

Technology as
Translation Strategy

AMERICAN TRANSLATORS ASSOCIATION
SCHOLARLY MONOGRAPH *Series*

Volume II 1988

EDITED BY

Muriel Vasconcellos

State University of New York at Binghamton (SUNY)

Copyright © State University of New York at Binghamton (SUNY) 1988

ISSN 0890-4111

Printed in the United States of America

Factors in the Evaluation of MT: Formal vs. Functional Approaches

MURIEL VASCONCELLOS

Early evaluations of MT focused largely on analysis of the output and on cost-effectiveness of the throughput. More recently there has been awareness that a system's value depends on the use to which it will be put. The focus has shifted to its suitability for the application in question. Consideration must be given to a broad range of factors in the system's environment before a judgment can be made about its overall effectiveness. These factors are emphasized differently by developers, managers, translators, end-users of the product.

The author, a translator of many years' experience, has guided the development and implementation of MT at the Pan American Health Organization since 1977. She also worked on the original MT project at Georgetown University.

The question of whether or not an MT system does an effective job tends to be viewed, depending on the party who asks it, from different perspectives. The developer is concerned with how well the syntactic structures and semantic representations have been handled. Dictionary errors, recognized instantly for what they are, will be irrelevant to him. The manager, on the other hand, wants to know if MT can save him some money. He will calculate the "break-even" point—i.e., the volume of translation that needs to be processed before there begins to be a return on the investment. The translator, in turn, will focus on how much work is left for him to do. The independent translator will give the system good marks to the extent that it saves him time and effort and therefore enhances his productivity. The in-house translator, on the other hand, looks not only at productivity

but also at the implications for his personal role within the working environment— how he will be able to adapt to the new procedures, and the impact that MT might have (but rarely does) on the security of his job. And finally, the end-user will want to have a text that he can use. The consumer of information, for example, wants mainly to know that the facts can be discerned from the text without distortion. He may also be interested in quick turnaround, since some information has value only in relation to its timeliness. The message “We attack at dawn” is stale and useless the following afternoon. The publisher of product manuals, on the other hand, is prepared to factor some degree of editing into his equation, and he will look at how easily the system fits into his production chain, especially if the same text is being translated into several languages.

Some evaluations of MT have combined more than one of these perspectives. This idea in itself makes sense, because it is important to take into account the interests of all who are directly concerned. But evaluators should be careful not to generalize about MT based only on the interests of one group of principals involved—be they developers or managers or translators or even the end-users. Since in the final analysis the success of MT depends on the cooperation of everyone, no one’s concerns can afford to be subsumed.

Formal evaluations of MT deal with parameters that are (or appear to be) susceptible to measurement. Specific, isolated aspects are under scrutiny—usually only one or two variables. Most commonly they focus on the machine output. Although there is no way that these studies can give a total picture of how MT works in a given environment, still they can be meaningful when undertaken in perspective. For example, studies of output quality have informative value for those end-users who will be dealing with the raw product directly, and they should not be discounted for such purposes when combined with other criteria. Formal results alone, however, do not provide an adequate basis on which to draw conclusions about the effectiveness of MT. Far more useful is a *functional* approach, in which account is taken of all the tasks and other factors involved in meeting the purpose that the system is intended to serve in the particular environment. It is necessarily a complex undertaking. Whatever the approach, the desire to come up with a “yes” or “no” answer, and hence to reconcile the different perspectives, should not be allowed to obscure the diversity of interests at stake. In actual practice, it is to be expected that most MT

evaluations will include elements from both the formal and the functional approaches.

Formal Approaches

ALPAC

The most famous evaluation of MT was the one that led to the “ALPAC Report,” published in 1966 by the U.S. National Academy of Sciences. The Report’s findings were based on a study initiated in April 1964 by an ad hoc Automatic Language Processing Advisory Committee (ALPAC) which had been appointed at the request of the National Science Foundation to advise the Department of Defense, the Central Intelligence Agency, and the Foundation itself on research and development in the field of machine translation.

As almost everyone knows who knows anything about MT, the results of this study were to cast a shadow on MT development in the United States that persists until the present day. The effect of the Report was to cut off almost all U.S. Government funding of MT development. MT got a bad name from which it is only now beginning to recover.

