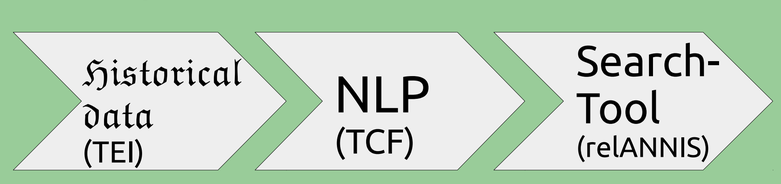
37. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft – Sektion CL, Leipzig 04.–06. März 2015

From TEI to linguistic corpora using Pepper

Florian Zipser, HU-Berlin IdSL
Martin Klotz, HU-Berlin IdSL
André Röhrig, HU-Berlin IdSL

Goal

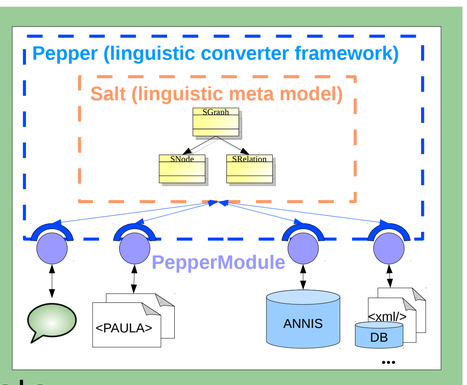
- Allowing TEI (Burnard & Bauman 2008) data to be processed with the Natural Language Processing tool WebLicht (Hinrichs et al. 2010)
- Bringing together analyses from different fields in a multi-layer corpus
- Visualizing the multi-layer corpus in the linguistic search- and visualization system ANNIS (Krause & Zeldes 2014)



Background

- Many research fields close to linguistics deal with textual resources e.g. historical sciences, literature and philology
- Sharing resources saves a lot of digitization effort
- A combination of analyses from different fields could lead to a better understanding of resources: for example named entity recognition or authorship detection could be helpful to classify historical texts
- A widely used format for the digitization especially of historical texts is TEI: huge set of meta data, text structure, codicological information etc.
- For linguistic annotations, formats like MMAX2, Tiger XML or TCF are more appropriate
- No simple transformation between formats 😢

SaltNPepper (Zipser & Romary 2010)



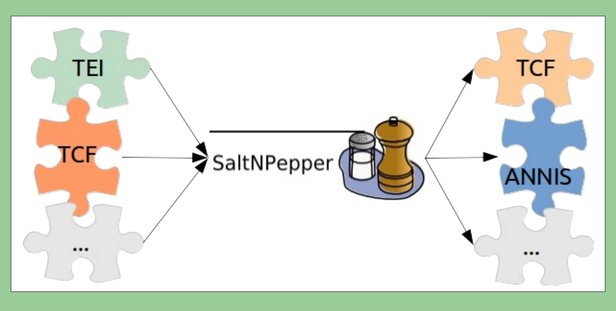
- Open source
  (Apache License, Version 2.0)
- OS independent (written in Java)
- http://u.hu-berlin.de/saltnpepper

Salt

- A graph-based meta model for linguistic data
- Abstraction of data: nodes, edges, labels ...
- Theory-neutral and independent of phenomena

Pepper

- A multi-format converter framework for linguistic data
- Easily extensible for further formats via Plug-in system
- Many existing modules:
  TigerXML, <tiger2/>, EXMARaLDA, MMAX2, rs3, PAULA, UAM, TreeTagger, CoNLL, Penn Treebank format, generic xml, ANNIS format, ...



New

- TEIModules (contains a customizable importer for a TEI subset)
  https://github.com/korpling/pepperModules-TEIModules
- TCFModules (contains im- and exporter for TCF)
  http://korpling.github.io/pepperModules-TCFModules/

Workflow

- Sample data: “Mannheimer Korpus Historischer Zeitungen und Zeitschriften” taken from the LAUDATIO repository

  (http://www.laudatio-repository.org)
  - Newspapers and magazines of the 18th, 19th and 20th century
  - Encoded in TEI: contains text structure like pages, linebreaks, tokenization, meta data etc.

```
<p>
  <w>
    <hi rend="ini2">
      <lb n="2" />J</hi>hro</w> Allerhöchste Kayserl. Majestät haben auf das
  <w>Rück=Schrei=
```

An excerpt of the corpus in its TEI representation containing a paragraph (<p>) a word (<w>) and a linebreak (<lb>)

