

Multi-Level Annotation with MMAX2



Christoph Müller
EML Research gGmbH
Heidelberg, Germany
`mueller@eml-research.de`

(Joint work with Michael Strube, EML Research gGmbH)

Outline



- A First Glimpse of MMAX2
- Annotation Tool Requirements
- Representing Data in MMAX2
- The MMAX2 GUI
- The Discourse API
- Conclusion / Claims

A First Glimpse of MMAX2



Annotation Tool Requirements



Annotation Tool Requirements



1. GUI Speed, i.e. short response times to user actions

Annotation Tool Requirements



1. GUI Speed, i.e. short response times to user actions
2. Customizability

Annotation Tool Requirements



1. GUI Speed, i.e. short response times to user actions
2. Customizability
 - (a) Annotation Scheme
 - (b) Display

Annotation Tool Requirements



1. GUI Speed, i.e. short response times to user actions
2. Customizability
 - (a) Annotation Scheme
 - (b) Display
3. Maximal (Re-)Usability of Annotated Data

Representing Data in MMAX2



Representing Data in MMAX2



- MMAX2 Document = Basedata + Levels of Annotation

Representing Data in MMAX2



- MMAX2 Document = Basedata + Levels of Annotation
- Basedata = Data to be annotated
 - Represented as sequence of `<word>` elements
 - Not to be modified by annotation

Representing Data in MMAX2: Basedata



```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE words SYSTEM "words.dtd">
<words>
  ...
  <word id="word_360">Eva</word>
  <word id="word_361">'s</word>
  <word id="word_362">got</word>
  <word id="word_363">a</word>
  <word id="word_364">laptop</word>
  <word id="word_365">,</word>
  <word id="word_366">she</word>
  <word id="word_367">'s</word>
  <word id="word_368">trying</word>
  <word id="word_369">to</word>
  <word id="word_370">show</word>
  <word id="word_371">it</word>
  <word id="word_372">off</word>
  <word id="word_373">.</word>
  ...
</words>
```

Representing Data in MMAX2



- MMAX2 Document = Basedata + Levels of Annotation
- Basedata = Data to be annotated
 - Represented as sequence of `<word>` elements
 - Not to be modified by annotation

Representing Data in MMAX2



- MMAX2 Document = Basedata + Levels of Annotation
- Annotations = Additional information
 - Represented as sequence of `<markable>` elements
 - Markables reference basedata elements in a stand-off fashion
 - Markables from different levels are stored in different files
 - Information is encoded in the form of normal XML attributes
 - Different *types* of attributes are supported
 - XML attributes receive their interpretation from an *annotation scheme*

Representing Data in MMAX2



- Annotations = Markables + Attributes + Relations

Representing Data in MMAX2



- Annotations = Markables + Attributes + Relations
- Supported Attribute Types
 - Freetext: Arbitrary string or numerical value
 - Nominal: One of a predefined list of values



Representing Data in MMAX2: Annotations

Segment level annotation scheme:

```
<?xml version="1.0" encoding="US-ASCII"?>
<annotationscheme>
  <attribute id="starttime_attribute" name="starttime" type="freetext">
    <value name="starttime"/>
  </attribute>
  <attribute id="endtime_attribute" name="endtime" type="freetext">
    <value name="endtime"/>
  </attribute>
  <attribute id="participant_attribute" name="participant" type="freetext">
    <value name="participant"/>
  </attribute>
</annotationscheme>
```

Segment level markable file:

```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml-research.de/ns/segment">
  ...
  <markable id="markable_46" span="word_360..word_373" starttime="52.900"
    endtime="55.706"
    participant="me003"/>
  ...
</markables>
```



Representing Data in MMAX2: Annotations

POS level annotation scheme:

```
<?xml version="1.0" encoding="US-ASCII"?>
<annotationscheme>
  <attribute id="tag_level" name="tag" type="nominal_list">
    <value name="none"/>
    <value name="cd"/>
    <value name="dt"/>
    <value name="jj"/>
    <value name="nn"/>
    <value name="nns"/>
    ...
  </attribute>
  ...
</annotationscheme>
```

POS level markable file:

```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml-research.de/ns/pos">
  ...
  <markable id="markable_219" span="word_360" tag="nnp"/>
  <markable id="markable_220" span="word_361" tag="vbz"/>
  ...
</markables>
```

Representing Data in MMAX2



- Annotations = Markables + Attributes + Relations
- Supported Attribute Types
 - Freetext: Arbitrary string or numerical value
 - Nominal: One of a predefined list of values

Representing Data in MMAX2



- Annotations = Markables + Attributes + Relations
- Supported Attribute Types
 - Freetext: Arbitrary string or numerical value
 - Nominal: One of a predefined list of values
- Supported Relation Types
 - Markable Set: Undirected relation between markables (set membership)
 - Markable Pointer: Directed relation from one to many markables

Representing Data in MMAX2



- Annotations = Markables + Attributes + Relations
- Supported Attribute Types
 - Freetext: Arbitrary string or numerical value
 - Nominal: One of a predefined list of values
- Supported Relation Types
 - Markable Set: Undirected relation between markables (set membership)
 - Markable Pointer: Directed relation from one to many markables
- Dependencies between attributes and relations also supported

Representing Data in MMAX2: Annotations



Coreference level annotation scheme:

```
<?xml version="1.0" encoding="US-ASCII"?>
<annotationscheme>
  <attribute id="exp_type_attribute" name="type" type="nominal_button">
    <value name="normal" next="coref_class_set"/>
    <value name="extrapos_it" next="postponed_element_pointer"/>
    ...
  </attribute>

  <attribute id="coref_class_set" name="coref_class" type="markable_set"
    style="straight"
    color="green">

    <value name="coref_class"/>
  </attribute>

  <attribute id="postponed_element_pointer" name="postponed_element" type="markable_pointer"
    max_size="1"
    color="red"
    style="lcurve"
    dashed="true">

    <value name="not_set"/>
    <value name="set"/>
  </attribute>
  ...
</annotationscheme>
```

