of the translation.

As part of the sublanguage MT research project at UMIST, mentioned above, we are investigating the possibility of using statistical methods to derive morphological and syntactic grammars from bilingual corpora. That the text-type is restricted should permit us to work with lower thresholds of statistical significance, and hence smaller corpora. We also intend to prime the system with a certain amount of *a priori* linguistic knowledge of a very basic kind, for example, what sort of morphological processes are likely (e.g. for English mainly non-agglutinative suffixes; stems should contain a vowel and be longer than their affixes), typical characteristics of phrase structure (notions of open- and closed-class words, headedness, and so on). We are resisting the idea of priming the system with a core grammar, but recognise that this may turn out to be a necessary step.

5. Conclusions

In this paper I have given a rather personal view of current research in MT. Of course there are probably numerous research projects that I have omitted to mention, generally because I have not been able to get information about them, or simply because they have not come to my notice. I am conscious that some readers of this paper will be relative newcomers to the field, and so I should stress that my coverage of the subject here has been from my own viewpoint rather than as a neutral reporter. For those readers, let me end by indicating some possible future sources of information on MT research.

Conference series such as this one and, to a lesser extent *Coling* and meetings of the ACL, provide a much-appreciated forum for publicising on-going research. There are typically one or two MT-related events in Japan each year, which generally repay the effort (and expense) of attending; the annual Aslib *Translating and the Computer* conference in London typically views MT from the translator's point of view. Finally, I should also mention two publications: the first is the bi-monthly Proceedings of the Information Processing Society of Japan, Natural Language Special Interest Group, which regularly contains information about MT research in Japan. Unfortunately you have to be in the fortunate position of either being able to read Japanese, or, like me, have a colleague (or a spouse) who can read and summarise the articles for you! The second publication is the journal *Machine Translation* is edited by Sergei Nirenburg and published by Kluwer which, as its name suggests, carries articles of direct relevance.

References

Yukiko Sasaki ALAM. A Lexical-Functional approach to Japanese for the purpose of Machine Translation. *Computers and Translation* 1 (1986), 199-214.

Valerio ALLEGRANZA, Erich STEINER & Steven KRAUWER (eds.) *Machine Translation* Special Issue on Eurotra (forthcoming).

Lisette APPELO, Carol FELLINGER & Jan LANDSBERGEN. Subgrammars, rule classes and control in the Rosetta translation system. Third Conference of the European Chapter of the Association for Computational Linguistics (Copenhagen, April 1987), Proceedings. 118-133.

István BATORI & Heinz J. WEBER (Hgg.) *Neue Ansätze in Maschineller Sprachübersetzung: Wissensrepräsentation und Textbezug* (Sprache und Information 13), Tübingen (1986): Niemeyer Verlag.

John L. BEAVEN & Pete WHTTELOCK. Machine Translation using isomorphic UCGs. In Vargha (1988), 32-35.

Ch[ristian] BOITET. Speech synthesis and dialogue based Machine Translation. ATR Symposium on Basic Research for Telephone Interpretation, Kyoto, Japan, December 1989. Proceedings. 6-5-1-9.

Christian BOITET & Nikolai NEDOBEJKINE. Recent developments in Russian-French Machine Translation at Grenoble. *Linguistics* 19 (1981), 199-271.

Peter F. BROWN, John COCKE, Stephen A. DELLA PIETRA, Vincent J. DELLA PIETRA, Fredrick JELINEK, Robert L. MERCER & Paul S. ROOSSIN. A statistical approach to French/English translation. Proceedings, Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, June 12-14, 1988, Carnegie Mellon University, Pittsburgh, Pennsylvania. (Page numbers not integrated). [a]

P[eter F]. BROWN, J. COCKE, S. DELLA PIETRA, V. DELLA PIETRA, F. JELINEK, R. MERCER & P. ROOSSIN. A statistical approach to language translation. In Vargha (1988), 71-76.

Jaime G. CARBONELL & Masaru TOMTTA. Knowledge-based Machine Translation, the CMU approach. In Nirenburg (1987), 68-89.

Jeremy J. CARROLL. Graph grammars: an approach to transfer-based M.T. exemplified by a Turkish-English system. PhD Thesis, Centre for Computational Linguistics, UMIST, Manchester, 1989.

Jacques DURAND, Paul BENNETT, Valerio ALLEGRANZA, Frank van EYNDE, Lee HUMPHREYS, Paul SCHMIDT & Erich STEINER. The Eurotra linguistic specifications: an overview. *Machine Translation* Special Issue on Eurotra (forthcoming).

J.S.G. ELLISTON. Computer-aided translation: a business viewpoint. In Barbara M. Snell (ed.) *Translating and the Computer*, Amsterdam (1979): North-Holland. 149-158.

Frank van EYNDE. The analysis of tense and aspect in Eurotra. In Vargha (1988), 699-704.

