[From: *Georgetown University Round Table on Languages and Linguistics 1989*, James E.Alatis, ed.]

Practice makes less imperfect: Users' needs and their influence on machine translation development

Veronica Lawson
London, England

Abstract. The problems which attract machine translation developers are not necessarily those that loom large in practice. Only by exercising a system on large quantities of normal text, as at Georgetown in the 1950s, can it be made to reflect the needs of customers.

After a brief historical summary, the relationship between practice and development is explored with reference to 'teaching', testing and technology. Ways are described in which systems were improved to take account of users' needs. The evaluation of machine translation is discussed, as is technological progress in the field.

> Practice is the best of all instructors.
> *Publius Syrus, ca. 42 B.C.*
> And practice drives me mad.
> *Elizabethan MS., 1570*

These two quotations in a sense sum up this paper on the relationship between practice and development in machine translation. Specifically, it deals with the needs of customers, and how those needs are reflected in systems. For this Georgetown University Round Table it is a particularly appropriate subject, for two reasons. Firstly, the Round Table is celebrating thirty-five years of machine translation. As you know, the first ever demonstration of machine translation on an electronic computer was the Georgetown/IBM experiment in 1954. The approach to machine translation here was pragmatic, firmly rooted in real text translated for real users. Georgetown dealt with the truly natural language found in 'translation as she is paid for', rather than the examples that occur to researchers. Not many people will pay for translations of *Time flies like an arrow* or even *The cat sat on the mat.*

Secondly, this Round Table is concerned with 'Teaching, Testing, and Technology'. All of these have been crucial in machine translation (MT): teaching because a computer must be 'taught' about language; testing because this is a major factor in MT's development and use; and technology because it not only permitted MT in the first place, but has shaped the course it has followed ever since.

History. Machine translation (MT) is translation generated by a computer, with or without human assistance (usually with it). In the mid

1950s, when machine translation began, the computer was still an intriguing new tool. It had mastered numbers with intoxicating speed, and researchers expected it to master words with something of the same alacrity. They were of course wrong, and in 1965 their innocent optimism gave way to disillusion when a committee set up by the U.S. National Academy of Sciences advised against further research into machine translation. This was the Automatic Language Processing Advisory Committee, whose famous report (ALPAC 1966) drastically cut government funding for MT research, not only in the United States but to a considerable extent elsewhere in the world. The study now appears questionable in a number of respects, as we shall see later.

Machine translation's second decade, therefore, was quiet. A few MT teams survived in the United States, the Soviet Union, France, and Germany, but they were usually small and on a low budget. The world at large, however, was changing. There was the continuing 'information explosion' (the great expansion of scientific and technical information which had begun in the 1950s). There were also large-scale increases in trade and international cooperation, and a rise in linguistic nationalism in bi- and multilingual regions. All combined to produce a 'translation explosion'. Some government institutions have used MT extensively since before 1970.

From 1975, therefore, in the third decade of MT, there was a renaissance of machine translation, first in Canada and Europe, then in Japan. Quality has improved considerably, and now even the United States government is investing in MT again. Whereas thirty-five years ago the pressure for MT had come from researchers with an intoxicating new tool, the driving force since the mid 1970s has been the users' need for faster and ever more translation. (For a scholarly and readable history, see Hutchins 1986.)

Users' needs affect MT systems in small ways as well as large. This paper sets out a few of the changes which have resulted. Some examples come from my own development work, others from users' experience of various practical MT systems. Since the changes affect teaching, testing, and technology, it may be useful to look at them under those headings.

Teaching. First, teaching: we must 'teach' language to the machine, as far as we can when our own knowledge is but sketchy, and a truly MT-oriented linguistics still eludes us. Linguistic software is powerful, and MT dictionaries include far more kinds of information than dictionaries for human use. A good machine translation system is therefore an Expert System, incorporating the linguistic insights of both its developers and the practitioners who exercise it.

If a machine translation system is to cope with the demands of practice, it is essential to do as Georgetown did in the 1950s: process large quantities of real text, and address the problems thrown up. The more the system aspires to be what my typology classifies as a 'try anything' system (Lawson 1982), the greater the quantity and variety of text must be.

In such a system the rules 'taught' to the software and dictionaries must be as general as possible, to be applicable to many sorts of text. My first MT study was a feasibility study on the machine-translatability of patents, performed in 1979/80 for the Commission of the European Communities on their English-French and French-English Systran systems, then young. One of my most surprising discoveries in that study was that there was no combined index of all the Systran dictionaries. I actually had to recommend that one be created for the development team. The Systran dictionaries were both a strength and a weakness: a strength, because they were powerful and flexible enough to deal with most of the difficulties which arise in natural language; but a weakness, because they were insufficiently classified and tended to grow excessively and in too random a manner. We went to great lengths to contribute dictionary entries as 'universal' as possible, but the lack of a combined index made it difficult and sometimes impossible to find groups of related expressions and substitute a more general rule, and to discover some sources of error.

