[From: *Translation and Communication: Translating and the Computer 6*. Proceedings of a conference ... 20-21 November 1984, ed. Catriona Picken (London: Aslib, 1985)]

Update on machine translation

Peter Wheeler

It is usual, as the more experienced conference-goers among you will have noticed, to start one's paper with unctuous disclaimers as to any attempts at completeness, or with an apology for inadequacy, or a patently insincere appeal for comments, or something similar. Unsure which of these gambits would find most favour, I have decided to go for broke: I start with an apology, a caution, a disclaimer and an appeal.

The apology is to all those of you who opened your September issue of a well-known language journal to read the caption that 'Peter Wheeler will address the Aslib conference in November, and jolly interesting it's going to be too', at the same time admiring the photograph above the caption - the high thinker's brow, the careless grin, the devil-may-care twinkle in the piercing eyes behind the intellectual-looking spectacles - that was of course a photograph of my colleague Peter Walker, who spoke here last year.

Or rather my ex-colleague. Which brings me to the caution. When I first received the flattering invitation to address this august body, I was a pampered Eurocrat of the European Commission, and thus could presumably be relied on to survey the machine translation scene with a degree of objectivity. Since then, I have joined the ranks of sordid commerce, but I will nevertheless attempt to perform this morning's task with the same lack of commercial bias.

Thus to the disclaimer. When you are asked to give this sort of paper, a review of the entire scene, what do you do? You write, of course, to everyone in the field and simply ask them what they have been up to. Which is what I did.

Whereupon I received replies from users of MT, from research institutes, from bodies who were still looking into the question of which system to choose, if any, and so on. I also received several sacks of mail couched in terms of 'Dear Dreamboat, loved your groovy pic in that high-brow mag. Could we get together and split a few infinitives some time?'

What I didn't get was any response from several of the mainline commercial companies - or, not to be coy about it, the competition. I had even added a special paragraph to them saying 'Look, you know me, Punctilious Pete, would you buy a second-hand software package from this man? You know I wouldn't sell you down the datastream, but if you don't let me have any info, from your side then my presentation is going to be simply a paean of Logos propaganda.' Reaction, nil. Which may make my presentation of what is happening on the commercial side, all my well-intentioned protestations notwithstanding, somewhat lopsided. In the absence of any received doctrine from some of the companies, I have to feel at liberty to rely on gossip, hearsay and my own prejudices. But I see several representatives of those worthy bodies in the audience, and no doubt they will speak up if what I say is too outrageously wrong.

That leads me to the appeal. I call it my 'Yes, but...' appeal. A paper like this one, a broad sweep over everything that is happening in such a fast-moving field, cannot hope to be complete. If I don't happen to mention your favourite Urdu-to-Tamil world knowledge case frame ATN parser compiler compiler system, therefore, do please pipe up with a 'Yes, but...' at the end.

One of the first features to strike anyone looking over the past twelve months in MT is the rise to prominence of German as a source or target language. Which is not to say that as the dust settled over 'Translating and the Computer 5' the assembled manufacturers looked at one another and spontaneously hollooed 'Let's all do German, so that Wheeler has a peg to hang his paper on next year.' But nevertheless the trend is there. While the pioneers of machine translation worked almost exclusively with Russian as their source language, for obvious historical reasons, and the next phase concentrated on French - for reasons which are less obvious, but which in the case of Systran, for example, arose out of the proximity of the supposedly bilingual Canadian market to the USA - nowadays everyone is doing German.

Where do we see this? Well, we see it in Logos, for a start. Arriving on the MT scene within the past couple of years as an apparent newcomer (despite its history stretching back to 1969), Logos has already established itself

as the one to beat in German-source MT. Quite apart from its growing base of commercial and military customers in Germany and Switzerland, Logos has also succeeded in gaining a toe-hold in that bastion of Systran support, the European Commission in Luxembourg, where it is currently undergoing a six-month trial.

Logos German-English at the Commission, Systran English-German at the Commission. In accordance with its long tradition of showing the way in practical MT development, the Commission is currently building a German synthesis module to graft on to its existing and highly developed English analysis. It will be interesting to compare this embryonic project with a similar foetus from Logos. Logos English-German is already installed at three customers. Installed, however, in an experimental form, and using the expertise of these customers in their own fields of technology, plus their suggestions and requirements on the purely linguistic front, as the basis for development towards a fully mature product.