Roy GREEN. The MT errors which cause most trouble to posteditors. In Veronica Lawson (ed.) *Practical Experience of Machine Translation*, Amsterdam (1982): North-Holland. 101-104.

Christa HAUENSCHILD. KIT/NASEV oder die Problematik des Transfers bei der maschinellen Übersetzung. In Bátori & Weber (1986), 167-195.

Natsuko HOLDEN & Harold SOMERS. Interference software for Japanese writers of English: Feasibility study. CCL/UMIST Report No. LX89-1, Centre for Computational Linguistics, UMIST, Manchester, October 1989.

Xiuming HUANG. Semantic analysis in XTRA, an English-Chinese Machine Translation system. *Computers and Translation 3* (1988), 101-120.

W. J[ohn] HUTCHINS. *Machine Translation: Past, present, future*. Chichester (1986): Ellis Horwood.

W. John HUTCHINS. Recent developments in Machine Translation. In Maxwell *et al.* (1988), 7-62.

Pierre ISABELLE, Marc DYMETMAN & Elliott MACKLOVTTCH. CRITTER: a translation system for agricultural market reports. In Vargha (1988), 261-266.

JEIDA (Japan Electronic Industry Development Association). A Japanese view of Machine Translation in light of the considerations and recommendations reported by ALPAC, U.S.A. Tokyo, July 1989.

R[oderick] L. JOHNSON. Parsing - an MT perspective. In Karen Sparck Jones & Yorick Wilks (eds.) *Automatic Natural Language Parsing*, Chichester (1983): Ellis Horwood. 32-38.

Roderick L. JOHNSON & Peter WHTTELOCK. Machine Translation as an expert task. In Nirenburg (1987), 136-144.

D. JONES & J. TSUJII. High quality machine-driven text translation. Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Austin, Texas, June 1990.

Ronald M. KAPLAN, Klaus NETTER, Jürgen WEDERKIND & Annie ZAENEN. Translation by structural correspondences. Fourth Conference of the European Chapter of the Association for Computational Linguistics, Manchester, 1989. Proceedings, 272-281.

R. KITTREDGE & J. LEHRBERGER (eds.) *Sublanguage: Studies of language in restricted semantic domains.* Berlin (1982): de Gruyter.

F. KNOWLES, G. JELINEK & M. McGee WOOD. Alvey Project. In M. Nagao (ed.) *Machine Translation Summit*, Tokyo (1989): Ohmsha. 45-49.

Bieke van der KORST. Functional Grammar and Machine Translation. In John H. Connolly & Simon C. Dik (eds.) *Functional Grammar and the computer,* Dordrecht (1989): Foris. 289-316.

Michiko KOSAKA, Virginia TELLER & Ralph GRISHMAN. A sublanguage approach to Japanese-English Machine Translation. In Maxwell *et al.* (1988), 109-120.

Ikuo KUDO & Hirosato NOMURA. Lexical-functional transfer a transfer framework in a Machine Translation system based on LFG. 11th International Conference on Computational Linguistics, Proceedings of Coling '86, Bonn, 1986. 112-114.

Jan LANDSBERGEN. Isomorphic grammars and their use in the ROSETTA translation system. In Margaret King (ed.) *Machine Translation Today: the state of the art*, Edinburgh (1987): Edinburgh University Press. 351-372. [a]

Jan LANDSBERGEN. Montague grammar and Machine Translation. In P. Whitelock, M.M. Wood, H.L. Somers, R. Johnson & P. Bennett (eds.) *Linguistic Theory and Computer Applications*, London (1987): Academic Press. 113-147. [b]

Dan MAXWELL, Klaus SCHUBERT & Toon WITKAM (eds.) *New Directions in Machine Translation* (Distributed Language Translation 4), Dordrecht, 1988: Foris.

Michael C. MCCORD. Design of LMT: a Prolog-based Machine Translation system. *Computational Linguistics* 15 (1989), 33-52.

David D. MCDONALD. Natural language generation: complexities and techniques. In Nirenburg (1987), 192-224.

Makoto NAGAO. *Machine Translation: How far can it go?* Oxford (1989): Oxford University press; transl. by Norman D. Cook of *Kikai Hon'yaku wa doko made kano ka,* Tokyo (1986): Iwanami Shoten.

Sergei NIRENBURG (ed.) *Machine Translation: Theoretical and methodological issues,* Cambridge (1987): Cambridge University Press.

Sergei NIRENBURG. Knowledge-based Machine Translation. *Machine Translation 4* (1989), 5-24.

Sergei NIRENBURG & Jaime CARBONELL. Integrating discourse pragmatics and propositional knowledge for multilingual natural language processing. *Computers and Translation 2* (1987), 105-116.

Sergei NIRENBURG & Lori LEVIN. Knowledge representation support. *Machine Translation* 4 (1989), 25-52.

Fujio NISHIDA, Shinobu TAKAMATSU, Tadaaki TANI & Tsunehisa DOI. Feedback of correcting information in postediting to a Machine Translation system. In Vargha (1988), 476-481.

