[*Translating and the Computer 22*: Proceedings of the Twenty-second international conference…16-17 November 2000. (London: Aslib, 2000)]

Evaluating Machine Translation: the Cloze Procedure Revisited

Harold Somers & Elizabeth Wild

Department of Language Engineering,
P O Box 88
UMIST,
Manchester M60 1QD
Harold.Somers@umist.ac.uk

Summary

This paper describes the use of the "cloze procedure" to make a comparative evaluation of some currently available MT systems. This technique, which involves masking some words in a text, and then asking subjects to guess the missing words, was used in the 1960s and 1970s as an evaluation method, but no more recent use has been reported, even though the methodology is simple to implement and provides an objective result. We report here two experiments in which we tested three MT systems against a human translation, with texts from three different genres, to see whether the procedure can be used to rank MT systems against each other. The paper discusses some details of the procedure which provide important variables to the test, notably what percentage of words are masked, and whether the scoring procedure should be right-wrong, or should differentiate between different degrees of wrong answer (crediting close synonyms and other plausible near-misses). We discuss other aspects of the procedure which may affect the test's usability. Especially of interest is the fact that there seems to be a lower quality threshold below which the procedure is less discriminatory: the translation is so bad that the subjects cannot make reasonable guesses at all.

Introduction

Amongst the early attempts to evaluate Machine Translation (MT) output was the often-cited work of Sinaiko & Clare (1972, 1973) who used Wilson Taylor's (1953) "cloze procedure" to evaluate the readability of English-Vietnamese MT output. Although the technique is very simple, the technique has not reportedly been used since then, even though MT quality in general has improved hugely. In this paper we report on experiments using the technique to make a *comparative* evaluation of some currently available MT systems.

It has been said that "machine translation evaluation is a better founded subject than machine translation" (Wilks, 1994, p.l). Much of what has been written in recent years discusses general issues in MT evaluation design (for example, Arnold et al. 1993; Vasconcellos, 1994; Sparck Jones & Galliers, 1995; White, forthcoming). For details of actual evaluations we often have to delve deeper. MT evaluation techniques are often classified along various parameters including who the evaluation is for (researcher, developer, purchaser, end-user), whether it makes use of information about how the system works (black-box vs. glass-box) and, above all, what aspect of the functionality of a system is evaluated (in the case of MT, cost, speed, usability, portability, readability, fidelity, and so on).

The cloze procedure

The cloze procedure was originally developed by Taylor (1953) as a measure of readability (and hence, comprehensibility) of human-written text, along the lines of

the well-known Flesch scale, and others similar, which grade texts to indicate their suitability for readers of different age ranges. The cloze procedure involves taking the text to be evaluated, masking some words in the text (every 5th word, say), and then asking subjects to guess the missing words. The name "cloze" comes from "closure", this being the term that

> ... gestalt psychology applies to the human tendency to complete a familiar but not-quite-finished pattern—to "see" a broken circle as a whole one, for example, by mentally closing up the gaps. (Taylor, 1953, p. 415)

The readability of the text is directly computed on the basis of the ability of the subjects to guess the missing words correctly. It compares favourably with other measures of readability used in MT evaluation such as rating scales (subjective) or error counts (difficult to implement), and with other more general measures of readability which involve computations based on average word- and sentence-length, such as the Flesch scale (Flesch, 1948) (familiar through its use in Microsoft Word), the Dale-Chall formula (Dale & Chall, 1948), Gunning's (1952) FOG Index, and so on.

We will not discuss here what exactly "readability" is in general, nor whether the cloze procedure accurately measures it. There is ample evidence that whatever it is measuring, the cloze test does so consistently. Results correlate highly with readability indices such as those mentioned in the previous paragraph. The cloze procedure has also been used as a tool in testing foreign-language learners (see Oiler, 1979; Hinofotis, 1987; Brown et al., 1999). Brown et al. stress the difference between using the cloze procedure for "norm-referenced purposes" such as admissions or placement testing and "criterion-referenced purposes" such as diagnostic, progress, or achievement testing, with the implication that the results are less delicate and therefore more reliable in the former case.

