GSCL 2011 – Jahrestagung der Gesellschaft für Sprachtechnologie und Computerlinguistik

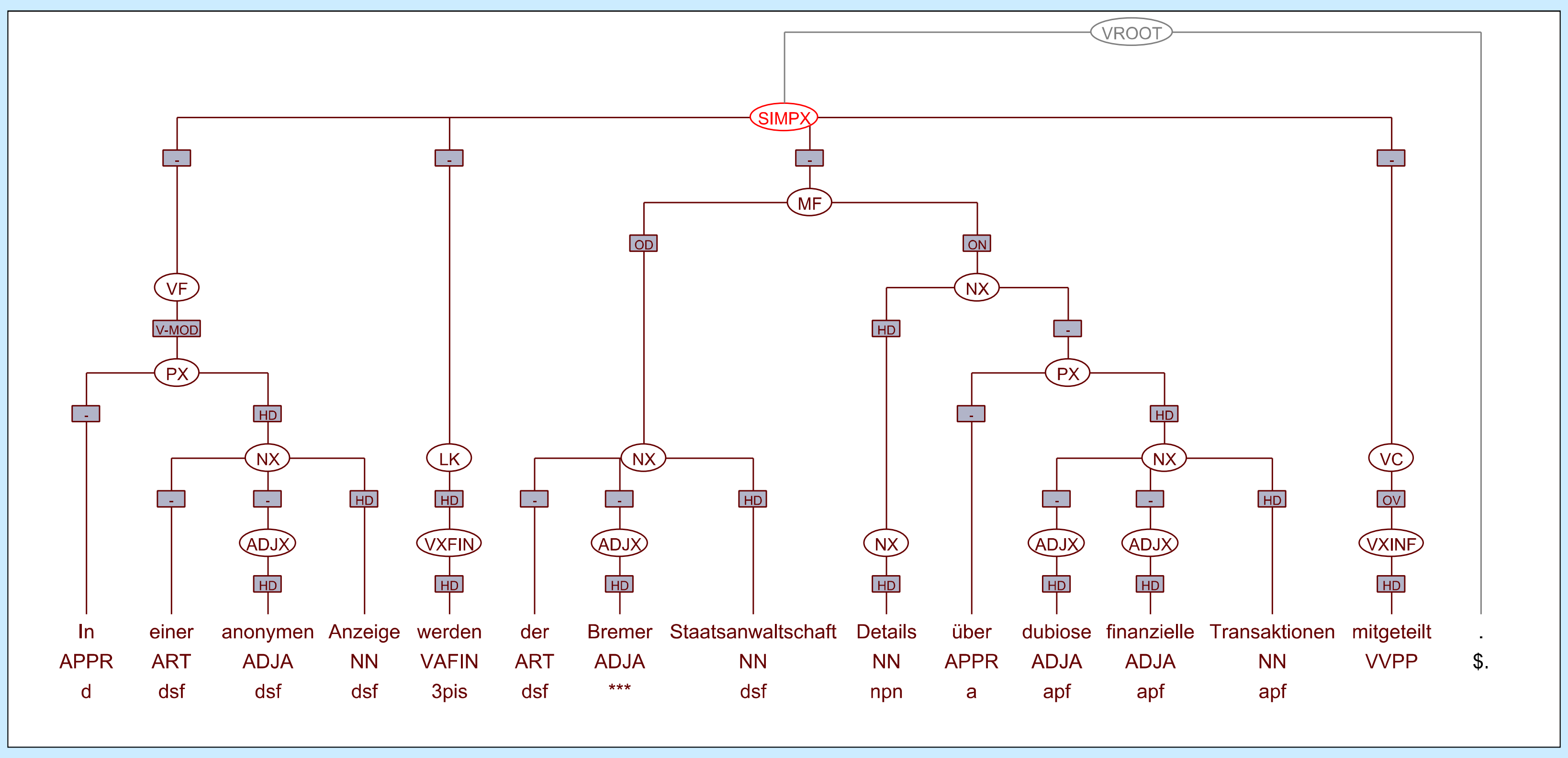
Topological Fields, Constituents and Coreference: A New Multi-layer Architecture for TüBa-D/Z

Thomas Krause*, Julia Ritz+, Amir Zeldes* and 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)

