[from *Survey of state of the art in human language technology*, ed. Giovanni Battista Varile & Antonio Zampolli (Pisa: Giardini, 1997; pp. 422-425]

# 13.5 Human Factors and User Acceptability

Margaret King
University of Geneva, Switzerland

It is quite astonishing how little attention is paid to users in the published literature on evaluation. To some extent, this can be explained by looking at who does evaluation and is prepared to talk about it. Essentially, we find three classes:

- Researchers or manufacturers concerned with system development: The researchers do not have the resources to carry out any systematic enquiry into what a group of users might actually want. The developers mainly come into contact with users through their customer support services. In both cases, when a user is taken into account, it is an abstract, ideal user, whose needs correspond to those the researcher or system developer thinks he would have.

- Funding agencies, especially, in this context, ARPA: Since what they are primarily interested in is the development of a core technology, evaluation is seen as an assessment of a system's ability to perform a pre-determined task taken to reflect the barriers the core technology should be attacking. In this perspective, thinking of an ultimate user is premature and irrelevant.

- Potential purchasers of commercially available systems: Here, of course, the user is directly present, but concerned only with his own needs.

## 13.5.1 State of the Art

One exception to the above comes from the area of machine translation. The Japan Electronic Industry Development Association's Machine Translation System Research Committee has a sub-committee, the Machine Translation Market and Technology Study Committee, which has recently published a report on evaluation criteria for machine translation systems. (A summary account can be found in Nomura & Isahara, 1992.)

The committee concentrated on three aspects:

- **User Evaluation of Economic Factors:** The aim is to support making decisions about what kind of system is suitable in those cases where introducing a machine translation system in the near future is being considered. Economic factors only are taken into consideration.

- **Technical Evaluation by Users:** The aim is to compare the users' needs with what is offered by a particular system, rather than to offer any abstract evaluation of the system per se.

- **Technical Evaluation by Developers:** The aim here is to support in-house evaluation of the technical level the system has achieved and of whether the system suits the purpose for which it was developed.

In what follows we shall concentrate on the first two aspects:

User evaluation of economic factors is essentially accomplished by analyzing the replies to two questionnaires, the first concerning the user's present situation, the second his perceived needs. The answers are evaluated in the light of a set of parameters relating the answers to what advantages a machine translation system could offer. The results of are presented graphically in the form of a *radar chart*, which provides a profile of the user.

In parallel, a similar exercise is carried out to produce profiles of typical users of types of machine translation systems. Seven types of systems are distinguished in all, which cover in fact the whole range of translators' aids. The committee members define a typical user for each type of system, and a profile for that user is constructed on the basis of the answers he would be expected to give to the questionnaire. This profile then becomes the profile of the system-type. Types of system can then be paired with types of users by comparing the radar chart profiles for user and for system and finding the closest match.

The validity of the procedure is confirmed by taking, for each system type, a further group of four (assumed) users, filling out the questionnaires on their behalf, and checking that the closest match is what is expected to be.

Two points are worth making about this procedure. The first is that what is being considered is not really systems but what Galliers and Sparck Jones (1993) call *setups*, that is, a system embedded in a context of use. This is important: from a real user's point of view, there is usually very little point in evaluating a system in isolation. The ISO 9000 series on quality assessment of software makes the same point, although from a rather different viewpoint:

> "The importance of each quality characteristic varies depending on the class of software. For example, reliability is most important for a mission critical system software, efficiency is most important for a time critical real time system software, and usability is most important for an interactive end user software."---ISO (1991)

The second point shades rather to the negative; the users considered in constructing the radar charts of the system type are not real users. It is important to be aware of the dangers involved in deciding on behalf of some third party what it is he really wants or needs.

This potential weakness is partially at least counterbalanced by the second type of evaluation, called in the committee's reports "technical evaluation by users." Here, an attempt is made to determine the user's real needs and to compare them with what can be offered by specific products in order to evaluate how satisfied the client is likely to be with what is offered.

