36. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft - Sektion CL, Marburg 05.-07. März 2014

Merging data, the essence of creation of multi-layer corpora

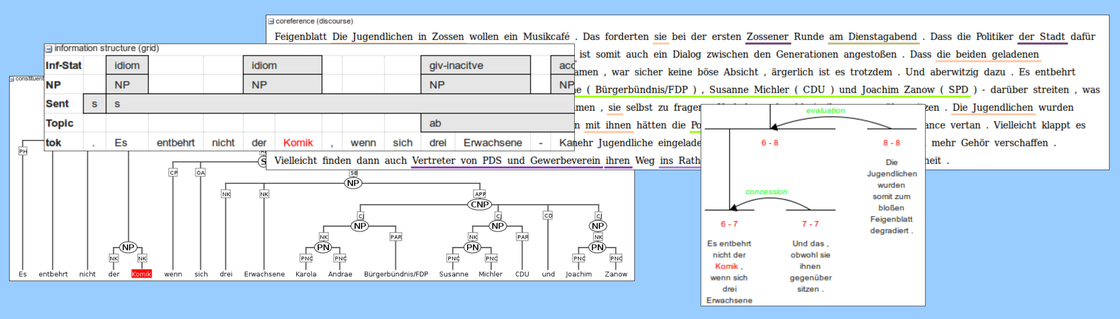
Florian Zipser, HU-Berlin IDSL
Mario Frank, University of Potsdam IICS
Jakob Schmolling, HU-Berlin IDSL

SaltNPepper

http://u.hu-berlin.de/saltnpepper

Motivation

- Multi-layer corpora allow analysis of phenomena spreading through multiple annotation layers
- More and more multi-layer corpora are available: TüBa-D/Z (Telljohann et al. 2009), PCC (Stede 2004), FALKO (Reznicek et al. 2012), ...
- Most annotation tools support the creation of single-layer corpora only, or limited sets of types of layers (trees, spans etc.)
- Annotation tools use different formats → corpora cannot be searched / displayed together
- **But:** Often these corpora share a common base (same primary data or even the same tokenization)



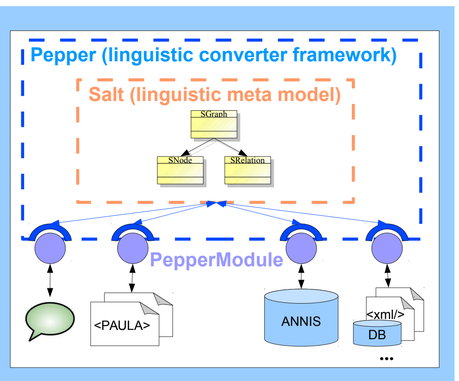
Goal

- We want to merge several annotation layers referring to the same underlying primary data or tokenization and make them searchable and displayable together

Approach

1. Convert data from different formats into a unified data model
2. Compare data and find their common base
3. Keep common base and merge different layers
4. Search / display / store multi-layer corpus

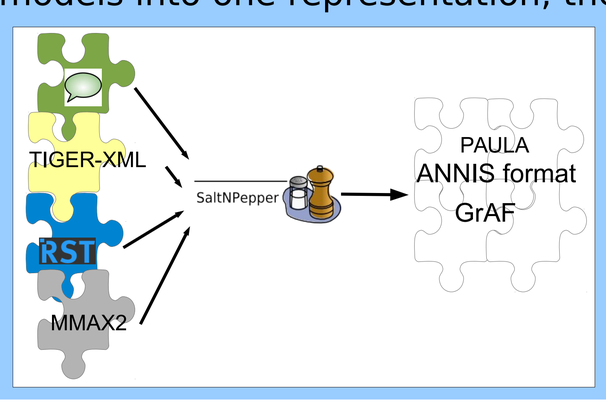
SaltNPepper



- SaltNPepper framework (Zipser et al. 2011)
  - Open source (Apache License, Version 2.0)
  - OS independent (written in Java)
- Salt is a graph-based meta model for linguistic data (Zipser & Romary 2010)
  - Abstraction of data: nodes, edges, labels, ...
  - Theory-neutral and independent of phenomena
- Pepper is a multi-format converter framework for linguistic data
  - Easily extensible via plug-in system
  - A lot of existing modules: TigerXML, <tiger2/>, EXMARaLDA, MMAX2, rs3, TreeTagger, CoNLL, Penn Treebank format, generic xml, ...