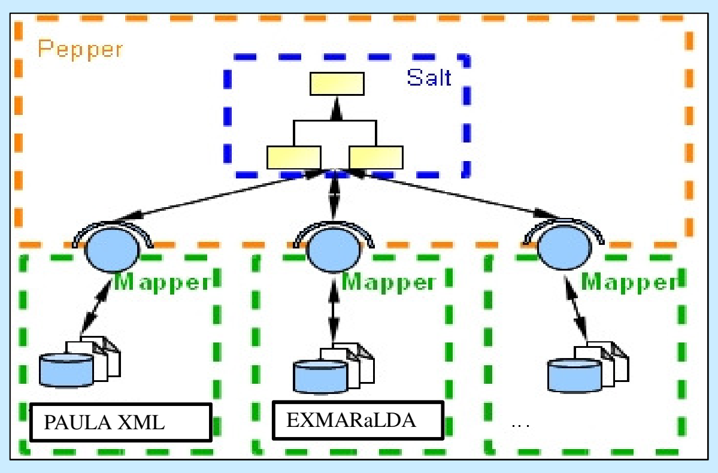


In einer anonymen Anzeige werden der Bremer Staatsanwaltschaft Details über dubiose finanzielle Transaktionen mitgeteilt .
APPR ART ADJA NN VAFIN ART ADJA NN NN APPR ADJA ADJA NN VVPP $.
d dsf dsf dsf 3pis dsf *** dsf npn a apf apf apf

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

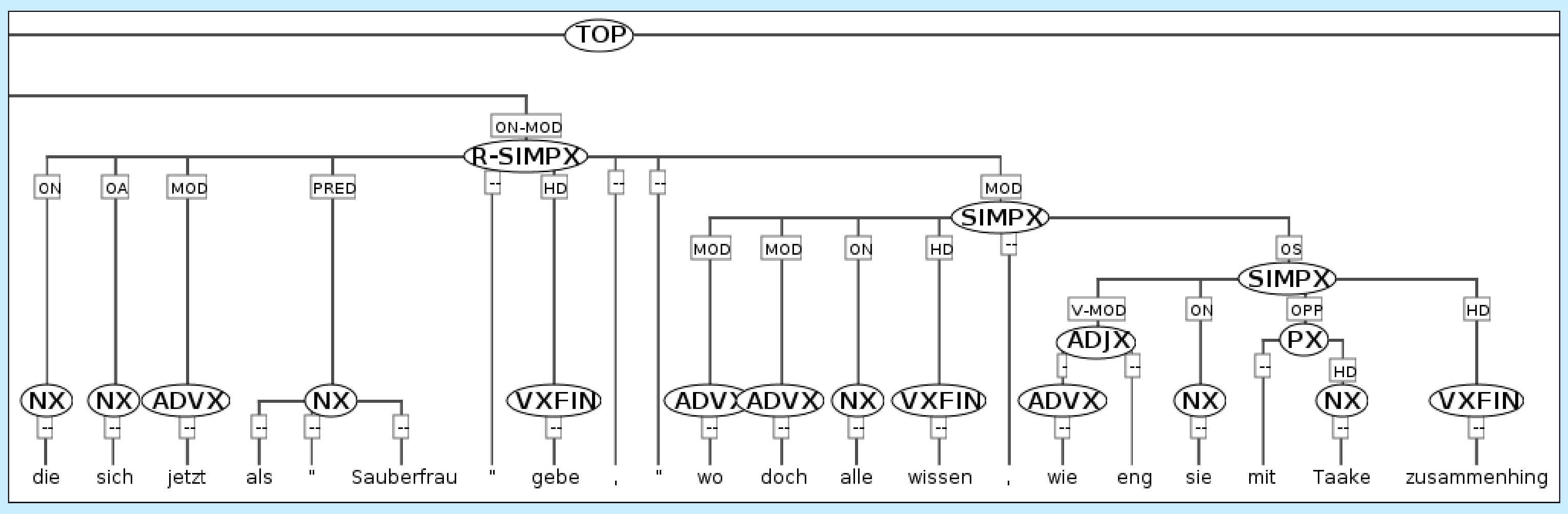
Reshaping TüBa-D/Z in PAULA XML with SaltNPepper