Nor can Systran German-English be overlooked, either. In a joint development project with the American Air Force and World Translation Center in La Jolla, California, the Festo Corporation, manufacturers of hydraulic and pneumatic equipment, is working on German-English; dictionaries to be built up by Festo, linguistic development work coming out of California. According to the partners in this joint development, there is still a lot of work to be done on Systran German-English because of the infinitely productive capacity of German to generate compound nouns. At least within the immediate future, Festo sees the system as useful only as a source of standardised terminology, rather than as actually producing usable translations. Interestingly, too, it sees the introduction of MT not so much as a major new step in its own right, but simply as an inevitable part of the increasing computerisation of office procedures themselves (Festo, 1984, private communication). [I love it when people say 'private communication' in a paper. It gives such an impression of mixing with the great and the famous. 'The translator workstation of the future will consist of 96 windows, each the size of a postage stamp.' (Melby, 1990, private communication). 'This paragraph has gone on far too long.' (Mrs Wheeler, Tuesday, private communication).]

Where was I? Precisely the opposite end of the same problem, the infinitely productive capacity of the English language to create noun groups, has been tackled in an interesting way at the Susy project of the University of Saarbrücken. As part of the creation of 'Susannah - Susy Anwender Nah', Saarbrücken has given birth to Betsy, a lexicon containing only complex technical noun groups, which the Susy system accesses almost as soon as processing

begins. The theory is that while the capacity for formation of noun groups such as 'radar signal processing equipment' is infinite, the components of such phrases consist of a notionally finite group of common nouns. While the target language equivalent of such expressions is searched for in Betsy, therefore, the morphology of the component parts is handled in the finite Susy dictionary of common nouns.

This has the added advantage that once a group has been recognised, homograph resolution is assisted, in that the possibilities of 'signal' being a verb or 'processing' being a gerund are excluded. Now where, I hear you cry, does all this differ from Systran LS expressions? Thus far, apparently not very much. Then come the clever bit and the surprising bit. The clever bit is that a subsequent stage of Susy can undo this recognition of groups if the resulting sentence is detected to be ungrammatical. Locating 'grease gun' in Betsy but subsequently finding Susy unable to make syntactic sense of the sentence 'grease gun and replace it in its holster', Susannah can decide that 'grease gun' isn't a noun group after all but an imperative and its object. The surprising bit, according to the proud parents, is that Susannah has never actually been obliged to change her mind in this way. Either she is a surprisingly constant young miss, surprisingly sure of the rightness of her choices, or else, I suspect, she hasn't been sufficiently exposed to ... [I'm sorry, I'll read that again] — or else she hasn't been sufficiently exposed to the real world of technical translation, with its syntactic delights such as 'check valve and seat cover' or 'light blue touchpaper'. However, Susy has been out in the big wide world. The Susannah project in fact grew out of the Susy-BSA project, a test of the applicability of the university's theories on language-processing to the demands and constraints of a real-life translation environment, namely that of the Federal German Bundessprachenamt (BSA). While it would presumably have to be said that the Susy-BSA project, viewed specifically, was a failure, in that the Bundessprachenamt found that they had no use for anything less than high quality translation - translation as the Lord intended it - it was an interesting sign of another strand in the pattern of recent events: the accelerating rapprochement between the ivory towers of academe and the commercial business of computerised translation at 55 Pfennigs a line. With Susy, for example, exposure to the hurly-burly of the real world does not stop with Susy-BSA or Susannah: Susy is also part of an operational abstracting system called CTX and a project confusingly called ITS. This is the Informative Translation System, and is intended to produce high-speed not-very-elegant German-to-English and English-to-German translations of large bodies of text such as databases.

First operational target date is early 1985. It is confusing, of course, because the old-stagers among us think that ITS stands for Interactive Translation System, a concept that was already confusing enough since it grew into both ALPS and Weidner.