The tremendous impact of this study is totally out of proportion to the quality of the data that went into it. Many of the Report’s conclusions were based on an evaluation of MT output from the point of view solely of one who would have to rely on the raw, unposted product (the “consumer of information” mentioned above). The machine’s output was subjected to a direct comparison with human translation, and the (somewhat irrelevant) conclusion was reached that fully automatic high-quality machine translation (FAHQMT) was impossible. Today, now that we know more about what happens when people actually use MT, it is interesting to look back and examine the methodology behind this study.

The approach was basically formal. The ALPAC Committee focused on the raw machine output and undertook to evaluate the product in terms of “the two major characteristics of a translation”: (1) intelligibility, and (2) fidelity to the sense of the original text (67).

A testing methodology was developed by ALPAC in order to accomplish this aim. In it, sentences were selected randomly from six different Russian-into-English translations of the same text, three of

them human translations and three of them done by machine.¹ These sentences, which came to a total of 144, were “interspersed in random order among other sentences from the same translation and also among sentences selected at random from other translations of varying quality” (67–68). In addition to the sentences being scrambled in relation to their order of occurrence in the original text, steps were taken to ensure that “no rater evaluated more than one translation of a given sentence” (71).

Using a nine-point scale of “intelligibility” and a similar one of “informativeness,” subjective ratings of these sentences were obtained from two sets of raters: 18 native speakers of English, Harvard undergraduates, who had no knowledge of the language of the original text (“monolinguals”) and 18 native speakers of English with a high degree of competence in the comprehension of scientific Russian (“bilinguals”), who were able to make direct comparisons between the translated texts and the original input. To assess informativeness (for determining fidelity), the monolinguals were provided with “carefully translated” sentences and asked to measure their informativeness *vis-à-vis* the sentences being tested, while the bilinguals were asked to make a judgment based on direct comparison with the original.

It quickly becomes evident that the methodology used by ALPAC was flawed.

To begin with, the study design flies in the face of what we know about discourse today. Only isolated sentences were judged. The sentences from a given text were intentionally presented in random order, effectively breaking up any cohesive ties or overall coherence that might have contributed to interpretation of the intended meaning. The scrambled order not only deprived the respondents of the original natural context but may also have created false contexts that were misleading.

Other flaws in the methodology had to do with the way in which differences between respondents were neutralized and with the size and characteristics of the experimental population.

In addition, the use of so many levels in the two rating scales (nine each) might be questioned. The subjects may have had difficulty being consistent when such fine-grained discriminations were required. Quite possibly the time charged to the “processing of information” may have included time spent in mulling over the possible choices at the nine different levels.² Furthermore, the criteria in the scale were vague,

judgmental, and semantically loaded. They included such terms as:

From the scale of intelligibility: “reads like ordinary text,” “masquerades as an intelligible sentence,” “stylistic infelicities,” “noise,” “grotesque” sentence arrangement, “bizarre,” “nonsensical,” “hopelessly unintelligible,” etc.

From the scale of fidelity: “makes all the difference in the world,” “puts the reader on the right track,” “gives a slightly different twist,” etc.

Finally, the procedure of having a monolingual rater measure the fidelity of the test translation by comparing it with a “carefully prepared” translation is fraught with uncontrolled variables, not the least of them being the fidelity of the “carefully prepared” translation itself.

Statistical results can only be meaningful to the extent that the original premises and the study design are solid. It is not unreasonable to conclude that the ALPAC results were tainted by mistaken premises, lack of knowledge about discourse and translation, and faulty procedures. And ALPAC focused its evaluation on the quality of machine output alone, which is only a small part of the total picture.