- 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



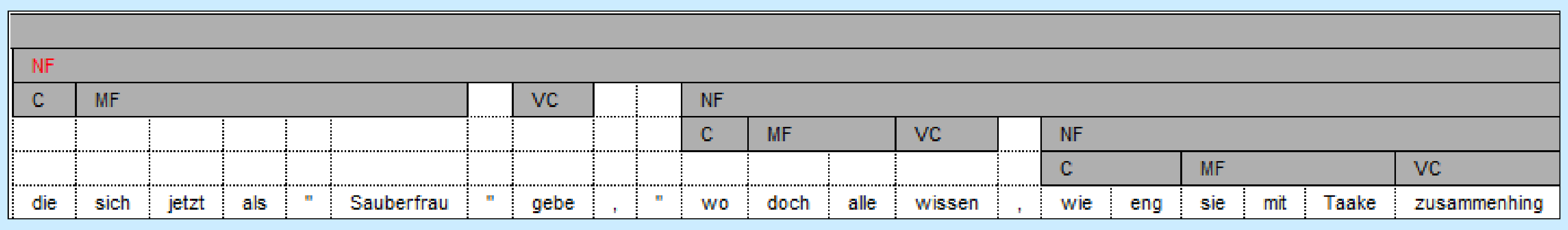
- We use SaltNPepper (Zipser & Romary 2010), a metamodel-based converter framework, to manipulate the input data structure as a Salt model in memory
- The hybrid syntax tree is split to create three structures:

1. A pure syntax tree containing only phrasal categories



die sich jetzt als " Sauberfrau " gebe , " wo doch alle wissen , wie eng sie mit Taake zusammenhing

2. A pure topological tree containing only topological fields (VF, MF, NF, LK, VC, C, LV)



NF																					
C	MF						VC			NF											
										C	MF		VC		NF						
															C		MF			VC
die	sich	jetzt	als	"	Sauberfrau	"	gebe	,	"	wo	doch	alle	wissen	,	wie	eng	sie	mit	Taake	zusammenhing

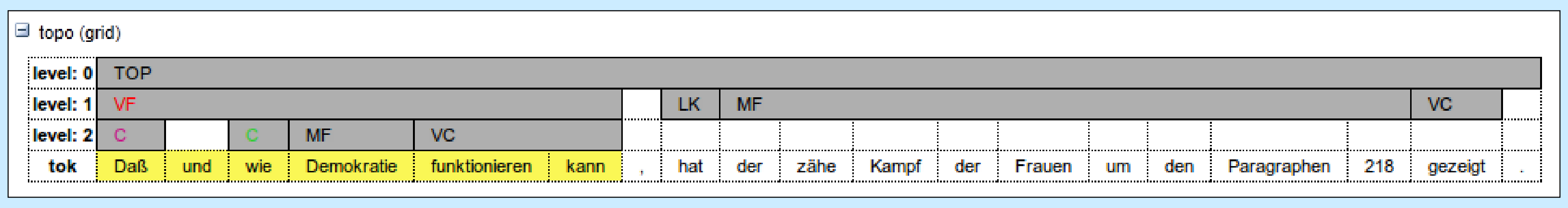
3. A copy of the original hybrid tree for backwards compatibility

- Coreference information is linked to the tokens of the trees to allow cross-layer queries (see on the right)