The problems which loom large in practice are often not those anticipated by the developers. My patent study got off to an inauspicious start when the translations would not run at all. After six months it was found that the handling program regarded 'unusual' characters and improbably long sentences as errors, and that it rejected entirely any corpus containing ten or more errors. Sentences of more than 105 words were regarded as improbable. Unfortunately, sentences in patents and other legal documents frequently run to half a page, sometimes to a page or even more. Similarly, 'unusual' characters included percent signs and mathematical symbols, very common in patents and other technical and even commercial documents. The maximum sentence length was therefore increased to 255, and 'unusual' characters were made acceptable.

Some of the more obvious problems—problems that one might feel any MT system must be able to solve-may be best left to the posteditor. Questions, for example, may be processed very badly in an MT system, and so researchers may spend much effort on them. However, they are surprisingly uncommon in many translated documents, and entirely absent from patents and some other text types. There is a limit too to the effort worth expending on the exceptionally complex area of the article, about which long books are written, and (short) wars fought: the excuse for the Six Day War between Israel and Egypt was the discrepancy between *territories* and *les territoires* in the parallel texts of a treaty. It is safer and more economic for the posteditor to check and amend articles than for the developer to deal with them exhaustively.

Numerals, on the other hand, are often troublesome, especially where mixed with text. Dates, for example, tend not to be addressed adequately, if at all, by developers before their system is implemented in a translation service. They then discover that dates are found frequently and in notable variety: *March 10* or *10th, 1989; 10* or *10th March 1989; 3.10.89* or *03.10.89,* or the European variants with transposed day and month; and so on. The Logos MT system will translate American-style *3.10.89* to European *10.3.89;* this can backfire, however, when a 'date' is detected in error.

Once early in the life of the European Communities' English-French Systran system, *U.S. Patent 1234 567* became *1 Brevet des Etats-Unis 234 567.* The translators, and hence the system, had had instructions to expand the premodifier *U.S.;* and the patent number in the source text contained spaces instead of commas to mark out groups of three digits. Usually, punctuation must be preserved, but sometimes, to convey the same information, it must be translated into the punctuation appropriate to the target language. Here French uses spaces; and at that time the commas in English numbers had proved so problematic that they were replaced by spaces when the source text was keyed in, as it then was—a rare and probably short-lived example of the MT tail wagging the language dog. At the time the problem was solved by translating *U.S. Patent* as *Brevet U.S.*

Like a space or a comma, a period, of course, can fulfill various functions. The software may need to interpret it differently according to whether, instead of marking the end of a sentence, it signals a decimal point (in which case it may need translation to a comma), an abbreviation, ellipsis (if one of three dots), another mathematical symbol, etc. If the software breaks up a sentence by mistake at an abbreviation, for example, it will produce nonsense. At least one fledgling system even used periods to mark the ends of lines. This could make for difficulties when the line did not contain a sentence and the software therefore could not find a verb form. My favorite MT error was in the address on a letter (Morgan-Girard, p.c. 1981):

ORIGINAL John Smith.
Managing Director.
RAW MT John Smith.
Directeur se debrouillant.

A director managing, or coping. Normally, the software translated the expression *managing director* correctly. On that occasion it missed it, looked for a verb, and found a job description.

Such howlers are fun, but it should be stressed that they are not the normal run of MT. Much more typical is the following, from a random European Community document which I fed to Systran:

ORIGINAL The problems we are to consider are difficult
ones and will not be easily resolved.
RAW MT Les problèmes que nous devons considérer
sont des difficiles et ne seront pas facilement
résolus.

A major problem faced by MT developers when they 'go public' is the preservation of format, which is often a vital component in the transfer of meaning. My first MT study ten years ago was criticized for regarding format and punctuation as part of language, but within a couple of years the critic was saying the same. Format is, I believe, a visual representation of the underlying structure of a text. It helps to make logical connections explicit. Thus we tend to slow down when faced with a text which is presented sentence by sentence instead in paragraphs (e.g. when source and target texts are displayed side by side); and many of us would prefer wordprocessor screens to be full page size. To tamper with format is unwise. Not only may it delay the reader's understanding. Worse, it may interfere with analysis, particularly when text is arranged in columns. Finally, it may seriously inconvenience the customer if, as in the computer manuals which are classic subjects for machine translation, the text is to be printed with illustrations.

