Granularity in MT Evaluation

Florence Reeder

MITRE Corporation
7515 Colshire Drive
McLean VA 22102
freeder@mitre.org

John White

Northrop Grumman Information Technology
4801 Stonegate
ChantillyVA 20105
John.White@northropgrumman.com

Abstract

This paper looks at granularity issues in machine translation evaluation. We start with work by (White, 2001) who examined the correlation between intelligibility and fidelity at the document level. His work showed that intelligibility and fidelity do not correlate well at the document level. These dissimilarities lead to our investigation of evaluation granularity. In particular, we revisit the intelligibility and fidelity relationship at the corpus level. We expect these to support certain assumptions in both evaluations as well as indicate issues germane to future evaluations.

Keywords

Evaluation; granularity, fidelity, intelligibility, DARPA corpus.

Abstract

This paper looks at granularity in machine translation evaluation. We start with work by (White, 2001) who examined the correlation between intelligibility and fidelity at the document level. His work showed that intelligibility and fidelity do not correlate well at the document level. These dissimilarities lead to our investigation of evaluation granularity. In particular, we revisit the intelligibility and fidelity relationship at the corpus level. We expect these to support certain assumptions in both evaluations as well as indicate issues germane to future evaluations.

Introduction

Machine Translation Evaluation has been costly to perform (e.g., White & O'Connell, 1994; Doyon, et al., 1998). Costs include corpus collection and vetting, arranging for human evaluators, controlling for human factors, etc. Therefore, for nearly as long as Machine Translation (MT) evaluations have existed, MT practitioners have sought less costly MT evaluation (MTE) techniques (e.g., van Slype, 1979). Two paths have arisen in the quest for reducing the amount of time and expense involved in MTE.

The first is to find metrics whose values correlate well with other quality judgments. That is, if one could find a correlation between the adequacy of MT output and the informativeness of it, one of these two metrics could safely be eliminated from testing, reducing the overall evaluation cost.

The second path, which gained prominence only recently, is to look for automated evaluation methods. The advent of automated evaluation methods represents a search for metrics which are similar to the Word Error Rate (WER) measure from speech transcription (Jurafsky & Martin, 2000). A single, agreed upon metric which correlates well with human quality judgments could do for machine translation (MT) what WER did for speech transcription. That is, provide an agreed-upon method for evaluating systems which also permits comparisons both horizontally (across systems) and vertically (across evaluations). The more frequent evaluations possible with automated metrics could facilitate large gains in MT development, by providing an accessible metric for constant system testing. Additionally, these metrics could even be embedded in MT system development algorithms to learn MT. The developers of these automated metrics seek ones which are straight-forward, relatively rapid and which correlate well with human quality judgments. One metric, BiLingual Evaluation Understudy (BLEU), reports a strong correlation with human judgments (Papenini et al., 2002b).

Evaluation Granularity

Examination of correlations along the two paths have yielded questions about evaluation granularity. The granularity of the evaluation is defined as the lowest amount of text for which a final score can be calculated. Early evaluations focused on the sentence level with scores given on a sentence by sentence basis. Regardless of the scale used (e.g., ALPAC, 1966; Wilks, 1992; Corston-Oliver et al., 2001), the judgment of the evaluators concerned the sentence itself. It was recognized even then (e.g., ALPAC, 1966; van Slype, 1979) that the sentence was not necessarily the right level of granularity. For adequacy, a sentence was often deemed as too long. For intelligibility, intra-sentential phenomena encouraged looking at something larger than a sentence or looking at the sentence in context (e.g., van

Slype, 1979; White & O'Connell, 1994). Often, evaluations were limited by the number and availability of raters and size of test corpus available. Evaluations designed for statistical relevance tended to be large scale, requiring hundreds of raters and large amounts of resources (e.g., White & O'Connell, 1994).

In the search for correlated human judgments, two metrics which intuitively should correlate to some degree - adequacy and fluency - have not in practice (e.g., White, 2001) correlated at the text level. This lack of correlation has caused closer examination at evaluation granularity to find the set points of these metrics.

