[*Translating and the computer 10*. Proceedings of a conference … 10-11 November 1988, ed. Pamela Mayorcas (London: Aslib, 1990)]

Machine aids for translators: what does the future betoken?

Francis E. Knowles

Aston University, Birmingham, UK

INTRODUCTION

I was very pleased and honoured when the Conference Committee asked me to give a paper in a session devoted to the memory of Margaret Masterman. Many readers will recall with pleasure Margaret's contributions to the Translating and the Computer Conference series and the intellectual vigour and panache with which she presented them. On those occasions she stood before us as a kindred spirit and as a pioneer in the field of computers and translation, the predominant theme in the portfolio of the Cambridge Language Research Unit. It is both heartening and fitting that Bill Williams, the new Director of Research at CLRU, has pledged to continue research and development work on this and allied topics in the unit's recently reconstituted infrastructure and programme of research. I presented, alongside Bill Williams, a personal appreciation of Margaret's unique contribution to the field of language and computers at Informatics 9[1], convened by Aslib in Cambridge. What I said on that occasion can be referred to in the published proceedings and therefore does not need to be repeated *in extenso* here but I would like to take the liberty of selecting a couple of brief points from that earlier address in the hope that it can set the scene for us today.

Machine translation (MT) was a major professional preoccupation for Margaret. As a philosopher she saw, quite correctly, that the main effort had to be directed towards the development of methods for identifying meaning and safeguarding it against corruption during the transformation process, in which the signifiers are replaced but the signifieds are not. She was fascinated by this problem and we often talked

about two particular axes which researchers and implementers have to align with each other. The first is perhaps best expressed by saying that in MT, and in natural language processing (NLP) in general, what cannot be computed has to be looked up, and vice versa. In some cases both options are available, each with its own pay-off or, conversely, its own overhead; in other instances only one option is realistic: but how can this be determined and reconciled with the abiding need to maintain an identity of sense as between input and output? All the various types of meaning have to be identified and transformable in MT: denotational, connotational, collocational, stylistic, rhetorical, and, not least, syntactic. One particular type of text which fascinated Margaret was translated text: it follows that she was just as much interested in human translation (HT) as in MT. She viewed human translation as the acme of skilled linguistic activity, rating it higher – I am tempted to say – than original creative writing because translators have the constraint of needing to fully reveal the brilliance of the original author's mind while totally concealing the brilliance of their own. However, Margaret was by no means oblivious to other highly-skilled linguistic activities, such as paraphrasing, summarising, stylistic transposition, all of them backed up by vital subsystems, such as deixis, comparison, analogy, enumeration, exemplification, generalisation or ellipsis. Margaret conceded readily that present-day MT is clumsy simulation relying on sleight of hand. It is not emulation. Even simulation practised by perfect prestidigitators would not have been sufficient for her: she wanted MT – and IT (Information Technology) for that matter – to actually emulate human behaviour. Nothing less would do.

In debates about machine translation Margaret almost without exception took up the cudgels on behalf of translators, asking 'where do translators fit in to this IT paradigm?' and 'when will hardware manufacturers and software designers really start to learn something from the accumulated experience of professional translators?' It is always tempting to speculate about MT/MAT but if we do that we should put ourselves in the shoes of translators, both those who are independent (freelance) and those who are employed by organisations in either the private or public sector. What then does the translator's microcosm look like today and what does the IT revolution (rather than just computers alone) offer, either in terms of facilities already available or of those very much in prospect?

WHAT DOES THE IT REVOLUTION OFFER?

Well, it is certainly true that IT devices and resources are continuing to develop very rapidly, bringing to the marketplace facilities of increasing sophistication and utility. Probably one of the best and simplest examples

of this is fax technology. It is fast and unencumbered by serious constraints appertaining either to technology or to the information structure of transmitted data. Simple bit-mapping does the job and presents professionals such as translators with an immediately usable and useful facility.

