PANEL: THE DARPA METHODOLOGY
Objective: *George Doddington*
Advanced Research Projects Agency

I would like to review the objective of the DARPA MT program with you in order to set the stage for the presentation and discussion of the evaluation. The objective is to develop revolutionary MT technology. In particular, the focus is on core MT technology—that is, the computer algorithms that take in source text in one language and output target text in a different language—rather than on technology to integrate the human translator's effort into this process, be it pre-, interactive, or postediting. The motivation for this program is that there have been tremendous advances in AI technology and in computer technology, and we feel the time is right to capitalize on the inventory of tools that we now have available to help us.

A key part of this program is evaluation—probably based in part on the perceived successes in the speech recognition program, wherein objective evaluation and common corpora have been perceived to have a tremendous positive impact on making progress. The whole program, in essence, is tailored around objective evaluation, which is used to focus and direct the R&D effort by defining what the problem is and by providing a mechanism for calibrating progress.

It was very interesting this morning to see that a very large part (perhaps 90%) of the effort in MT evaluation is at the systems level. Basically the question being asked is, Is a system an effective tool in production translation use? But that is not at all the direction of the effort for the DARPA program. DARPA is sort of a futuristic program. It's looking to develop the technology to *support* such systems in the future, and the evaluation is intended to assess the core technology directly and not to evaluate the effectiveness of MT systems. Of course, there's much more that goes into an MT system than just the core technology.

The key requirements for the evaluation are: (1) sensitive diagnostics that measure the effectiveness of the algorithms, and (2), equally important, a cost-effective evaluation methodology that does not cripple the R&D effort by directing most of the effort toward *evaluating* technology rather than developing it. Finally, I think that the evaluation should be targeted toward fully automated MT systems rather than human-interactive MT, and I would like to give five reasons for this.

• First, as I see it, the computer algorithms that have to do with machine translation are the big question mark. They are the high-risk factor—the "How do we get from here to there?"—as opposed to some of the other factors that are more like translator workstation issues. While these latter issues are certainly very important and require a lot of creative effort, they are far less uncertain in terms of strategic risk.

• My second reason for evaluating fully automated MT is that putting a human in the loop tends to dull sensitivity to the MT algorithms. Essentially, the quality of the resulting translation reflects more the skill of the translator than the MT algorithms, and what one is left with, perhaps, is a measure of the effort required on the part of the translator—the most common one that we have talked about being simply translation time.

• The third, and perhaps most important, reason for evaluating fully automatic MT is that I believe the judgments of quality or adequacy of fully automated MT will strongly correlate with the effectiveness of the algorithm in general for a variety of applications.

• The fourth reason is sort of a corollary of that—namely, the productivity of the human translator will also correlate with the quality of the output from a fully automated machine translation system.

• And, finally, the fifth reason is that the evaluation measures for human-aided MT are further confounded by the variance of the human translators and by the implementation of the interface system between the machine and the translator.

I hope this has helped to clarify the objective of the DARPA MT program.