As part of this same rapprochement, no university congress on MT, however academic and rarefied, can be considered complete nowadays without the presence of at least some of the commercial systems. The emphasis of this two-way traffic is definitely lopsided: it seems evident that the academics are adopting some of the pragmatic insights of the 55-Pfennig-a-liners rather than the other way round. Reading the universities' current project literature, one is struck by the frequency with which the systems have recourse to safety nets to allow a failed translation to be automatically rerun under less strict grammatical parameters. An academic adoption of the axiom long current in the commercial MT world that a translation, any translation, however flawed, is better than no translation.

I quote from Christian Boitet's excellent paper on the work currently being done in Grenoble - an archetype, if one will, of a Gallic tour d'ivoire:

> 'We don't want [our second-generation systems] to stop and produce nothing if they encounter an ill-formed clause in the unit of translation. Rather, we want them to produce the best translations they can, under all circumstances, annotating them with special marks, analogous to error messages, to be used later during postediting.'

One brings in Grenoble to counteract the impression which may have been given that everything, but everything, is happening in German. Not so; French is still in there and still active. While Grenoble continues its academic work on French-English, and German-French, and its translation into French of Russian technical abstracts - 5,000 to 7,500 running words per month (and incidentally also its work on English to Malay - laboratory prototype just over the horizon - and on English to Thai - early stages of a co-operative project with three universities in Thailand) - perhaps of most significance is the selection of the Grenoble Ariane 78 system as the basis for the French government's national machine translation project, known, to those in the know, as the Projet National.

Here's another trend. After a government report dealt a crippling blow to government sponsorship of MT research, leaving the flame to be kept alive by commercial interests, dreamers, and pure research freaks, suddenly almost 20 years later we see the emergence of not one but several

national or at least supra-commercial machine translation projects.

In France, there is the Projet National, with a staff of nine or ten from the university and a similar number from industry, with a target date of October 1985 for a demonstration system running on IBM coupled to a text-processing system on a French micro, taking as its corpus the translation of aircraft manuals for French-English and computer manuals for English-French.

In America, we have the MCC project, or Microelectronics and Computer Technology Corporation, which is a grouping together of several major hardware manufacturers such as Sperry and Texas Instruments on the one hand, and software systems houses on the other, for co-operative development of advanced systems. Within the project area of artificial intelligence, one section will be dealing with natural language processing.

In the European community, of course, there's Eurotra. The news from Eurotra is that everything is taking even longer than expected - which was itself to be expected. Completion date, for a 20,000-word prototype, is put at mid-1988. To date, only two of the ten countries of Europe have signed the contracts of association under which the work is to proceed, and while the plan approved and signed and sealed and generally applauded over by the Council of Ministers makes provision for a strong co-ordinating team of eight persons to be located in Luxembourg, the same Council of Ministers hasn't actually got round to approving the funds for these eight posts. As the Eurotra interim report to the Council drily puts it: 'the lack of a central team in Luxembourg may constitute some risk for success'. This is not to conclude, as some Jeremiahs might, that the project was doomed from the start. What should be concluded, perhaps, is that the fundamental concept - gloriously European - of research being carried out in ten different locations and all the work matching up to a defined interface and all the interfaces being pulled together in one central location, is going to be even more demanding to implement than had been anticipated. Nevertheless, work is proceeding. Considerable effort is going into making sure that all the ten teams of researchers mean the same thing by the same expression: coming from ten different academic backgrounds and traditions, one team may say 'bottom-up parser' and not understand what another means by 'data-driven parser'.

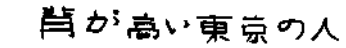
Work is continuing, too, on the specifications of the linguistic models and strategies and on the choice of a domain within which the first experimental version of Eurotra will operate. Binding specifications for the software design are being laid down, from which a contractor will work in

actually building the software itself which the linguists will then use to write their linguistic rules. Meanwhile, the experimental software assembly is now ready. The plan now is to build an experimental body of software on which preliminary linguistic work can start, without the need to wait for the production of the industrial body itself (not projected to be ready until 1986). Grammar writing on the basis of this experimental body of software will start in January 1985.

