

PERSPECTIVES ON THE ASSESSMENT OF MACHINE-TRANSLATED OUTPUT

Muriel Vasconcellos
Pan American Health Organization
Washington, D.C. 20037

Home address:
1739-1/2 Corcoran Street, N.W.
Washington, D.C. 20009

Daytime telephone: (202) 861-4338
Home telephone: (202) 667-7781
Facsimile: (202) 223-5971

FIT Miscellany on Translation Criticism, ed. Milan Hrala.
Amsterdam: John Benjamins. In press as of June 1989.

PERSPECTIVES ON THE ASSESSMENT OF MACHINE-TRANSLATED OUTPUT

Muriel Vasconcellos

1. Overview

Few if any of the usual measurements of translation quality are meaningful when it comes to the output of a machine. A machine translation (MT) is in reality no more than the tracks made by whirring cogs and spinning wheels. It is not a direct product of human creativity, and it does not owe its existence to principles of cooperative communication between people.¹ For these reasons alone, there is no point in subjecting it to the approaches used for evaluating human translation.

On the other hand, criteria do exist for assessing MT. They are concerned above all with what use can be made of the machine's output. This judgment is closely linked to the "competence" of the system that generated it. The question of usefulness can be viewed from several perspectives. In terms of the product per se, we must know, first of all, that it is reproducible. Once this is established, the main issue can be addressed, namely serviceability. With regard to the system that generated the output, we are concerned about its capacity for improvement and about its extensibility to the translation of similar texts, different but related texts, and other applications entirely. We also need to know about its compatibility with the particular setting envisioned, as well as the relative ease with which the different supporting tasks can be performed. These are not watertight categories, however. We cannot judge output from any of the perspectives mentioned without looking at the capabilities of the system as a whole, and we cannot know how well a system works without reference to its output. The divisions that follow,

therefore, are somewhat artificial. They are merely intended to point up the angles from which the evaluation task can be viewed.

2. The MT Product

2.1 Reproducibility

In the evaluation of MT output, as with any assessment of a mechanical process, reproducibility is an essential condition. The entire evaluation effort will be for naught unless we have the assurance that the object of our study can be reproduced by the same system a second time. Only when this is established are we in a position to make judgments.

Sometimes the quality of MT output is controlled, either directly or indirectly, by the person who is demonstrating the system. Direct control involves introducing changes manually in the completed computer translation. The indirect approaches to controlling MT output--"overprotective pre-editing," "overcoding," "undercoding," and "overprotective demo text selection" (Bédard 1988)--involve strategies applied specifically to a particular translation that will show it off to best advantage. The output, though reproducible in the strict sense, is of little meaning for purposes of evaluation, because there is no basis for knowing how effective the MT system will be when it is extended to other texts, even very similar ones (see Section 3.2 below).

2.2 Serviceability

Once the MT output has been determined to be reproducible and therefore a legitimate object of study, the next question to be addressed

is how serviceable it is for its intended purpose. Depending on the application, it may already be acceptable in its raw state; it may be used nearly raw with only a very light review; or it may serve as a draft to be more extensively postedited by a translator or a subject specialist. In the second and third cases we want to know how much intervention will be required and whether or not this intervention can be accomplished efficiently in the particular circumstances. In such an evaluation, "errors" as such are not the issue.² The definition of "error" will vary depending on the purpose of the application and the values of the user community. What is important is how easily the output can be fixed up to meet the standard that will be applied to it.

2.2.1 Raw output

Totally raw MT output is mainly useful for "gisting," or getting a general idea about the original text. Up to now there have not been too many MT applications in which raw MT is delivered directly as the final product. One of the problems has been that the source material, given its random nature, is seldom available in machine-readable form. The input texts have to be re-keyed, often following special conventions for the foreign-language characters, before they can be presented to the computer for translation, and this greatly detracts from the cost-effectiveness of MT.

The direct use of raw MT was the subject of the often-cited study undertaken by the Automatic Language Processing Advisory Committee (ALPAC) under a 1964 mandate from the U.S. National Research Council. One of the main questions addressed by ALPAC was whether it was possible to achieve

a level of "fully automatic high quality translation" (FAHQT) that did not, or in the future would not, require any human intervention at all. The conclusion reached was that MT was impossibly far from attaining such a goal.³ For this and other reasons, the Committee recommended in its final report (ALPAC 1966) that research on MT be abandoned.