Much the same can be said with respect to OCR (optical character recognition) although typographical variety can operate as a nuisance factor here. On the networking scene, however, problems arise, provoked by constraints on the interconnection of devices or data-binding and bundling difficulties as regards character set equivalences and mappability. This hits translators hard. How ironic this can sometimes be, given that one of the purposes underlying networks is the accessibility of public domain or subscriber-restricted facilities such as term banks, or financial and legal information, both domestic and international. One of the catchwords in word processing over the last few years has been WYSIWYG (what you see is what you get) a prime requirement for translators who are increasingly involved in document finishing and printing. There are no standards currently in force or even beginning to emerge which might represent good practice for highly-formatted documents embodying typographical variety and interspersed graphics, a perfectly normal input/output situation for the translation profession. WYSIWYG, of course, can be abandoned in favour of a document mark-up system such as SGML (standard generalised mark-up language) or FORMEX (formalised exchange) but all too often the real interest is and has to be in document specifics rather than in a generic coding scheme such as SGML. If translated documents are to be electronically archived, for instance, choices have to be made: filter into straight ASCII text, thereby abandoning macrostructural information, or preserve 'as is' and offer conversion facilities to the end-user? When was the translation profession last surveyed about this and similar matters? The mention of graphics serves to remind us of the vital role graphical information plays in professional documentation: translators so often need, and can obtain at a cost, IT systems for capturing and generating graphics. This is something of a royal road which leads all the way to the sort of DTP systems, incorporating good page make-up tools, which are now amply available on the market.

In professional contexts and in educational circles at post-secondary level it seems increasingly to be accepted that word processing facilities are insufficient on their own. They need to be supplemented and complemented by database and spreadsheet facilities – hence the rapid growth of integrated software. I submit that the same needs are in evidence in the translation environment: the internal relativities may well differ but the facilities themselves are needed. At the lowest level, spreadsheets can be used for recalculating columns of figures in a

different currency – more seriously they might be used for expressing quantities in a different system of units. In the documentation put out by automobile manufacturers, for instance, tyre pressures are expressed for English speakers in 'pound-force per square inch', but German or French speakers now expect to find them designated in 'bars'. Similarly, English 'miles per gallon' needs to become 'litres per 100 kilometres' in French or German. However, of all the add-on facilities, the database is the most important to translators. For example, personalised dictionaries are not really feasible without database facilities. Such dictionaries may have a simple two-column structure or they may be typified by much more complex data structure involving, for example, domain tags, cross-references and several languages. There are, relatively speaking, quite a lot of these systems on the market now which are devoted to technical terminologies of one sort or another and readers will undoubtedly be familiar with many of them. These systems represent in some cases, although all too rarely, a microcosmic reflection of major national facilities which exist in certain foreign countries with government finance and other support, or in supranational institutions such as the EC Commission. Unfortunately, no large-scale funding has so far been forthcoming from private industry, from the public sector, or from the research councils for a national facility in the UK.

It is also possible to purchase dictionaries in CD-ROM (Compact Disc-Read Only Memory) format. This offers some advantages: easy updating via the periodic issue of new CDs and easy navigation through the dictionary materials. However, many of the deficiencies of hand-held dictionaries and other reference books are perpetuated by the CD-ROM approach. We must hope, therefore, that flexible database management will soon be applied to materials of this sort, thereby offering us all a quantum leap in intellectual, functional and operational terms.

OTHER USEFUL/DESIRABLE FACILITIES

Of course, databases can be used for other purposes too: one potential attraction, where the pattern of translation work is repetitive or affected by rigid informational constraints, is the use of database-implemented archives to boilerplate translations (constantly recurring legal phrases, for instance, are manageable in this way). It is also time that the translation profession had a greater choice of other text handling software, in the form of utility programs to assist the translation process. A brief selection of such utilities should, I suggest, include:

1. Proper dehyphenation software to cater for those cases where the source text arrives via a network as a fully-formatted document, or where the source document is captured by OCR (optical character

recognition) on site. Note that this software is not naive since only soft hyphens must be removed and hard hyphens can, of course, occur in an end-of-line position.

2. Software to spell-check efficiently and rapidly in all major source and target languages: this software should allow a whole constellation of dictionaries to be consulted or generated. Once again, the task is not simple: proper names present a special problem, as do capitalised acronyms and even sentence initial words can play tricks because of their volatile capitalisation. As far as English is concerned, Saxon genitives and plurals simply have to be handled intelligently.

3. Software to highlight, at the inspection phase of the source text, all potential technical terms. Taking English (the awkward case) as an example, this would involve scanning the text for what have sometimes been called pentads, that is, groups of five words which conform to a certain structure. In rough and ready terms criteria such as the following would most probably be invoked:

   - no sentence boundaries
   - no major clause or phrase boundaries
   - no function words apart from 'of'
   - no verbs (?).

   A great part of the output from this utility could then be lexiconised, either permanently or temporarily. Separate, descending passes of the utility would, of course, then need to be made for tetrads, triads and dyads – the whole basis of this stratagem is the well-known axiom that the longest match offers the highest probability of correct identification.

