[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland; ed. Kirsten Falkedal (Geneva: ISSCO).]

Evaluating Commercial MT Systems

Elliott Macklovitch
Canadian Workplace Automation Research Centre

1 Purported productivity gains

Vendors of commercial machine translation systems will often claim that their system can increase translator productivity *x*-fold. In order to verify such claims, we need to answer the following two questions: First, how is translator productivity generally measured? And second, precisely how does one go about comparing human translator (HT) productivity with MT productivity? The answer to the first question is relatively straightforward, at least for translators that are part of a translation service: productivity is generally measured in terms of the number of words translated per unit of time. In fact, translators frequently have to meet production quotas – 1300 words per day, for example – and their promotion may be contingent upon producing a certain number of words per year.[1]

The answer to the second question is slightly more complicated and involves, I would suggest, the comparison of two production chains: one in which the human translator works in tandem with the MT system; and another in which he works alone, without the aid of the system. Now there are many ways for a human translator to actually produce his texts: he can write them out, or type them, dictate them or use a word processor. Most commercial MT systems, on the other hand, come bundled (or at least interface with) a word processor. My intuition six years ago, when I was asked to participate in a trial of the Weidner MicroCat system at the Canadian government's Translation Bureau, was that the purported productivity gains reported by the vendor were at least partly attributable to the introduction of a word processor in place of more traditional modes of production. Be that as it may, it is surely important, when designing an MT trial, to attempt to isolate the contribution of the machine translation module to overall productivity, since this is what costs the most to develop and what justifies the hefty price tag, not the word processor.

[1] This is less obviously the case for free-lancers, who are generally paid a fixed fee to translate a given text within some time limit, regardless of the number of words they actually produce per hour or day.

1.1 The Weidner trial

The trial of the MicroCat system that was organized at the Translation Bureau in 1985 involved six translators from three federal government departments, and lasted four and a half months.[2] The translators selected to participate were autonomous (i.e. their translations were not systematically revised), eager to take part in the trial, and all had previous experience with word processors. Every effort was made to ensure the success of the trial and to provide the translators with optimum working conditions. Weidner Corporation sent a representative to Ottawa to give them a week of on-site training. Weidner also participated in the selection of simple, repetitive texts to be translated during the trial. Furthermore, it was decided that each participant would concentrate on only one text type for the duration of the trial. For each text type, Weidner was supplied with a list of domain-specific terminology, which was loaded into a separate sub-dictionary before the trial began. The trial was conducted in a quiet isolated room at Translation Bureau headquarters, where I was available to offer the participants on-site technical support.

An independent management consultant was hired to organize the collection of data and write up the final evaluation report. He prepared a log form which the participants were asked to complete for each text they processed during the trial, and on which they noted the time required for each phase of the machine translation process. (A copy of this form may be found in Appendix A.) By totalling the times for the various phases and dividing this figure into the number of words in the text, we obtained a production ratio of words translated per minute. After an initial learning period of about ten weeks, these figures showed – somewhat to our surprise – that four of the six participants were consistently surpassing the figures their managers had given us for their productivity in their sections. Were we to conclude that the vendors claims about increased productivity were in fact correct?

Our view was that given the many differences between the translators' situation on the trial and their working conditions in the sections, such a conclusion would be premature. In their sections, the translators handled many types of texts, of varying difficulty. The incitement to produce as many words as possible was not as intense in the sections as it was during the trial. Moreover, the daily production figures that the section chiefs gave us for each translator were calculated on the basis of a rough yearly average, and did not accurately reflect the translators' other duties, such as revising the work of junior translators, providing linguistic advice to clients, attending meetings, etc. In an effort to control all these variables and clarify this

[2] For more details on the Weidner trial, see E. Macklovitch, "MT Trial and Errors", *Proceedings of the International Conference on Machine and Machine-Aided Translation,* Aston University, Birmingham U.K., April 7-9, 1986. Also available from the author.

question of comparative productivity, we therefore decided to have the participants translate a number of texts using only Weidner's word processor. Just as with the MT production chain, the times required for each human translation were recorded, and divided into the total number of words in the text, giving us a production ratio we could compare with the MT times, because the two sets were produced under strictly comparable conditions.