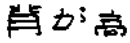
What has sparked this flurry of national projects? Naturally, someone else's national project - the Japanese Fifth Generation project. A three-year project, due to be concluded early next year, with the objectives of developing English-to-Japanese and Japanese-to-English high-quality machine translation of scientific and technical abstracts, 'with emphasis placed on accurate translation of the contents (information) of the abstract'. I am quoting here from a description of the objectives of the Fifth Generation project by Professor Nagao of Kyoto, who goes on: 'In this sense the question of proper selection of syntactical form and other linguistic features is not always the first priority. Unlike conventional university research, this project is characterised by the requirement to yield a system that is linked firmly to practical use'. A view of 'conventional university research' which is not, as I suggested earlier, any longer totally consonant with reality. The technical dictionary compiling is largely in the hands of the Japan Information Center of Science and Technology, with target size for the dictionary being projected as one million words. Development of the machine translation software and the grammars is being carried out by the ETL Electrotechnical Laboratory and by Kyoto University. At the same time as the national Fifth Generation project, however, and beavering away in parallel to it, are various projects from various commercial companies. Some of them are still in the design stage - or may even have been abandoned, the Japanese being as reluctant to talk about what they are not doing as about what they are - but several of them have come on to the market during the course of 1984.

Given the marked and well-known reticence of the Japanese, I let out a whoop of joy when I came across a translation of an article from a Japanese electronics magazine which promised to reveal all. This was what I needed for this bit of the paper. It started by listing at least ten different projects currently underway in Japan, all of them with appealing names like Lute, and Kate, and in one case a pair called Venus for Japanese-English and Trap for English-Japanese. An intriguing combination of names. And all those names which to you or me mean cameras or hi-fi are up there too: Toshiba, NEC, Hitachi. This is great stuff,

I thought, Wheeler gets through the inscrutability barrier. The more so as the article was then clearly leading towards detail of how the systems work. 'Now a typical problem with translating Japanese,' this English version went on, 'is that given the sentence:



it is impossible to tell from part-of-speech information alone whether



modifies

or

Just so. At this point I stopped reading - there is more than one way of being inscrutable!

What can one say about concrete Japanese developments, however? New from Fujitsu are the Atlas-I English-Japanese and Atlas-II Japanese-English systems, first displayed publicly late in 1984 and intended to be commercially available in the spring of 1985. These systems aim firmly at the commercial technical market of manuals, product literature and technology transfer contracts. Results are said to be '80-90 per cent satisfactory', which is one of those delightful descriptions which can mean everything or nothing. No doubt nudged by the appearance of the Fujitsu product, Hitachi have also revealed their own English-to-Japanese system, while freely admitting that it is not yet ready for marketing. At present it has a dictionary of 70,000 words, translates at 20,000 words per hour and uses something called quasiphrase structure.

Revealed within a few days of each other, both of these products were presented as Japan's first machine translation system. They may be the first MT systems actually developed in Japan, but they are at best ambiguous in overlooking the appearance several months earlier of Weidner's Japanese-English system which claims 92 per cent accuracy. Just what that figure means, if anything, can be demonstrated at the Weidner stand during the exhibition sessions. Why Weidner Japanese-English? Logical enough, as Weidner is now largely Japanese-owned.

Which brings me neatly to the next of my strands in the 1984 MT Tapestry ('You can tell this guy's written conference papers before!'). If, as has been estimated, half a million pages worldwide were translated by computer in 1984, then clearly machine translation is here to stay. But

it seems to me to be entering a phase of consolidation, of retrenchment, of looking around and wondering where to go next. The phase of its being laughed out of the boardroom is well and truly over, the phase of its being gleefully accepted as the next corporate toy is also waning, and instead it is reaching a phase where its existence as an industrial product is taken as a matter of course, to be evaluated according to the same ruthless criteria as the new office copier or next year's replacement for the salesmen's cars. This will be a phase in which the purely commercial competition is likely to hot up, and which will bring changes, regroupings and casualties. The buying-up of Weidner has already been mentioned. The once hoped-for worldwide Systran empire of harmony and co-ordination appears to be a dead dream. Faces from well-known MT companies are popping up at other well-known MT companies. One major project for a machine translation service bureau, to be run by a massive American corporation in co-operation with a very well-known MT company has been abandoned. Challenge Systems has gone bankrupt.

