Digital Hummanities Tag an der Humboldt-Universität zu Berlin, Berlin 21. Mai 2015

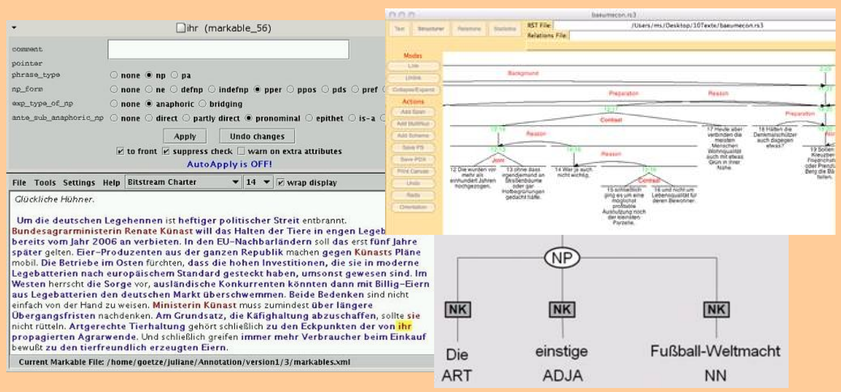
Institut für deutsche Sprache und Linguistik: Korpuslinguistik und Morphologie

SaltNPepper, ANNIS & Atomic: Eine Infrastruktur für Mehrebenenkorpora

Florian Zipser, André Röhrig, Anke Lüdeling, Martin Klotz, Thomas Krause, Stephan Druskat & Vivian Voigt
Humboldt-Universität zu Berlin, IdSL

Motivation

- In Sammlungen von Textdaten und dazugehörigen linguistischen Annotationen lassen sich empirisch linguistische Phänomene untersuchen.
- Sprachdaten können auf vielen Ebenen klassifiziert und in einem Korpus annotiert werden: Wörter können Wortarten zugeordnet werden, Sätze können syntaktisch annotiert und mit rhetorischen Strukturen angereichert werden, in Lernertexten können grammatische Fehler angegeben werden etc.
- Für die Annotation und Analyse einzelner Ebenen existieren unterschiedliche Werkzeuge: MMAX2, RSTTool, @nnotate, EXMARaLDA, Elan, TiGerSearch und viele weitere.

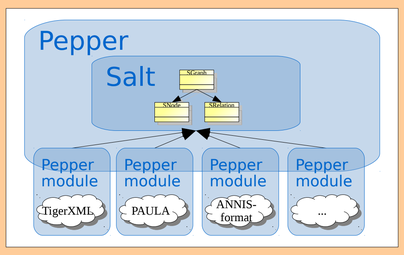


- Einige linguistische Phänomene wie bspw. Informationsstruktur lassen sich nur über mehrere Ebenen (Betonung, Wortstellung, Definitheit, Gegebenheit etc.) hinweg untersuchen (Lüdeling et al., erscheint).
- Die Werkzeuge verlangen unterschiedliche Eingabeformate und produzieren verschiedene Ausgabeformate. Dadurch ist eine Analyse über verschiedene Ebenen hinweg schwierig.
- Einige Werkzeuge werden nicht weiter gepflegt.

Ziele

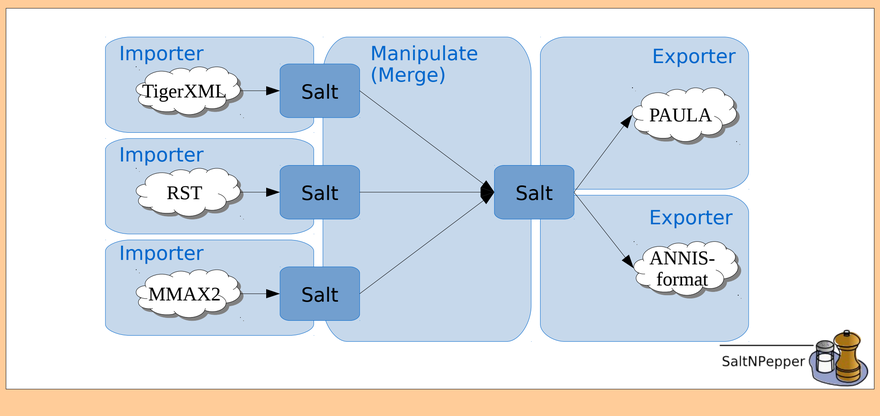
1. Zusammenführen der Ebenen zu Mehrebenenkorpora → Pepper
2. Ebenenübergreifende Analyse der Daten → ANNIS
3. Erstellung von Mehrebenenkorpora in einem Tool → Atomic

1.Pepper (Zipser & Romary 2010)



Pepper ist ein Konvertierungstool für linguistische Daten.

- Mit Pepper soll ein beliebiges linguistisches Format in jedes andere linguistische Format überführt werden können.
- Es wird eine graphbasierte Zwischenrepräsentation genutzt, die theorieneutral Daten darstellen kann.
- Einzelne Module übernehmen die Überführung der Daten in die Zwischenrepräsentation, dabei werden Daten zunächst aus dem Quellformat importiert, können dann manipuliert werden und werden schlussendlich in das Zielformat exportiert.

