LAUDATIO - Eine Infrastruktur zur linguistischen Analyse historischer Korpora

Carolin Odebrecht, Humboldt-Universität zu Berlin
Florian Zipser, Humboldt-Universität zu Berlin, INRIA

Linguistische Analyse historischer Korpora - Fragestellung

Fragestellung:

Was heißt es, (korpus-)linguistisch zu arbeiten?

Welche Benutzerszenarien ergeben sich aus der Datenaufbereitung und -analyse (historisch-)linguistischer Forschungsdaten?

Welche Anforderungen stellen sich dann für eine Infrastruktur, die diese Nutzer ansprechen will?

Linguistische Analyse historischer Korpora - Forschung

Forschungsfrage bedingt

- Datenauswahl
- Annotationen
- Art der Annotationen
- Auswertung der Annotationen

→ Folge: heterogenes Feld an Daten (Editoren + Formate) und Auswertungen (Analysetool + Formate)

→ Kombination dieser verschiedenen Annotationen und Analysen

→ Infrastrukturen müssen damit umgehen können

Linguistische Analyse historischer Korpora - Nutzerperspektive

Nutzerperspektive Linguistik:

- (Weiter-)Nutzung dieser heterogenen Forschungsdaten in Abhängigkeit von Forschungsfragen
- grundsätzliche Anwendungen

/1/ Daten annotieren
/2/ Daten speichern
/3/ Daten durchsuchen
/4/ auf Daten zugreifen

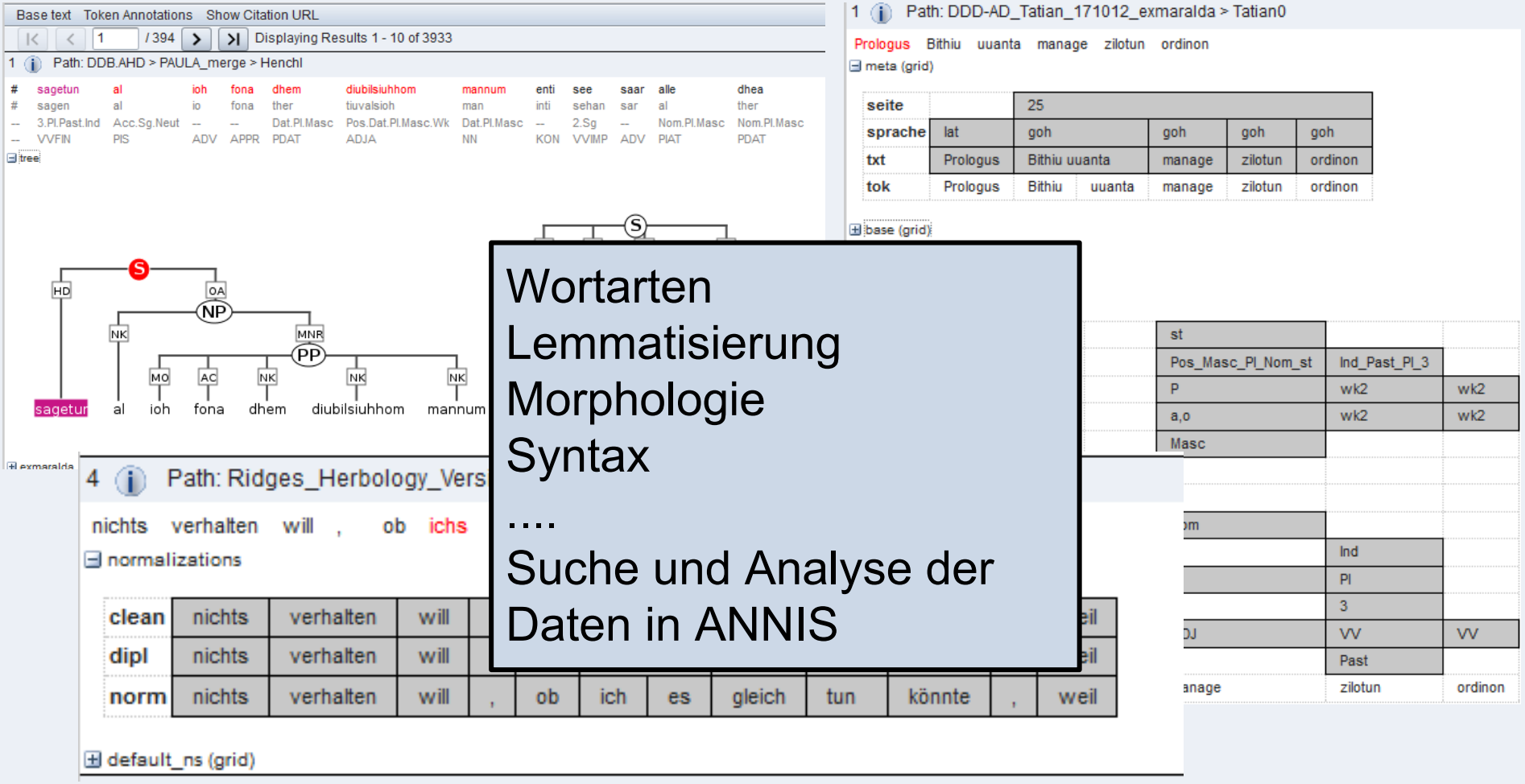
Linguistische Analyse historischer Korpora - Nutzerperspektive