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" & field="C" &
#1 > #2 & #1 > #3 & #2 .* #3
```

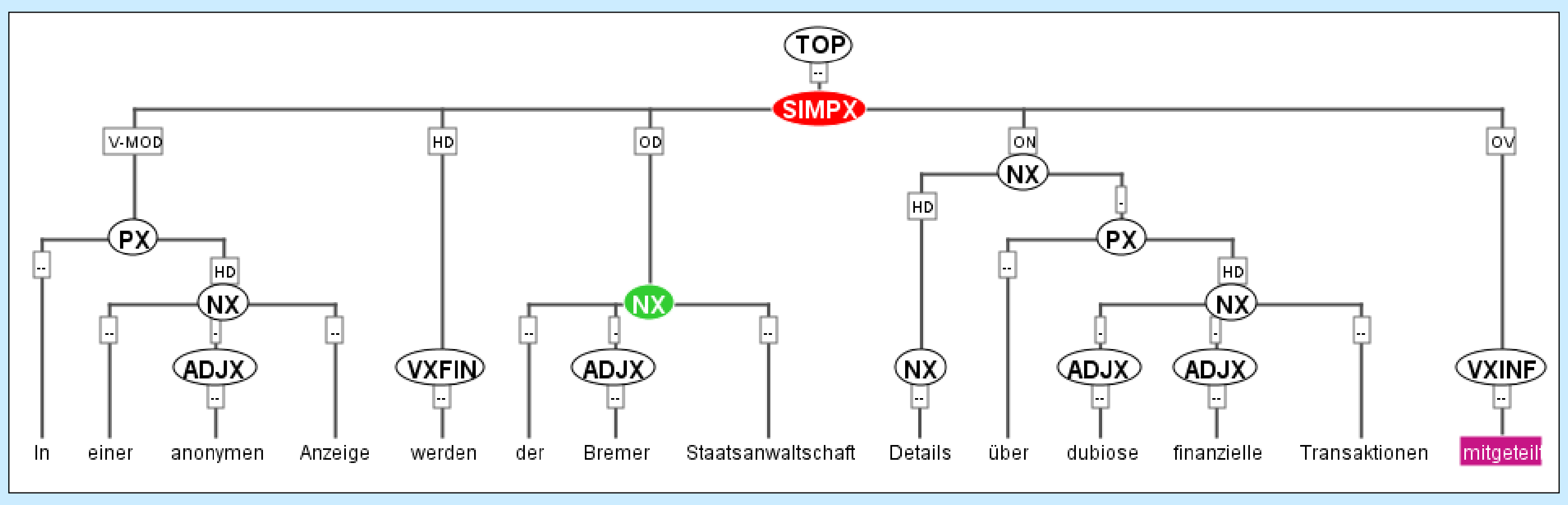


⊟ topo (grid)

level: 0	TOP																		
level: 1	VF							LK	MF									VC	
level: 2	C		C	MF	VC														
tok	Daß	und	wie	Demokratie	funktionieren	kann	,	hat	der	zähe	Kampf	der	Frauen	um	den	Paragraphen	218	gezeigt	.

  - Get all dative arguments of sentences with *mitteilen* 'inform':

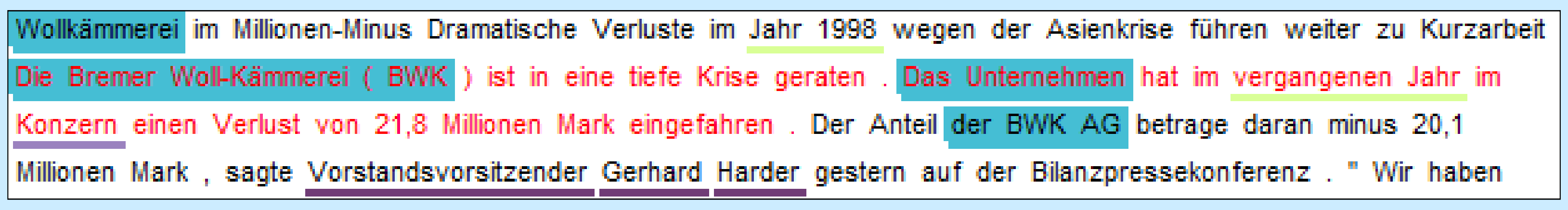
```
phrase="SIMPX" & lemma="mit#teilen" & phrase="NX" &
#1 >2 #2 & #1 >[func="OD"] #3
```



In einer anonymen Anzeige werden der Bremer Staatsanwaltschaft Details über dubiose finanzielle Transaktionen mitgeteilt

  - Look for adjacent sentence antecedents of definite NPs in VF:

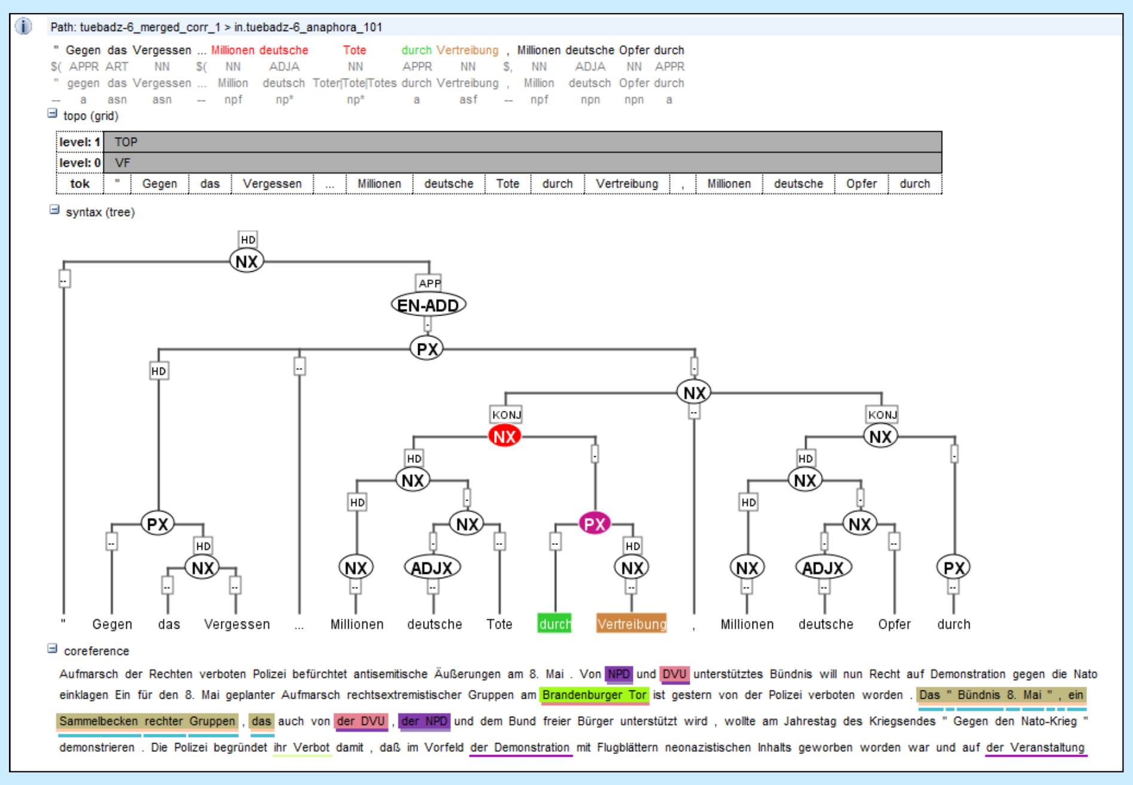
```
//VF covers NX:
     field="VF" & cat="NX" & #1 _=_ #2 &
//A definite article is at left edge of NX:
     pos="ART" & #2 >@l #3 & lemma="der" & #3 _=_ #4 &
//Two coreferential nodes, one covering the NX:
     node & #5 _=_ #2 & node & #5 ->coreferential #6 &
//Two adjacent sentences containing the coref nodes:
     cat="TOP" & #7 _i_ #1 & cat="TOP" & #8 _i_ #6 & #8 . #7
```



Wollkämmerei im Millionen-Minus Dramatische Verluste im Jahr 1998 wegen der Asienkrise führen weiter zu Kurzarbeit
Die Bremer Woll-Kämmerei (BWK) ist in eine tiefe Krise geraten . Das Unternehmen hat im vergangenen Jahr im