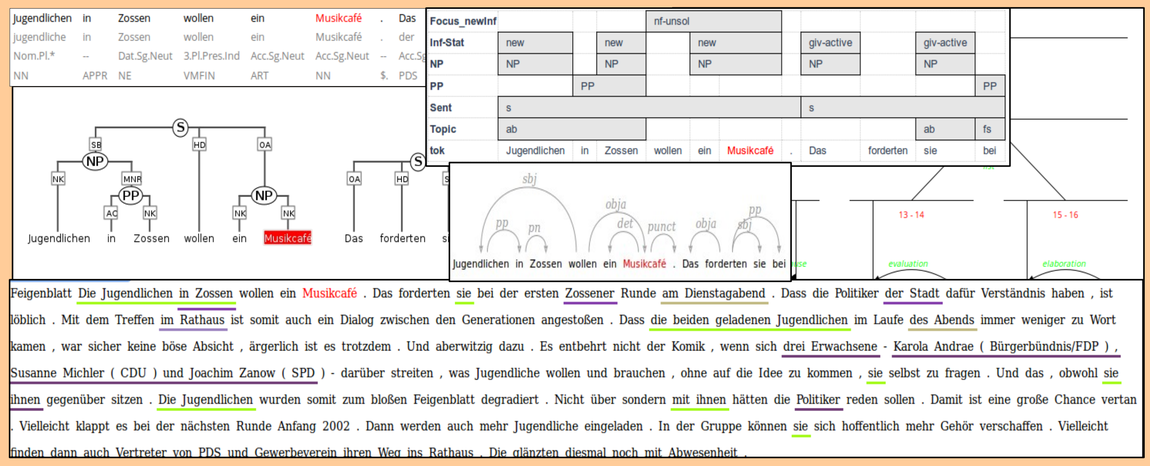


- Bisher werden folgende Formate unterstützt:
  Elan, CoNLL, MMAX2, ANNIS, Gate, RST, TCF, CoraXML, TreeTagger, Aldt, UAM, EXMARaLDA, generic XML, PTB, PAULA, TEI (subset), txt, SaltXML, …
- Durch einen Plugin-Mechanismus können weitere Module integriert werden.

2.ANNIS (Krause & Zeldes 2014)

ANNIS ist ein Such- und Visualisierungssystem für linguistische Daten und insbesondere Mehrebenenkorpora.

- Beliebige Korpora können mit ANNIS dargestellt werden, da es nicht für ein bestimmtes Korpus entwickelt wurde.
- Mit der Anfragesprache AQL können unterschiedliche Phänomene über alle Annotationsebenen hinweg in ANNIS gesucht werden.
- Ebenen- und annotationsspezifische Visualisierungen sind durch einen Plugin-Mechanismus in ANNIS integriert.
- ANNIS unterstützt UTF-8 und ist dadurch sprach- und alphabetunabhängig.

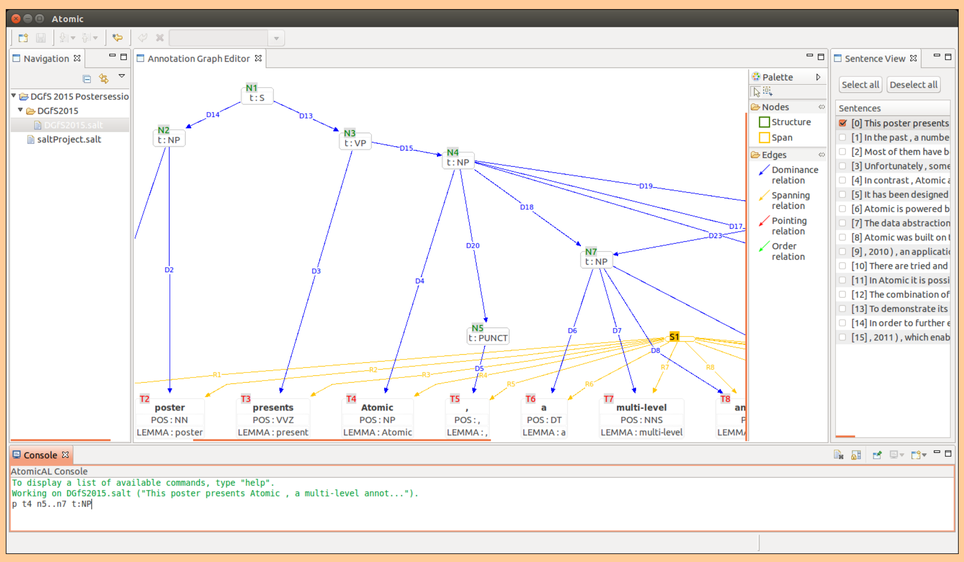


(excerpts of PCC in ANNIS)

3.Atomic (Druskat et al. 2014)

Atomic ist ein Mehrebenenannotationstool für linguistische Daten.

- Atomic ist graphbasiert und erlaubt die Erstellung sehr unterschiedlicher Annotationen und Ebenen.
- Es gibt einen allgemeinen Grapheditor und verschiedene auf bestimmte Arten der Annotation angepasste Editoren.
- Atomic basiert auf Eclipse und ist somit durch den integrierten Plugin-Mechanismus erweiter- und anpassbar.
  - Weitere linguistische Komponenten (z.B. Editoren, NLP-Plugins) können entwickelt und integriert werden.
  - Bestehende (nicht linguistische) Plugins wie bspw. Versionsverwaltung, kollaboratives Arbeiten, unterschiedliche Editoren (XML, TEI, HTML) etc. können problemlos integriert werden.



Referenzen

- **S. Druskat, L. Bierkandt, V. Gast, C. Rzymski & F. Zipser (2014).** Atomic: an open-source software platform for multi-level corpus annotation. In J. Ruppert & G. Faaß (eds.): Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014), October 2014 (pp. 228–234). ISBN 978-3-934105-46-1.
- **A. Lüdeling, J. Ritz, M. Stede & A. Zeldes (erscheint).** Corpus Linguistics. In: C. Fery, & S. Shinishiro (Hrsg.) OUP Handbook of Information Structure. Oxford University Press, Oxford.
- **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/

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

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