4. Output from the previous program could then be transferred to another utility to search for, plug in (if they exist) and view translation equivalences, given that the cognitive units in the source language by definition constitute the translation units required. This would also prevent one nuisance factor from rearing its head in sentence-by-sentence processing: the same lexical unit acquiring different translations. A stricter control over terminological consistency could be guaranteed in this way and pseudo-synonymic variation avoided. The only deviancy permitted would be the replacement of a lexical item by abbreviated place-holders, such as acronyms, but even in this case orthogonality between source and target text would be preserved.

5. A further utility of some considerable size and power, in the translator's eyes, is easily constructible for the case of a one-to-

many relationship between source and target lexical items. A very simple concordance, which could fairly easily be enlarged as a by-product of the ongoing translation process, needs to be generated. For example, for each source item, in English, a batch of 100 examples of, let us say, contextually-arrayed German equivalences could be retrieved for inspection and could thus help to prime a judicious choice. It matters not whether the source lexeme is polysemous – that would be properly reflected in the display. Various other options are also available: the frequencies alone of a particular translation equivalence might be more eloquent in settling choice than the concordance itself. With only a slightly different configuration, encyclopaedic information could also be retrieved in order to assist the translation. The German *Bundesforschungsministerium* (Federal Research Ministry) has no infrastructural equivalent in the UK and it could be pertinent to point this out.

One optional, but often very welcome opportunity in the activity of computerised translation is to collect and update word frequency information: adding to global frequency lists is relatively easy and useful; enhancing special-subject glossaries or other thematic holdings in this way is less easy but proportionately much more useful.

The online consultation of domain thesauri is another highly desirable facility: translators need access to various technical terminologies presented in an onomasiological fashion, reflecting real-life linkages rather than the vagaries inherent in alphabetically-organised compendia. Alphabetic entry into such subject microcosms is, of course, a welcome convenience but the fundamental value of such thesauri is to verify knowledge structures and context. The ideal case, of course, is a fully (about 95 per cent) co-ordinated multilingual reference system, encyclopaedically arranged and incorporating copious illustration. Unfortunately, such works are not yet really available – even in traditional printing – either as professional or pedagogical compendia. What little does exist offers precious little in terms of amenability for translating purposes. Only computerisation can release the full power of such a concept and translators must encourage such developments and, preferably, participate in them.

Everything that I have said so far has implied a working context involving machine aids to translation. This is, after all, the topic of this paper! A strong implication, however, has been made *vis à vis* hardware even though my explicit remarks have concentrated on software. Let us now, for the sake of complementarity, if not completeness, dwell for a few moments on the *desiderata* for hardware and how the various factors can be reconciled without becoming compromised. In doing this we begin to form a mind's eye view of something that might well be called the

'translator's workstation': let us give a thumbnail sketch of this workstation's functionality. We must be talking about a 32-bit processor with at least 1Mb of RAM (Random Access Memory), backed up by 100Mb of spinning store itself backed up by tape-streamed archives. The screen should be high-resolution colour and of A3 size so as, firstly, to permit a good synoptic view of enlarged text and graphics and, secondly, to subtend the multiple-window facility of the highly-versatile operating systems which are normally used on modern workstations. The operating system's functionality is sometimes referred to, counter-intuitively, as WIMP: which stands for 'windows, instructions, menus and pointers'! The UNIX operating system, now available on a number of machines, is often held to be the acme of desirability and functionality: it certainly incorporates many facilities which translators would find irresistibly useful, such as pattern-matching facilities and a transparent programming facility called AWK to call on if need be. Integrated operations or even the integratability of operations is facilitated by the filters and pipes approach which can designate the output of one program to be the input of the next.

In this account, so far, mention has been made of a number of facilities which translators either have welcomed or could be expected to welcome in due course. Most of these facilities, however, are aimed at removing logistic constraints from the backs of translators and on to the machine. Ever since the advent of computers this has been an important aim often, however, subtly phrased in terms of 'removing drudgery and giving massive scope for creativity and intellectual effort!' The bottom line, however, has always been that personal productivity shows a marked improvement: in other words, output rises. Whatever the area of work, the involvement of computers has had this effect and in most cases only this effect. The search for computational methods of achieving qualitatively new insights and working methods has, as we all know, been going on for many years and it has not yet brought any radical change of paradigm or obvious fundamental success in pragmatic terms. Working practices have been affected commensurately as a better division of labour between person and machine becomes visible or visibly necessary. All sorts of attempts have been made, and with some success, to delimit the MT environment, to bypass the pseudo problems and to expose the true nature of the outstanding task. If the institution of practices such as pre-editing, inter-editing (in MAT) and post-editing has achieved anything, it has reconfirmed in an oblique way that human beings can and normally need to save the machine from disaster: in this way therefore the two distinct 'parties' may work in concert with a reasonable amount of success. *Suum cuique* is the phrase which encapsulates this wisdom in Latin.