Despite the negative findings of ALPAC, information scanning remained an important priority for the U.S. Government. With the door closed to public funding, the U.S. Air Force turned to the private sector and bought into the development of Russian-English MT produced by Systran, the first of the commercial MT ventures. The use of this system for mainstream translation at the Air Force's Foreign Technology Division (FTD) in Dayton, Ohio, is described in 2.2.2 below. Apart from this activity, since 1987 the FTD has been producing raw translations on-line from Russian, German, and French into English for information specialists who access the service via 1,400 networked PCs. The on-line facility is recommended mostly for tables of contents, titles, keywords, abstracts, and isolated paragraphs. The rate of consultation has been rising steadily,⁴ which is the ultimate test of serviceability.

In addition to this type of application, in which input text is keyed in and submitted from the individual workstation, increasing consideration is being given to harnessing MT to data bases that are already in machine-readable form and producing raw output for direct consumption. Traditionally the contents of such information banks have been largely in English. Today, however, scientific and technical information is being published in some 70 languages,⁵ and whole data bases are being devel-

oped in languages other than English--e.g. French, German, Japanese. Although there has always been some interest in the translation of information from data bases, now, with the availability of material in many different languages and the greater need for access thereto, the demand is being felt more acutely. As we see the confluence of several trends--quantum expansion of computer technology, the proliferation and widespread use of data bases, the growing demand for information translated into languages other than English and vice versa, and steady improvement in the technology of machine translation--it can be expected that MT will be increasingly used to tap into information sources of this kind.

With the emergence of such applications it is important to be able to evaluate the direct serviceability of raw MT output. The best way of doing this is to gather reactions from actual or potential users. A specific study on the acceptability of raw MT was conducted by Newman (1988), who questioned 58 scientists at Sandia National Laboratories on their reactions to Systran output from German into English. The 41 respondents, confronted with raw output for the first time, unanimously agreed that it was acceptable as a screening mechanism--i.e. for the identification of texts to be translated more carefully. Only seven of the 41 thought that it was unacceptable for information purposes, "no matter what the savings in cost or delivery time," although many added that postediting would be essential for "crucial information" (181). Asked to rate the translation on a scale of 1 (poor) to 10 (excellent), 60% of them gave the machine a score of 5, 6, or 7, and the average and median grades were 5. Newman speculates that the differences in the

ratings may be attributed to such factors as familiarity with the subject field, knowledge of the source language, and expectations about quality. Another factor is that over time the reader becomes accustomed to MT style and overlooks problems of expression which in the beginning seemed daunting.

2.2.2 Nearly raw output

"Nearly raw" MT output will have been scanned and lightly post-edited by a professional translator or subject-field expert. In terms of purpose of the translation, there are two basic applications for such a product. The first is similar to the "gisting" mentioned above. This is the case of MT at the Translation Directorate in the Air Force FTD, where full-length articles and books have been being processed with only a light review since 1969. In assessing the suitability of the raw product for this type of application, the evaluator needs to ask such questions as: Does the output present problems for comprehension? What are the risks of misinterpretation? Are there too many source words not being found in the dictionaries? In an application of this kind, not-found words should represent less than 1% of the output.

At the FTD the review of Russian-English is semi-automatic. The output is passed through a postprocessor, called EDITSYS, which identifies seven types of target errors and brings them to the translator's attention by means of a flashing line across the screen (Bostad 1987). EDITSYS picks up: (1) not-found words, for the translator to provide the English equivalent; (2) acronyms, for the expansion to be checked; (3) rearrangement, for adjustments to be made (typically one sentence in 10);

(4) slashed entries; (5) flawed input erroneously matched to valid dictionary entries; (6) uncertainties not resolvable by homograph routines; and (7) "problem words." Corrections, if needed, are introduced manually. The work of EDITSYS affects about 20% of the output (Bostad 1987:438 and p.c. June 1989). While automatic postediting reduces the cost of MT, presumably it is only used when the raw output can be counted on to be comprehensible and free of misconstructions. There should be a reasonable expectation that the 80% of output not affected by EDITSYS will be of a quality that is acceptable.

An MT system used for the purpose information-gathering will be challenged by texts from a wide range of sources, earning it the name "try-anything" system (Lawson 1982). To be general in its coverage, it has to have very large and deeply coded dictionary files. This condition is met, for example, by the Russian-English installation at FTD: the dictionaries, built up over a period of 20 years, by now have some 365,000 entries, comprising 210,000 stems and 155,000 expressions (Bostad p.c. June 1989).

