Final Conference of the SFB 632 Information Structure: Advances in Information Structure Research 2003 - 2015, Berlin 08th.–09th. of May 2015

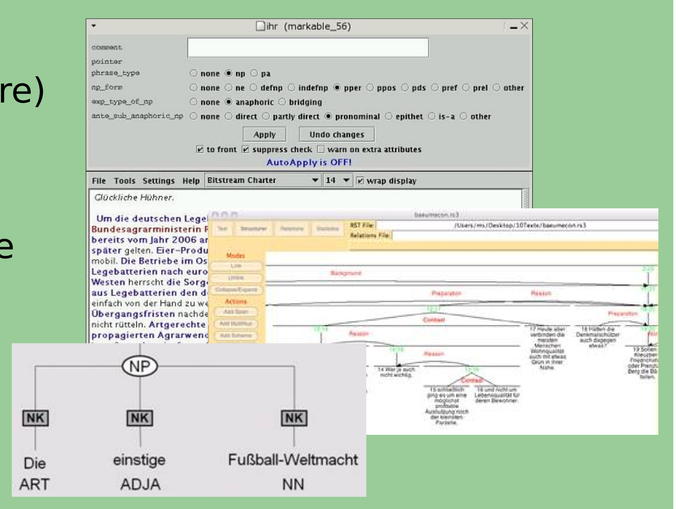
Project D1: Linguistic Database for Information Structure: Annotation and Retrieval

ANNIS, SaltNPepper & PAULA: A multilayer corpus infrastructure

Florian Zipser*, Thomas Krause*, Anke Lüdeling*, Arne Neumann+, Manfred Stede+ & Amir Zeldes~
*Humboldt-Universität zu Berlin, IdSL + Universität Potsdam ~Georgetown University

Motivation

- Information structure, like many other linguistic phenomena, influences different linguistic levels at the same time (stress, word order, definiteness, etc.).
- Corpus-based research on information structure therefore needs access to different types of annotation (Lüdeling et al., to appear).
- There are now many multi-layer corpora with annotations of linguistic phenomena on several levels (see, e.g. Tüba-D/Z (Telljohann et al. 2009), Falko (Reznicek et al. 2012) or PCC (Stede & Neumann 2014)).
- The annotation of different types of information often require different tools. The PCC corpus, for instance, used the following tools:
  - MMAX2 (co-references)
  - RSTTool (rhetorical structure)
  - @nnotate (constituencies)
- Tools have different formats which may not be interoperable
  → No data exchange between tools
  → No analysis on multiple layers

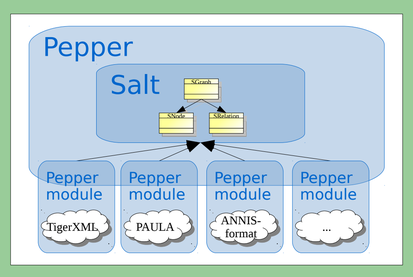


Goals

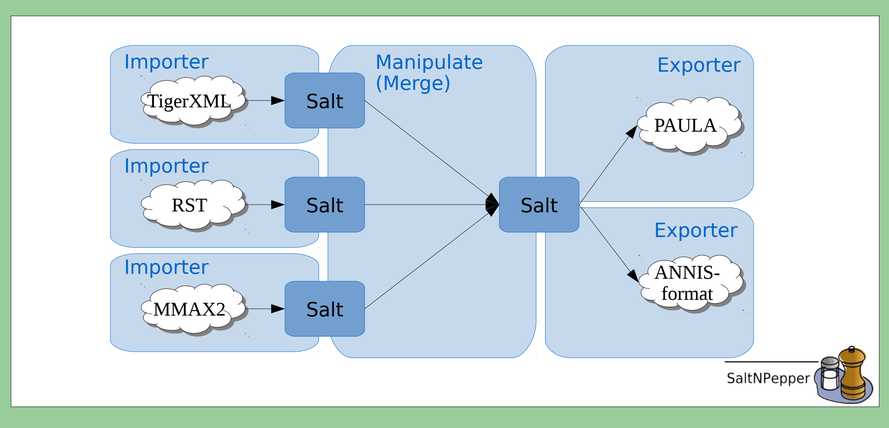
1. Merging different types of annotations of the same primary text to a single corpus → Pepper
2. Storage of different types of annotations in only one format → PAULA
3. Search in different corpora and different phenomena in one single system → ANNIS

1.Pepper (Zipser & Romary 2010)

Pepper is a universal converter framework for linguistic data to convert data between many different formats.



- Pepper uses an intermediate model to reduce the number of mappings (to implement) from n^2-n to $2n$.
- The graph-based intermediate model Salt is theory neutral and not limited to a specific set of annotations.
- The workflow is divided into steps 1) import, 2) manipulate and 3) export. That allows to manipulate e.g. to merge the data during the conversion.



(workflow for merging of PCC)

- Pepper supports many formats:
  Elan, CoNLL, MMAX2, ANNIS, Gate, RST, TCF, CoraXML, TreeTagger, Aldt, UAM, EXMARaLDA, generic XML, PTB, PAULA, TEI (subset), txt, SaltXML, ...
- It can be extended for further formats or manipulations via Plugin-mechanism.