For our application (MT evaluation) the term "readability" is used with a quite specific meaning. In general usage, "readability" is supposed to correlate with ease of reading for a reader of a given age, whereas in MT evaluation, the term is used as a quasi-synonym for "intelligibility", which in turn is just one aspect of the generally vague notion of translation "quality".

There are a number of variables in the application of the cloze procedure. The first is the **masking rate and interval,** i.e. the percentage of words to be masked and whether they are chosen at random or at a regular interval, and where the masking starts. A typical case would be every 5th word, starting at the beginning. But one could just as easily decide to keep the first paragraph intact, and then mask 20% of the words at random. Note also that what counts as a "word" needs to be determined. Another major variable is **what counts as a correct answer.** In the extreme case, one would accept only the exact word that has been masked. But, considering that the masked word may be a proper name or a figure or date, then it may be impossible for the subject to guess the missing item correctly, in which case a plausible but wrong guess may be acceptable. Taking this one step further, a near-synonym that preserves the original meaning might be deemed acceptable. In our experiments we investigated the effect of this parameter.

Using cloze to evaluate MT

The cloze procedure was first proposed as a method for MT evaluation by Crook & Bishop (1965), as cited by Halliday & Briss (1977). Extensive efforts to obtain a copy

of Crook & Bishop's report have revealed that the original was probably lost in a fire at Tufts University, so we can only go on Halliday & Briss's summary. Crook & Bishop ran two tests of MT output: in each case every 8th word was eliminated, the two tests differing in that in one only the exact word was accepted as correct, while in the other meaning-preserving synonyms were allowed. In both cases, the score for readability correlated with "quality of translation" (though it is not clear how this was independently measured). The scores also correlated with reading time, measured as an independent variable.

Sinaiko & Klare (1972, 1973) reported two experiments to compare human and machine translations of English into Vietnamese. Klare et al. (1971) had previously used the technique also to compare two different human translations against the English original. In the 1972 paper, they compared an expert human translation with two versions translated by the Logos MT system, one post-edited, the other not. Three passages each of approximately 500 words representing three different levels of technical complexity were tested with two groups of subjects: 88 English speakers read the English-language version of the text to provide a baseline against which to compare the translations. There were 168 Vietnamese-speaking subjects. Two scores were derived from the cloze tests: proportion of correct responses, and proportion of blanks left unfilled. The masking interval was every 5th word. Subjects were also asked to make a subjective judgement about the intelligibility of the texts. The time it took the subjects to read the passage, complete the test and make the judgment was noted. They found that the overall results were as expected: the English original and human translation got the best scores for reading time, subjective rating and cloze test. The machine-translated outputs scored significantly lower, with the less technical material scoring better than the more technical material. One interesting finding was that post-editing the less technical material did not significantly improve the scores obtained. The main significance of this result of course is not that human translations are better than machine translations, which was then (and remains) fairly obvious, but that here was a procedure that could accurately quantify that difference.

A second experiment, reported in Sinaiko & Klare (1973) was very similar. This time there were 141 Vietnamese-speaking subjects, and 57 English-speaking controls. The text used for then test was a single 500-word passage, and the cloze tests were supplemented again by a subjective intelligibility rating and, in addition, a multiple-choice reading comprehension test. The masking interval was again every 5th word, starting with the 2nd word. The results were much the same as in the first experiment, with cloze accuracy score averages of 55% for the human translation, 41% for the post-edited MT output, and 27% for the raw output, these differences being statistically significant.

Despite the simplicity, objectivity and apparent accuracy of this evaluation technique, strangely, we have not been able to find any report of its use for evaluating MT since the early 1970s. Important and large-scale MT evaluations of recent years, such as the JEIDA and DARPA evaluations (Nagao, 1989; White et al., 1994) have used many other techniques, but not the cloze test. The 1996 EAGLES survey does not mention it.