1.2 The Weidner findings

For the purposes of this comparison, some twenty-four texts were retained, which were translated using only the word processor. When the consultant calculated the average words-per-minute figures for these texts and compared them with the average MT figures for the four best translators[3] during the final weeks of the trial, he found that the HT production chain was significantly faster than the MT production chain. How much faster depends on which phases of the MT chain are counted. If we count all the steps on the log form, human translation was nearly twice as fast as machine translation. If we discount the time that the machine actually takes to translate (on the assumption that the participants could use this time to do other useful tasks), as well as the time for the second dictionary update (on the grounds that these new or modified entries are not intended for the current text), MT remains 27% slower than HT. If, in addition, we discount the time for text entry, assuming that source texts arrive in machine readable form that Weidner could import,[4] MT still remains 5% slower than HT for all the texts translated during the operational phase of the trial.

What accounted for this disappointing performance of the MT production chain? Clearly, one very important factor was the poor quality of the raw translations which the MT system produced. Even though the text types selected were relatively simple, extensive post-editing was required in order to turn the raw machine output into a product that could be delivered to the client. In fact, on average, this post-editing effort amounted to over 45% of the total time needed to produce a text with the MicroCat system. Not only was post-editing time-consuming; the translators also found it to be quite arduous. When asked by the consultant if they would like to continue working with Weidner on the same texts after the end of the trial, not a single participant accepted.

There is another less obvious factor that may partially account for the comparatively poor performance of the MT chain, and which could be re-

[3] Two of the participants took less well to the system, never managing to attain their section production figures. Had they been included in the comparison, the differences between HT and MT would have been even greater.

[4] In fact, this assumption is rather implausible, seeing that Weidner's word processor was wholly incompatible with every other word processor on the market.

ferred to as the tortoise and hare syndrome. As a cursory examination of the log form shows, there are many steps which have to be carried out *before* the MT system actually begins to translate. The source text needs to be keyed in; segments that are not to be translated have to be specially marked; a vocabulary search must be run and the missing words added to the system's dictionary. While all this is being done, the human translator continues to translate, certainly not as quickly as the machine but without losing any time on unessential tasks. However fast the system churns out its raw translation, it must then be post-edited, and as we have seen, this can be a slow and painstaking process. Meanwhile, the human translator continues to plod on. The upshot of all this is that for relatively short texts – and for some reason, Weidner's Canadian representative insisted on selecting texts of under 1,500 words – the tortoise may well arrive at the finish line before the speedy hare. One other finding of the trial also merits mention. The vendors of commercial MT often harp upon the importance of dictionary updating as the key to the successful operation of their systems. On the MicroCat trial, we found that the best production times did not correlate with the domain in which the most new dictionary entries were added. Over a twenty week period, more than 1400 new entries were added for plant variety descriptions, compared to about 450 entries for reports on the Labour Force. And yet production times were far better on the latter texts than the former – probably because the Labour Force reports were composed of simple, straightforward sentences that corresponded more closely to Weidner's general grammar of English than did the lists of complex noun phrases that made up the crop descriptions. This is not to deny the importance of dictionary updating, but simply to put it in perspective. Since commercial MT systems rarely allow users to modify their grammars, the *only* way users have to customize a general purpose system is via the lexicon. Clearly, however, there are limits to the extent that a system can be customized or modified by means of dictionary entries. In fact, a summary error analysis that I performed on samples of post-edited output following the trial suggested that at least half of the corrections made by the translators could not be implemented by modifying the system's dictionaries.