Representing Data in MMAX2: Annotations



Coreference level markable file:

```
<?xml version="1.0" encoding="US-ASCII"?>
<!DOCTYPE markables SYSTEM "markables.dtd">
<markables xmlns="www.eml-research.de/ns/coreference">
  ...
  <markable id="markable_954" span="word_363..word_364" type="normal" coref_class="set_4"/>
  <markable id="markable_956" span="word_371" type="normal" coref_class="set_4"/>
  ...
  <markable id="markable_8323" span="word_1963" type="extrapos_it" postponed_element="markable_8610"/>
  ...
</markables>
```

The MMAX2 GUI



The MMAX2 GUI



- First Requirement: Speed
 - Calls for highly optimized, hard-coded methods for display updates

The MMAX2 GUI



- First Requirement: Speed
 - Calls for highly optimized, hard-coded methods for display updates
- Second Requirement: Customizability
 - Calls for flexible, "scriptable" display behaviour

The MMAX2 GUI



- First Requirement: Speed
 - Calls for highly optimized, hard-coded methods for display updates
- Second Requirement: Customizability
 - Calls for flexible, "scriptable" display behaviour
- Potential conflict between those requirements

The MMAX2 GUI



- Solution: Distinguish between display *content* and display *style*

The MMAX2 GUI



- Solution: Distinguish between display *content* and display *style*
- Content: *Which text* is displayed
 - Specified by user-modifiable XSL style sheet
 - Several style sheets can define different *views* of the data
 - Changes to the display content require style sheet re-run
 - But: Changes to content do not occur frequently

The MMAX2 GUI



- Solution: Distinguish between display *content* and display *style*
- Style: *How* text is displayed (color, size, bold, underlined, etc.)
 - Specified by the user by means of *customizations* like

```
<?xml version="1.0" encoding="US-ASCII"?>
<customization>
  <rule pattern="{all}" style="background=green bold=true"/>
  <rule pattern="!coref_class={empty}" style="underline=true"/>
</customization>
```

- Changes to the display style are *very* frequent
- Supported by hard-coded and optimized methods

The Discourse API



The Discourse API



- Annotated corpus is a resource, not an end in itself

The Discourse API



- Annotated corpus is a resource, not an end in itself
- Use of this resource requires some form of programming

The Discourse API



- Annotated corpus is a resource, not an end in itself
- Use of this resource requires some form of programming
- MMAX2 Discourse API provides Java wrappers for all corpus elements

The Discourse API



- Annotated corpus is a resource, not an end in itself
- Use of this resource requires some form of programming
- MMAX2 Discourse API provides Java wrappers for all corpus elements
- High-level access to annotated corpus, no need to interact on the XML level

The Discourse API



```
MMAX2Discourse discourse = MMAX2Discourse.buildDiscourse("Bed017.mmax");

MarkableLevel corefLevel = discourse.getMarkableLevelByName("coreference", false);

/* Access Markable attributes */
ArrayList mList = corefLevel.getMarkablesAtDiscourseElementID("word_360", null);

Markable corefMarkable = (Markable) mList.get(0);

String typeValue = corefMarkable.getAttributeValue("type", "none");

/* Access Markable relations */
MMAX2AnnotationScheme annoScheme = corefLevel.getCurrentAnnotationScheme();

MMAX2Attribute corefAttribute = annoScheme.getMMAX2AttributeByName("coref_class");

MarkableRelation corefRelation = corefAttribute.getMarkableRelation();

MarkableSet corefSet = corefRelation.getMarkableSetContainingMarkable(corefMarkable);

ArrayList corefMarkables = corefSet.getMarkables();
```

Conclusion / Claims



Conclusion / Claims



- GUI Speed / Display Customizability
 - MMAX2 offers a good balance between both requirements
 - Use of XSL makes it possible to adapt the tool for the annotation of diverse phenomena

Conclusion / Claims



- GUI Speed / Display Customizability
 - MMAX2 offers a good balance between both requirements
 - Use of XSL makes it possible to adapt the tool for the annotation of diverse phenomena
- Annotation Scheme Customizability
 - Merely formal definition of *markable*, *attribute* and *relation*
 - Can be associated with various semantic interpretations, as required



Conclusion / Claims

- GUI Speed / Display Customizability
 - MMAX2 offers a good balance between both requirements
 - Use of XSL makes it possible to adapt the tool for the annotation of diverse phenomena
- Annotation Scheme Customizability
 - Merely formal definition of *markable*, *attribute* and *relation*
 - Can be associated with various semantic interpretations, as required
- Annotation (Re-)Usability
 - Arbitrarily many independent levels of annotation
 - Identical stand-off representation format for all levels
 - Clean file-level data separation, allows *distributed* annotation
 - No problem even with potentially overlapping markables

More Info



- MMAX2 research licenses have been issued to several groups, e.g.
 - National Institute of Education, Nanyang Tech. Univ., Singapore
 - Toyota R&D, Japan
 - Potsdam University, Germany, Dept. of Linguistics (teaching)
 - upcoming: MeLLANGE project (<http://mellange.eila.jussieu.fr/>)
- Also being used for individual (Ph.D. / Master) research purposes, e.g. in Stuttgart, Essex, Edinburgh
- Additional info (download, publications):
<http://mmax.eml-research.de>
- The development of MMAX2 is funded by the Klaus Tschira Foundation
<http://www.ktf.villa-bosch.de>