One exception to this observation is the very recent work by Miller (2000), who uses the cloze procedure to evaluate the translation of prepositions by MT systems. Miller took 13 different French texts of between 250 and 400 words in length from the Canadian Hansard corpus, and had them translated into English. Test subjects were

two English monolinguals, and two bilinguals. The monolinguals completed the test in the usual way, whereas the bilinguals were provided with both the cloze texts and the original French text, and asked to complete the test as if they were filling in information missing in a translation of the source text. We should also mention Koutsounouris (2000), who conducted a parallel experiment to the one reported here, and which we will discuss briefly below.

Our own experiment

In our experiment, we tested three MT systems against a human translation, with texts from three different genres, to see whether the procedure can be used to compare MT systems with each other. We selected texts of approximately 500 words in length from three French web pages and had them translated by three MT systems which, subjectively, we judged to be "good", "average" and "poor". We also used the high-quality (presumably, human) translations from the official English versions of the websites. We were interested to see if the cloze procedure would rank the translations in the same order as our subjective judgment.

The three texts chosen were a semi-technical text about cell-phones, a text about cooperation between airlines on route sharing, and a piece about a UNESCO programme to promote women in science. We had the texts translated into English by a professional translator and by three MT systems: the version *of Systran* available via AltaVista's Babelfish service, MicroTac's *French Language Assistant,* and Transparent Language's *EasyTranslator 2.0.* We judge the output quality of these three MT systems subjectively to be ranked in the order given.

Before conducting the full experiment, we ran a small pilot study to practise the experimental procedure and check the methodology. The pilot ran with 8 second-year students of computational linguistics: these were not ideal subjects, given their background knowledge, but were suitable for testing certain aspects of the procedure. This was a useful procedure, as we found that there were certain aspects of the methodology which were unsatisfactory.

The most important aspect of the methodology that we found unsatisfactory, and changed for the full experiment, was the masking interval. As in all previous applications of the cloze test for MT evaluation, we blanked out every 5th word. This gave the subjects 300 blanks to fill in, and several of the subjects complained that the test was too difficult and/or too boring. Others resorted to filling in the blanks with nonsense, e.g.

"... one ... two ... three ... four ...", or
"... ham ... and ... green ... eggs ...", or even
"... sorry ... but ... I ... don't ... know ... what ... to ... write ..."

One can sympathize with the subjects, especially faced with the translations of lower quality, where the frequent blanks really make the task quite daunting - cf. Figure 1. For this reason we decided in the full experiment to change the masking interval to 10 words, starting with the first word. Our pilot experiment also revealed some other problems, such as collusion between subjects, and a lack of clarity in the instructions (subjects had filled some of the blanks with punctuation marks, or left them blank). We also decided to address the problem of low motivation by rewarding subjects with a present on completion of the test.

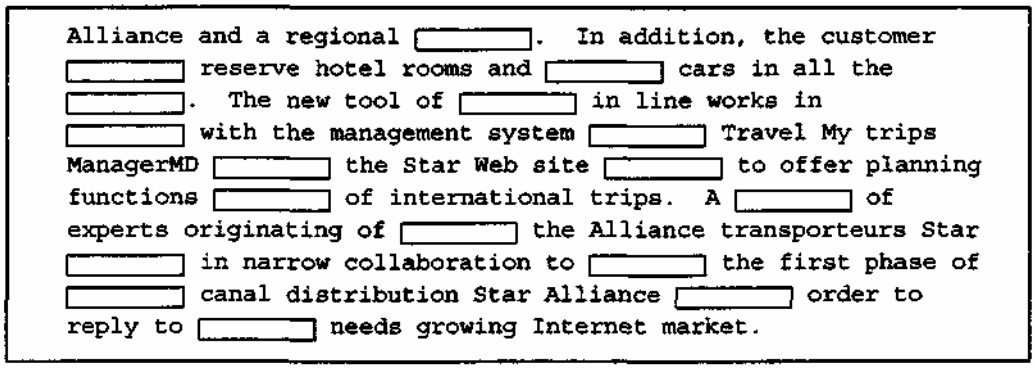

Figure **1.** *Short excerpt of text translated by* EasyTranslator, *with every 5th word blanked out.*

