Approaches to Black Box MT Evaluation

John S. White
PRC Inc.
1500 PRC Drive
McLean, VA 22102 USA
white_john@prc.com

Abstract. In the course of four evaluations in the Advanced Research Projects Agency Machine Translation series, evaluation methods have evolved for measuring the core components of a diverse set of systems. This paper describes the methodologies in terms of the most recent evaluation of research and production MT systems, and discusses indications of ways to improve the focus and portability of the evaluation.

0. Introduction. Over the past four years, a set of evaluation methodologies have evolved within the MT initiative of the U.S. Advanced Research Projects Agency (ARPA). The ARPA program has faced unique challenges for evaluation, because the systems participating differ radically in the linguistic approach, their level of maturity, and the languages translated.

The differences among these systems have made a black-box orientation to evaluation inevitable. While such an orientation differs from the methods that might be employed in the evaluation of a particular system by that system's developers, there are nevertheless certain advantages to the black box approaches in determining the focus and metrics of evaluation.

This paper describes the methodologies of the ARPA program, in terms of their objectives, results, and evolution, and discusses analyses of the methods themselves to continue to improve the process.

1.0. Background. The ARPA initiative in machine translation began in 1991 as part of the Human Language Technologies Program. Three projects in MT research were sponsored under the initiative, with voluntary participation by several commercial and institutional MT organizations. The sponsored projects were: Candide (IBM Watson Research Center), a statistical modeling approach, translating French to English (Brown et al., 1993); Pangloss (Center for Machine Translation, Computing Research Laboratory, and Institute for Information Science), using knowledge-

based approaches, translating Spanish and Japanese to English (Frederking et al., 1993); and Lingstat (Dragon Systems, Inc.), using a combination of modeling and rule-based approaches, translating Japanese and Spanish to English (Yamron et al., 1994).

Many organizations have provided production MT systems for these evaluations, principally to assist ARPA's mission by helping to determine industry/discipline benchmarks. A significant goal of the ARPA MT Evaluation program, in return, is to provide a useful set of evaluation processes for a general standard. In the most recent test-evaluation cycle of August - November 1994 (the "3Q94" evaluation), the following production systems participated:

- the Sietec METAL system (French - English);
- the Nippon Electric PIVOT system (Japanese - English);
- Globalink Power Translator (French and Spanish - English);
- the Pan American Health Organization SPANAM system (Spanish - English);
- Systran (French, Japanese, and Spanish - English);

  the SOCATRA XLT system (French - English).

2.0. Testing. ARPA's mission and focus has been toward advancement of the "core" technologies underlying MT. Even though it is fully recognized that all MT systems in the foreseeable future will use human-assisted techniques as part of their translation processes, testing the "core" must distill the performance of human interaction and peripheral support tools from the performance of the internal algorithms that render strings of one language into strings of another. There is of course a difference between algorithms that do all of the translation processing prior to (or after) the human participation, and those which use mixed-initiative interaction during the course of processing. But even here, the reasoning is that the essential approach to translation is encoded in the automatic functions that accomplish the rendering, regardless of the point or periodicity of human interaction. From this reasoning the test methods have moved toward fully-automated outputs, which preclude test-specific lexical update, pre-editing, post-editing, and interaction (the aspects of these human-assisted methods that are pertinent to system training are discussed further in section 5.3; see also White et al., 1994.).

In the 3Q94 test, each system translated 100 general newspaper articles of approximately 400 words, derived from standard news retrieval services. As

noted, no human processing of these translations was done, including lexical update.

The collection of translations was augmented with two expert human translations of the same material from professional translation agencies. One set served as a reference for two of the three evaluation methods described in Section 3, and the other as a control in the evaluation itself. A set of translations from the previous evaluation (January 1994; White et al. 1994, 1995) were included in the collection to assist in factoring evaluator effects from the evaluation results.

3.0. Evaluation. The ARPA MT evaluation methods are based on human judgments. This is because assessment of the correctness of any translation - even by professional humans - will vary widely among experts and novices alike. The best approach to handling this highly variable subjectivity is to use it as the basis for measurement, decomposing these judgments into relatively small units, focused and controlled by separate evaluations for adequacy, fluency, and informativeness of outputs.