2 Evaluating MT improvability

Once they have sold their systems – perhaps by promising fabulous increases in productivity – commercial MT vendors will often respond to users' complaints by claiming that this or that problem will be corrected in the next (or some upcoming) version. In this way, user satisfaction becomes like a mirage on the horizon: it keeps receding as the user advances with the system, forever remaining out of reach. How can the user respond to such claims by the vendor? How can he objectively evaluate, in other words, whether the linguistic performance of the system he is using is indeed improving with each successive release? One thing he can do is present the developer with "bug lists", and then verify whether those bugs have been corrected in the next release. We all know, however, that in programs as complex as MT systems, improvability does not follow a linear progression, and that the changes introduced to correct one problem often produce unforeseen side effects in other areas. I now want to present two related ways of evaluating the linguistic performance of MT systems which I have had occasion to test, and which can provide the user with some indication of linguistic improvability. When the former TAUM-Météo system was rewritten to run on a microcomputer, one of the contract requirements was that there be no deterioration in linguistic quality or performance. In order to verify this condition, samples of raw output were collected from both versions of the system, and compared in each case with the final, post-edited versions. Each change that the post-editor made to the raw output – every word that had to be modified in form, deleted, displaced or inserted – was counted as one "error".[5] To illustrate with a simple example, (ii) below is a raw machine translation of (i), and (iii) is the post-edited version:

(i) [. . . for the week ending February 18 . . .]

(ii) [. . . pour le terminer de semaine le 18 février . . .]

(iii) [. . . pour la semaine se terminant le 18 février . . .]

To get from (ii) to (iii), the post-editor has to make five changes: *le* must be changed to *la*; *semaine* must be shifted backward; the pronoun *se* must be inserted; *terminer* changed to *terminant*; and the preposition *de* deleted.[6] The basic idea is to count primitive post-editing operations rather than keystrokes or commands, since the latter counts may vary with different editors or word processing packages. Comparing the error count (say, per 100 words of source text) for version A with the error count for version B gives us a straightforward indication of whether the system's linguistic performance has improved, deteriorated or remained stable. The technique works particularly well for weather bulletins, where the texts are highly uniform. In other domains, if we wanted to ensure that sample A was no more difficult than sample B, we could resubmit the same text to the two versions of the system.

[5] I put "error" in quotation marks because not all of the post-editor's corrections may appear to be warranted. Under this approach, however, the analyst does not question the post-editor's decisions; he simply takes them at face value.

[6] Note that the final count is the same if, instead of inserting *se* and deleting *de,* the post-editor changes *de* to *se* and then places it before *terminer.*

Similarly, if we were concerned that one post-editor may introduce more non-essential (or stylistic) changes than another, we could ask the same person to revise the output of the two systems.

The advantage of this methodology is that it is relatively objective and straightforward to apply. In this respect, it is preferable, I would maintain, to scalar judgments of intelligibility, fidelity or clarity like those that John Carroll employed in the (in)famous ALPAC report. As several speakers at the Forum pointed out, it is not obvious that a respondent can meaningfully distinguish such closely related parameters for a given sentence. How does one rate an unintelligible sentence for fidelity, for example? Moreover, in order for such judgments to be reliable, the translations must be submitted to a fair number of respondents. From a strictly practical point of view, it may be less trouble for an evaluator to analyse a few samples of machine output in order to arrive at an error count. On the other hand, this type of analysis does not tell us anything about the *kinds* of improvements or deterioration the system has undergone, or the areas in which there has been no improvement from one version to the next. For this, we need to analyse (or at least classify) the errors that the revisor corrects at post-editing.