To address the need for automated evaluation, recent MT evaluations have tended towards using techniques which are best served by large bodies of data. Evaluators in this vein (e.g., Papenini et al., 2002a; Papenini et al., 2002b; Melamed et al., 2003) have relied on large corpora of reference translations. The basis of score acceptance is often its correlation to aggregated human judgments. For instance, Papenini et al. (2002a) show the correlation of the BLEU scores to human judgments at the corpus level; where the corpus consists of over 100 texts (per system) with each text at roughly 400 words. At this level of granularity, the metric correlates well with the human judgments. Since BLEU and metrics like it rely on multiple reference translations or large document collections to ensure statistical reliability, it tends to work at the document or corpus level.

Much of BLEU's strength derives from the fact that it was shown to correlate well (R^2 ~0.95) with human judgments on the corpus level (Papenini et al., 2002a). At the sentence level, BLEU exhibits some anomalies, particularly for poorly translated sentences. Unless a four-gram can be found, the algorithm as distributed gives a zero score to the sentence[1]. Therefore, the question of the lowest practical granularity for evaluation arises here as well.

DARPA 1994 Data Set

The goals and results of the DARPA Machine Translation Initiative of the early 1990's have been described in numerous publications (e.g., Doyon et al., 1998, White 1995, White 2001). The initiative yielded an evolving evaluation methodology culminating in 1994 (known as "3Q94") with:

- a large corpus of multiple machine (and control) translations of hundreds of newspaper articles in French, Spanish, and Japanese;
- A methodology for capturing the *adequacy, fluency, and informativeness* of a translated passage; and
- Measures captured at the sub-sentence, sentence, text, and system levels, comprising over 200,000 decision points scored by non-specialist human evaluators.

Scoring for these metrics were defined along these lines:

- *Adequacy,* where evaluators were given texts arranged with an expert translation on one column, MT output (or control) on another column, and a space for scoring, on a 1-5 anchored scale. The evaluators determined the extent to which meaning conveyed in a segmented portion of the expert translation (generally sub-sentence) was conveyed in the MT output text.
- *Fluency,* in which evaluators looked at output texts and scored on an anchored 1-5 scale each sentence, on the extent to which the sentence was intuitively acceptable to a native speaker, was well formed, grammatically correct, and makes sense in the context of the overall text.
- Scores were combined by averaging the 1-5 scores and dividing the average by 5.

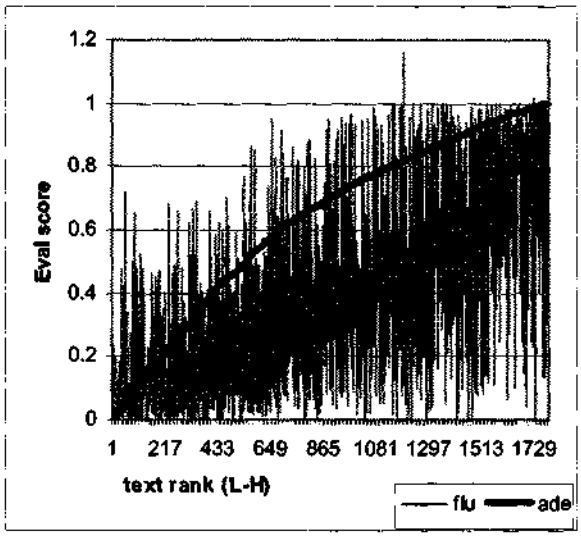

Figure 1: Fluency scores compared to adequacy curve.

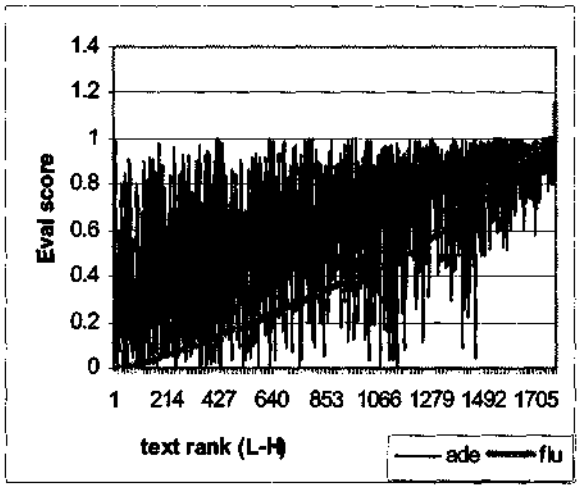

Figure 2: Adequacy compared to fluency curve.