The second type of MT application that is effective with only minor postediting is the opposite, in scope, of the type just described. The domain is highly restricted, and the system is said to be specialized. Either because the input is self-limited by the subject matter itself or because it has purposely been constrained prior to processing, the output can be relied on to be consistently problem-free. In fact, it is well known to students of MT that there is "an inverse relation between coverage and quality" (Sigurd & Werngren 1989). The quality of MT

output can only be guaranteed when the input texts do not offer any challenges beyond the grammar and lexicon that have been specified. Probably the best example of a specialized system is METEO, which since 1977 has been translating Canadian weather forecasts from English into French (Chandioux 1988), and now since October 1988 from French into English as well. The output of METEO 2 is highly acceptable for the purpose intended; postediting affects only about 3% of the words (Chandioux p.c. 1988).

2.2.3 Postedited output

For the other more typical kinds of translation--the run-of-the-mill "translation as she is paid for" (Lawson 1989)--the evaluator's attention should be focused less on serviceability of the output in its raw form and more on its manageability as the object of human interventions designed to make it faithful to the original text, grammatical, and readable. These interventions occur most often in the form of postediting. But postediting is not the only way of improving the quality of output. Changes can also be introduced either before the translation is submitted to the computer (e.g. Smart 1988) or on-line while it is being processed (e.g. Tomita 1986, Weaver 1988). Known respectively as pre- and interactive editing, these alternatives should not be ruled out. However, only rarely do they obviate the need for postediting. Usually two passes are required, one before and one after. For this reason the upstream approaches tend to be more cost-effective when the original text is being translated into multiple targets.

Strategies for efficient and effective postediting have been

described elsewhere (Löffler-Laurian 1986, Vasconcellos 1986, 1987a, 1987b, 1989, McElhaney and Vasconcellos 1988). Ideally, the posteditor works on-screen taking maximum advantage of word-processing features such as selective and global replacement, as well as macros for frequently used operations (Vasconcellos 1987b, Kingscott 1988, Datta 1989). Working on-screen is without question the fastest mode (Vasconcellos 1987b).

There are other ways in which time can be saved. For example, SPANAM and ENGSPAN reduce the time spent on terminological research by flagging in the output those terms that are considered to be reliable (Vasconcellos and León 1988).

There are also some linguistic strategies that are time-savers. To the extent possible, work should proceed steadily from left to right with a minimum of rearrangement. Reordering of the output is discouraged not only because it slows down the process but also, and more important, because it breaks up the information structure--i.e. the order in which the elements of information were presented in the original text (Clark and Haviland 1977, Vasconcellos 1985, 1986). Where the grammatical patterns of the two languages are different, it is often possible to alter the syntactic function of a given information element--e.g. change a verb to a noun--in a way that will keep the original concepts in approximately the same order. For example:

- (1-S) En el proceso de promoción y aplicación de políticas nacionales en investigación, se publicaron y distribuyeron ampliamente los documentos y conclusiones de la Primera Conferencia Panamericana sobre Políticas en Salud, que en esencia señalan los criterios para diseñar o redefinir políticas en este campo.

- (1-MT) In the process of promotion and application of national policies on research, **there** were **published** and **distributed** widely the documents and conclusions of the First Pan American Conference on Policies in Health, which in essence point out the criteria for designing or redefining policies in this field.
- (1-PE) In the process of promotion and application of national policies on research, **steps** were **taken to** publish and distribute widely the documents and conclusions of the First Pan American Conference on Policies in Health, which in essence point out the criteria for designing or redefining policies in this field.

This approach can be coupled with linguistic strategies for making the output more cohesive (Vasconcellos 1989). These include reference, substitution/ellipsis, lexical cohesion, and conjunction (Halliday and Hasan 1976).

It may in fact be possible to establish two levels of postediting in addition to the light review described under 2.2.2 above, with standards for each, so that for certain types of applications there is less investment of time and effort (Wagner 1985, Löffler-Laurian 1986, Vasconcellos 1989).

2.2.4 A typology of postediting corrections

Once the output has been postedited, the changes that were made can be examined and classified. It is preferable to initiate the analysis at this point rather than provide raters with pre-set categories of problems to look for. The most realistic results are obtained when the postediting has been done in an operational situation by a professional translator or reviser, following which the linguist-evaluator undertakes the analysis.