For the full experiment, the subjects were 12 science and engineering students (6 male, 6 female) aged between 19 and 22. All were native English speakers, and had no knowledge of French or linguistics. The instructions for the test began with the following statement in a deliberate attempt to hide the purpose of the test:

> Thank you for volunteering to take part in this experiment. I am conducting it in order to look at what happens when parts of a text are missing and how a human compensates for that. This test simulates such a situation. Words have been removed from the texts and replaced by blanks. We are going to ask you to fill in these blanks. It is not a test of your ability but a test of this scenario.

Subjects were instructed to write one word in each blank, the size of which was not to be taken as indicative of the length of the missing word. They were urged to fill all the blanks, and told that they should make a reasonable guess if they were not sure (or, in the case of proper names, numbers and so on, had no way of being sure). A "word" was defined as any sequence of characters surrounded by blank spaces or punctuation marks. The preparation of the texts and insertion of blank boxes was done with a simple program which included the recognition and separation of punctuation marks. Commas within numbers (e.g. *2,200)* were not treated as punctuation; nor were apostrophes (including liaison apostrophes in untranslated French words), but hyphens were.

The 3 texts x 4 translation modes gives 12 test combinations. Each of the 12 subjects was given three tests to complete, such that each subject saw one version of each text, each one translated in a different mode. This also meant that the 36 tests were evenly distributed over the texts and translation modes, and no two subjects had the same combination of texts and modes. Each subject was given a "pack" of three tests, which they could complete in their own time, and in any order.

Results

The texts contained between 49 and 61 blanks. For simplicity, all scores have been normalised as scores out of 50. We first analyse the results scoring the answers as right (exactly the same as the missing word) or wrong (anything else).

The scores range from 28.57 to 6.57. An analysis of the scores grouped by text show that the scores are fairly evenly distributed (the average scores for the three texts are between 15.23, 17.34 and 19.09). This means that we can say that the texts were equally difficult, and can concentrate on the distribution of scores by translation mode.

Figure 2 shows the average score for each translation mode: as expected, the human translation is the best, and the ranking of the three MT systems reflects our *a priori* subjective judgment.

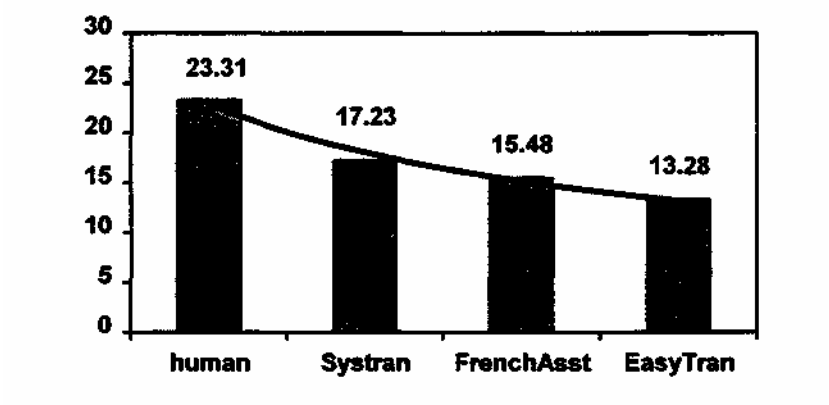

Figure *2. Average scores for each translation mode*

Although the sample size is very small, the differences between the first three modes are statistically significant at the 5% level: only the difference between *French Assistant* and *Easy Translator* is not significant.

If we look at the average scores for each text-mode combination (Figure 3), we see that the scores for the human and *Systran* translations are more "stable" than for the other two modes.

