[MT Summit IV, July 20-22, 1993, Kobe, Japan]

Panel Contribution on MT Evaluation

L. Rolling
E.C. Commission, Luxembourg

In 1975 the Commission of the European Communities considered that the emerging technology of computer-assisted translation might help overcome the language barriers hampering the Europe-wide information market and increase the productivity of its own translation department. A technology watch showed that over 50 systems were in development or experimental use, and a comparative evaluation of two operational systems led to the acquisition of a license for use of the Systran system by the European institutions and the government agencies of the E.C. Member States.

TIMETABLE

1975 - 76	Technology watch, comparative evaluation and acquisition of a Systran license
28.2.78	First workshop on MT evaluation in Luxembourg
1978 - 80	Systematic evaluation of Systran English-French and French-English by Bureau M. van Dijk
1981 - 91	Pragmatic, corpus-based progress assessment
1981 - 85	Specific text-type evaluations
1984	On-site testing of Systran and Logos
1986	Comparative assessment of Japanese-to-English MT systems
1991	Audit of the Commission's Multilingual Action Plan incl. Systran
1992 - 93	Comparative evaluation of Systran, Logos and Metal for German-English translation by Rinsche and Blatt
1993	Introduction of a periodic benchmark mechanism

In the initial phase (1975 - 80) the usefulness of the MT system was assessed using "oldtimer" criteria such as intelligibility, consistency, correctness, style and acceptance by potential users.

In the development phase (1981 - 88) the relative importance of the quality, rapidity and cost criteria was assessed and the result was the selection of a single criterion representative of global usefulness: REVISION RATE.

Revision rate is measured as the percentage of words in a text that must be replaced, shifted or modified, added or deleted during the revision following raw translation. A revision rate of 18% means that 82% of the text was correctly translated, which means a quality rate of 82%.

Unfortunately the revision rate depends to a large extent on the background and attitudes of the person in charge of post-editing. A trained translator is likely to make stylistic modifications that a subject specialist would not consider indispensable. A reliable, credible evaluation would therefore require parallel evaluation by two or three persons with different backgrounds.

Revision rate is the best criterion for a one-off punctual evaluation.

The measurement of progress in output quality is a different action, which requires the use of representative text corpora in a benchmark procedure. The Commission is presently preparing such an activity, which raises a number of questions and problems.

SYSTRAN BENCHMARK CORPUS

CORPUS DEFINITION: size, modularity, tagging, text types, subjects and languages

PERMANENT, EXPANDING CORPUS for periodic identification of improvements and degradations resulting from system development

TASK-ORIENTED CORPORA for quality assessment for new text types and new subjects

---

Benchmark corpora can also be used for comparative evaluation of competitive systems. It requires relative important expenditures for staff and computer capacity.