The experience of reviewing postedited Spanish-English output produced by SPANAM (Vasconcellos 1986, 1989), as well as that of direct postediting, suggests that the corrections made reflect five basic types of deficiencies in the output: (1) dictionary errors (e.g. source words not found, target entries missing), (2) syntactic problems, (3) semantic problems, (4) need for improved cohesion, and (5) need for improved coherence.

"Syntactic problems" may be considered to include any difficulties that can be addressed by applying syntactic rules, including subcategorization. "Semantic problems" have to do with the semantic field(s) of a specific lexical item and also with the requirements of verbs considered as states, processes, and actions. "Need for improved cohesion" refers to the linking relations between elements manifest in the surface structure of the discourse, calling for the introduction of cohesive strategies. Many of these strategies are formulable, or at least learnable by the posteditor. "Need for coherence," on the other hand, involves the interpretation of connectedness in the underlying text (Vasconcellos 1989). In all these areas, problems can sometimes be solved by substituting a word or a phrase, but it is important not to conflate the different types of problems under the catchall heading of "lexical." Use of this category obscures both the nature of the difficulties and the motivation behind the solutions.

The advantage of a three-tiered classification of postediting is that it correlates, up to a point, with the different applications for which MT can be used. "Gisting" would be expected to require a bare

minimum of changes unless the reviewer had access to a semi-automatic postprocessor such as EDITSYS. In a regular translation service, the typical first-pass postedit would deal not only with dictionary shortcomings but also with syntactic problems, semantic problems, and cohesion. The full-scale postedit would introduce interpretations for coherence (Vasconcellos 1989). Posteditors of technical texts can be trained to reserve their interpretations as much as possible for the most demanding level of quality, and in this way a certain amount of time can be saved.

2.2.5 Benchmarking

When it comes to judging output from different systems on a competitive basis, evaluators should be always mindful that the primary concern is serviceability for the intended application. There are a few simple ground rules that are important to follow.

The first rule is that the input text should always be supplied by the prospective user. Machine translations of texts from other domains are not relevant, and in fact they can be seriously misleading. Thus, near-perfect raw weather reports from METEO 2 would in no way predict the system's performance in a different subject area and discourse genre--for example, manuals on the maintenance of copier machines. Similarly, if problems are encountered in output from a general system being tested on random text, it is still possible that the same system could be tailored to handle a specific domain quite respectably. On the other hand, if the output in the random demonstration comes close to meeting the user's standards, this level of performance is significant

because it indicates that the translation program and the dictionaries have already been developed in the domain of interest, and that therefore the new user will have less of an initial investment to make.

The second rule is that the output should be generated in the presence of the evaluator.

Finally, the prospective user should go through the exercise of actually postediting the different outputs. It is not sufficient to merely look at them; they need to be road-tested. It could be that what seems like a gross error is in fact correctable with a single stroke of a macro (e.g. the case of changing its to of their), whereas the task of fixing up an awkward construction that is not too offensive may turn out to be unjustifiably time-consuming. The following sentence, for example, would be passable for some purposes, fixable with minor alterations for others, or subject to a major overhaul in the case of a publication:

(2-S) En los últimos años, la estructura de la población se ha visto continuamente afectada por el aumento de personas de edad avanzada.

(2-MT) In recent years, the population structure has been seen continuously affected by the increase of persons of advanced age.

The value of the output will depend on the user's level of tolerance.

Benton (1989) used an eminently effective approach to benchmarking in a study conducted for McGraw-Hill Information Management Company. Each of several competing MT suppliers was provided with English word lists from a set of eight technical articles to be translated into Spanish and/or French, plus excerpts from the articles themselves. Benton acknowledges that this approach provided the MT suppliers with a word list in advance of the test but not a phrase list. However, since the

suppliers had also been provided with excerpts, they could prepare their systems to handle these excerpts, at least, in the best possible manner.

The competitors were given several weeks in which to prepare for the test. At the end of that period, Benton and his colleagues traveled to the demonstration site with disks containing the full texts of the eight articles, and the translations were performed on the spot. The results were immediately transferred to disk and handed to Benton, who in turn passed them on for evaluation by the prospective end-users.

The evaluators were translators and would-be posteditors in the relevant subject fields, and they were the final judges of whether or not the output would be helpful to them in their production process. For each of the eight articles they first ranked the competing outputs in ranked order of preference. After this was done, they identified the candidates they felt would be adequate for postediting. They then marked up each of the usable outputs with the changes they considered necessary. They coded their changes according such categories as "missing term," "incorrect term," "incorrect word order," etc. Finally, the marked-up and categorized outputs were tabulated, and the results gave a profile of the kind of output that was considered acceptable.