Contacting the bodies which I happened to know to be looking into the possibilities of MT a year ago, and asking them where they had got to, without exception I found that they are still looking. However, their reasons for hesitation were far from consistent. The commonest reason given was that the language combinations needed were not available. Amongst the combinations whose non-availability was lamented were English to Korean, English to Samoan, and English to Tongan. Market opportunity for someone there.

On a more serious note, this body, the Mormon Church, whose involvement in machine translation of course goes right back to the early days at Brigham Young University, makes the point that while they are indeed currently translating from English into 96 languages, the actual volume in any one language is very low and the quality required unusually high. They are currently putting more effort into the utilisation and development of word processors, and are looking to develop word processing in languages where it is currently not available, such as Thai, Hebrew and Persian. Other people's reasons for not yet buying an MT system were that the hardware configuration did not match what they already had, and in some cases, of course, that the quality of the output was not regarded as high enough. Others, and this is perhaps a more hopeful sign, felt that developments were happening so quickly that it was not yet the moment to make a choice. But back to gloom.

The documentation centre of the French Centre National de la Recherche Scientifique (CNRS), hailed a year or so ago as a potentially massive Systran customer, has after all carried out only a few tests with Systran English-French and

Titus French-English, for the familiar reason of budgetary restrictions. The one case of fully automatic high quality translation known in the world is fully automatic no longer. It used to be the case that the TAUM Météo System would either translate perfectly or else flag the 20 per cent of sentences that it had failed to analyse, with the human translators having such confidence in the machine that they would not bother to review the machine-generated 80 per cent. This is no longer so, and virtually all sentences are now human-checked. However, the fact that the translators now can carry out such a review is a step forward.

This list of setbacks perhaps sits uneasily on the lips of a speaker known in MT circles for radiating a boundless and essentially gormless optimism, but, as I have learned to say since starting to work for an American company, 'I tell it like it is, man'. Nor must these setbacks be seen as the whole story, or even a major part of it. The hopeful sign is that nowadays it is the tellers of the 'Spirit is willing...' joke, rather than the joke itself, who are laughed at. While some companies do occasionally lose a customer, this is no longer regarded as a portent that the beginning of the end of MT is nigh - it is just a normal commercial occurrence. With those companies and bodies who have installed machine translation early, the picture appears to be 'MT means More Translation' (he said, purloining the original title of Veronica Lawson's recent article, and coyly not saying which MT system was described in the article under that optimistic heading). The American Air Force, for example, long pioneers in Russian to English MT, are now experimenting with French-English, German-English and Japanese-English. They are also using their MT system in an experimental project on automatic abstracting of Russian articles, and in another, less purely experimental, on the automatic pre-processing of Russian patent journals ready for machine translation.

Things are moving on the Spanish front, too. General Motors of Canada, while still machine-translating large volumes of English-French car and truck manuals, has now gone into English-Spanish. At the Pan-American Health Organization (PAHO) in Washington, whose Spanam system is currently translating 80,000 words a month from Spanish to English, for a to-date total of 2,115,000 words, the Engspan system, translating in the opposite direction, is now in use on a pilot project to translate 234,000 words of text on pesticides, namely data sheets and users' manuals for distribution in Latin America. So far, including part of this project, Engspan has translated over 150,000 words for official use within PAHO. Another development reported from PAHO is that the machine and human translation departments are currently being merged into one.

Still in America, the Siemens-sponsored METAL system at the University of Austin, Texas, is expected to be in a production-ready phase by the end of this year. At present they are achieving correctness rates of up to 85 per cent at a speed of around 2 seconds per input word. With the end of their work on German-English thus in sight, the METAL team have already started work on other source and target languages. English to German is already experimental, English to Chinese is imminent, and both English and German into Spanish are projected for the near future.

