Optional extensions - a proposal for a flexible annotation system

Sven Strömqvist
Dept of linguistics
Helgonabacken 12
SE-223 62 Lund
fax +46 46 2224210
e-mail Sven.Stromqvist@ling.lu.se

Over the past decades, researchers in linguistics, psycholinguistics, cognitive science and related fields have taken an increasing interest in placing language in its socio-cultural and cognitive contexts. The interplay between verbal language and other information resources in a communicative situation is approached with a revitalized interest, often with the means of documentation offered by on-line methodologies. In this situation, the importance of an efficient and flexible annotation system for Meta-Descriptions for Multimodal/Multimedia Language Resources is a key to future research cooperation and success.

The notion of flexibility must be stressed. Researchers from different traditions have arrived at partly different taxonomies for describing their data, and the need for descriptive detail varies from one research team to another. We therefore propose that the common system be constructed in such a way that there is a "backbone" of basic categories and of support and search mechanisms and then a possibility for special interest groups/experts in particular research fields to plug in their more fine-grained search categories or search macros. That is, the annotation system and the future search engine should support optional extensions. And new extensions can be made known through a bulletin board. We believe that this type of solution would not only offer a sufficient degree of flexibility, but that it would also stimulate future developments and cooperation between researchers working with Multimodal/Multimedia Language Resources.

As an illustration of resources, needs and possibilities, consider the situation at the faculty of humanities, University of Lund. A survey of corpus resources at the faculty shows that there is ample usage of text corpora in both education and research, and that several on-going research projects could result in proper corpus resources. Available text corpora, include, for example, large reference corpora (e.g., the Brown corpus) learner corpora (e.g., International Corpus of Learner English) longitudinal corpora (e.g., longitudinal case studies in Childes/Chat format) parallell corpora (e.g., English-Swedish; French-Swedish).
The longitudinal corpora are extendable to multimedia corpora. The original recordings were made on video, but so far neither the sound track nor the video signal has been digitized and linked to the machine readable transcripts. A first approach of this kind would probably focus on descriptions of episodes and events and link the audio and video tracks to the transcripts in terms of episode/event onsets and offsets. Further annotation would then detail the internal structure and content of these episodes, in principal down to micro events, such as, for example, gestures.

On-going projects at the faculty further include potential corpora from phonetic studies and studies of on-line writing. The writing activity is computer-logged and the writer is both audio recorded (sound track of speaking to oneself during writing) and video recorded. This means that all the layers of recorded information are on-line and so the synchronization of the layers is reduced to an alignment problem (no need for manual specifications of onsets and offsets of episodes and events).

We further propose that user studies might be a useful way to elicit a basis for meta descriptions. Let different categories of users (e.g., students, teachers, researchers) specify search categories according to their needs and interests, or let them participate in so-called wizard-of-oz type of experiments with non-perfect search engines. The latter technique might be a good way both to study the expectations on a multimedia search engine that different user categories bring to the experiment and to study the users reactions to the options offered by the machine.