2.2.6 Final observations on the assessment of MT output

Of course there are no definitive criteria for assessing a translation. The discussion above has not addressed the problem of how to test for conceptual fidelity, or how to determine whether or not the original author's desired response has in fact been elicited in the

receptor (Nida and Taber 1974). Rather, it has emphasized that the worth of MT output is measured in terms of the amount of effort that needs to be invested relative to the application for which it is intended.

Often the purpose of an evaluation is to make judgments about the usefulness of a given MT system. While the quality of MT output can sometimes be predictive of future products generated by the system in question, on the other hand a particular output may not necessarily be indicative of the system's potential. The perspective must shift to the system itself in order to see how easily it can be improved and extended to other domains.

3. The MT System

Up to now it has been emphasized that the defects leading to human intervention in MT output need to be weighed in terms of the time and effort they represent. Importance was also placed on the motivation for the changes as a basis for stratifying the editing task. While these criteria apply to the judgment of output as such, when it comes to testing the limits of an entire MT system, we need to look at the software elements of which it is comprised: the dictionary/ies, the grammar(s), and the computer's algorithm that acts on them. It is the performance of these elements that will indicate the investment needed in order to bring the system up to par for the application envisaged and determine its extensibility to new domains.

3.1 Improvability

We have already seen in the evaluation of MT output that "errors"

are looked at largely in terms of the effort required to fix them. While an analysis of these lapses has only marginal value for assessing the usefulness of the output, on the other hand it can be quite helpful in predicting the system's potential for improvement. By telling us something about the the relative maturity of the system and about which functions it can and cannot perform, the "error analysis" will help us to know whether in time it can be expected to produce output of better quality.

The dictionary is a very valuable part of the MT system, not to be trivialized. In hard cash, it represents, both for the potential purchaser and for the developer, at least half the investment at stake. The status of the dictionary can be detected, up to a point, through the types of problems that appear in the output.

If there are many words not being translated at all, it stands to reason that the dictionary needs massive buildup. A working system is one with a dictionary that is both large and adequately coded. Conversely, when the dictionary is small or sparsely coded the system cannot be said to be operational unless it is for a specialized purpose. For general translation the minimum dictionary size is about 25,000 stem forms.⁶ A rate of not-found words in excess of 5% (not counting repeated occurrences of the same word) is cause for concern. It can mean that the system is specialized in a domain unrelated to the test translation; the system is still quite new, possibly even released prematurely; or the supplier, as a marketing strategy, may have intentionally delivered only a small dictionary and left the tailoring to the user.

Whatever the reason for the deficiency, the need for dictionary-building cannot be dismissed lightly. It calls for a sizable investment of time and money.

On the other hand, when the words are found in the dictionary but the equivalents are inappropriate, there are various causes to be investigated, some of them minor and some quite serious.

One of the most frequent difficulties has to do with part-of-speech homographs. The Spanish word médico, for example, could function in the source text as the noun 'physician' or the adjective 'medical'. If the dictionary offers only the noun translation, the system will find the noun and might translate servicio médico as *'service physician' instead of 'medical service'. Usually the dictionary lacks entries for part-of-speech homographs simply because it has not been fully developed. Another possibility, however, is that it has been selectively and intentionally "undercoded" in order to ensure correct choices for a particular demo translation (see Bédard 1988). Or it may be that the syntactic analysis is weak. To determine the extent of the problem, one could challenge the system with a list of about 30 sentences that contain homographs. They should be common everyday words. The same homograph should be tried in its different functions. The sentence structure should be varied. If the results are reasonably successful, then the problem may not be generalized.⁷ The system is probably "improvable." For specialized and semi-specialized applications, the time and money invested in adding alternate part-of-speech translations in the new domain(s) may well be worth it.

Another frequent problem is that the gloss found by the dictionary is not appropriate for the context. For example, the Spanish word medio could be translated as 'means' 'environment' or 'medium'. What needs to be found out in this case is whether the system has adequate resources for triggering context-sensitive translations. It should offer a range of options--microglossaries, codes for selectional subcategorization, rules for discontinuous idioms, etc. Moreover, we would need to know, in the event that it does offer several such capabilities, the extent to which the dictionary has already been coded to utilize them. To start the task from scratch is an overwhelming proposition. It was seen above, for example, that Systran's general-purpose Russian-English dictionary has a total of 155,000 idiomatic entries, which corresponds to three for every four stems in the dictionary. These entries can be very complicated to code, and they may be three to five times more costly to add than univocal entries. The scope of the job will depend on how generalized the system is expected to be. For a specialized system, again the effort is probably worthwhile.