1. Convert data from different formats into a unified data model

- Import data from different formats into one Salt model each
- A specific module for merging in the Pepper workflow merges all of these Salt models into one representation, the “head model“



2. Compare data and find their common base problem

- Text *b* is contained in *a* but the common part differs slightly

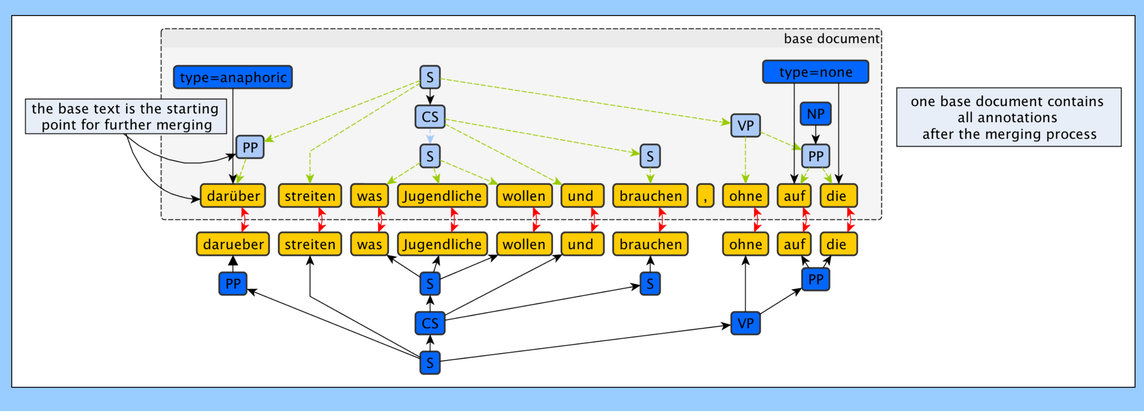
Text a = […] wenn__ sich drei Erwachsene darüber streiten […]
Text b = wenn sich drei Erwachsene darueber streiten

Solution

- Find containment of one text in the other and align them
  - Normalize texts by for instance removing whitespaces and unfolding umlauts (tokens „*darüber*“ and „*darueber*“ are recognized as equal)
  - Align position of token „*wenn*“ in *b* to position of „*wenn*” in *a*
- Identify tokens in data model which are identical in both texts (i.e., if they occur in both texts at the same position after aligning texts)

3. Keep common base and merge different layers

- Reduce problem to a graph matching task
  - Find isomorphic nodes and edges bottom-up based on tokenization, and move their annotations into head model
  - Find non-isomorphic nodes and edges and move them into head model



4. Search / display / store multi-layer corpus

- Export consolidated Salt model to output formats:
  - For searching / displaying, export data to ANNIS format (Zeldes et al. 2009)
  - For archiving, export data to PAULA (Dipper 2005) or GrAF (Ide & Suderman 2007)

References

- **Dipper S.** (2005). XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: Eckstein R., Tolksdorf R. (eds.) Berliner XML Tage.
- **Ide N.& Suderman K.**(2007). GrAF: A Graph-based Format for Linguistic Annotations. In: Proceedings of the Linguistic Annotation Workshop, Prague, Czech Republic.
- **Schmid, H.** (1995). Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.
- **Stede M.** (2004). The Potsdam commentary corpus. In Proceedings of the 2004 ACL Workshop on Discourse Annotation (DiscAnnotation '04), Bonnie Webber and Donna Byron (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 96-102.
- **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.
- **Reznicek, M.; Lüdeling, A.; Krummes, C.; Schwantuschke, F.; Walter, M.; Schmidt, K.; Hirschmann, H.; Andreas, T.** (2012). Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01
- **Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Chiarcos, Christian** (2009). "ANNIS: A Search Tool for Multi-Layer Annotated Corpora". In: Proceedings of Corpus Linguistics 2009, July 20-23, Liverpool, UK.
- **Zipser F., Romary L.** (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. Malta. URL: http://hal.archives-ouvertes.fr/inria-00527799/en/
- **Zipser F., Zeldes A., Ritz J., Romary L. & Leser U.** (2011). Pepper: Handling a multiverse of formats 33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft. Göttingen, 23.- 25. Februar 2011