Annotation Tools

What is important?

- should be able to do your task
- speed, stability, and practical usability
- customizability (with respect to the annotation task and the user)
- data generated by the tool should adhere certain standards which foster use and reuse

MMAX

- light-weight and highly customizable annotation tool (Müller & Strube (2001a, 2001b, 2003);
- supports the multi-level annotation of (potentially multi-modal) corpora;
- based on the concept of *markables* carrying *attributes* and standing in certain *relations* to each other;
- focus on speed, stability, and practical usability.

Motivation I

- most of the recent tools handle well the phenomena on the levels they are supposed to handle (e.g., coreference, dialogue acts, discourse structure, . . .);

- however, there are still problems:
  - the annotations on these levels exist independently of each other and can be combined only with difficulties;
  - combining the levels, however, is necessary for simultaneous browsing and annotating on several linguistic levels;
  - also, distributing the annotation task to several groups with individual expertise is impossible;

- multi-level annotation solves these problems.

Motivation II

None of the existing annotation tools did what we wanted them to do:

- the MATE workbench had a nice concept but did not work at all (slow and unstable);
- most other tools were based on inline annotation (DAT, DTT, Alembic, . . .);
- platform dependent (DTT, Alembic, . . .).

Even none of the current tools do what we want them to do:

- the NITE workbench has an ever better concept than MATE, but still doesn't work;
- ATLAS is low-level annotation on the signal level;

Concepts: Base Data

- *word* elements;
- groupings of word elements:
  - *sentence* elements for written text; or
  - *turn* elements for spoken dialogue;
- basic MMAX document: *words*-file plus *sentences*- or *turns*-file (not to be modified).

Concepts: Markables

- carry the annotation information;
- a markable is a formally defined entity which aggregates an arbitrary set of elements from the base data (list of word element *IDs*);
- markables can be defined on arbitrary levels of linguistic annotation (e.g., for coreference *referring expressions*, for dialogue act tagging *utterances*, . . .);
- markables on each level are stored in their own files.

Concepts: Attributes

- markables can have arbitrarily many attributes (name-value pairs);
- *nominal* attributes which have a closed set of possible values;
- *freetext* attributes which have an arbitrary string value.

Concepts: Relations

- *relations* between markables;
- *member-relation* and *pointer-relation* currently supported;
- e.g. for coreference *coref_class* attribute of type *member* and *antecedent* attribute of type pointer.

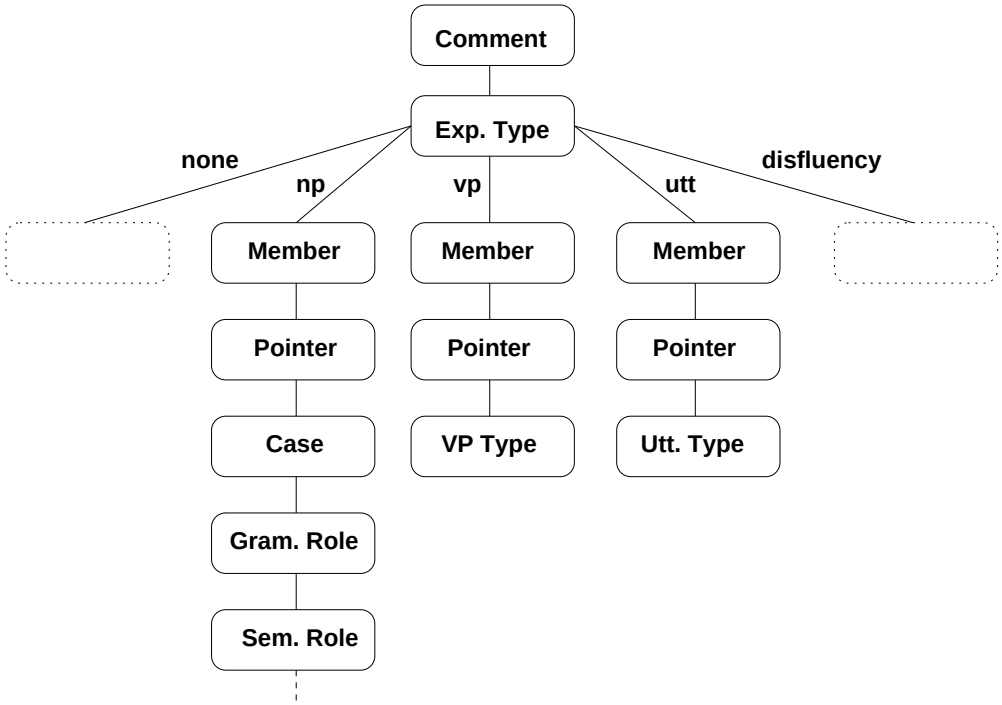
```
<markable id="markable_75"
          span="word_169"
          coref_class="set_7"
          antecedent="markable_69"
          npform="prp" ... />
```