Summary of Formal Approaches

As a reaction to ALPAC, other approaches began to be sought for the evaluation of MT. A conference on the subject was held at the University of Texas in 1971,³ and in 1978 experts from around the world were convened for an international workshop in Luxembourg.⁴ By that time Systran was running at the U.S. Air Force from Russian into English and other Systran language pairs had been installed at the Commission of the European Communities in Luxembourg. The meeting reviewed all the major approaches that had been applied to the evaluation of MT up to then. Probably the earliest experiment was that of Miller and Beebe-Center. Sinaiko’s survey mentioned a study of the Air Force’s Mark II system done in the early 1960s, as well as further testing of Mark II output later in the decade (Orr and Small). There was also a formal study of output done by Pfafflin in 1965. Other studies reviewed at the meeting included the evaluation of Systran by Van Slype, subsequently published, in which both the quality of the output and the cost of throughput were assessed. In the Van Slype study, quality was determined by asking professional revisers to make

their corrections and then weighting the changes according to their importance as perceived by the investigator.

Such *error analysis* of MT output was then, and continues to be, a fairly frequent exercise, and understandably so, since it is the first possibility that comes to mind when one is initially confronted with translations produced by machine. But direct assessments of quality will always vary greatly between raters, as will the relative weights assigned.

The more important point about direct error analysis is that strangers to a system may not be able to distinguish between a problem created by a single dictionary error, fixable in a twinkling, and a major deficiency in conceptualization of the system. Also, some errors, such as those associated with anaphora resolution (e.g., choices between 'its', 'his', 'her', 'their', 'your' as a translation for Spanish *su*) may look silly at first sight but are easy to correct in the postedit and prohibitively expensive to deal with at the level of the algorithm. In other words, the developer knows that they are there but has deliberately assigned them low priority. Thus the important thing about MT errors is not their manifestation in the output but rather how easily they can be fixed, both on an ad hoc basis in the postedit and on a long-term basis in the dictionary or the algorithm.

The *rating technique* used by ALPAC was a notable improvement over the direct analysis of errors. There was no attempt to identify or assign values to specific words or constructions; the sentences were judged in their entirety as messages. In addition, the use of many subjects gave statistical validity to the results. The problems with ALPAC stem not from the rating concept as such but rather from features of the study design that were introduced later in the process.

There are other means of measuring the quality of MT output that are also more objective than error analysis. In the *comprehension test* the respondents' understanding of a text is ascertained through a series of questions that they answer after reading it through. The *performance test*, which might be applied to the translation of manuals, for example, requires that the subject act out a set of instructions as understood from the machine output. Still another method, which has been suggested by Brislin, is *back-translation*. The respondent, an experienced translator, manually back-translates the text to the original language (which presumably is his native tongue). It has been suggested, also, that machine translation be used in both directions. The difficulty in either case, and particularly the latter, is that the second

translation must be assumed to be of the highest quality – which creates a vicious circle. Of all the approaches, the most promising appears to be the *Cloze procedure*: in a sample of text, every *n*th word (typically every fifth word) is deleted, and the reader-subject is asked to fill in the blanks. This method has been used with MT by Sinaiko in a series of carefully constructed experiments (Sinaiko 4–5; Sinaiko and Klare).

In addition to quality of the output, there are other parameters that can be treated formally. A paper by Hofstetter proposed, as the basic measurement, the time required to bring the machine-translated text up to the desired level of quality. Sager also proposed speed of revision as a parameter, combining it with a larger set of criteria that included intelligibility and acceptability of the output. And, of course, the Van Slype studies and others have measured the total cost of throughput.

Functional Approaches

End-User Evaluation

None of the foregoing methods addresses the question of whether MT is serving its intended purpose. The first investigator to do so was Henisz-Dostert. She takes as the theme for her work a statement made by Bar-Hillel in 1971:

Every program for machine translation should be immediately tested as to its effects on the human user. He is the first and final judge, and it is he who will have to tell whether he is ready to trade quality for speed, and to what degree (76).

Starting from this premise, she conducted a survey among users of unedited machine translations from Russian to English (the Georgetown Machine Translation System, installed at the EURATOM Research Center in Ispra, Italy, and at the Atomic Energy Commission in Oak Ridge, Tennessee). Responses to a questionnaire were received from 58 scientists and engineers who had used raw output from the system during the period 1963–1973.