Correlating Adequacy and Fluency

As reported earlier (White 2001, White 2000), it appears from statistical analyses of the 3Q94 data that a

[1] The gram profile and the combination of the n-gram scores can be changed to yield non-zero results.

predictable correlation could be found among the measures fidelity (adequacy) and intelligibility (fluency). To find the nature of the correlation, and whether it was consistent through the range of measures, we performed a simple comparison of the text-by-text fluency scores, mapped onto the sequence of adequacy scores, lowest-to-highest (Figure 1). The result of this mapping, as well as the opposite one (adequacy mapped onto fluency, Figure 2) is difficult to discern in its raw form.

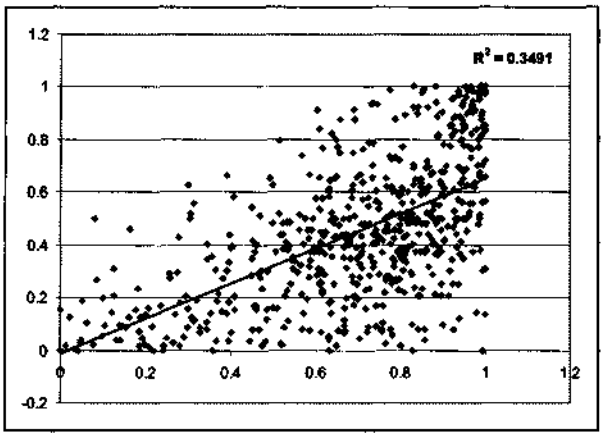

Figure 3: Adequacy versus Fluency for French

The hypotheses were that the correlation, if there was one, can be determined at the text level, and will show either a roughly parallel mapping between adequacy and fluency (i.e., the texts with better fluency scores uniformly have better adequacy scores); or, the two could diverge at some point on the slope, and then re-converge, symmetrically (i.e., the worst adequacy texts are the worst fluency, the best the best, with some divergence in between whose correlation can be captured). We see from Figure 1 and Figure 2 that neither of these hypotheses is supported by the raw, text-by-text comparison of intelligibility and fidelity.

In looking at the scatter plot of fluency versus adequacy, we see a low correlation between fluency and adequacy in the French-English corpus on a text by text basis (Figure 3). The R^2 value of 0.35 is very low, showing little correlation between the fluency scores and the adequacy scores. We now aggregate the documents to see whether the correlation can be improved. Taking the French-English results, we average for five document chunks and compute the adequacy and fluency correlation of this (Figure 4). Documents were taken in order of appearance so that the first five texts for the Candide system represent the first data point in the graph and so on. By averaging, we increase the granularity of the text size to roughly 2000 words per data point. The correlation here improves to 0.67, although it is still a weak correlation.

Next, we average across groups of ten documents (Figure 5). At this point, both clustering along the correlation line and correlation improve. Note that by this point, the granularity is at 4000 words. In averaging across 20 documents, the correlation shows a stronger correlation (Figure 6) at 0.74. Finally, we look at correlations at the corpus level (Figure 7) where the correlation between fluency and adequacy is improved at $R^2 = 0.85$, although the number of words involved is roughly 40000. One word of note, that will be discussed later is that removing the scores for the human translations improves the correlation significantly at the corpus level to an R^2 value of 0.92 (Figure 8).