And what of ALPS? ALPS has a new slogan, 'Technology for Human Translation', which in practical terms seems to mean 'Dear Dreamboat, we are sure that you have the technology for continual updating of your paper, and that you will therefore be human and understanding about it if our reply to your August request for information reaches you on November 19!' Whether or not I have the technology, they certainly do, as they have this year added full word-processing capabilities in Russian and Arabic, and keyboard facilities in several further European languages as well as, again, Russian and Arabic. Printers, too, to cover Arabic and Cyrillic scripts. On the language side, Alps has added English to Italian in CTS, the top-level interactive computer translation. One level lower, at the ADL automatic dictionary look-up level, the source languages French, German, Spanish and Russian have been added, with Italian and Dutch just around the corner. On the linguistic side, Alps is claiming significant progress in its two major releases of 1984. ALPS also claims greater simplicity in dictionary creating, with a reduction in the amount of information which the lexicographer must provide. Furthermore, CTS and ADL dictionaries are now compatible, and it has become easier to switch from one mode to another.

On the hardware side, interfaces have been developed with several word processing systems including Wang and CPT; with several typesetting systems including Penta with whom they have a joint marketing agreement; and with the ubiquitous Kurzweil OCR. Things are happening on the ALPS corporate side, with the opening of a new regional office in Switzerland and the expansion of the regional offices in the States. Earlier I referred to 'familiar faces popping up at other MT companies'. ALPS calls it 'growing in breadth with the hiring of former employees of MT companies X and Y'. X and Y may well call it something else.

ALPS replied, as I said, yesterday. But they seem indecently premature by comparison with BSO (Buro voor Systeemontwikkeling) who gave me their response in the last coffee break! However, there was a reason for the lateness - BSO was waiting for a response from a government

ministry. And the news, while late, was good. Having carried out a feasibility study on the DLT system, BSO has now heard that the system will be supported by the Dutch Ministry of Economic Affairs to the tune of £6 million over four years. Ultimately, the system will be multilingual, but in the first two-year phase it aims to translate simplified technical English into French.

Things are moving with Weidner, too. While its system was originally marketed for mainframe use, it is now available on the IBM PC, marketed in this country by the Software Connection. (California has Silicon Valley, the East Coast has Route 128, Britain has Fareham, Hampshire!) Weidner on the PC runs at upwards of 1,600 words per hour, roughly comparable with the 1,500 words per hour of a word-processor-based system such as Logos. Post-editing is reckoned to be possible at 600 to 800 words per hour and the pairs available are English to French, German, Spanish and Portuguese; and French and Spanish to English.

Reference to Weidner reminds us that 'MT means More Translators', too. At ITT Europe, for example, use of Weidner has been so successful on company-internal documentation for ITT worldwide that the translation department has expanded into taking on contract work from outside, at the same time increasing their translation staff up to about ten, with the translators each producing something like 5,000 to 6,000 words of finished text a day as against 3,000 or so before the introduction of MT. As its own brochure puts it, ITT's configuration includes printers, terminals, all that electronic stuff, plus 'a number of translator/editors with square eyes, enthusiasm and a friendly disposition towards computers'. A similar story is reported from one of Logos' customers, Eppendorf Gerätebau, which has also taken on new staff to help cope with the increased workload generated as the company became aware internally of the improved throughput of the translation department. So much for the cry that the machine would take the bread out of the mouths of starving translators - not that the cry has been heard for a couple of years now.

On the other hand, translators of English to Arabic had better watch out, because 5,000 of them are for the chop. I read it in a French magazine, so it must be true! Systran English-Arabic has arrived, in a development funded by Robinetterie Gachot S.A. of Paris. It runs at the well-known Systran speed of 300,000 words per hour, and can perform the work, said the magazine, of 5,000 human beings. One has a delightful image of all those French translators, each no doubt equipped with a quill pen, carefully writing one word per minute! More seriously, while the system is now available, Gachot is still intending to

carry out major improvements, in particular in the handling of semantic categories, and will be totally recasting the dictionary structure. They also intend to open a translation bureau in Paris in late 1984 or early 1985, offering not only English-Arabic but also the various other Systran language combinations. It is to be hoped that the long-running dispute about who has rights to what will by then have been solved.

It's almost time for the 'Yes, but...' session, in which those of you who felt your system got rather short shrift can leap up and put the record straight. Of course, the system that probably got shortest shrift of all was Logos, as I bent over backwards to keep this presentation objective. In fact, so far did I go in keeping the name Logos out of it that I am not even going to mention that we are also developing English to French!

AUTHOR

Peter Wheeler
Logos Computer Systems GmbH, Lyoner Strasse,
Frankfurt, West Germany.