THE TECHNOLOGICAL AND COMMERCIAL ENVIRONMENT

In the environment within which we now live, factors are at work with which we need to come to terms. Firstly, the elaboration of MT/MAT software is a vastly expensive business which may, in the future, net a gigantic return for investors' or taxpayers' money: this is not happening at the moment. In the meantime, organisations in the field naturally feel a compulsion to keep a strict proprietary control over developments such as they are; moreover they still appear compulsively to exaggerate claims about the functionality of nascent or *beta-test* software. Furthermore, and for less accountable reasons, they often keep their primary clientele, the translators, at bay to the extent that they give the impression of being impervious to the accumulated experience of and genuinely charitable advice and help from translators. The development of MT/MAT software is tantamount to a juggernaut's foray, lurching from one financial deadline to another and demanding many sacrifices of people as well as money. The principal reason for this is that, even 40 years on from the start of the race, we are still dealing with something that is essentially research. MT/MAT has not yet reached the stage 'development' of being largely based on known 'research' results. Even R (research) plus D (development) is insufficient to enter the marketplace with: commercialisation must follow and this process tends to cost an order of magnitude more than the original R & D. The R & D process should result in a prototype which can demonstrate that satisfactory, comprehensive and feasible intellectual answers have been found to the particular R & D problem tackled.

This is the trap into which many MT/MAT projects seem to fall: even welcome projects such as EUROTRA are apparently prone to this. I quote from a recent issue of the European Commission *Bulletin*.[3] 'The aim of the seven-year EUROTRA programme (1983-89) is to create a prototype translation system of advanced design, working with a limited vocabulary and limited text types, that can handle all the official languages of the EC. It will form the basis for the development of a working system that will be of major benefit to the institutions of the Community and to industry. Following a preparatory phase (1983-84) the current (research) phase (1985-87) is concerned with the development of a small-scale translation system between all languages covering a vocabulary of 2,500 words. A third (development) phase (1988-89) is planned to extend this small system to a precompetitive prototype system with a vocabulary of 20,000 words. There are also plans for industrial implementation after 1989. The EC has contributed 27 million ECUs to the current budget of 45 million ECUs. It pays the cost of the central team in Luxembourg and shares the cost of the work undertaken in member states.'

I certainly found it worrying to hear, in a recent public lecture, a EUROTRA team member glossing over arguably significant changes in the project's original targets and chronological anchor points. To be told, in addition, that Eurotra has actually turned out to be a way of establishing a European computational linguistics community, and that we ought to be as satisfied with this return on investment as the European Commission apparently is, was quite disconcerting! There should be no need for such casuistry but it always emerges when there is a mismatch of expectations in the triangle of contractor, supplier and customer. The last thing we need now is another ALPAC (Automatic Language Processing Advisory Committee) type report – the British climate is ripe for such reports, the European climate less so, fortunately. What we do need, however, is less politicking, more *glasnost* from all parties and more honesty and forthrightness about the intellectual problems which still need to be solved in MT/MAT. Only one act of faith and bravery is needed: to conduct a sustained programme of research in the total awareness that it is essentially a feasibility study.

Only bureaucratic error or political disingenuousness could lay down in advance, in 'blue skies' research, that a research phase *will* be complete in two years. Only an honest assessment of work achieved – rather than 'milestones' specified *a priori* – can determine progress and the managerial options surrounding it. This is just as true for the venture capitalists who have shown a lot of interest in MT/MAT in the United States, as it is for the European Commission – or indeed, for the Alvey Directorate and its successor. If a large R & D effort is to be distributed across several centres then a proper assessment of the contingent financial, logistic and 'personal chemistry' overheads should be conducted. If it is proposed to add languages to developing systems, what evaluation of start-up effort is even attempted? Is it easier to lexicographically codify the first 5,000 words or the last five per cent in a restricted vocabulary? Cannot the aggregation – even at a seemingly glacial pace – of lexicographical material provoke new insights into lexical structurality? The point of these largely rhetorical questions is really to suggest that developing a multilingual, multi-functional system such as EUROTRA is really akin to NASA trying to launch a space shot to Mars. The true nature of the complexity is not known beforehand and those funding it have to accept that quite openly at the beginning. At that moment the system is amorphous and elastic – given good management the system, that is the people involved, can develop a more purposeful and goal-seeking style, leading, sporadically, to really useful results.