If the system does not offer different types of context-sensitive coding, its basic conceptualization is likely to be too primitive for serious translation purposes. It probably cannot be considered improvable; without basic changes made by the developer, the user would be faced with the same corrections to make forever.

It should be kept in mind that some word choices are not easily made by any system. For example, the gloss 'its' for Spanish su--instead of 'his', 'her', 'their', or 'your'--may be distracting, but the problem

cannot be easily solved at the level of the algorithm. The most efficient solution may be to fix it in the output.

When it comes to difficulties that appear to be syntax-related, it is well to try to classify them in terms of whether they involve source analysis or target synthesis. The problem is on the analysis side if the machine has failed to resolve a homograph correctly, misconstrued relationships within a noun phrase, or completely "misinterpreted" a sentence. The fault may lie with the grammar, the parsing strategy, the dictionary coding that the parser addresses, or a combination of these. Synthesis problems, on the other hand, are more local and tend to be less disruptive in their overall effect, though they can be annoying and even misleading. Typical examples have to do with resolution of a verb string, prepositional government, the elements of a discontinuous idiom. The root cause is usually at the level of the algorithm and the grammatical rules, but the dictionary may also be involved. Whether or not the algorithm and the rules can be improved, be it on the analysis or the synthesis side, will depend on the developer and the type of system. Only the developer can indicate the ease with which such changes can be introduced.

3.2 Extensibility

The extension of a system to domains for which it has not already been developed necessarily involves dictionary-building. This fact is a given. Many of the same considerations apply that were discussed in the previous section with regard to whether or not a system can be improved. There should be a core of basic high-frequency vocabulary. The size of

the general dictionary will depend on how specialized or general the application is to be. Of course, this dictionary can be built up even if it is still rather small, but the effort will represent a major expense. What is indispensable, on the other hand, is that the system have several different ways of introducing context-sensitive expressions.

Most crucial in the case of extensibility, however, is the power of the grammatical rules. These constitute the most important aspect of the system to know about, and paradoxically they are also the most difficult to test. But before any judgment can be made about extensibility, it must be determined that the linguistic rules contained in the source, transfer, and target components are adequate for dealing with the type of input expected.

To arrive at this determination, Lehrberger and Bourbeau (1988) have proposed that the rules be "reconstituted" based on their effects on passages from the proposed domain. We would visualize a procedure along the following lines. The first step would be to identify a series of representative texts for translation, which should contain at least 30 different types of structures or linguistic phenomena. The texts would then be submitted to the computer, the output studied, and lists made of the linguistic phenomena that are (1) successfully processed, (2) unsuccessfully processed, and (3) not processed at all. After each type has been examined, any missing vocabulary would be added to the dictionary and further codes and idiomatic expressions would be provided where needed. The texts would then be run again and the output studied once more. An approach of this kind would yield a reasonable understanding of the system's capacity to handle the new domain in question.

There is little point in speculating on the extensibility of a system to domains not tested in this way. As a rule, however, it can be assumed that a general-purpose system with a large existing dictionary and past experience in translating many text types is maximally extensible, and that specialized systems are just the opposite. Sublanguage systems, dedicated as they are to a specific text type and subject field, are not readily extensible to other domains (Shann 1987:89). MT history has repeatedly demonstrated this fact.

3.3 Manageability and Compatibility

Just as with the output we looked at its serviceability, with the system itself we need to know how easily it can be managed and how well it fits into the proposed environment.

The factors of manageability include: the hardware on which the system and its peripherals run; the actual operation of the system; the facilities for updating the dictionaries and, if applicable, the associated lexical rules; and the ongoing support that can be expected from the MT supplier.

If the system does not run on the hardware already available at the prospective installation, thought needs to be given to the desirability of purchasing the new equipment. In any case, there should be an efficient interface both at the input and the output ends. At the input end there should be a means of easily capturing existing machine-readable text. If manual keyboarding is the sole form of input, the operation will be costly. While optical character recognition (OCR) reduces that cost to a certain extent, its usefulness for the application in question

needs to be carefully investigated. By far the most preferable form of input is a flat ASCII file or a word-processing document. Conversion from one word processor to another may be more or less difficult. At the output end, the word processor should be one that the translators feel comfortable using. In many environments it makes sense for the MT system to be part of a larger production chain: if the text is to be published, it should go from postediting to photocomposition with a minimum of steps in between.