Annotation Schemes: Basics

- depending on the task, users define the annotation scheme by themselves;
- attributes and/or relations may depend on each other;
- attributes and/or relations can be turned on/off depending on the values of other attributes;
- it is possible to formulate *constraints* on which attributes can occur together or which are mutually exclusive.

Annotation Schemes: Branching vs. Non-Branching Attributes

- if an attribute is branching its current value influences which other attributes are available.



Levels

- simultaneous representation of multiple levels of linguistic description;
- realized by the concept of *markables*;
- since markables are not embedded in the base data, but reference them by means of the *span* attribute, the simultaneous application of several levels is possible;
- also *overlap* and *discontinuous markables* are possible (this should not be possible by a straightforward implementation of inline annotation);
- rigorous implementation of the principle of *stand-off annotation*.

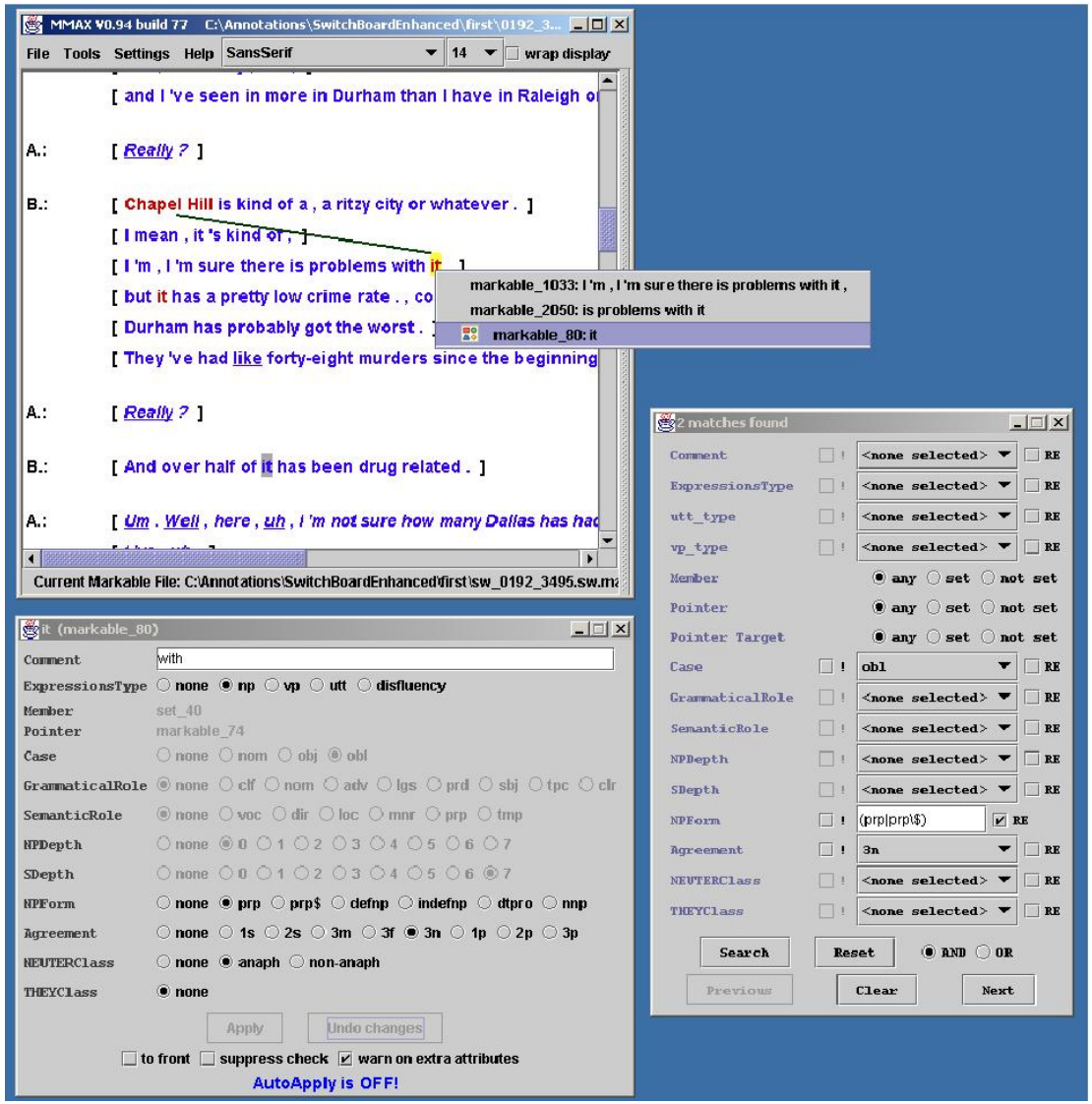
MMAX: The Annotation Tool

- MMAX written in Java, XML and XSL functionality is supplied by Apache Xerces and Xalan engines;
- GUI consists of main annotation window, *Search Window*, *Attribute Window*;
- because of performance considerations we did not use HTML display but a text-only display (standard SWING component).

MMAX: The Annotation Tool

- distinction between *content-bearing* and *layout* information;
- content-bearing information is conveyed by markables and their properties; requires frequent updates, hence hard-coded in Java; (e.g. selection of markables, display of anaphoric chains);