Hopefully, during the final decade of the 20th century, we will learn which way the EC chose to play it or found itself playing it. There will be as much of value in that account about management and leadership as about the conquest of the unknown. It will also be interesting to see what

sort of competition is provided by the private sector, what scoops they can carry off. One thing is certain: whatever their ecology the researchers will not be loners but members of closely-knit teams. The model which they develop may well rub off from the providers of MT/MAT software to those who use it: after all, lots of customisation will be involved, lots of machine dictionaries will need to be constructed, lots of market and product research will be needed, lots of financial shrewdness will be required to cope with shortening timescales for equipment depreciation, with continually enhanceable software and with add-on language pairs. It is to be hoped that a coalescence of interest, firstly among providers, and subsequently extending into the customer base for MT/MAT, will make it possible to integrate effort, avoid duplication, set pragmatic but professional standards and serve to create a climate in which MT/MAT research and development will be seen by all parties not as expenditure but as an indispensable investment.

CONCLUSION

Who, however, can say with any confidence today what the inside story of MT/MAT software will turn out to be? What is likely to be the ultimate basis of disambiguating text by machine for translating purposes? Is it going to be achieved by a symbol-processing or by a connectionist approach? Will it involve codified world, or at least microcosmic, knowledge, or will it rely on minute details of linguistic mechanisms, on fine juxtapositional analyses of segments of discourse? I am not even sure which of these approaches should be called the direct method, and which the oblique method! The former would in a way be a victory for language universals, the latter a re-affirmation of the variety and pluralism in human languages which engenders the need for both MT and HT! We must, clearly, reconcile ourselves to the need for a lot more research into this question and that research must rely on a very pragmatic philosophy. De Beaugrande's eloquent exposition of the inadequacy of modern linguistic theory to define what translation is and, *a fortiori*, what machine translation might be, force us more into the business of searching for new methods.[2] These methods will be based almost exclusively on contextually-dependent factors and on environmental analyses of textual segments which can be agglomerated in order to produce computer-driven synopses of messages the length of a paragraph or greater – and ensure that we leave behind the highly-damaging restriction of the sentence-by-sentence approach which is still (pre)dominant in MT/MAT R & D.

Only context and co-text, to use Werlich's terminology[5], can guide us on the trickiest problem of all: disambiguation. De Beaugrande's formulation runs thus: 'to "understand" a word, sentence, or any other

linguistic unit is to limit the number of things it can be significant in relation to, within the current context'. The process of understanding, by humans, appears to consist of three phases, according to Kintsch[4] and his colleagues: phase one is the sense activation phase, during which all lexical or encyclopaedic information is activated, whatever the particular context of its trigger; the sense selection phase follows, during which knowledge of the particular context permits, forces even, the human processor to accept or discard discrete packets of this internally-generated information (this is the crux of understanding); the third phase, the sense elaboration phase, merely consolidates (Kintsch's word is 'enriches') the process by pursuing it to its conclusion. It is clear where the focus of MT/MAT effort should continue to be placed and it is equally clear that experimental work of no mean order needs to be conducted on the middle phase of this three-phase process. Until and unless the whole process can be successfully modelled by machines acting alone or in concert with humans, venture capital will be required rather than gilt-edged finance and the stoic patience and forbearance shown over many years towards MT/MAT by its well wishers will need to be redoubled yet again!

REFERENCES

1. Jones, K.P. (ed.) *Meaning: the frontier of informatics.* Proceedings of Informatics 9. London: Aslib, 1987.
2. De Beaugrande, R. Translation as text processing. *ALSED/LSP Newsletter*, 10(1)1987, 2-22.
3. European Commission Bulletin. L222/1 88/445/CEC.
4. Kintsch, W. and Mross, E. Context effects in word identification. *Journal of memory and language*, 24, 1985, 336-349.
5. Werlich, E. *A textual grammar of English.* Heidelberg: Quelle/Meyer, 1982.

AUTHOR

Professor F. E. Knowles, Aston University, Modern Languages Department, Aston Triangle, Birmingham, B4 7ET.