2.1 The Logos evaluation

The Translation Bureau of the Secretary of State has been running a trial of the English-to-French version of the Logos machine translation system since March 1987. One of the goals of this pilot project is to determine to what extent the contractor can customize its linguistic software to meet the particular needs of the Bureau. In this regard, Logos is to deliver to the Secretary of State three new releases of the software per year, based on requests for improvements or modifications provided by the Bureau's project team. In February 1988, the project leader at the Informatics section in Montreal asked me to conduct a linguistic evaluation of the system. In particular, Mr. Levy was interested in the following questions: First, what sorts of problems had the developer managed to rectify over the three versions thus far delivered, and what sorts of errors persisted? Second, what types of errors were most important, both in terms of absolute frequency and in terms of their impact on post-editing? And finally, among the persistent errors, which in my opinion could Logos reasonably be expected to correct, and which would the users have to live with? Mr. Levy's concern was that the linguistic improvements in the latest version of the system had not been as marked as in previous versions, and his hope was that my report might help him in his discussions with Logos' development team.

To conduct this evaluation, I obtained from Mr. Levy four texts that had already been translated by Logos, two by each of the previous versions of the system (versions 3.1 and 4.2), along with the source texts and the final, post-edited copy. I then had these four texts retranslated by what was then the current version of the system (5.3). This allowed me to verify whether the errors which the post-editor had corrected in the earlier raw translation reappeared or had been eliminated in the latest version, and whether any new errors had been introduced. I also attempted to classify each post-editor correction in an error taxonomy, which is reproduced in Appendix B.[7]

I do not make any strong claims for the universality of this classification scheme, which is an adaptation of one that was used during the TAUM-Aviation project. Ideally, we want a taxonomy that is fine enough to highlight the principal and recurrent errors of the texts under analysis,[8] without being so fine as to make classification choices difficult. Nor do we want a schema that is overly general, to the point that it conflates errors that we would prefer to keep distinct. The role of the taxonomy is to help the analyst articulate the user's intuitions about the system, in a way that the developer will also find useful. The taxonomy in Appendix B is divided into three large categories: (i) morphology, which is extended to include unknown words and delimiting symbols, as well as certain problems of text layout; (ii) source language analysis; and (iii) transfer and generation. As should be obvious, all the post-editor corrections which the analyst must classify appear as modifications to the target language output; it is unlikely, moreover, that he will have access to the contents of the system's grammars, or even to its dictionaries. The analyst's job is to attempt to deduce the source or cause of each error that the post-editor has flagged. It goes without saying that many of his hypotheses will be highly tentative.

2.2 The Logos findings

A simple error count of post-editing operations, comparing the translations produced by the two older versions of the system with the translations of the same texts produced by the most recent version, revealed that the linguistic improvements which had been implemented in version 5.3 were by and large minimal. Put another way, our numbers tended to support the project leader's impression that the linguistic quality of the system had in some sense peaked. Most of the improvements observed in the second machine translation of the four texts were localized in the lexical and morphological

[7] For a more detailed discussion of each of the classes in this taxonomy, see E. Macklovitch, "*A Linguistic Evaluation of the Logos English-to-French Computer-Assisted Translation System*", May 1988. Available from the author.

[8] This implies that the taxonomy will need to be tuned to different MT systems. To take a simple example, suppose another system handled upper and lower case flawlessly (instead of simply reproducing the source language case in the target language output); this category could then be dropped. The focus, in other words, is on the system's weaknesses, not its strengths.

components; for example, some missing compounds and other unrecognized forms had been entered in the dictionary. Such changes, however, are more indicative of a bias in the methodology than of a general improvement in linguistic performance, since they are sure to show up when the same text is retranslated. Some non-negligible improvements were also made in the area of target language generation; for example, English imperatives were now translated as French infinitives, and the appropriate syntactic complements had been specified on a number of target language equivalents.