Another area of some importance is the postediting of inflections. It may be the work of a moment to change an ending, but the moments soon mount up, affecting the economics and acceptability of MT. One text type for which MT is suitable is minutes of meetings, but French and English write these in different tenses—past in English, present in French—and changing all the verbs is extremely time-consuming. The European Commission has therefore developed an algorithm to change the tenses of minutes automatically between these two languages. It works for almost all tenses, and for all common ones. (It also influences vocabulary, so that, for example, *chair* comes out as *présidence* instead of *chaise.)*

Changes of synonym can also mean extra work for the posteditor. If, in a translation into an inflected language such as French, a noun is changed to a synonym of a different gender, endings in associated adjectives and verbs must also be changed. If the synonym has been offered by ALPnet's interactive MT software (which asks the editor to choose between alternatives), the software will correct the endings. Other systems too may generate the correct inflections automatically once a noun has been changed.

Testing. Testing has determined the level of support for machine translation, whether it is funding for research or implementation in the field. Most spectacular, of course, was the ALPAC report, now increasingly recognized as of dubious quality. In particular, the ALPAC committee assumed that machine translation must not be edited, although much human translation is edited as a matter of course. What it rejected was therefore only Fully Automatic High-Quality Translation (FAHQT), and not, as is often thought, machine translation in general.

To take only three of the other defects, the test passages consisted of sentences taken from six translations (three human and three machine) and jumbled at random, destroying cohesion; the sentences were judged in isolation; and half of the evaluators were Harvard undergraduates instead of real translation users. The ignorance of discourse was general at the time, but the committee's ignorance of translation was less excusable, for it consulted translators little.

In fact, the evaluation of MT is not an easy matter. The Commission of the European Communities, concerned to evaluate its own systems, gathered experts from many countries for a symposium on evaluation in 1978, but no consensus was reached (Van Slype 1979). The Commission's own evaluations had tried numerous criteria (Van Slype 1980). Only two had shown any correlation: postediting ratio (the proportion of words amended) and intelligibility.

My patent feasibility study in 1979/80 was therefore to use these two criteria. Neither, however, was satisfactory. Patents, notoriously, are often imperfectly intelligible in the original. Translations of them may therefore be less than intelligible and yet still accurate, desired, and consequently useful. The postediting ratio was not entirely reliable or suitable, particularly at that early stage in development, and was therefore replaced by an accuracy evaluation.

From my experience of examining professional translators, I felt that the evaluation of translation was inevitably subjective; and that it was better to acknowledge this subjectivity and concentrate on reducing it. One way to reduce it is to specify the use for which the translation is required. This, as it were, gives a restricted definition of *translation* for the particular evaluation concerned. I therefore asked my evaluators to assume that the text was wanted only for scientific or technical information.

In addition, a further criterion of 'usefulness' evolved. Patents can be translated for several different purposes and to very different standards. Evaluators were asked whether the MT was suitable for other uses than for information.

The usefulness criterion as applied in that study was somewhat crude, but correlated well with accuracy. However, 'usefulness' is probably somewhat too weak a criterion. An evaluator's surprise that the machine can produce anything helpful may make him overenthusiastic, and a stricter criterion such as 'usability' should be applied if the public is not, as in the past, to be disappointed at an early stage.

At the request of the Commission, evaluators were also asked, 'Would the text be useful for postediting?' However, one can postedit to any standard, and the question is unanswerable unless it specifies the purpose for which the postedited translation is required. One evaluator actually suggested that the answer could be obtained by means of the formula:

L x 2
2
where L = length of piece of string.

As a translator, I perceived the standard of the MT as very low. My evaluators, however, were translation users (patent attorneys or research chemists), and were more lenient. In a sense both user and translator are right, for users know exactly what they want from a translation, whereas the translator must try to anticipate the needs of numerous, unknown users by supplying a foolproof and 'multivalent' translation. Some users, to save time or money, can accept lower standards than a translator dare supply. Certainly the goal that ALPAC insisted on but rejected--the combination of 'fully automatic' and 'high quality'—is now seen as unrealistic for the present, except perhaps by inexperienced researchers. Instead, human assistance and/or low quality (by our standards) are not only expected, but acceptable, such are the speed and volume which MT can offer.

Even now there is no widespread agreement on how to test. Developers tend to evolve their own mixture of methods. Vasconcellos (1988), after discussing the ALPAC report in detail and surveying other approaches, recommends a mixture of formal and functional criteria.

Perhaps it should be emphasized that MT varies immensely in quality. Firstly, raw MT varies with the inherent suitability of the text for MT, the similarity of the source and target languages, the quality of the system, its experience with the domain and text type, and consequently the size, relevance and depth of coding of its dictionaries. Secondly, postediting varies from rapid postediting of only the most glaring errors to full postediting, possibly to a standard not perceptibly lower than human translation. I once read a whole paper in Georgetown's *Jerome Quarterly* without recognizing that it was a postedited machine translation (Santangelo 1986).