The three evaluation measures were administered in 200 sets of evaluation materials, representing 100 collections of output designed to: avoid occurrence of more than one translation of any source text; avoid more than one occurrence of any system's output in any one evaluation; limit halo and learning curve effects by avoiding repetition of certain output sequences; and limit fatigue and other human factors.

3.1. Adequacy. The objective of the adequacy is to determine the extent to which all of the content of a text is conveyed, regardless of the quality of the English in the output. In this evaluation, expert reference translations were divided along syntactic constituent lines into meaningful sentence fragments. The evaluator was asked to look at each fragment, and judge (on a 1-5 scale) the extent to which the information in the fragment is present in the side-by-side translation. The results were computed by averaging the judgments over all of the decisions in the translation set, and then mapping this result onto a 0-1 scale.

3.2. Fluency. The objective of the fluency evaluation is to determine how much like "good English" a translation appears to be, without taking into account the correctness of the information. In this evaluation, evaluators made intuitive judgments on a sentence by sentence basis for each translation (on a 1-5 scale again), without referring to any reference text. The results of the fluency evaluation were compiled in the

same manner as those for the adequacy.

3.3. Informativeness. The objective of the informativeness evaluation is to measure a system's ability to produce a translation that conveys sufficient information so that someone can gain necessary information from it. It is in the form of a multiple choice test, rather like an reading comprehension examination except that it is the content of the reading, and not the person answering the questions, that is being tested (cf. Church et al., 1991). The informativeness evaluation in 3Q94 consisted of six questions for each text used in the tests. Developed from the reference set of expert translations, the six questions each had six possible answers, including "none of the above" and "cannot be determined". The relevant question set was appended to each translation (including the expert human control). The results were computed as the number of right answers for each translation, averaged for all outputs of each system, and mapped onto a 0-1 scale.

4.0. Results of the evaluation. Figures 4-1 through 4-3 show the results for French-English, Japanese-English, and Spanish-English respectively. The figures compare the 3Q94 results with the previous evaluation (January, 1994, hereafter "1Q94"). As we will discuss in section 5.0, the more mature systems, those which probably have a more comprehensive coverage of lexical or other types of knowledge, seem to do better than newer systems (especially Pangloss Japanese and Lingstat Spanish, which were new for this evaluation). Candide, which has been in development longer than the other research systems, does better, comparable to the mature production systems.

Comparison with the 1Q94 evaluation results show an ambiguous result, seeming to do worse for fluency in the 3Q94 evaluation. However, this issue was resolved by comparison of the scores for the 1Q94 translations that were included in the 3Q94 evaluation. Since these translations are the same as were evaluated in the 1Q94 evaluation, the 3Q94 scores for them serve as a normalizing factor for making a truer comparison of current results with the previous evaluation. This comparison concludes that a 0.6 increment in fluency scores across the board is a closer indication of the true comparative score, and this difference indicates that the systems have in fact improved.

5.0. Analysis of the methodology. The evolution of the evaluation methods has been oriented toward segregating the measurement of the core approaches from other variables,

