ÖAW AUSTRIAN ACADEMY OF SCIENCES

ACDH – AUSTRIAN CENTRE FOR DIGITAL HUMANITIES

MTAS @ ACDH

Hannes Pirker hannes.pirker@oeaw.ac.at
Matej Durco matej.durco@oeaw.ac.at

Main Use Case: AMC - Austrian Media Corpus

- ACDH‘s main „showcase“ corpus
- almost complete coverage of Austria‘s media landscape of the last two decades
- 10 Billion token, 40 Mio documents

- Metadata for <doc>: date, source, department, …
- Structures: <doc>, <p>aragraphs, <s>entences
- Annotations
  - Lemma
  - PoS (2 versions)

Current Solution: SketchEngine

- Basically we are happy with it, but ...
- Input format restricted to „verticals“:
  - tabular format
  - only strictly hierarchical structures
- Limits reached e.g.
  - NER: possibly M results from N different engines
  - Syntactic structures
- Offset annotation seems like most „natural“ representation

Why MTAS?

- Meertens was interested in testing MTAS on bigger corpora

- ACHD was interested in evaluating other corpus query tools apart from the SketchEngine

Procedures & Progress

- Path 1: „**Bring MTAS closer to AMC**“ i.e., make MTAS fit for SKE-verticals
  - Meertens provided mtas.analysis.parser.MtasSketchParser
  - But using verticals leaves us stuck with same old problems
  - (also AMC verticals turned out to be not valid XML)
  - => not further pursued

Procedures & Progress

- Path 2: „**Bring AMC closer to MTAS**“ i.e., encode AMC data in FoLiA format
  - Meertens adapted FoLiA parser
  - provided mapping + Demo page / Demo broker

Procedures & Progress

- ACDH produced test corpus:
  - 50M token, ca. 200.000 documents (articles)
- Test case: NER & entity linking
  - Using multiple engines in Stanbol
    - gnd, geonames, wikipedia...
  - Multiple (possibly richly structured) results per engine

FoLiA: <entity> e.g. „Paris“

<entity xml:id="APA_20100127_APA0701:urn:enhancement-7051d764-d35d-62fe-b89b-fdfcc7a7bd9b.9" class="NOTYPE" set="stanbol-all" abs_start="190" abs_end="195" t="Paris">
<wref id="APA_20100127_APA0701.p.2.s.1.t.26"/>
<feat class="http://dbpedia.org/resource/Paris" subset="entity-reference"/>
<feat class="dbpedia-fst-linking" subset="enhancer"/>
<feat class="dbp-ont:Place" subset="type"/>
<feat class="dbp-ont:PopulatedPlace" subset="type"/>
<feat class="dbp-ont:Settlement" subset="type"/>
</entity>

<entity xml:id="APA_20100127_APA0701:urn:enhancement-21255887-2de3-0f12-939c-c59df89508b0.16" class="Person" set="stanbol-all" abs_start="190" abs_end="195" t="Paris">
<wref id="APA_20100127_APA0701.p.2.s.1.t.26"/>
<feat class="http://d-nb.info/gnd/104261994" subset="entity-reference"/>
<feat class="gndPersons" subset="enhancer"/>
<feat class="http://d-nb.info/standards/elementset/gnd#DifferentiatedPerson" subset="type"/>
</entity>

Sample CQL Queries

<entity/>

<entity=”Person”/>

<entity=”Location”/>

<entity/> containing [t=\"Paris\"]

<entity/> containing [pos=\"ART\"]

“Der Spiegel” “Der Ring der Nibelungen”

* (<entity="Person"/> containing [entity.feat.enhancer="gndPersons"])

Thank you.