The major finding was that 90% of the respondents judged the quality of the translations to be “good” and “acceptable” for their purposes;

moreover, 93% found them to be informative, 81% felt that they were complete, and 59% said that they were readable. Getting used to reading MT style did not present a problem. (Still, MT output was considered to take nearly 100% longer to read than natural text, and human translations were also reported to take about one-third more time to read.) Only 19% found that the machine translations could give misinformation, while 80% had not had this experience.

Despite delays in throughput (due mainly to problems of keyboarding the input text), 96% of the respondents had either already recommended MT to their colleagues or would not hesitate to do so; moreover, 87% of them actually preferred MT to human translation.

The respondents considered that semantic factors (vocabulary, linguistic context) were more important for understanding the MT output than syntax, but they felt that the most important factor in understanding the texts was their own familiarity with the subject matter (extralinguistic context).

Here, then, is the proof of the pudding. The technical personnel who were end-users of raw MT output were apparently not very concerned about the type of error being produced by the machine as long as they could understand the text and glean the information that they were looking for.

The Broad Functional View

A fully functional evaluation of an MT system will take into account not only end-user satisfaction but also the concerns of management (i.e., return on the investment, productivity, service to the organization) and of the postediting translators (manipulability of the output, ease of building the dictionaries, etc.). Since a dynamic environment includes interaction with the developer, consideration should also be given to the needs, priorities, and constraints of the latter. But more than anything else, what must be determined is whether or not the MT system has the capacity to grow and that the human principals who will be interacting with it will be able to make effective contributions to that growth.

An MT system's potential hinges on four main factors: the dictionary structure, the daily working environment, the translators who use it, and the ongoing support provided by the developer.

The *dictionaries* should be large to begin with—20,000 stem entries

of general vocabulary at the very least, with the possibility of adding subject-specific and user-defined subdictionaries that override the default translations. If the system is to make lexical choices based on context, the dictionary entries must necessarily have syntactic and semantic codes that can be supplied by the user as well as the developer. The system should also offer several different ways of coding for contextual associations. Updating should be quick and mechanically easy to perform, and, without sacrificing the linguistic specificity just described, it should be easy to learn.

The *working environment* should include, at the front end, effective means of capturing the input text in machine-readable form. If manual keying is the sole source of input, the operation will be costly. In addition, the tools should be user-friendly. Submission of the texts for translation should be a simple operation, and the output should be available in word-processing form such that the translator can work with it readily on-screen. From the standpoint of management, it makes sense that the MT system 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 *translators* should be prepared to make a long-term commitment to building the system. In a typical installation, as opposed to the few that deliver raw MT directly to end-users, the translators make their contribution at three levels (see Ryan and Santangelo in this volume). First, through their everyday postediting they exercise the system, they sharpen their own skills, and they generate feedback both for updating the dictionaries and for improving the algorithm. Second, they participate directly in dictionary-building. It is only through their contribution, fertilized by daily contact with text, that the dictionaries can become tuned to the types of discourse being translated. And third, in some settings at least, translators interact directly with system developers, providing useful suggestions about recurring patterns. The system will grow faster and better to the extent that translators have positive attitudes, innovative responses, and a can-do spirit; to the extent that they are willing to use the word processor for long periods; and to the extent that they postedit resourcefully and come up with creative solutions at every point in the production chain.

It is important to have *ongoing support* from the developer, or vendor, in the form of regular software upgrades and adjustments that may be required by the particular application.

The functional evaluation should be designed to determine the existence of the conditions just described. If these can be assured, then the system can be counted on to flourish in its new environment.

Conclusion

There is no single ideal model for the evaluation of MT. Always, the need to be met is the bottom line. It is crucially important to look at a system from the standpoint of all parties concerned, since MT is necessarily a common endeavor in which everyone contributes to the fullest. But even though there is no single “right” way to go about an evaluation, it is safe to say that formal data should be eyed from their inherently limited perspective and that priority should always be given to the functional factors that shape the future of the system.

