[*Proceedings of the Evaluators' Forum*, April 21st – 24th, 1991, Les Rasses, Vaud, Switzerland; ed. Kirsten Falkedal (Geneva: ISSCO).]

Evaluation of MT Systems at Union Bank of Switzerland

Doris Albisser
UBILAB (UBS Informatics Laboratory)
Union Bank of Switzerland

At Union Bank of Switzerland (UBS), we have been active in the field of computer-assisted translation (CAT) and, in particular, the evaluation of MT systems for more than two years. The overall goal of this project is to provide a decision basis for management as regards new technologies relevant to the language industry. Eventually, the project aims at providing translators and other staff members with effective and efficient translation tools (terminology database, MT system(s), grammar and style checkers, etc.). As for MT systems, we have evaluated commercially available products and observed projects still in the research stage. The outline below refers to the strategies employed for the evaluation of MT systems.

An MT system must be evaluated as an overall system and not only for the quality of the MT output. Thus, the aspects to be taken into account comprise the linguistic capabilities of a translation system, the technical environment provided and the organizational changes involved. All three aspects are equally important. As for the procedure, our evaluations are generally carried out inhouse using company-specific texts in both economics/banking and information technology (e.g. user manuals). This approach has proven worthwhile since it allows testing of an MT system in the actual environment and not in a demo-room outside the company.

Before illustrating the three categories mentioned above, there are some general aspects of MT evaluation to be pointed out. First, the evaluation criteria and, in particular, their weighting are company-specific and thus subjective to some extent. Second, quality issues should not be quantified; they should be rated according to their degree of importance for the company. And third, the corporate situation of the MT supplier plays an important role when selecting an MT system. Issues raised in this respect include the size of the company (resources for development), the importance of MT software within the overall product range, the market share, management, the financial situation, and – very importantly – customer support. Once these general issues are clarified, the MT system is evaluated in linguistic, techni-

cal and organizational terms as outlined below. As for the linguistic part, we have tried to assess the quality of the raw MT output using the following method: First, the sentences of a given text are categorized according to their degree of complexity (ranging from I to IV). Category I refers to simple sentences, whereas category IV defines highly complex and involved sentences (commonly found in German), which sometimes make even a human translator's mind pace furiously back and forth. Second, the mistakes found in the raw translation are scored. The underlying principle for scoring the mistakes is whether a mistake can easily be corrected, whether it seriously hampers understanding, and whether it violates basic grammatical structures. Thus, a mistranslated article is considered less serious than, for instance, an incorrect verbal construction. Dictionary errors are counted separately. The same applies to non-translated sentences. As a general remark, it should be noted that the linguistic evaluation is largely language-dependent and to some extent even specific to the text type. Our linguistic evaluation model was designed for translations from German into English.

Furthermore, the technical environment offered by MT software suppliers has to meet certain requirements so as to fit into a company's overall strategy in information technology. This may include portability, interfaces to sophisticated word processors and desktop publishing systems, access to terminology from the word processor mode, import/export of terminology, options for information retrieval (e.g. for recurring texts, updates), single vs multi-user systems, and – most importantly – user-friendliness. A system requiring five keystrokes to finally delight the user with a German umlaut on the screen must be considered 'user-fiendish'. Also important are the system's capabilities for further enhancement. Since commercially available systems lend themselves more to specific text types, the question arises as to how far a system could be customized to meet the customer's needs to optimum effect. In this respect, customization might be facilitated with MT systems designed for integration into an open system architecture.

Third, the organizational changes involved have to be analyzed in detail, a factor which is often neglected. An evaluator has to determine the required user profile. Questions arising in this context: Are terminologists needed? Who administrates the system? Can presently employed translators be trained (and if so, what is the learning curve)? Another important factor comprises the cost/benefit analysis. Thus, what is the price of the system, what is the minimum translation volume to justify MT, and what is the throughput per day, including both the volume of MT output and the time required for dictionary coding and pre-/post-editing? The latter can only be estimated. As a consequence, the increase in productivity can be assessed during the evaluation phase to a limited extent only. Nonetheless, close attention has to be paid to seemingly minor tasks such as preparing a text for the translation process and putting the finishing touches to it when

it leaves the translation system. Very often these tasks are claimed to be almost fully automatic, with the emphasis preferably set on 'almost'. Routine work of this type may end up being very time-consuming, thus affecting the productivity gained in translation time. Referring to the dictionary coding, it must be emphasized that testing a translation system without prior updating of the dictionary is 'unfair' to the system. As the examples below illustrate, it contributes, however, to lighten an evaluator's job.

Example 1: Der Preiseinbruch wirkte sich wesentlich auf das Konsumentenverhalten aus.
MT before dictionary update: The price burglary influenced the consumption duck behavior significantly.
MT after dictionary update: The drop in prices influenced the consumer behavior significantly.

Example 2: Der Dollar kam unter Druck.
MT before dictionary update: The dollar came under print. (wishful thinking even with the correct preposition!)
MT after dictionary update: The dollar came under pressure.

In all three categories of evaluation, close cooperation with the MT supplier is imperative. It is indispensable for evaluators to specify and communicate their corporate requirements to MT suppliers if future systems are to be enhanced and tailored to individual needs.

Finally, in view of future integration of MT systems into a corporate environment, two general questions might be worth a moment of reflection. First, what is the potential of an MT system to be integrated into a translators' workstation? Second, does the MT supplier take into account that translation is only part of the entire document production process or does he offer the MT system as an isolated component?

In conclusion, I should like to emphasize that all three parts (linguistic, technical and organizational) are of equal importance. Thus, an MT system must be evaluated as an overall system. Furthermore, subjectivity cannot be avoided in an evaluation because each company has its own needs and priorities. What might be generalized to some extent is the evaluation criteria as such, but not their weighting. Finally, with more potential MT users communicating their needs to MT suppliers, we might come one step closer to what may be termed an 'ideal translators' workstation', namely a station that ideally works for the translator (and not vice versa).