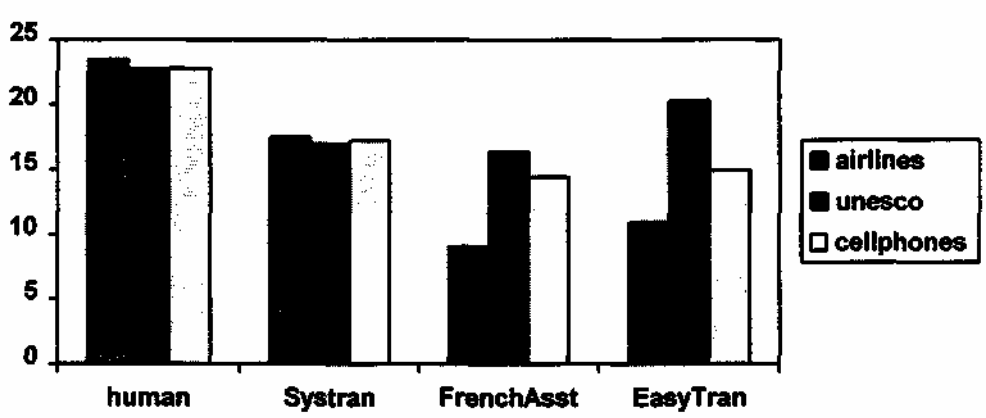

Figure 3. *Average scores for each text-mode combination*

Our conclusion is that the cloze technique does seem to reflect accurately our expectations where the translations are quite good. But as the quality of the output deteriorates, random factors such as the text-genre and even the accidental placement of the gaps to be filled have a much more significant effect on the score.

There is much discussion in the literature on cloze testing about whether to accept as a correct answer only exactly the missing word or whether close synonyms or, in the case of proper names, figures, dates and so on, plausible but incorrect suggestions. It is obvious that acceptable-word scoring will produce higher raw scores, but the debate is whether this impinges on the use of the cloze procedure as an objective measure to rank subjects (or, in our case, text sources). If the improvement on score is uniform, it

should not matter. Mobley (1980) felt that exact-word scoring was more appropriate for scientific texts, while acceptable-word scoring was suitable for fiction. A major drawback with the latter is that it introduces a subjective judgment (i.e. which answers are "acceptable") into the scoring procedure. There is ample evidence (e.g. Anderson, 1976; Oller, 1979; Hinofotis, 1980) that the two scoring methods provide the same rank-order of subjects, and, as Hinofotis states, "the exact word method is the preferred grading procedure in practical terms".

We estimate that between 3% and 6% of the blanks fell in positions where it would be more or less impossible to guess the missing word. Figure 3 illustrates a case in point, though this passage has a higher than normal percentage of such cases. In some cases, a figure or date could be calculated from the surrounding text: a nice example was the following:

...and *a* turnover of 21,3 [ ] of euros (25,0 billion dollars), ...

As well as proper names and dates, for some of the versions the blanks coincided with untranslated words, so it would obviously be unreasonable to require the subjects to provide the original French word, since ignorance of French was a prerequisite of the subjects.

The network [ ] Alliance gathers Air Canada, Air New Zealand, Nipponese All [ ] (ANA), Ansett Australia, Lufthanasa, Scandinavian Airline Systems (HOPPER), Thai [ ] International, United Airlines and Varig, Brazilian conveyor.
Mexicana de [ ] will become member at summer 2000, while the group [ ] Airlines Aviation, which includes Austrian Airlines, Lauda Air and [ ] Airways, recently announced its intention to adhere to the [ ] Star Alliance in time for the estival peak period [ ] year 2000.

Figure 3. *Extract from* Systran *translation of "airlines" text*

We re-scored the tests using a more complex scoring system which gave a half mark for wrong but plausible answers, including near synonyms, translations of untranslated words, and unguessable literals. Figure 4 shows the averages for the four translation modes next to the results already seen in Figure 2. The difference is negligible.