- layout information conveyed by line breaks and indentation but also font style properties; does not require frequent update, hence done via XSL stylesheet processor; (e.g. utterance segmentation);
- good balance between customizability and performance.

MMAX: The Annotation Tool



The Discourse API (Müller & Strube, 2002)

- platform for exploitation and reuse of annotated data;
- maps elements of the base data and markables to Java classes and defines operations on them;
- allows to access the annotated data without worrying about XML parsing and transformations;
- MMAX format not only annotation format but also target format for NLP systems.

Conclusions I

- theory: simplification of annotations to a set of simple concepts based on the notion of *markable*;
- versatility: almost any kind of annotation can be expressed through markables;
- multiple levels: different types of markables can refer to base data without interfering with each other (overlap, discontinuity);
- customizability: MMAX can express and enforce highly customizable annotation schemes.

Conclusions II

- performance: simple markable concept and restrictions in the display made it possible to implement a tool with short response times;

- MMAX used in the real world: creation of several annotated corpora, uni-modal coreference (among others, Salmon-Alt and Viera, 2002; Müller et al., 2002, Strube and Müller, 2003), multi-modal coreference (Müller and Strube, 2001; Rapp and Strube, 2002; Elting et al., 2003), dialogue act tagging (Traum);

- is compatible to upcoming ISO standard (ISO TC37 SC4).

Download MMAX

- `http://www.eml-research.de/nlp` (follow Download link)

Further Information (Papers, . . .)

- `http://www.eml-research.de/nlp`
- `http://www.eml-research.de/english/homes/strube`