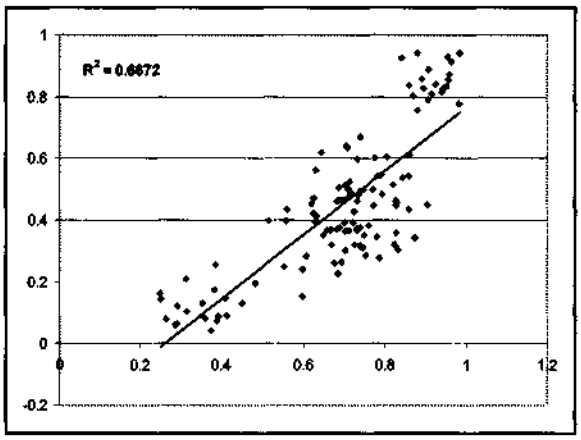

Figure 4: Adequacy versus fluency average of five scores

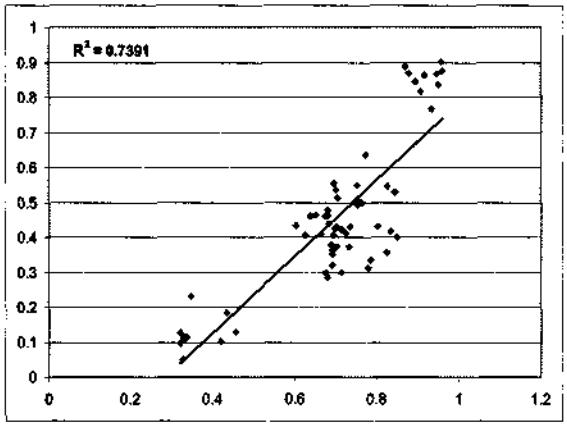

Figure 5: Adequacy versus fluency average of ten scores

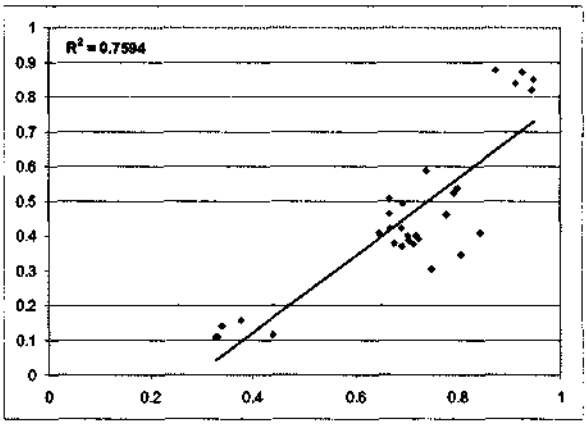

Figure 6: Adequacy versus fluency average of 20 scores

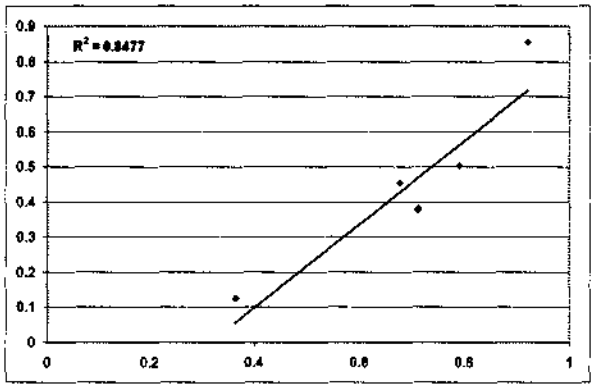

Figure 7: Correlation for System and Human Scores

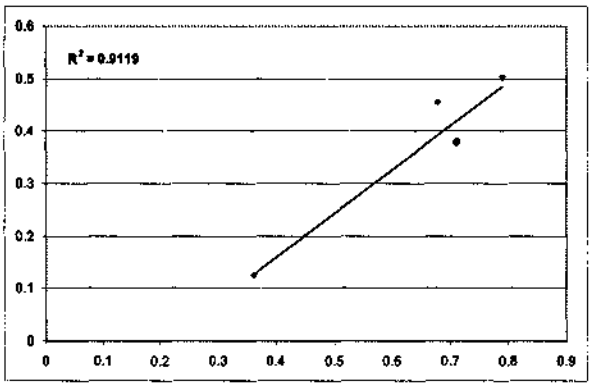

Figure 8: Systems without Human Translations

Post-evaluation analysis-of-variance performed at the time of the 3Q94 series (White et al. 1994) suggest that beyond 20 documents in a corpus, the rate of reduction in the extraneous variation slows significantly. Meanwhile, the aggregate measure performed on averages of 20 texts shows a value of $R^2 = .76$, a strong predictive indicator. These stochastic aggregations thus appear to agree in general with the 3Q94 post-evaluation findings that a corpus of at least 20 documents is sufficient.

Automated Metric Correlations

Recent work in automated MT evaluation metrics have shown the correlations between the metric and human judgments (e.g., Papenini et al., 2001) for the DARPA 1994 corpus. These correlations were calculated at the corpus or document collection level rather than by individual documents. With the results from the previous section on adequacy and fluency, we now revisit the BLEU metric.

The DARPA 1994 French data was run through the BLEU scorer with scores prepared on a document by document basis. These scores were then compared to the DARPA-1994 scores of fluency and adequacy (Figures 9, 10). As can be seen, at the document level, the correlation is not very good, although equivalent to that of intelligibility and adequacy on a document by document basis. The correlation of BLEU to adequacy is 0.33, based on roughly 600 data points. The correlation of BLEU to fluency is better, at about 0.45 for the same number of data points.