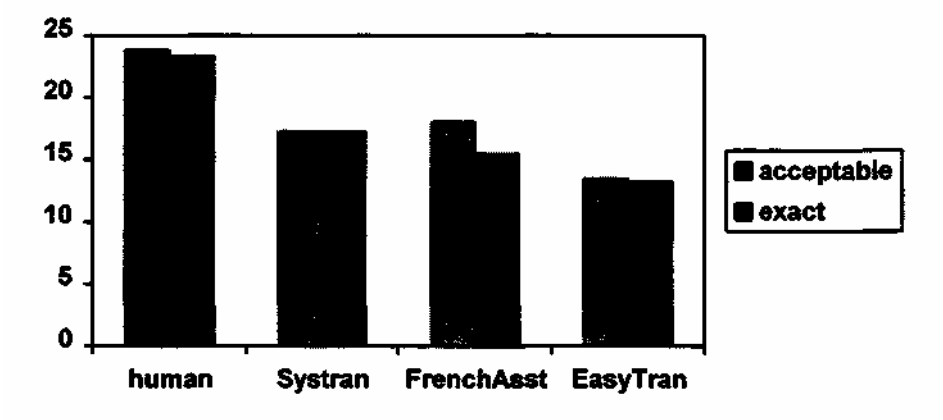

Figure 4. *Average scores using both scoring methods for each translation mode*

In fact, this more complex scoring method did not affect the overall results at all, which suggests that the procedure as originally conceived is entirely adequate.

Another experiment

It is worth discussing briefly here a further experiment conducted by Koutsounouris (2000) under the present first author's direction. Like our own experiment, Koutsounouris set out to see if the cloze procedure could be used to give a comparative evaluation of different MT systems. The same four translation modes were used, though this time into rather than out of French. An unusual aspect of this test was that the subjects were *not* native speakers of the target language. Four 600-word texts were taken from the May 2000 issue *of National Geographic*, as found on their web site. French translations were provided by a human and the three systems already named. The masking interval was again 10. The subjects were 16 Greek nationals who were teachers or advanced students of French, with (self-assessed) only basic knowledge of English. The 4 texts x 4 modes gives 16 combinations; each subject was given 4 different texts each translated in a different mode (so each subject saw one example of each text and one example of each mode), giving us 64 test results. Unlike in our own experiment, subjects were restricted to 30 minutes to complete the test, and so were allowed to leave blanks.

Three sets of scores were derived from the tests: exact-answer, acceptable-answer (as above) and also number of blanks left unfilled. The overall average scores taking exact answers only are shown in Figure 5 (note that in this experiment scores are shown as percentages). Compared to Figure 2 above we see the effect is even more marked, with the human and *Systran* translations clearly better than the other two, which are rated about equal (if anything the supposedly inferior *Easy Translator* scores better, as is seen even more clearly in Figure 6).

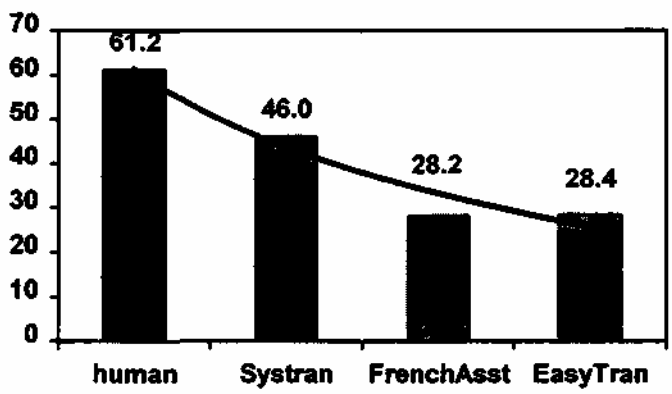

Figure 5. *Second experiment: average scores for each translation mode*

We get the same picture as before if we look at scores including acceptable answers; and the percentages of spaces not left blank also follows a similar pattern (Figure 6).