The system itself should be user-friendly and easy to operate. Submission of the texts for translation should be a simple operation. Any requirement for translator interaction and on-line assistance should be carefully evaluated, since professional time spent in front of a screen is costly, especially if the final product will also have to be postedited. All the steps involved in operating the system should be factored into the overall cost, particularly the time of the translator.

In addition to the dictionary requirements discussed above, the system should offer the possibility of creating and maintaining user-defined subdictionaries that override the default translations. It should be possible for syntactic and semantic coding to be provided by the user as well as the developer. Updating should be a relatively quick and user-friendly process, and it should be reasonably easy to learn--bearing in mind, however, that if it is too quick or too easy, this facility will undoubtedly be at the expense of linguistic power and specificity.

Ongoing support from the supplier is another important factor.

The supplier should regard the client as a long-term partner and be committed to providing regular software upgrades as well as to making any adjustments in the system that are essential to proceeding with the particular application.

4. A Melding of the Perspectives

The viability of the MT system and its output is a multifaceted question. We have looked at a number of the issues from different perspectives. At the same time it is important to understand that these viewpoints overlap and are difficult to tease out as separate avenues of investigation. Ultimately, in the final judgment of an MT system, the different perspectives will converge.

NOTES

¹Discourse analysts consider that the properties which tell us a text exists, as opposed to merely a string of words and sentences, are derived from principles of cooperation between human beings, whether the addressee is present, as in a conversation, or distanced by time and place with the written form serving as the link.

²See Vasconcellos (1988) for a detailed discussion on the shortcomings of error analyses applied to MT output.

³The systems tested were IBM's Mark I and II (the latter installed at the U.S. Air Force) and Georgetown University's GAT.

⁴According to Bostad (p.c. 1989), consultations of Russian-English have averaged 185 per month in the last 20 months, with recent peaks of over 300. French-English has been consulted an average of 25 times a month, and German-English, 27 times.

⁵Figure cited by Deanna Hammond in "The Demand for Translations in Scientific and Industrial Research--The Challenge," talk given at the dedication of the National Translations Center (Washington, D.C., 15 June 1989).

⁶The statistics in Section 3.1 are based on experience with SPANAM/ENGSPAN and on reports by other MT developers.

⁷The test could be further refined so as to pinpoint whether the problem lies with the dictionary of the linguistic rules.

ACKNOWLEDGMENTS

The author is grateful to Peter Benton, Dale Bostad, and Marjorie León for their review and comments on portions of this article.

REFERENCES

- Automatic Language Processing Advisory Committee. 1966. Language and Machines: Computers in Translation and Linguistics; A Report by the Automatic Language Processing Advisory Committee (ALPAC). Washington, D.C.: National Academy of Sciences, Division of Behavioral Sciences. National Research Council Publication 1416.
- Bédard, Claude. 1988. "You Trust Your Mother, But YOU Cut the Cards." Language Technology 7:26-27, May-June.
- Benton, Peter M. 1989. "Report on the Machine Translation Study at McGraw-Hill." In: Proceedings of the 30th Annual Conference of the American Translators Association (Washington, D.C., 11-15 October 1989), ed. Deanna L. Hammond. In press.
- Bostad, Dale A. 1987. "Machine Translation: The USAF Experience." In: Proceedings of the 28th Annual Conference of the ATA (Albuquerque, 8-11 October 1987), ed. Karl Kummer. Medford, N.J.: Learned Information, Inc. pp. 435-443.
- Chandioux, John. 1988. "METEO: An Operational Machine Translation System." Presentation at RIAO 88 (Massachusetts Institute of Technology, March 1988).
- Clark, Herbert H., and Susan E. Haviland. 1977. Comprehension and the Given-New Contract. In: Discourse Production and Comprehension, ed. by Roy. O. Freedle. Norwood, N.J.: Ablex.
- Datta, Jean. 1989. Homespun Term Help for Translators. Jerome Quarterly 4(3):3-5.
- Halliday, M.A.K., and Ruquaya Hasan. 1976. Cohesion in English. London: Longman.
- Kingscott, Geoffrey. 1988. "Translator Strategies for Getting the Most out of Word Processing." In: Technology as Translation Strategy, ed. M. Vasconcellos. pp. 14-17.
- Lawson, Veronica. 1982. "Machine Translation and People." In her: Practical Experience of Machine Translation. Amsterdam: North Holland. pp. 3-9.
- Lawson, Veronica. 1989. "Practice Makes Less Imperfect: Users' Needs and Their Influence on Machine Translation Development." In: Georgetown University Round Table on Languages and Linguistics 1989. Washington, D.C.: Georgetown University Press. In press.
- Lehrberger, John, and Laurent Bourbeau. 1988. Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation. Amsterdam/Philadelphia: John Benjamins.