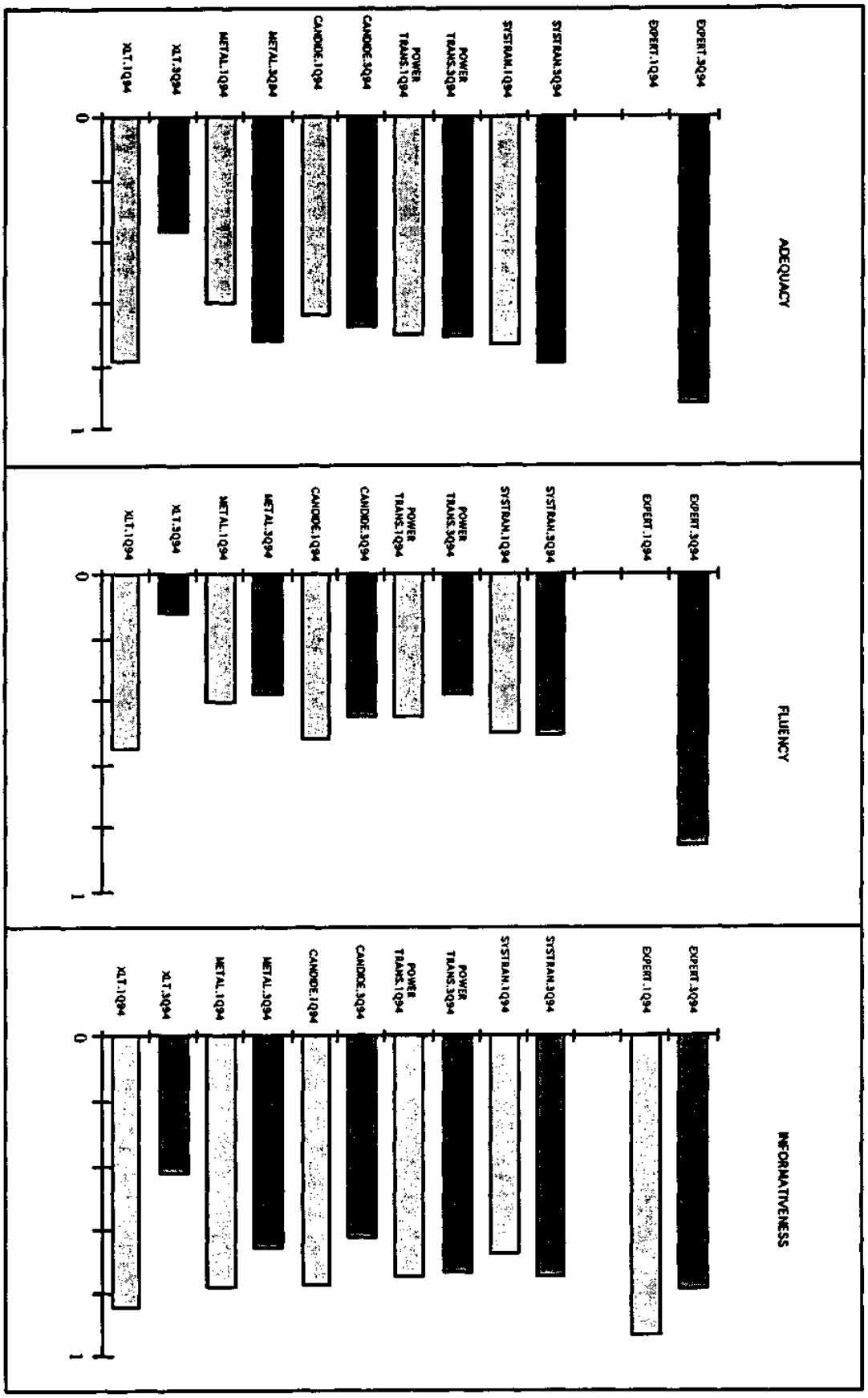

Figure 4-1. Results of 3Q94 French-English, compared to 1Q94

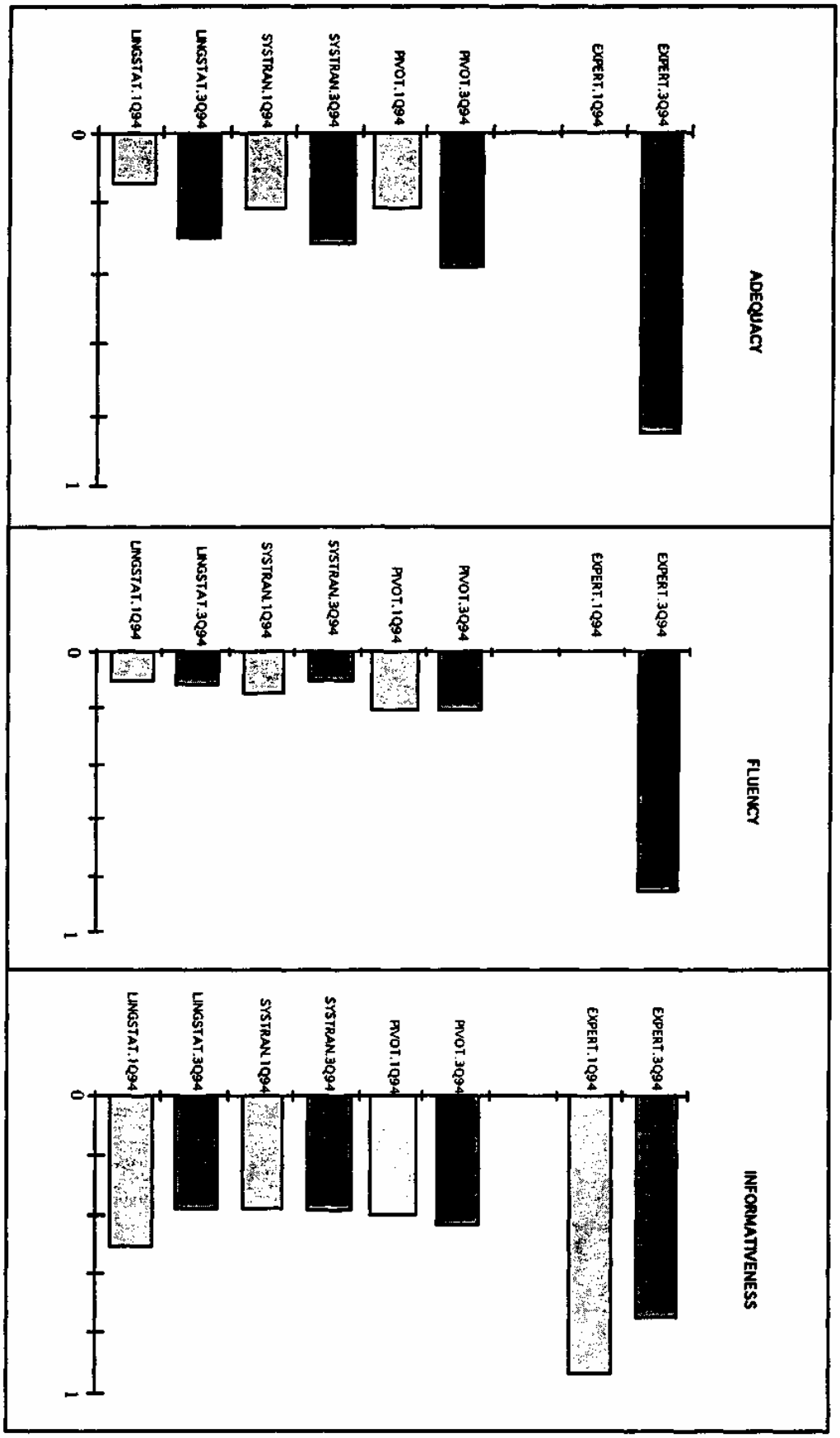

Figure 4-2. Results of 3Q94 Japanese-English, compared to 1Q94

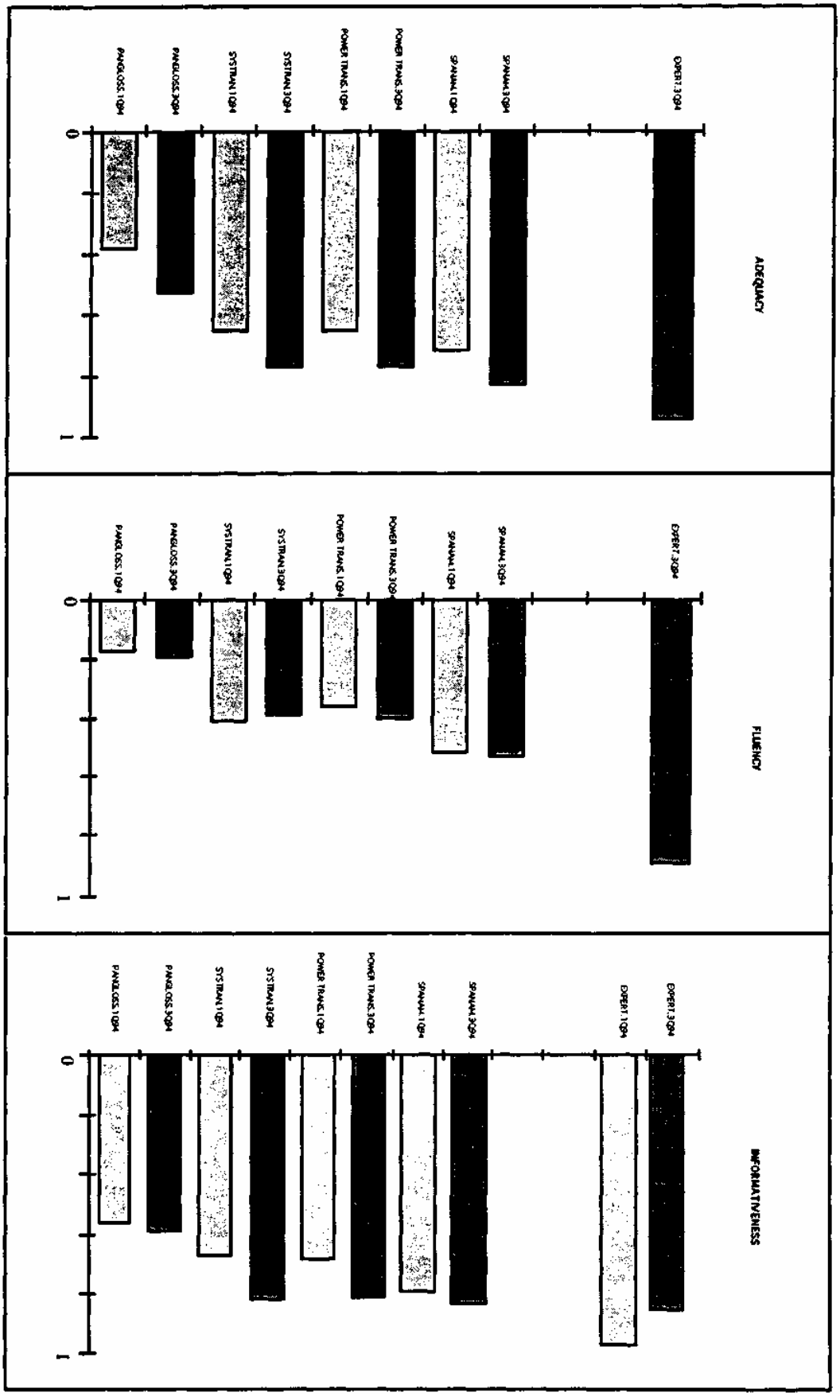

Figure 4-3. Results of 3Q94 Spanish-English, compared to 1Q94

increasing the sensitivity of the measurements, and continuing to enhance the portability and timeliness of the evaluation. To those ends we analyzed the methods used in the 3Q94 test and evaluation, and have been able to draw some conclusions that will improve future evaluations.

5.1. Number of texts. The 3Q94 test used 100 texts instead of the 20 texts used in the past, in order to reduce the variance. While pooled standard deviation did decrease (meaning that the relative rankings of the systems are likely to be correct) individual system standard deviations did not (meaning that there was still great variation among individual translation scores for each of the systems). It appears that the system standard deviations reflect inherent variation of the population of evaluators, and that there may be no statistical reason to use more than 20 texts per language pair for these tests.

5.2. Redundancy of measures. Pearson product moment correlations conducted on the results of each of the evaluations indicate that there is a common kernel of "quality" that all three methods measure (O'Connell et al. 1995). The two that behave most alike are adequacy and informativeness, which is expected because of the focus in both on the content rather than the form of the output English. From this it appears that we can enhance the timeliness and portability of the evaluation by eliminating either adequacy or informativeness for the evaluation suite.