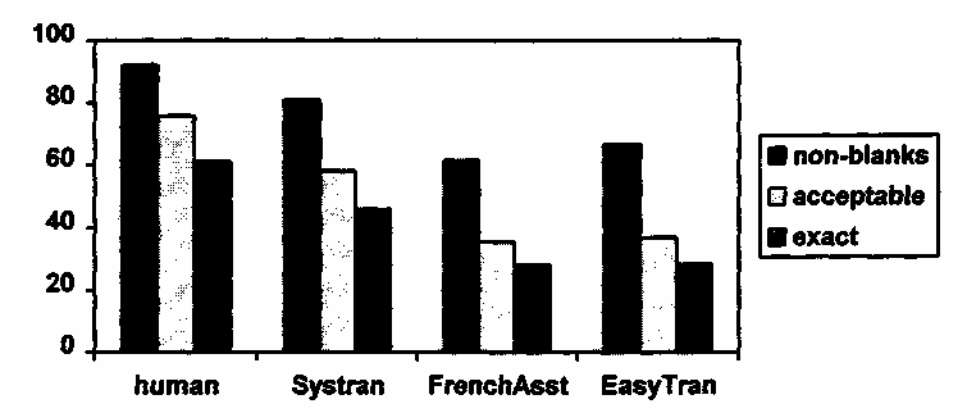

Figure 6. *Average scores using both scoring methods, and average number of spaces not left blank, for each translation mode*

Conclusions

Both the experiments reported here were with very small samples - too small to permit reliable tests of statistical significance for instance. So we would need to confirm our results with a bigger sample. However, tentatively we can confirm that the cloze procedure seems to be a reliable indicator of relative quality of translation, at least as far as readability goes.

We found that the **masking interval** should not be too small: the interval of every 5th word suggested in most of the literature on cloze testing may be suitable for tests with real texts written by humans, but on top of the distortion caused by machine translation, a larger interval seems more appropriate. If the interval is too large, then we need a correspondingly larger text to get a sufficient number of blanks to counter the effect of random masking which can lead to blanks which are either too easy or too difficult to fill in. An interval of 10 means that 10% of the words in a text are masked, and a text of 500 words seems to provide a reasonable sample within the framework of a test that can be completed **in a reasonable time.** All these are important considerations in evaluation methodology.

Another aspect which we have not yet mentioned, but which is fairly obvious is the **choice of subject matter:** we took care that the texts chosen dealt with topics more or less unfamiliar to the subjects, so that for the most part they had to use linguistic rather than factual knowledge to fill the blanks. This is not entirely avoidable, as in the example where a list of countries variously (depending on the version) appeared as follows:

The United _____ and Canada,
Australia and New _____,
Hong _____, _____ Africa, ...

Another text included names of airlines.

Importantly, we feel confident that the **exact-answer scoring method** is adequate, and that allowing near synonyms and so on does not give a different result. This means that the test can be administered and scored quite objectively. One could even envisage mechanising the entire process (see for example Pilypas, 1997) so that all the experimenter would have to do would be find some texts and some subjects!

On a less positive point, we note that the discriminatory power of the method diminishes with the quality of the MT system. We could say that there seems to be a lower threshold below which the test is so difficult - the text is so garbled - that we cannot really rank systems. All we can say is that they fall below a kind of minimum level for readability. This in itself may be a useful finding, if, experimentally, we can establish what that threshold is.

Acknowledgments

Thanks to Jackie Ellis (Tufts University), John Hutchins (UEA), Ed Hovy (USC) and John White (Litton PRC) for help in tracking down archive material, and to Antonis Koutsounouris (UMIST) and Keith Miller (Mitre Corp) for background material.

References

Anderson, J. (1976) *Psycholinguistic Experiments in Foreign Language Testing.* St Lucia, Qld.: University of Queensland Press.

Arnold, D., R.L. Humphreys and L. Sadler (eds) (1993) Special Issue on Evaluation of MT Systems. *Machine Translation* 8.1-2.

Brown, J.D., A.D. Yamashiro and E. Ogane (1999) Three strategies for "tailoring" cloze tests in secondary EFL. *TUJ Working Papers in Applied Linguistics* 14, published online by Temple University, Japan. http://www.tuj.ac.jp/tesol/press/papers0014/brownetal.html

Crook, M.N. and H.P. Bishop (1965) *Evaluation of Machine Translation, Final Report.* Institute for Psychological Research, Tufts University, Medford, Mass.