As for the errors that persisted in the more recent translation of each text, our study showed that little progress had been made on the classical stumbling blocks of English analysis, i.e. coordinate structures, stacked nominals, categorial homography, "-ing" forms, etc. While not the most frequent of the errors tabulated, these often had a devastating effect on the intelligibility of the output, and would require considerable effort to correct during post-editing. In this respect, they contrast with the most frequent of the errors, the class of articles, which rarely affect the general intelligibility of the output. Another extremely frequent but more serious problem was target terminology. At first sight, this may appear somewhat paradoxical, since terminology is in principle amenable to lexical correction. What we discovered, however, was that the majority of these errors were in fact due to the polysemy of the source language terms, and hence were not so easily corrigible. That is, while an equivalent which is incorrect or inappropriate for a given text may indeed be modified in the system's dictionary, there is no guarantee that the new transfer will be appropriate for the next occurrence of that term. In the course of our analysis, we came across numerous examples of vague or polysemous source language terms that require multiple target language equivalents. Indeed, one of the sobering effects of this sort of analysis is that it underscores the pervasiveness of polysemy in natural language, even within a restricted technical domain like data processing. Now Logos is one of the few commercial MT systems that allows for the selection of a TL term among alternate equivalents by means of rules that verify the term's syntactic and semantic environment. Judging from the number of terminological corrections in our samples, however, it would appear that this limited class of context-sensitive rules is simply not sufficient for the enormous task of source language lexical disambiguation.

At the Evaluators' Forum, some questions were raised regarding the size of the samples that were used to arrive at these conclusions. From the four previously translated texts, I selected segments of between five and seven hundred words long; in total, then, about 2500 words of source text, for which two parallel machine translations were obtained. Conducting an error analysis of the sort described above is an extremely labour-intensive and time-consuming endeavour. Perhaps it would have been preferable to analyse a larger corpus. I would maintain, however, that the small size of the corpus does not in any way invalidate our findings, mainly because these findings were almost entirely predictable. For anyone who has any experience in natural language processing, it is surely no surprise to learn that Logos stumbled on the same well-known set of analysis problems that continues to baffle all NLP systems. Nor is it astonishing for anyone with the least experience in translation to be told that the great majority of a language's vocabulary is polysemous, and that polysemous terms in two languages rarely match meaning for meaning – a fact that is sure to cause major headaches for any general purpose machine translation system. And the same goes for the system's problems in correctly delimiting the textual units to be translated when this text is embedded in a wide variety of complex layouts and formats. The point is that none of these conclusions is really unexpected; and therefore, it should not be necessary to analyse massive samples to arrive at findings that are by and large a reflection of the state of the art.

A final word on the methodology itself. It is, as we mentioned, long and arduous; furthermore, it can only be properly applied by someone who knows the two languages in question, and who has a minimal understanding of how the particular MT system functions. This is because the analyst is in a complete "black box" situation. He has the input to the system, in the form of the source language text; its output, in the form of the raw machine translation; and an indication of the "errors" in the output, in the form of the post-editor's final, corrected version. His job is to use these corrections to tease out the system's linguistic specifications,[9] and in particular the weaknesses in its grammars and dictionaries. To illustrate how difficult it can be to correctly classify corrections made by the post-editor, consider the following authentic example of machine output (which, incidentally, was not produced by Logos): for the French input *maux de dos,* the system produced *promenades of back,* instead of the obvious *backache* or the more literal *back pains.* Has someone for some obscure reason entered *promenade* as a possible translation of *maux,* or is something else going on here? As it turns out, the source of the problem is an incorrect morphological analysis: *maux* is incorrectly taken to be the plural of *mail* instead of *mal*, on the model of *bail-baux.* One possible translation of *mail* is *mall* or shopping *promenade;* whence *promenades of back.* Though most cases are not this baroque, the example does serve to illustrate the point: it is often far from obvious how a given error should be classified. In other instances, two categories of the taxonomy may overlap, as for example with a misanalysed "-ing" form, which could be classed as such, or as a case of categorial homography. Despite all these pitfalls, I would still contend that if the analyst manages to do his job properly, the findings will often prove to be revealing, both for the system

[9] For an opposing view on the reasonableness of this undertaking, see the paper presented at the Forum by Steven Krauwer and Ton van der Wouden.

developer and for the system user.