2.PAULA (Zeldes et al. 2013)

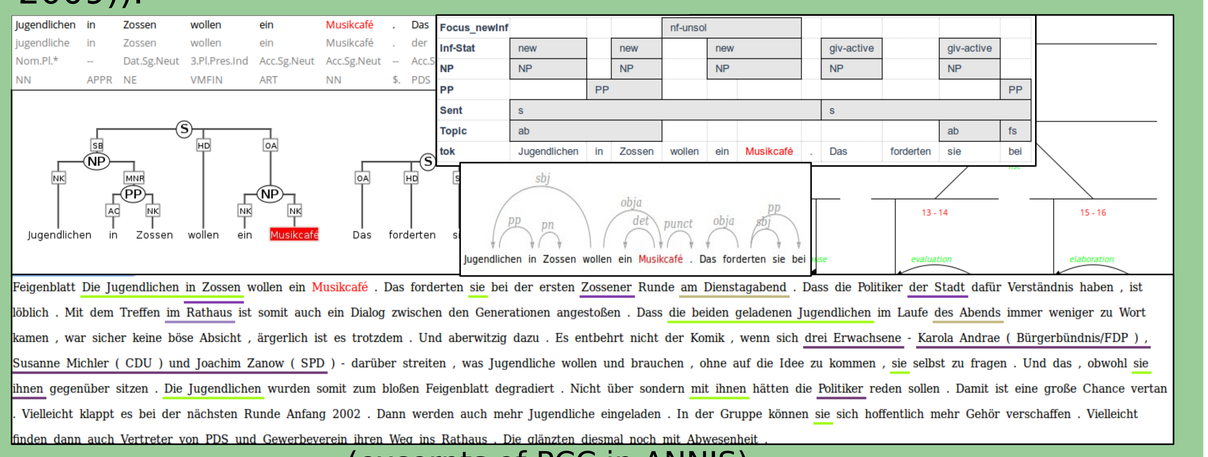
PAULA is a human and machine-readable XML format to store linguistic data which are annotated on multiple layers.

- PAULA is graph-based, which makes it theory neutral and not limited to specific annotation.
- With it's standoff mechanism, discontinuous and overlapping annotations are possible.
- Annotation layers are separated into files. One layer can be added or removed without influencing the other layers.
- PAULA stores an unlimited number of annotation layers and is not restricted to specific corpora.

3.ANNIS (Krause & Zeldes 2014)

ANNIS is a search and visualization system to query linguistic data, especially corpora annotated on multiple layers.

- ANNIS is not limited to specific annotations or a single corpus.
- The query language AQL enables a uniform search for different corpora.
- ANNIS comes with specialized and configurable visualizations for different annotation layers.
- Results can be exported for further statistical analysis and (statistical) evaluation: CSV, plain text or ARFF (WEKA (Hall et al. 2009)).



(excerpts of PCC in ANNIS)

Corpora

- Support of 24 SFB corpora (some are archived in CLARIN repositories for sustainability)
  - Different languages and periods of languages (French, Aja, Dagbani, Hindi, Modern German, Old High German, ...)
  - Different types of data (spoken vs. written; historical vs. synchronic, etc.)
  - Corpora with different types and depth of annotations (pos, lemma, information structure notions, syntax, rhetorical structure, etc.)
- Widely used outside the SFB by more than 20 different projects like Coptic SCRIPTORIUM (USA), Perseus (USA), DDD (Germany), PROIEL (Norway), The Language Archive (Netherlands), ...

References

- **A. Lüdeling, J. Ritz, M. Stede & A. Zeldes (to appear).** Corpus Linguistics. In: C. Fery & S. Ishihara (Hrsg.) OUP Handbook of Information Structure. Oxford University Press, Oxford.
- **M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, M. Walter, K.Schmidt,H. Hirschmann & T. Andreas (2012)**. Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01.
- **M. Stede & A. Neumann (2014).** Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Island.
- **H. Telljohann, E. W. Hinrichs, S. Kübler, H. Zinsmeister & K. Beck (2009).** Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Universität Tübingen Seminar für Sprachwissenschaft.
- **A. Zeldes, F. Zipser & A. Neumann (2013)** PAULA XML Documentation: Format Version 1.1.
- **T. Krause & A. Zeldes (2014)**. ANNIS3: A new architecture for generic corpus query and visualization. in: Digital Scholarship in the Humanities 2014
- **F. Zipser & L. Romary (2010)**. A model oriented approach to the mapping of annotation formats using standards. In: Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010. Valetta, Malta. URL: http://hal.archives-ouvertes.fr/inria-00527799/en/

funded by:
SFB 632 "Information Structure"
DFG Deutsche Forschungsgemeinschaft
CLARIN-D
Inria INVENTORS FOR THE DIGITAL WORLD

https://www.linguistik.hu-berlin.de/institut/Professuren/korpuslinguistik