Panelists

Eduard Hovy, Information Sciences Institute, University of Southern California: We were part of the PANGLOSS project. I suppose you're all wondering why PANGLOSS is way out on the time axis there. I would like to say something about that. When you have so many differences between the systems you're evaluating, you're forced to step back and treat everything as a black box. Eventually, in order to take into account the human aid, or lack or it, what you're measuring is the amount of time it takes to perform a translation. Instead of trying to look at how much interaction the human actually does in order to help the system over its little humps, all you measure is time. Our interpretation of this was, "Well, when the human sits down to work with the machine, we switch the clock on, and when the translation comes out, we switch the clock off." Other projects were smarter: they did the *automatic* part of the test (the parsing and so on) first and buffered up those results, and then they switched on the clock when the human actually sat down. But a lot of the work had already been done. So their times were obviously much shorter than ours.

Elke Lange, SYSTRAN: Everybody knows SYSTRAN, I guess. It is a very large, generalized system and, unlike the research systems being tested, it is not trained on any specific domain—certainly not on acquisitions and mergers. Also, we would like to point out that systems in different stages of development were used. We're wondering why a pilot Spanish-English system was used as a benchmark when there was also a more mature Spanish system available for this purpose.

Mark Mandel, Dragon Systems: Ours is the LINGSTAT system, and I'd like to mention a couple of things that we encountered. First we discovered that the Japanese-English language pair is very distant linguistically from Spanish-English and French-English. This came up during the exercise in several ways. First, we had to modify the evaluation protocol (the order in which our Level 2 translators translated using machine output and human-alone) because the overall translation time was so slow (especially human-alone) that we were afraid we wouldn't be able to complete the test in the 48 hours allotted. Also, as is evident if you think about it, any kind of user interface that deals with Japanese has to handle the character set problem, whereas this is not an issue for Spanish or French. The syntax of the language is also something that needs to be seriously considered. After working with Japanese, one comes to regard the syntaxes of French and Spanish as similar enough to English to be almost its dialects. If you line up the words in a sentence from left to right, they pretty much match up. Japanese syntax, on the other hand, is almost the exact reverse of English word order, which makes for an interesting problem to deal with.

Marjorie León, Pan American Health Organization: The Pan American Health Organization participated in this program because we were asked to. We never quite understood how comparing the raw translation of machine-alone output against human-assisted output could really yield any interesting results, but maybe it has. Our system is normally used in a postediting mode. In order to get comparable results, if the quality of our output is going to be judged against human-assisted machine translation, we should have postedited our output and found out how long it took to do that. On a time scale, our translation took no time to produce, but that's not the translation quality that we usually deliver to our clients.

Lynn Carlson, Department of Defense: I helped with establishing the quality panel criteria and conducting that part of the exercise. I would like to make a few observations about the quality panel. We found that high-quality translators are not necessarily familiar with evaluation criteria. We had a two-hour training session to get them used to the criteria that you see in your packets, but we found that this was

probably not enough. People didn't quite follow directions, and sometimes they modified things to fit their own perspective. We're looking for ways to simplify that problem in the future.

Peter Brown, IBM: I want to address two issues that came up in Yorick Wilks' talk. He was concerned about the DARPA evaluation in two respects. One was the sense it makes to evaluate a combination of machine-aided and machine-alone translation; the other was about computer time. Let me offer you some assumptions, which I think are conservative, and see if you buy them. Because if you do, you're going to be in trouble. First of all, human-alone translation costs 25C a word (I think this is conservative). Speak up if you object. Second, humans can translate 20% faster by postediting today's machine translations (we observed a 35% speed up in the DARPA evaluations, so this is certainly conservative). Third, postediting speed increases with the quality of the input translations. If it was perfect translation, at the far end of the spectrum, it would be very quick to postedit. Fourth, today's machines translate at 25 words per hour. Now, that *is* conservative. I think we probably have the slowest translation system every built, and ours is 25 words per hour. People are laughing—that's why I say it's conservative. Fifth, in 1995 a 10- nanosecond PC will cost about $5,000 a year (I can tell you from the hardware people at IBM, that's quite conservative). And sixth, it's much easier to evaluate machine-alone translation than machine-aided translation—no disagreement there.