3 Conclusion

In the course of this discussion, we have mentioned four different ways of evaluating machine translation systems. No one method is absolutely better than any other; rather, each addresses different questions, and hence may be more or less appropriate in a given situation. Counting post-editor operations provides a rough-and-ready way of comparing the linguistic performance of successive versions of an MT system – or, if the same test corpus is used, two competing systems. A point in its favour is that it is fairly straightforward to apply, and actually can be automated.[10] On the other hand, the bald results tell us nothing about the system's strengths and weaknesses, and must be interpreted with care, since the number of post-editing operations does not necessarily correlate with post-editing time.

Classifying post-editor corrections in an error taxonomy provides a natural complement to the simple counting of post-editor operations. As discussed in the previous section, the methodology is not easy to apply. Nevertheless, for the system developer, some such approach would seem to be essential if he is to establish priorities among the improvements that can and need to be made to the MT system.

In our view, the best way of determining whether a system is cost-effective is to compare human translation and machine translation production times. Of course, care must be taken to ensure that the two production chains are evaluated under identical conditions, as we attempted to do in the Weidner trial. If, working in tandem with the MT system, the human translator can turn out more words per hour or day than he can without it, then the system *may* be cost-effective, depending on a number of other factors such as the current cost of human translation, the system's purchase and maintenance costs, hardware and training requirements, etc. But unless there is an increase in translator throughput, there is no way the system can be cost-effective.

We have had relatively little to say about the ranking of MT output on various subjective scales, like those used by Carroll in the ALPAC report. Still, if the scales are clearly defined and the number of evaluators high enough to ensure reliable results, this technique could be profitably applied to the comparative evaluation of the quality, not of raw machine output – no one seriously entertains such illusions today – but of post-edited machine translations versus human translations of the same texts. As MT

[10] See the paper presented at the Evaluators' Forum by Harri Jäppinen. This is particularly good news, since applying the error-counting methodology manually can be rather tedious and time-consuming.

gains wider acceptance, more and more clients will want to know whether its cost-effective use inevitably entails a deterioration in linguistic quality. This is an extremely important question, and one which has not received of late the public attention it deserves. On such a sensitive issue, which tends to spark emotional debate, it would be helpful to have some dependable data.

Appendix A

Weidner Trial

LOG FORM

Date: ____________________ **Translation No:** ________

Translator: ____________________ **No of words:** ___________

1. **Text entry**

   a) mode: diskette or manual

   b) time begun:________ finished:________

   c) entered by: _______ d) date: _______

2. **Pre-editing**

   a) time begun:_________ finished: _______ b) initials:_______

3. **Vocabulary search**

   a) time begun:_______ finished: _______ deferred: _____

   b)submitted by:_____

4. **Dictionary update 1**

   a) time begun:_______ finished: ______

   b) number of entries:_____ c) initials: ____

5. **Machine translation**

   a) time begun: ________ finished: ________ deferred: ____

   b) submitted by: ______

6. **Post-editing**

   a) time begun: ________ finished:_________ b) initials: ____

7. **Dictionary update 2**

   a) time begun: ________ finished:_________

   b) number of entries: ____ c) initials:_____

Appendix B

ERROR TABULATION

Text #259-3727: Customs and Excise

		Version 3.1	Version 5.3
I.	Morphology, graphology & layout		
I.1.	Number, inflection & agreement		
I.2	Upper/lower case		
I.3	Hyphens, slashes & quotes		
I.4	Layout; underlining		
I.5	References; place names		
I.6	Unknown words/symbols		
II.	Analysis		
II.1	Categorial homography		
II.2	'-ing forms		
II.3	'-ed forms & passives		
II.4	Coordinate structures		
II.5	Stacked nominals		
II.6	Anaphora & ellipsis		
II.7	Articles		
II.8	Gibberish		
III.	Transfer and generation		
III.1	Incorrect/incomplete TL equiv.		
III.2	Polysemy		
III.3	Restructuring		
III.4	Inappropriate/incorrect form generated		
III.5	Stylistic changes		