Attempts to take user needs into consideration were also made within the Esprit Translators' Workbench projects (ESPRIT project 2315, TWB I and 6005, TWB II). Catalogues were developed for describing user requirements, term banks, translation memories, machine translation, machine assisted terminology work and for checkers. The catalogues were intended to serve a double purpose, first as a way of setting up requirements specifications, and secondly as a way of evaluating to what extent a particular tool corresponds to a given user's needs. In general terms, each catalogue comprises facts relevant to the software and related to a certain quality characteristic, such as task adequacy, error tolerance, execution efficiency, ease of use, ease of learning, etc. Users can tick items which are relevant to them, give items an individual priority and rate each priority by specifying its relative importance compared to other items of the same type (Höge, Hohmann, et al., 1992; Höge, Hohmann, et al., 1993).

## 13.5.2 Current Work

In this section, we look at the efforts of the EAGLES Evaluation Group to build on these and other efforts in order to define an evaluation methodology where the users' views and needs are systematically taken into account.

The overall aim of the Evaluation Group is to define a common general framework within which specific evaluations can be designed. In this work it has also been influenced by the discussions reported in Thompson (1992), by the work of Galliers and Sparck Jones (1993) and by the work on evaluation within the ARPA/DARPA community.

The group distinguishes three types of evaluation: progress evaluation, where the aim is to assess the progress of a system towards some other ideal state of the same system, diagnostic evaluation, where the aim is to find out where things go wrong and why, and adequacy evaluation, where the aim is to assess the adequacy of the system to fulfill a specified set of needs.

User-centered evaluation is clearly adequacy evaluation. The first problem becomes evident at this point. Adequacy evaluation involves finding out whether a product satisfies the user's needs. But users are very numerous, and have widely differing needs. It would be out of the question to work in terms of individuals. However, on the basis of surveying what a sufficiently large number of individual users say, it should be possible to identify classes of users and to construct profiles of each one of these classes. These profiles can then be used as the basis for determining what attributes of particular classes of products are of interest to particular classes of users. Then, for each such attribute, a procedure can be specified for discovering its value in the case of any particular product.

The appropriate analogy is with the kind of reports published by consumer associations, where different products of the same general class are compared along a number of different dimensions. Consumer reports typically are concerned with products based on a relatively stable technology. Transferring the paradigm to the more sophisticated products of the language industry can require a great deal of work, and sometimes a considerable degree of ingenuity. In the interest of producing concrete results in the short term, while at the same time checking the validity of the general framework, the EAGLES group, together with an associated LRE project, TEMAA, is concentrating on designing evaluation packages for market or near market products in two areas, authoring aids and translation aids. These areas are of particular interest partly because the market is large, and therefore the results are likely to be of interest to a large number of potential users, partly because at least some of the products in these areas are based on a fairly stable technology.

If it proves possible to produce evaluation packages for a range of language industry products, they can be expected to constitute a de facto standard for such products. Working on how this can be done for the more modest products of the language industry lays the foundation for extending the enterprise to more sophisticated products.

## References

Galliers, J.R. and Sparck Jones, K. (1993). Evaluating natural language processing systems. Technical Report 291, University of Cambridge Computer Laboratory. To appear in *Springer Lecture Notes in Artificial Intelligence*.

Höge, M., Hohmann, A., and Mayer, R. (1992). Evaluations of TWB: Operationalization and test results. Final Report of the ESPRIT I Project 2315 Translators' Workbench (TWB).

Höge, M., Hohmann, A., van der Horst, K., Evans, S., and Caeyers, H. (1993). User participation in the TWB II project: The first test cycle. Report of the ESPRIT II Project 6005 Translators' Workbench II (TWB II).

ISO (1991). Information technology—software production evaluation, quality characteristics and guidelines for their use. Technical Report 9126, International Organization for Standardization.

Nomura, H. and Isahara, H. (1992). JEIDA's criteria on machine translation evaluation. In *Proceedings of the International Symposium on Natural Language Understanding and AI*, Kyushu Institute of Technology, Iizuku, Japan. Part of the International Symposia on Information Sciences

Thompson, H., editor (1992). *The Strategic Role of Evaluation in Natural Language Processing and Speech Technology*. Human Communication Research Centre, University of Edinburgh.