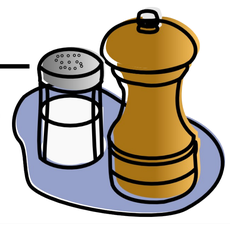
33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft - Sektion CL, Göttingen 23.-25. Februar 2011

Pepper:
Handling a multiverse of formats

Florian Zipser, INRIA, HU-Berlin IDSL &
Amir Zeldes, HU-Berlin IDSL &
Julia Ritz, Universität Potsdam &
Laurent Romary, INRIA, HU-Berlin IDSL &
Ulf Leser, HU-Berlin

SaltNPepper



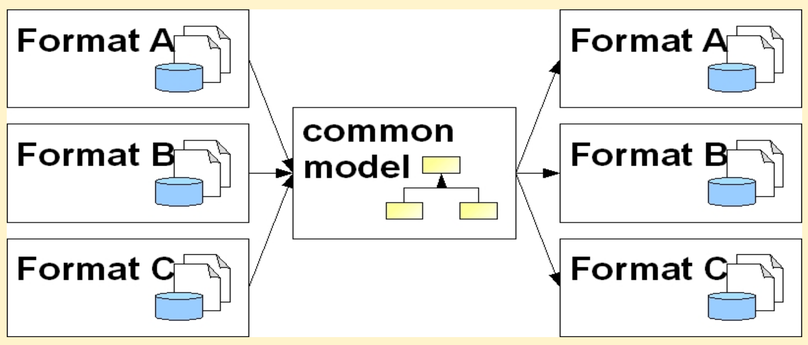
Motivation

- A lot of tools only for specific levels of linguistic annotations for instance:
  - •Syntactic structures (e.g. TIGERSearch, Lezius 2002; TrEd, Pajas & Štěpánek 2008)
  - •Dialogue structures (e.g. EXMARaLDA, Schmidt 2004)
  - •Anaphoric structures (e.g. MMAX, Müller & Strube 2006)
- Tools deal with their own proprietary formats
  - •data cannot be exchanged between tools
  - •data can only be annotated with a small set of annotations corresponding to the possibilities of the tool
- only a few multi-level-annotation formats exists (PAULA, Dipper 2005, ISO standard candidate GrAF, Ide & Suderman 2007)
  - •Currently only for persistence, but no tool support

What we need: the possibility to convert the data into several formats with a minimum of information losses.

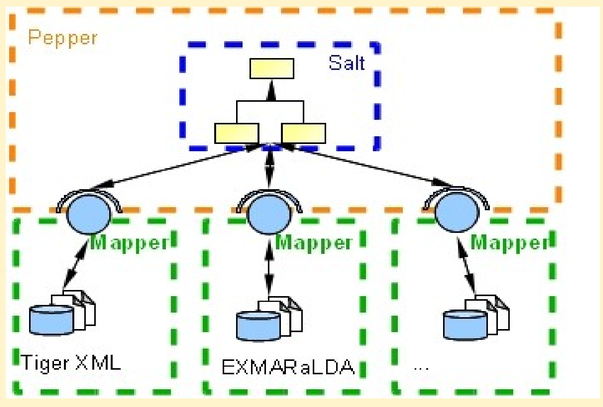
General purpose of Pepper

- Conversion framework for linguistic data
- Based on the linguistic meta model Salt (see Zipser & Romary 2010)
- converts data from *n* formats into *m* formats
  - •minimal number of mappings (*2n*) in contrast to 1:1 mappings (n^2-*n*) with fixed number of mapping steps (*2*)



Open source (http://korpling.german.hu-berlin.de/saltnpepper/)

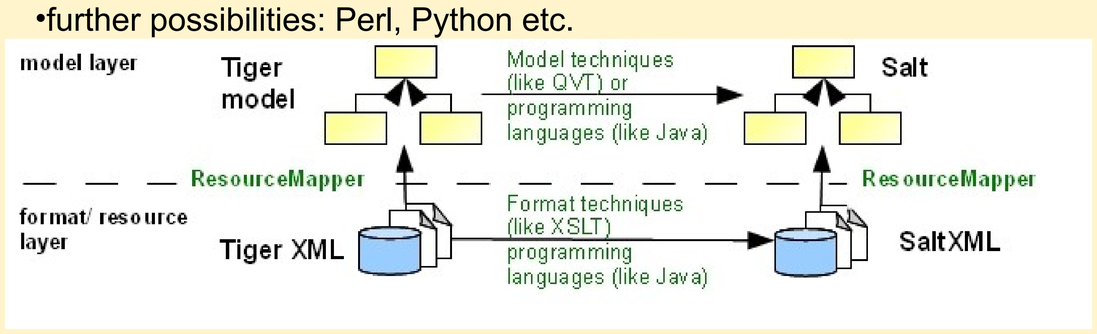
- java based
- extensible to new plugins to integrate further formats (uses OSGI for plugin mechanism), plugins can also be integrated by the users



- advantage of the framework approach: everyone can use what others have already developed

Several techniques on several layers

- Salt provides several layers for processability: model layer, java layer and persistence layer (XML)
- several techniques can be used, only a java container is necessary
  - •already in use: Java, XSLT and QVT
  - •further possibilities: Perl, Python etc.



- Several kinds of data persistence are supported (e.g. XML, SGML, tabular formats (e.g. RDBMS dumped data), bracketing formats or mixtures thereof)

Importers, exporters and manipulators

- Pepper can integrate three kinds of modules
  - •Importers: Format X → Salt (Imports data from format X into Salt)
  - •Exporters: Salt → Format Y (Exports data from Salt into format Y)
  - •Manipulators: Salt → Salt (Manipulates data in Salt, for instance for further processing like : merging of corpora etc.)
- Already supported formats:

  Treetagger, EXMARaLDA, Tiger XML, GrAF, PAULA, CoNLL, RST, relANNIS

Outlook

- a web-service to make Pepper available via the www
- a graphical user interface
- further importers and exporters, for new formats
- further manipulators, for example merging of corpora on several levels, consistency checks

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.

Lezius W. (2002) Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Ph.D. thesis, Stuttgart University.

Müller C. & Strube M. (2006). Multi-Level Annotation of Linguistic Data with MMAX2. In: Braun S. & Kohn K. & Mukherjee J.(eds.), Corpus Technology and Language Pedagogy. Frankfurt: Peter Lang, 197–214.

Pajas P. & Štěpánek J. (2008). Recent Advances in a Feature-Rich Framework for Treebank Annotation. In: Proceedings of the 22nd International Conference on Computational Linguistics. Manchester, 673-680.

Schmidt T. (2004). Transcribing and Annotating Spoken Language with Exmaralda. In: Proc. of the LREC-Workshop on XML Based Richly Annotated Corpora, Lisbon 2004. Paris: ELRA.

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.