Fujio NISHIDA & Shinobu TAKAMATSU. Automated procedures for the improvement of a Machine Translation system by feedback from postediting. Forthcoming in *Machine Translation.*

P.J. PYM. Pre-editing and the use of simplified writing for MT: an engineer's experience of operating an MT system. In Pamela Mayorcas (ed.) *Translating and the Computer 10: The translation environment 10 years on,* London (1990): Aslib. 80-96.

Anthony RAW, Bart VANDECAPELLE & Frank van EYNDE. Eurotra: an overview. *Interface* 3 (1988), 5-32.

Anthony RAW, Frank van EYNDE, Pius ten HACKEN, Heleen HOEKSTRA & Bart VANDECAPELLE. An introduction to the Eurotra Machine Translation system. Working Papers in Natural Language Processing 1, TAAL Technologie, Utrecht & Katholieke Universiteit Leuven, 1989.

Christian ROHRER. Maschinelle Übersetzung mit Unifikationsgrammatiken. In Bátori & Weber (1986), 75-99.

C.J. RUPP. Situation Semantics and Machine Translation. Fourth Conference of the European Chapter of the Association for Computational Linguistics, Manchester, 1989. Proceedings, 308-318.

Hiroaki SAITO & Masaru TOMITA. On automatic composition of stereotypic documents in foreign languages. Presented at 1st International Conference on Applications of Artificial Intelligence to Engineering Problems, Southampton, April 1986. Research Report CMU-CS-86-107, Department of Computer Science, Carnegie-Mellon University.

Jonathan SLOCUM. A survey of Machine Translation: its history, current status, and future prospects. *Computational Linguistics* 11 (1985), 1-17.

Harold [L.] SOMERS. Where will MT be in the next 20 years - honestly? (Position paper for Panel Session "Paradigms for MT" Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages, Carnegie Mellon University, Pittsburgh PA, June 12-14, 1988. CCL/UMIST Report No. 88/5, Centre for Computational Linguistics, UMIST, Manchester, May/September 1988.

Harold [L.] SOMERS & John MCNAUGHT. Proposal for a Malaysian national MT project. CCL/UMIST Report No. MT90-2, Centre for Computational Linguistics, UMIST, Manchester, February 1990.

Harold L. SOMERS, Jun-ichi TSUJII & Danny JONES. Machine Translation without a source text. 13th International Conference on Computational Linguistics (Coling 90), Helsinki, August 1990.

Loong-Cheong TONG. The engineering of a Translator Workstation. *Computers and Translation 2* (1987), 263-273.

Sami TRABULSI. Le système SYSTRAN. In *La Traduction Assistée par Ordinateur: Perspectives technologiques, industrielles et économiques envisageables à l'horizon 1990*, Actes du Séminaire international (Paris 17-18 mars 1988) et dossiers complémentaires. Paris (1989): DAICADIF. 15-27.

Jun-ichi TSUJII & Makoto NAGAO. Dialogue translation vs. text translation - interpretation based approach. In Vargha (1988), 688-693.

Dénes VARGHA (ed.) *COLING Budapest: Proceedings of the 12th International Conference on Computational Linguistics*, Budapest (1988): John von Neumann Society for Computing Sciences.

Muriel VASCONCELLOS & Marjorie LEON. SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization. *Computational Linguistics* **11** (1985), 122-136.

Bernard VAUQUOIS. The approach of Geta to automatic translation: comparison with some other methods. Paper presented at International Symp osium on Machine Translation, Riyadh, March 1985; in Christian Boitet (ed.) *Bernard Vauquois et la TAO: Vingt-cinq ans de traduction automatique - Analectes*, Grenoble, 1988: Association Champollion, 631-686.

Bernard VAUQUOIS & Christian BOITET. Automated translation at Grenoble University. *Computational Linguistics* **11** (1985), 28-36.

VSYESOYOUZNYI CENTR PEREVODOV. *Meždunarodnyi Seminar po Mašinomu Perevodu "EVM i Perevod 89" (Tbilisi, 27 noyabra • 2 dekabrya, 1989 g.) Tezisy Dokladov*. Moskva (1989): VCP.

Yorick WILKS. Machine Translation. In Stuart C. Shapiro (ed.) *Encyclopedia of Artificial Intelligence*, New York (1987): John Wiley. 564-571.

Yorick WILKS. More advanced Machine Translation? International Forum for Translation Technology IFTT '89 "Harmonizing Human Beings and Computers in Translation", Oiso, Japan, April 1989. Manuscripts & Program, 59.

WORLD SYSTRAN CONFERENCE. *Terminologie et Traduction* 1 (1986) Numéro Spécial.

This paper first appeared in the Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language, 11-13 June 1990, published by the Linguistics Research Center, University of Texas at Austin, who have kindly agreed to allow us to reproduce it here.