From these assumptions, I think there's a fairly airtight argument that we should evaluate machine-aided translation by evaluating machine-alone translation. *That's* what we should evaluate from now on. I can give you the argument quickly. If a machine can translate for 5C a word, you're breaking even, because that's 20% of the cost—and even with a 25-word-an-hour system, it costs 2C a word. That means that the most you're going to get out of making a faster system is another 20 a word; all the cost is on the human side. So, the aim of economical machine-aided translation is to increase this 20% savings. That means you have to concentrate on the quality of machine-alone translation. If we're after a quality machine-alone translation, that's what we should evaluate.

Robert Berwick, MIT: We need to watch out for the LCD hazard, which can stand for a lot of things. First of all, it stands for Lowest Common Denominator: simple scoring can go absolutely awry. It also stands for Lamplight Can Do-it: you always look where the lamp can shine, so we look for methods we know can work, even if they don't apply to the problem at hand. And finally, it stands for Lucky that Concatenation Does anything—in other words, just pure luck.

General Discussion

• *(Henry Thompson,* University of Edinburgh) What, if any, of the results that you showed on your graph are statistically significant? With three people on each quality panel, I would judge that none of the quality panel evaluations are statistically significant.

• *(White)* I am very sensitive to your comment, and it's something we intend to work on and fix.

• *(Loll Rolling,* European Commission) I was entirely in agreement with what Marge León said earlier. In fact, this project is overly ambitious, trying to simultaneously evaluate various systems for various languages running on various operational pilot systems, including MT, MAT, and human translation, for various different criteria.

I want to add just one more comment. When machine translation attains a high standard—let's say, a quality level of 85% or 90%—postediting is no longer necessary. In fact, of the 100,000 pages that we translate per year at the European Commission, less than 15% still require postediting. The end users

accept raw translation as it is without postediting. The possibility of no postediting at all should be included in such a general benchmark exercise.

• *(David Farwell,* New Mexico State University) That makes an assumption which must always be analyzed—namely, that you have an audience that will accept it. I remember one of the previous speakers divided the goals into three different purposes—the users of machine translations (people who want to scan something to see if it's worth translating); people who want to understand the meaning; and people who want to publish. If you're looking to produce a professional publication, or if you're looking, perhaps, to publish a product manual to be used by installers and end users, you want to be very careful about the exact linguistic and stylistic content of your end product. The kind of machine translation product you speak of, which may be totally adequate in terms of meaning to an educated reader already familiar with its content, will not be sufficient in such an environment.

• *(Muriel Vasconcellos,* PAHO) I'd like to mention a couple of areas that we should be thinking about. One of them is the idea of using back-translations. With back-translations you get an additional remove from the original text. Back-translation as such has always been questionable as a measure of quality, and yet in this case you're one more generation removed; you have a back-*back*-translation. To my knowledge, second-generation back-translations have never been used before in an evaluation exercise. We need to think about what happens when you compound some of the particular problems involved in translating from the different languages. The problems translating from Japanese to English are very different from those from Spanish or French to English. Are we looking at ways of capturing the potential of the system to deal with these translation problems?

Also, we need to be looking at the performance levels described in the packet. Are they realistic for an evaluation exercise? Are they useful? And should be thinking about the field-testing of comprehension tests. I mean, if your PRC management personnel had problems with the comprehension tests, maybe the test-takers had problems with them, too. And then there are the quality criteria. Do they tell us what we need to know about the translation? Are the criteria that were applied in this exercise appropriate for machine translation, and could they be further developed for human translation? We should be thinking about these things when we evaluate our own Warm-up Exercises.

• *(Margaret King,* ISSCO) The thing that strikes me when I listen to all of this is that it must have cost rather a lot. Can anybody tell me how much it did cost?

• *(White)* A lot.

• *(Jack Benoit,* MITRE) How do you factor out the bias that's introduced by the leading multiple choice questions? In other words, someone reads something and says, "Gee, I don't know what that means"; then they look at the answer and say "Ah ha! That's what it means!" Can you factor that out in any way?

• *(White)* With the number of people that we had taking the test (12 test-takers saw 12 articles), it's difficult. The person who developed the test was experienced in test theory, and we had it reviewed by a number of people who were similarly interested and similarly experienced. The person who developed the test raised a number of concerns along the same lines, particularly the ability to guess the right answer from looking at the title. The title seems to be "content-ful" enough that some of the questions could be answered by knowing what the title was, and this person suggested that we leave the title out next time. We're aware of this.