MATE – Multilevel Annotation and Tools Engineering

Laila Dybkjær, Natural Interactive Systems Laboratory (NIS), Odense, Denmark

The MATE project (http://mate.nis.sdu.dk) was launched in March 1998 in response to the increasing need for standards and tools in support of creating, annotating, evaluating and exploiting language resources. The production of enriched corpus data is time- and cost-intensive and re-use of annotated data would seem very attractive. However, so far re-use has usually required a painstaking, time-consuming and often inefficient adaptation process due to the lack both of standards and widely used tools. MATE aims to facilitate the re-use of spoken language resources.

On the basis of results from projects world-wide on spoken dialogue annotation and tools, MATE has developed a standard framework for the annotation of spoken dialogue corpora at multiple levels, including prosody, (morpho-)syntax, co-reference, dialogue acts, and communication problems, as well as issues concerning the interaction among these levels. MATE proposes state-of-the-art best practice coding schemes for its annotation levels and has completed the first version of a java-based workbench in support of the annotation framework and best practice schemes. The basic file format used by the workbench is the widely accepted XML format. The workbench is currently being tested by consortium members and by the 70+ members of the MATE Advisory Panel. The MATE consortium includes NIS (Odense, Denmark) (coordinating partner), CSELT (Torino, Italy), DFE (Barcelona, Spain), DFKI (Saarbrücken, Germany), HCRC (Edinburgh, UK), ILC (Pisa, Italy), IMS (Stuttgart, Germany), and TID (Madrid, Spain).

Each proposed best practice coding scheme follows the same standard structure (called a coding module) which basically includes what is needed to markup spoken language corpora, such as a markup declaration, examples, and a coding procedure.

The best practice coding schemes are included in the workbench but the user is also offered the possibility of adding new coding schemes via an easy-to-use interface. On the basis of the entered markup declaration, a DTD for the coding scheme is automatically generated that defines which tags are available and how they can be used during markup of a corpus.

An audio tool offers the user the possibility of listening to speech files while making a transcription. The transcribed file may then be annotated according to a selected coding scheme.

A number of default style sheets define how output to the user is visually represented. For instance, phenomena of interest in the corpus may be given a certain colour or shown in boldface. The user may modify a style sheet or define new ones. However, for the moment no style sheet editor is available, so a fairly detailed understanding of XML concepts and structure is required.

The workbench enables information extraction of any kind from annotated MATE corpora. Using a powerful query language the users can specify their query. The result is given as a set of references to the queried corpus. The query mechanism also supports extraction of statistical information from corpora (e.g. the number of marked-up nouns). Moreover, computation of important reliability measures, such as kappa values, is enabled.

Import of files from XLabels and BAS Partitur to XML format is supported. Other converters can easily be added to the workbench. Export to file formats other than XML may be achieved by using style sheets. For example, information extracted by the query tool may be exported to HTML in order to serve as input to a browser.

Improvements to the workbench will continue throughout 1999 based on feedback from testers. These are encouraged to use their own corpora and tasks when evaluating the usability and functionalities of the workbench, but a small set of dialogues can also be downloaded from the MATE web site. Moreover, the web site offers an example dialogue annotated at all MATE levels, several other examples addressing one or more levels, and documentation of the workbench, including how to use it!