NOTES

1. The provenance of the three machine translations, at that time in the history of MT, is not documented. In its only reference to the texts used for the study, the Report says:

The measurement procedure was tested by applying it to six varied English translations—three human and three mechanical—of a Russian work entitled *Mashina i Mysl'* [*Machine and Thought*], by Z. Rovenskij, A. Uemov, and E. Uemova (Moscow 1960). These translations were of five passages varying considerably in type of content. (All the passages selected for this experiment, with the original Russian versions, have now been published by the Office of Technical Services, U.S. Department of Commerce, Technical Translation TT 65-60307.) The materials associated with one of these passages were used for pilot studies and rater practice sessions; the experiment proper used the remaining four passages.

2. In addition to collecting ratings for each of the sentences, ALPAC also asked the raters to clock, with stopwatches, the time they took to come up with their responses. Processing time, all other factors being equal, would seem to be a measurement of interest to consumers of information. It has the advantage of claiming no more than exactly what it is—i.e., the time required to grasp the sense of the text, which for intelligence-gathering purposes translates into budgetary considerations.

3. December 1971. Proceedings published by Lehmann and Stachowitz.

4. Workshop on Evaluation of Machine Translation Systems, Commission of the European Communities (Luxembourg, 28 February 1978).

REFERENCES

- Automatic Language Processing Advisory Committee. *Language and Machines: Computers in Translation and Linguistics; A Report*. . . . National Research Council Publication 1416. Washington, D.C.: National Academy of Sciences, Division of Behavioral Sciences, 1966.
- Bar-Hillel, Yehoshua. "Some Reflections on the Present Outlook for High-Quality Machine Translation." *Feasibility Study of Fully Automatic High-Quality Translation*. Ed. Lehmann and Stachowitz. Austin: University of Texas at Austin Linguistic Research Center, 1971. Cited in Dostert 1979 (151).
- Brislin, Richard W. "Introduction." *Translation: Applications and Research*. New York: Gardner Press (John Wiley & Sons), 1976. 1-45.
- Henisz-Dostert, Božena. "Users' Evaluation of Machine Translation: Georgetown MT System, 1963-1973." Paper presented at the Workshop on Evaluation of Machine Translation Systems (Luxembourg, February 1978). Typescript. Luxembourg: Commission of the European Communities, 1978. Results subsequently published in extenso under the same title as Part III of *Machine Translation*. The Hague, Paris, New York: Mouton, 1979. 147-244.
- Hofstetter, A. "Methodological Concept for the Evaluation of Translation Quality." Paper presented at the Workshop on Evaluation of Machine Translation Systems (Luxembourg, February 1978). Typescript. Luxembourg: Commission of the European Communities, 1978.
- Lehmann, Winfred P., and Rolf Stachowitz. *Feasibility Study of Fully Automatic High-Quality Translation*. Austin: University of Texas at Austin Linguistic Research Center, 1971. Report No. RADC-TR-71-295.
- Miller, G.A., and J.G. Beebe-Center. *Mechanical Translation* 3 (1958): 73. Cited in ALPAC 1966.
- Orr, D.B., and V.H. Small. *Comprehensibility of Machine-Aided Translations of Russian Scientific Documents*. Washington, D.C.: American Institutes of Research, 1966. Cited in Sinaiko 1978.
- Pfafflin, S.M. *Mechanical Translation* 8 (1965):2. Cited in ALPAC 1966.
- Sinaiko, H.W. 1978. Some Thoughts about Evaluating Language Translation. Paper presented at the Workshop on Evaluation of Machine Translation Systems (Luxembourg, February 1978). Typescript. Luxembourg: Commission of the European Communities, 1978.
- Sinaiko, H.W., and G.R. Klare. "Further Experiments in Language Translation: Readability of Computer Translations." *ITL* 15 (1972). Cited in Sinaiko 1978.
- Sinaiko, H.W., and G.R. Klare. "Further Experiments in Language Translation: A Second Evaluation of the Readability of Computer Translations." *ITL* 19 (1973). Cited in Sinaiko 1978.
- Van Slype, G. 1979. Critical Study of Methods for Evaluating the Quality of Machine Translation. Final Report. Brussels, Luxembourg: Commission of the European Communities, 1979. See also his summary of the Second Evaluation of Systran published as: Systran: Evaluation of the 1978 Version of the Systran English-French System of the Commission of the European Communities *Incorporated Linguist* 18.3 (1980): 86-89.