Konzern einen Verlust von 21,8 Millionen Mark eingefahren . Der Anteil der BWK AG betrage daran minus 20,1
Millionen Mark , sagte Vorstandsvorsitzender Gerhard Harder gestern auf der Bilanzpressekonferenz . " Wir haben

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



Path: tuebadz-6_merged_corr_1 > in.tuebadz-6_anaphora_101

" Gegen das Vergessen ... Millionen deutsche Tote durch Vertreibung , Millionen deutsche Opfer durch

⊟ topo (grid)

level: 1	TOP														
level: 0	VF														
tok	"	Gegen	das	Vergessen	...	Millionen	deutsche	Tote	durch	Vertreibung	,	Millionen	deutsche	Opfer	durch

⊟ syntax (tree)

- 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

- Brants, T./Plaehn, O. 2000. Interactive Corpus Annotation. *Proceedings of LREC 2000*. Athens.
- Dipper, S. 2005. XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin, 39-50.
- Telljohann, H./Hinrichs, E. W./Kübler, S./Zinsmeister, H./Beck, K. 2009. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Universität Tübingen Seminar für Sprachwissenschaft.
- Zeldes, A./Ritz, J./Lüdeling, A./Chiarcos, C. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. *Proceedings of Corpus Linguistics 2009*, Liverpool, July 20-23, 2009.
- Zipser, F./Romary, L. 2010. A Model Oriented Approach to the Mapping of Annotation Formats using Standards. *Workshop Language Resource & Language Technology Standards, LREC 2010*. Malta, 7-18.