A Tool for Multi-Level Annotation of Language Data

Christoph Müller and **Michael Strube**
European Media Laboratory GmbH
Villa Bosch
Schloß-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
{Christoph.Mueller, Michael.Strube}@eml.villa-bosch.de

1 Motivation

We present MMAX, an XML-based Java tool for the annotation of language data on multiple levels of linguistic description.[1] Most currently available annotation tools are restricted to single levels of linguistic description, e.g. coreference, dialog acts, or discourse structure. Annotations produced for individual levels cannot easily be combined or applied to the same language data. This, however, would be highly desirable because it would allow for simultaneous browsing and annotating on several linguistic levels. In addition, annotation tasks could be distributed to several research groups with different expertise, with one group specializing in e.g. dialog act tagging, another in coreference annotation, and so on. After completion of the individual annotation tasks, the annotations could be combined into one *multi-level* annotation that a single group could not have produced. The annotation tool MMAX is intended as a lightweight and flexible implementation of multi-level annotation of potentially multi-modal corpora. It is based on a simplification of annotations to sets of markables having attributes and standing in relations to each other. Due to its simplicity, the tool is fast, robust, and highly usable.

2 Underlying Concepts

1. Base Data We use the term *base data* to refer to the language data to which annotation is added. MMAX supports the annotation of both written text and (transcribed) spoken dialog. Written text is simply modelled as a sequence of sentence elements, each of which spans a sequence of word elements. For spoken dialog, turns are used instead of sentences.[2] Turn resp. sentence and word elements are stored in separate files and linked by means of *span* attributes, which serve as pointers:

```
<turn ID="turn_12"
  span="word_163..word_169" speaker="A"/>

   <word ID="word_163">What</word>
   <word ID="word_164">'d</word>
   <word ID="word_165">you</word>
   <word ID="word_166">do</word>
   <word ID="word_167">with</word>
   <word ID="word_168">them</word>
   <word ID="word_169">?</word>
```

2. Annotation The MMAX tool is based on the assumption that annotations can be simplified to sets of markables having attributes and standing in certain relations to each other.

A **markable** serves as the carrier of information. It aggregates an arbitrary set of elements from the base data. In the case of coreference annotation, markables would represent *referring expressions*, in the case of dialog act tagging, markables would represent *utterances*, and so on. In order to add information to the base data, markables have to associate some **attributes** with them. In MMAX, markables can have arbitrarily many name-value pair attributes. In dialog act tagging, when markables represent utterances, one relevant attribute could be *dialog_act*, with possible values like *initiation*, *response*, and *prepa-*

[1]This abstract is based on a longer version in (Müller & Strube, 2003). The current release version of MMAX can be downloaded at http://www.eml.org/nlp.

[2]For multi-modal dialogs, turns can contain not only words but also gestures.

ration. While markables and their attributes are sufficient to add information to sequences of base data elements, they cannot *relate* these to each other to model structural information. For this, **relations** between markables are required. *Member* and *pointer* are among the relations currently supported by MMAX: member relations express undirected relations between arbitrary many markables. For coreference annotation, a member relation like *coref_class* could be used to mark sets of coreferring markables. Pointer relations express directed (1-to-1 or 1-to-n) relations. The following is an example markable from a coreference annotation. Not all actual attributes are shown.

```
<markable ID="markable_75" span="word_165"
   coref_class="set_7" npform="prp" ... />
```

There is one set of markables per annotation level in a MMAX document. Different annotation levels are kept in separate markable files. Multi-level annotation is made possible because markables are not directly embedded into the base data, but reference base data elements in a *stand-off* fashion (Ide & Priest-Dorman, 1996) by means of their *span* attribute. Since markables on different levels are related only indirectly by virtue of shared base data elements, issues like overlap or discontinuous elements do not arise.

3. Annotation Scheme The annotation scheme contains the user's specification of which (user-defined) attributes and relations are permitted on a markable under which circumstances. These specifications are enforced during the annotation process, thus ensuring annotation consistency and at the same time guiding the human annotator. In a multi-level annotation, markables on different annotation levels (e.g. referring expressions and utterances) require different attributes and relations. Therefore, one separate annotation scheme can be defined for each annotation level.

3 The Tool

For performance reasons, the MMAX tool has a text-only display, but it can visualize relations between markables graphically by means of lines drawn on the text. Installing the tool (under Windows or Linux) is done by simply extracting a directory structure to the local hard disk; no further installation is required. A MMAX project file contains references to all files comprising a MMAX document. For each annotation level, there is one markable file and one annotation scheme file. The project file also contains a list of customizable XSL style sheets which allow different *views* of the same document during a MMAX session. Among other things, these style sheets support the insertion of *markable handles* (usually brackets) in the display, which allow the direct selection of a markable even in cases of multiple embedding. Annotation levels can be activated or deactivated. If required, a popup menu is displayed containing all active markables in a clicked display position. The attributes of the currently selected markable are displayed (and can be modified) in a separate window. If the markable has relations to other markables, these relations are visualized graphically. Creating and deleting markables and relations between them is done by means of context-dependent popup menus. After each modification, the display is refreshed in order to reflect changes to the selected markable's attributes.

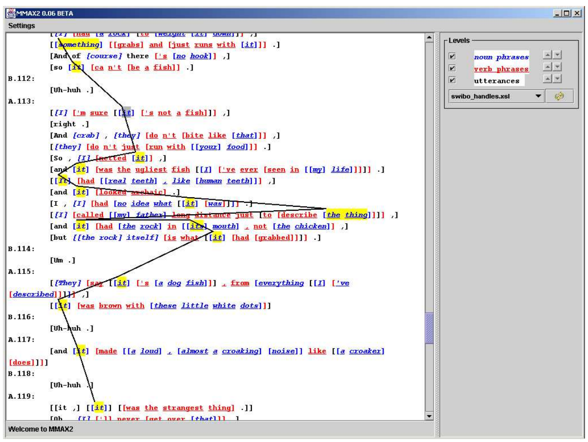

Figure 1: Selected coreference set in MMAX

References

Ide, Nancy & Greg Priest-Dorman (1996). *The Corpus Encoding Standard*. http://www.cs.vassar.edu/CES.

Müller, Christoph & Michael Strube (2003). Multi-Level Annotation in MMAX. In *Proceedings of 4th SIGDial Workshop on Discourse and Dialogue,* Sapporo, Japan, 5-6 July 2003, pp. 198–207.