Even raw machine translation may be suitable for some users. Motivated subject specialists can decode a message even if half of it is lost. A public notice in Cornwall had lost more than half of its characters:

A
AID A D
DI A ED?

I understood it at once. So would you, if you saw it while walking out of a parking lot-provided that you were familiar with our 'pay and display" parking: an open car park, where you buy a ticket from a machine and display it on your car. As a motivated subject specialist, I knew instantly what the notice said.

HAVE YOU
PAID AND
DISPLAYED?

I bought a ticket. We are glad of the redundancy in language when we are not paying full attention, but we do not need it when we are properly motivated-and I did not want to pay a fine.

Technology. Progress in technology and machine translation go hand in hand. The new tool which had brought MT into being was at first slow and crude. Text input was on punched cards, painfully tedious and so expensive in the United States that the cards were punched in Germany and flown across the Atlantic.

Users pushed continually for improvements in quality, speed, and cost. Gradually, matters advanced. Computers became faster, with much larger memories, so that MT dictionaries could be larger and more powerful. Translation was now quicker and better, but input and pre- and postediting were still difficult. Postediting was performed by making handwritten amendments on the printed translation--another very tedious task, which produced a dog's dinner for the typist or customer.

Then came the wordprocessing revolution. Not only are pre- and postediting now far easier, but many source documents are prepared with wordprocessing and can therefore be input without being rekeyed. This cuts both the cost and the time involved in processing a translation. According to one international organization, any human intervention adds at least three days to turnaround time. MT services increasingly, therefore, insist on source texts being submitted in machine-readable form.

I could see from the start that, because typographical and grammatical errors in the source text interfered with analysis, MT needed spell checkers. Soon I heard of the first report-writing program, and knew an automatic pre-editor would follow. This was to be what we now know as text critiquing software. Then the US Air Force developed a semiautomatic posteditor.

Nor were these the only developments to speed the cleaning up of input and output. A more recent tool, the optical scanner, has made another dramatic difference. At the Nuclear Research Center in Karlsruhe, in Germany, a 500-page translation which could take a month to input now takes only a short time. The translation itself takes only minutes to run, and is then ready: their scientists use raw MT, and so there is no delay for editing.

The next breakthroughs, we are told, will come from artificial intelligence and parallel processing. To date AI has been applied only on a very small scale. MT systems are very large, and it may prove prohibitively expensive to incorporate sufficient AI in them to make a major difference. The greater speed and power offered by parallel processing are attractive and will be useful, but will do little to help with the central task, the analysis of language.

For that, there is no substitute for the process of testing the machine on large quantities of real text, seeing where it fails, and teaching it to do better. That process is a fascinating one. For just as the damaged brain tells us about the healthy brain, the failures of the computer to process language tell us about language. The path pioneered by Georgetown thirty-five years ago is still the way forward.

Practice is the best of all instructors.

References

ALPAC. 1966. Languages and machines: Computers in translation and linguistics; A report by the Automatic Language Processing Advisory Committee. National Research Council Publication 1416. Division of Behavioral Sciences, National Academy of Sciences, National Research Council, Washington, D.C.

Hutchins, W. John. 1986. Machine translation: Past, present, future. Chichester Ellis Horwood, distr. Wiley.

Lawson, Veronica. 1980. Final report on EEC study contract TH-21. Feasibility study on the applicability of the Systran system of computer-aided translation to patent texts. CETIL/205/80. Luxembourg: Commission of the European Communities.

Lawson, Veronica. 1982. Machine translation and people. Practical experience of machine translation. Proceedings of a conference, London, 5/6 November 1981, ed. Veronica Lawson. Amsterdam: North-Holland. 3-9.

Morgan-Girard, Danièle. 1981. Personal communication.

Santangelo, Susana. 1986. Machine translation: Where to? / Traducción automática: Rumbos. Jerome Quarterly 1.4:4-5.

Van Slype, Georges. 1979. Critical study of methods for evaluating the quality of machine translation. Final Report. Brussels, Luxembourg: Commission of the European Communities.

Van Slype, Georges. 1980. Systran: Evaluation of the 1978 version of the Systran English-French system of the Commission of the European Communities. Incorporated Linguist 18.3:6-69.

Vasconcellos, Muriel. 1988. Factors in the evaluation of MT: Formal vs. functional approaches. In: Technology as translation strategy, ed. Muriel Vasconcellos. Binghamton: SUNY. 203-13.