Repeating the process of averaging over documents, and thereby increasing the size of the sample, we see that the correlation improves as the sample size increases (Figures 11-14). Even an aggregation across five documents, increasing the sample size to 2000 words shows a much stronger correlation for the two metrics, particularly for fluency with $R^2 = 0.69$ (adequacy is 0.45). By twenty documents or 8000 words, the correlation is strong enough to claim BLEU as a good predictor of adequacy or fluency with R^2 of 0.82 and 0.95 respectively. Much like fluency and adequacy, the corpus-based correlation is very strong, particularly if human scores are removed (Table 1).

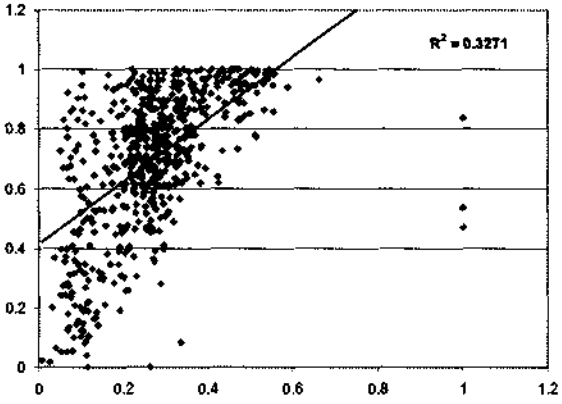

Figure 9: Adequacy versus BLEU for French-English

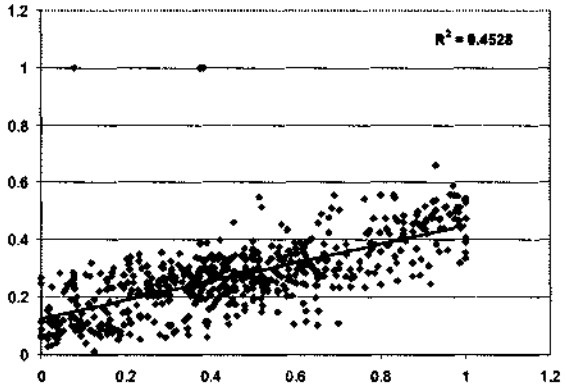

Figure 10: BLEU versus Fluency for French-English

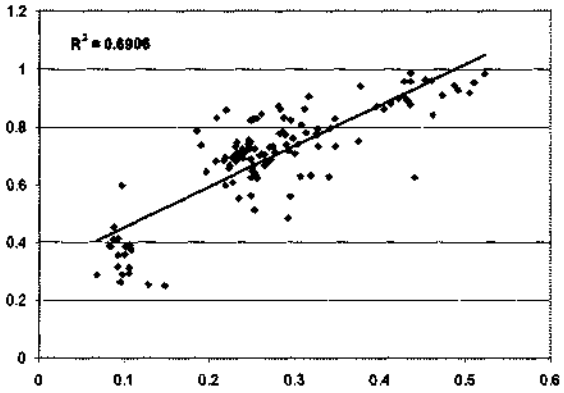

Figure 11: Adequacy versus BLEU for 5 documents

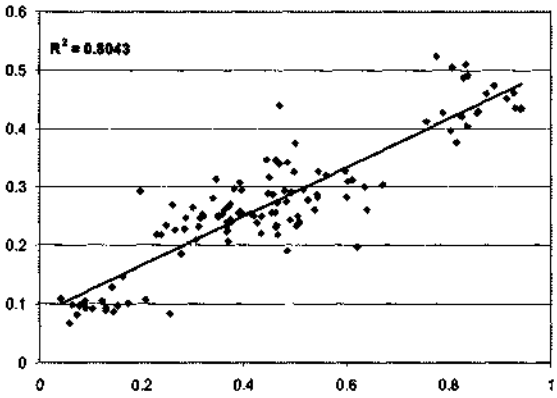

Figure 12: BLEU versus Fluency for 5 documents

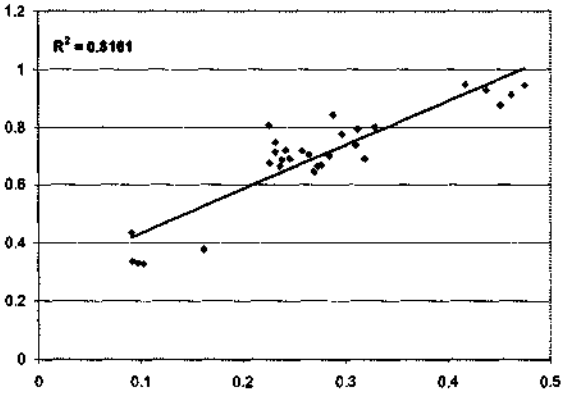

Figure 13: Adequacy versus BLEU for 20 documents

	Adequacy	Fluency
With HT	0.89	0.99
Without HT	0.91	0.99

Table 1: BLEU scores for Corpus

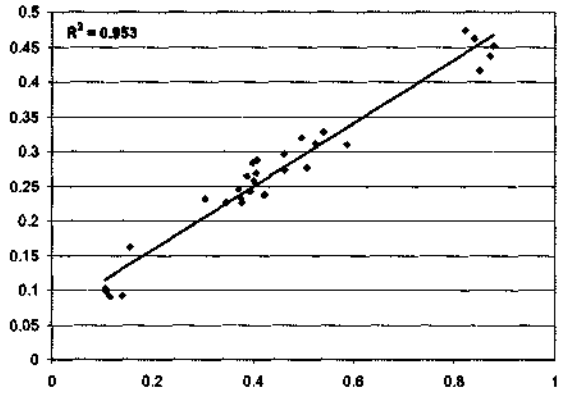

Figure 14: BLEU versus Fluency for 20 documents

Conclusions

Intelligibility measures on fragments, sentences, or single documents will not show the correlation between intelligibility and fidelity. Intelligibility and fidelity do correlate at the multi-document, corpus level. It may thus be possible to use automatic measures of how human-like a translation appears to be (thus intelligibility; eg., e.g., e.g.,) to predict the correct capture and representation of the information conveyed (thus fidelity). However, several cautions must be administered at this time. First, there must be preventions against gaming (outputting the same fluent output for every input, for example). Second, meaningful correlations should be based on a corpus size of no less than 20 documents and roughly 4000 words. This is not a surprise, as these metrics advertise that they are dependent on the law of numbers to be meaningful, but does serve as a caution to those who try to use the metric for a purpose other than intended. The evaluation issue is not a solved one as we need finer-grained metrics for smaller data sets.