- Löffler-Laurian, Anne-Marie. 1986. Post-édition rapide et post-édition conventionnelle. Part 1, Multilingua 5:81-88. Part 2, 5:225-229.
- McElhaney, Terrence, and Muriel Vasconcellos. 1988. "The Translator and the Postediting Experience." In: Technology as Translation Strategy, ed. M. Vasconcellos. pp. 140-148.
- Newman, Patricia E. 1988. "Information-Only Machine Translation: A Feasibility Study." In: Technology as Translation Strategy, ed. M. Vasconcellos. pp. 178-189.
- Nida, Eugene A., and Charles R. Taber. 1974. The Theory and Practice of Translation. Leiden: E.J. Brill.
- Shann, Patrick. 1987. "Machine Translation: A Problem of Linguistic Engineering or of Cognitive Modelling?" In: Machine Translation Today: The State of the Art. Proceedings of the Third Lugano Tutorial (Lugano, 2-7 April 1984), ed. Margaret King. Edinburgh: Edinburgh University Press. Information Technology Series, 2. pp. 71-90.
- Sigurd, Bengt, and Barbara Gawronska-Werngren. 1989. The Potential of Swetra--A Multilanguage MT System. Computers and Translation 3:237-250.
- Smart, John M. 1988. "Getting Smart in Many Languages: MT with an Option of Preprocessing." In: Technology as Translation Strategy, ed. M. Vasconcellos. pp. 124-126.
- Tomita, Masaru. 1986. "Sentence Disambiguation by Asking." Computers and Translation 1(1):39-52.
- Vasconcellos, Muriel. 1985. "Theme and Focus: Cross-Language Comparison via Translations from Extended Discourse." Unpublished Ph.D. dissertation, Georgetown University. Washington, D.C.
- Vasconcellos, Muriel. 1986. "Functional Considerations in the Post-editing of Machine-translated Output: Dealing with V(S)O versus SVO." Computers and Translation 1(1):21-38.
- Vasconcellos, Muriel. 1987a. "A Comparison of MT Postediting and Traditional Revision." In: Proceedings of the 28th Annual Conference of the American Translators Association (Albuquerque, 8-11 October 1987), ed. Karl Kummer. Medford, N.J.: Learned Information, Inc. pp. 409-416.
- Vasconcellos, Muriel. 1987b. "Postediting On-Screen: Machine Translation from Spanish into English." In: Translating and the Computer 8: A Profession on the Move, ed. Catriona Picken. London: Aslib. pp. 133-146.

- Vasconcellos, Muriel. 1988. "Factors in the Evaluation of MT: Formal vs. Functional Approaches." In her: Technology as Translation Strategy. pp. 203-213.
- Vasconcellos, Muriel, ed. 1988. Technology as Translation Strategy, American Translators Association Scholarly Monograph II. Binghamton (N.Y.): University Center at Binghamton (SUNY).
- Vasconcellos, Muriel. 1989. "Cohesion and Coherence in the Presentation of Machine Translation Products." In: Georgetown University Round Table on Languages and Linguistics 1989. Washington, D.C.: Georgetown University Press. In press.
- Vasconcellos, Muriel, and Marjorie León. 1988. "SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization." In: Machine Translation Systems, ed. Jonathan Slocum. Cambridge, etc.: Cambridge University Press. pp. 187-235.
- Wagner, Elizabeth. 1985. "Rapid Post-Editing of Systran." In: Tools for the Trade: Translating and the Computer 5, ed. Veronica Lawson. London: Aslib. pp. 199-213.
- Weaver, Alan 1988. "Two Aspects of Interactive Translation." In: Technology as Translation Strategy, ed. M. Vasconcellos. pp. 116-123.