5.3. **Maturity of training**. If there is a unifying generalization of the results of the ARPA series, it is that the systems that have the most training have the better results in the evaluations. This appears to be a large generalization to make, since training means very different things to the different systems, including adaptive modeling (Candide), ontological articulation (Pangloss), and classical lexicon development (Systran, SPANAM, etc.). The generalization is meaningful, however: each method of training inculcates a characterization of a discourse universe, the relationships among the concepts in the universe, and a representation of those relationships in the source and target languages. In some cases these characterizations are indirect (classic MT lexicons) and in some cases implicit (statistical models). But to the extent that the MT system has access to a well-articulated large body of knowledge, the performance of the system is superior in a black box evaluation. Systran French and SPANAM have lexicons so highly developed that they may do little or no new lexical update in the course of production work. Candide uses a statistical model of the languages and the translation between them, de-

rived from a gigantic parallel corpus. To what extent are the higher scores of these systems based upon the core approach rather than effectively articulated and comprehensive knowledge?

MT evaluations should take the training differential into account, even while avoiding favoring a particular method of training. Perhaps a training set for each source language, with some sort of ground truth data, would help to show the true potential of particular core approach without adding undue human-factors bias. It is common in most varieties of MT to conduct internal tests with ground truth lexicon sets (that is, lists of lexical items in the test texts, with their translations). There are barriers to providing this for the ARPA-style tests: the different approaches cannot take equal advantage of lexicon lists, and even relatively innocuous decisions about word/phrase lexical items have theoretical implications that can bias against an approach. However, providing a subset of the texts with their expert translations could provide accommodation of training while avoiding bias. An adaptive approach may be able to enhance the models with these parallel texts, a classical MT approach may lexicalize the constituents of the texts in any way that best fits its particular approach, and a knowledge based system may be able to incorporate conceptual relations presupposed by collocations or other phenomena in the texts. And a mature system may choose to do nothing at all with the subset. Evaluation could then compare the results of each system against both training texts and new text.

The improvements to the evaluation methodologies will serve to provide a more focused and portable evaluation suite, measuring of the merit of an MT approach separate from the maturity of the system that uses the approach. These streamlined methods should also maintain a significant level of comparability with the previous evaluations to show the progress in both the research systems and the other participant systems.

6.0. Conclusion. The ARPA MT evaluation is the first to maintain a black box evaluation suite over a series of four test cycles. The challenges of heterogeneity, maturity levels, languages, etc. have led us to lessons about the methodology which are as valuable as the evaluation results themselves. The evaluation series should become sufficiently stable, focused, and portable to enable it to become a standard means of evaluating all MT systems.

References

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". Computational Linguistics vol. 19:2. pp. 263-312.

Church, Kenneth, and Eduard Hovy. 1991. "Good Applications for Crummy Machine Translation." In Jeannette G. Neal and Sharon M. Walter (eds.) Proceedings of the 1991 Natural Language Processing Systems Evaluation Workshop. Rome Laboratory Final Technical Report RL-TR-91-362.

Frederking, R., D. Grannes, P. Cousseau and S. Nirenberg. 1993. "An MAT Tool and Its Effectiveness". Human Language Technology. March 1993, pp. 196-201.

O'Connell, Theresa, Francis O'Mara, and Kathryn Taylor. 1995. "Sensitivity, Portability and Economy in the ARPA Machine Tranlsation Evaluation Methodology". To appear in the 1995 ARPA MT Workshop Proceedings.

Yamron, J., J. Cant, A. Demedts, T. Dietzel, Y. Ito. 1994. "The Automatic Component of the LINGSTAT Machine-Aided Translation System." Proceedings of the ARPA Workshop on Human Language Technology Workshop. Plainsboro NJ: March 1994.

White, John S., and Theresa O'Connell. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. Proceedings of the 1994 Conference, Association for Machine Translation in the Americas.

White, John S, Theresa O'Connell, and Francis O'Mara. 1995. Evaluation Methodologies in the ARPA Machine Translation Initiative. Proceedings of AIPASG95. Tyson's Corner, VA., March 1995: Automatic Information Processing Association Steering Group.