Topological Fields, Constituents and Coreference: 
A New Multi-layer Architecture for TüBa-D/Z

Thomas Krause*, Julia Ritz†, Amir Zeldes* & Florian Zipser‡

* Humboldt-Universität zu Berlin, † Universität Potsdam, ‡ INRIA

TüBa-D/Z – A Multi-layer German Treebank

- Manually annotated newspaper corpus developed at the University of Tübingen, Seminar für Sprachwissenschaft (Telljohann et al. 2009)
- Contains deep syntactic annotation with topological fields, constituents and grammatical functions, as well as coreference (since Version 5) and lemma information (Version 6)
- Valuable resource for the study of German word order, but:
  - Topological field information, constituents and grammatical functions built into the same syntax tree (limitation of format TigerXML and annotation tool "annotate", Brants & Plaehn 2000)

Reshaping TüBa-D/Z in PAULA XML with SaltNPepper

- To make the corpus more accessible to multi-layer queries, we convert and merge the data in PAULA XML (Dipper 2005), a standoff XML format for multi-layer corpora
- We use SaltNPepper (Zipser & Romary 2010), a metamodel-based converter framework, to manipulate the input data structure as a Salt model in memory

The hybrid syntax tree is split to create three structures:
1. A pure syntax tree containing only phrasal categories
2. A pure topological tree containing only topological fields (VF, MF, NF, LV)
3. A copy of the original hybrid tree for backwards compatibility

Coreference available in separate format (Tiger limited to sentence graph)
No interface to query and visualize coreference information in conjunction with syntactic annotation

Querying Layers with ANNIS2

- Using ANNIS2 (Zeldes et al. 2009) we can now run queries on each layer separately without interference, or jointly:
  - Find preverbal fields (VF) containing 2 subordinate complementizers:
    ```
    field="VF" & field="C" & #1 > #2 & #1 > #3 & #2 > #3
    ```
  - Get all dative arguments of sentences with mittelen 'inform':
    ```
    phrase="UDHFX" & lemma="mittelen" & phrase="NX" & #1 > #2 & #1 > [func="OD"]
    ```
  - Look for adjacent sentence antecedents of definite NPs in VF:
    ```
    //VF covers NX:
    field="VF" & cat="NX" & #1 __ #2 &
    //A definite article is at left edge of NX:
    pos="ART" & #2 > #3 & lemma="der" & #3 __ #4 &
    //Two coreferential nodes, one covering the NX:
    node & #1 __ #2 & node & #5 >coreferential & #6 &
    //Two adjacent sentences containing the coref nodes:
    cat="TOP" & #7 __ #8 & cat="TOP" & #6 __ #5
    ```

Conclusion and Possible Directions

- Improvements in query complexity, visualization and performance in ANNIS:
  - First graphical interface for query and visualization of all annotations in corpus
  - Separate visualization of hybrid and phrase only trees
  - Dedicated grid visualization for topological fields
  - Discourse view for coreference in context of entire document
  - Infrastructure for the manipulation of corpus structure with SaltNPepper
  - Splitting, duplicating, renaming and welding trees back together for better partitioning and searchability
  - Possible creation of additional annotations, e.g. explicit dependencies between tokens extracted from grammatical functions in hybrid tree

We are always interested in more ideas / use cases / corpora!

References