TEI to TCF with Pepper

1. TEI to Salt with TEIModules
2. Salt to TCF with TCFModules

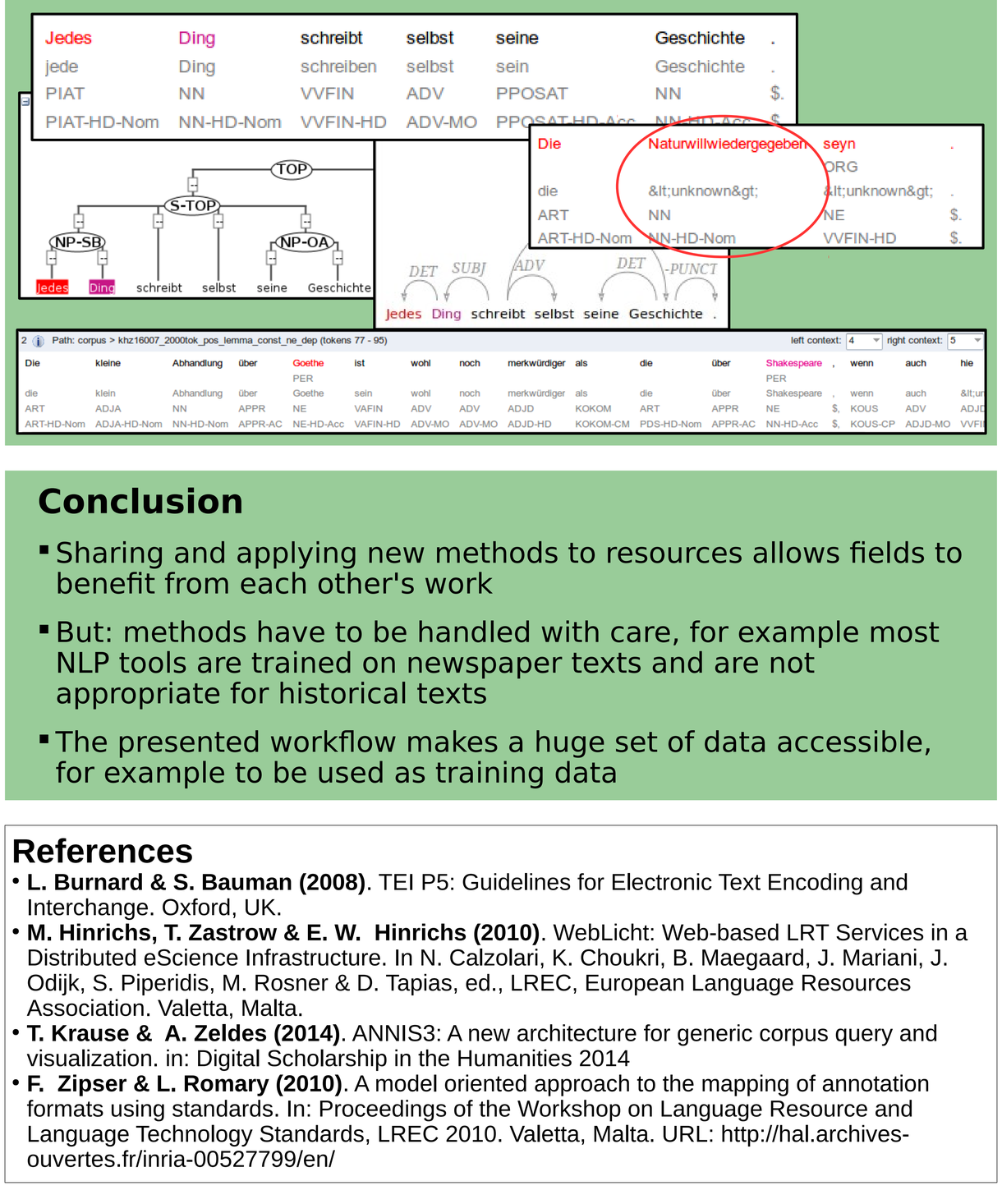
Processing data with WebLicht

http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php

- POS tagging
- Lemmatization
- Named Entity Recognition
- Syntax Parsing

Getting data into ANNIS

1. TCF to Salt with TCFModules
2. Salt to ANNIS format with RelANNISModules



Conclusion

- Sharing and applying new methods to resources allows fields to benefit from each other's work
- But: methods have to be handled with care, for example most NLP tools are trained on newspaper texts and are not appropriate for historical texts
- The presented workflow makes a huge set of data accessible, for example to be used as training data

References

- **L. Burnard & S. Bauman (2008)**. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford, UK.
- **M. Hinrichs, T. Zastrow & E. W. Hinrichs (2010)**. WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias, ed., LREC, European Language Resources Association. Valetta, Malta.
- **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/

SFB 632 "Information Structure"
CLARIN-D
DFG Deutsche Forschungsgemeinschaft
Bundesministerium für Bildung und Forschung

web: http://u.hu-berlin.de/saltnpepper
e-mail: saltnpepper@lists.hu-berlin.de