Bibliographical References

Carroll, J. (1966). An experiment in evaluating the quality of translations. In Pierce, J. *Language and Machines: Computers in Translation and Linguistics.* Report by the Automatic Language Processing Advisory Committee (ALPAC). Publication 1416. National Academy of Sciences National Research Council

Corston-Oliver, S., Gamon, M. & Brockett, C. (2001). A machine learning approach to the automatic evaluation of machine translation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-01).*

Doyon, J., Taylor, K., & White, J. (1998). The DARPA Machine Translation Evaluation Methodology: Past and Present. *Proceedings of AMTA-98.* Philadelphia, PA.

Jurafsky, D. & Martin, J. H. (2000). Speech and Language Processing. Prentice-Hall.

Melamed, I. D., Green, R. & Turian, J. (2003). Precision and Recall of Machine Translation. Proteus technical

report #03-004, a revised version of the paper presented at NAACL/HLT 2003, Edmonton, Canada

Papineni, K., Roukos, S., Ward, T. & Zhu, W-J. (2002b). Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of ACL-2002,* Philadelphia, PA.

Papineni, K., Roukos, S., Ward, T., Henderson, J., & Reeder, F. (2002a). Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results. In *Proceedings of Human Language Technology 2002,* San Diego, CA.

Van Slype, G. (1979). *Critical Methods for Evaluating the Quality of Machine Translation.* Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management. Report BR-19142. Bureau Marcel van Dijk.

White, J. (1995). Approaches to Black-Box Machine Translation Evaluation. Proceedings of MT Summit 1995. Luxembourg.

White, J. (2000). Toward an Automated, Task-Based MT Evaluation Strategy. Proceedings of the Workshop on Evaluation, Language Resources and Evaluation Conference, LREC-2000. Athens, Greece.

White, J. (2001). Predicting Intelligibility from Fidelity. Proceedings of the Workshop on Evaluation, MT Summit VI, Santiago, Spain.

White, J., & O'Connell, T. (1994). The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. *Proceedings of the 1994 Conference, Association for Machine Translation in the Americas*

Wilks, Y. (1992). Systran: it obviously works, but how much can it be improved? Newton: 166-188.