Dale, E. and J.S. Chall (1948) A Formula for Predicting Readability. *Educational Research Bulletin* 27, 11-20, 37-54.

EAGLES (Expert Advisory Group on Language Engineering Standards) (1996) *Evaluation of Natural Language Processing Systems, Final Report.* Report for DGXIII of the European Commission, available at http://issco-www.unige.ch/projects/ewg96/ewg96.html.

Flesch, R. (1948) A new readability yardstick. *Journal of Applied Psychology* 32, 384-390.

Fotos, S.S. (1991) The cloze test as an integrative measure of EFL proficiency: a substitute for essays of college entrance examinations? *Language Learning* 41, 313-336.

Gunning, R. (1952) *The Technique of Clear Writing.* New York: McGraw-Hill.

Halliday, T.C. and E.A. Briss (1977) *The Evaluation and Systems Analysis of the Systran Machine Translation System.* Report RADC-TR-76-399, Rome Air Development Center, Griffiss Air Force Base, NY.

Hinofotis, F.B. (1980). Cloze as an alternative method of ESL placement and proficiency testing. In J.W. Oiler and K. Perkins (eds), *Research in Language Testing,* Rowley, MA: Newbury House, pp. 121-128.

Hinofotis, F.B. (1987) Cloze testing: An overview. In M. Long and J. Richards (eds) *Methodology in TESOL,* Rowley, MA: Newbury House.

Klare, G.R., H.W. Sinaiko and L.M. Stolurow (1971) *The Cloze Procedure: A Convenient Readability Test for Training Materials and Translations.* Report DAHC15 67 C 0011, Institute for Defense Analyses Science and Technology Division, Arlington, Va.

Koutsounouris, A. (2000) *Readability Metrics for MT Evaluation.* MSc dissertation, Department of Language Engineering, UMIST, Manchester.

Miller, K. J. (2000) *The Lexical Choice of Prepositions in Machine Translation.* PhD dissertation, Georgetown University, Washington, DC.

Mobley, M. (1980) The readability of school textbooks. *Language for Learning* 2, 11-19.

Nagao, M. (ed.) (1989) *A Japanese View of Machine Translation in Light of the Considerations and Recommendations Reported by ALPAC, U.S.A.* Tokyo: Japan Electronic Industry Development Association (JEIDA).

Oller, J.W. (1979) *Language tests at school.* London: Longman.

Pilypas, H. (1997) *The Use of the Computer as a Tool for Testing Reading Comprehension,* BEd dissertation, School of Education, Flinders University of South Australia, Adelaide; see http://wwwed.sturt.fluiders.edu.au/edweb/programs/bedtheses/helen/online1.htm

Sinaiko, H.W. and G.R. Klare (1972) Further experiments in language translation: readability of computer translations. *ITL* **15,**1-29.

Sinaiko, H.W. and G.R. Klare (1973) Further experiments in language translation: a second evaluation of the readability of computer translations. *ITL* **19,**29—52.

Sparck Jones, K. and J.R. Galliers (1995) *Evaluating Natural Language Processing Systems: An Analysis and Review.* London: Springer.

Taylor, W. (1953) "Cloze Procedure": a new tool for measuring readability. *Journalism Quarterly* **30,**415—433.

Vasconcellos, M. (ed.) (1994). *MT Evaluation: Basis for Future Directions.* Proceedings of a workshop sponsored by the National Science Foundation, San Diego, California.

White, J.S. (forthcoming) How to evaluate Machine Translation systems. To appear in H. Somers (ed.), *Computers and Translation: a Handbook for Translators,* to be published by John Benjamins.

White, John S., Theresa O'Connell and Francis O'Mara (1994) The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas,* Columbia, Maryland, pp. 193-205.

Wilks, Y. (1994) Keynote: Traditions in the evaluation of MT. In Vasconcellos (1994), pp. 1-3.