Nutzerperspektive Linguistik:

- (Weiter-)Nutzung dieser heterogenen Forschungsdaten in Abhängigkeit von Forschungsfragen
- grundsätzliche Anwendungen

/1/ Daten annotieren
/2/ Daten speichern
/3/ Daten durchsuchen
/4/ auf Daten zugreifen

Linguistische Analyse historischer Korpora

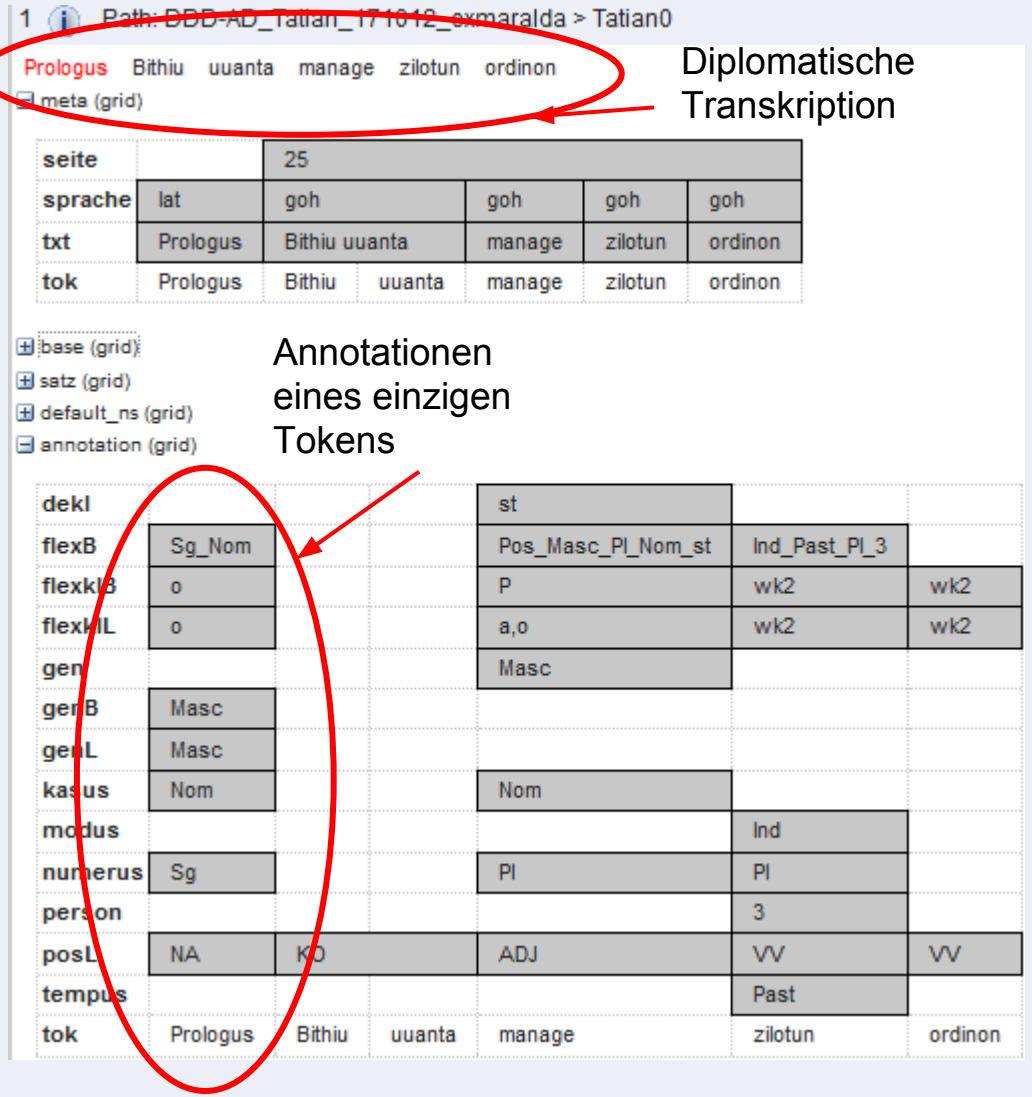


Linguistische Analyse historischer Korpora - Multikey-Annotationen

Tatian DDD-AD

Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsgegen-stand
- Token und/oder Spannen-annotationen erhalten auf unterschiedlichen Ebenen Werte:
  - Numerus
  - Kasus
  - Genus
  - Modus
  - Part of Speech

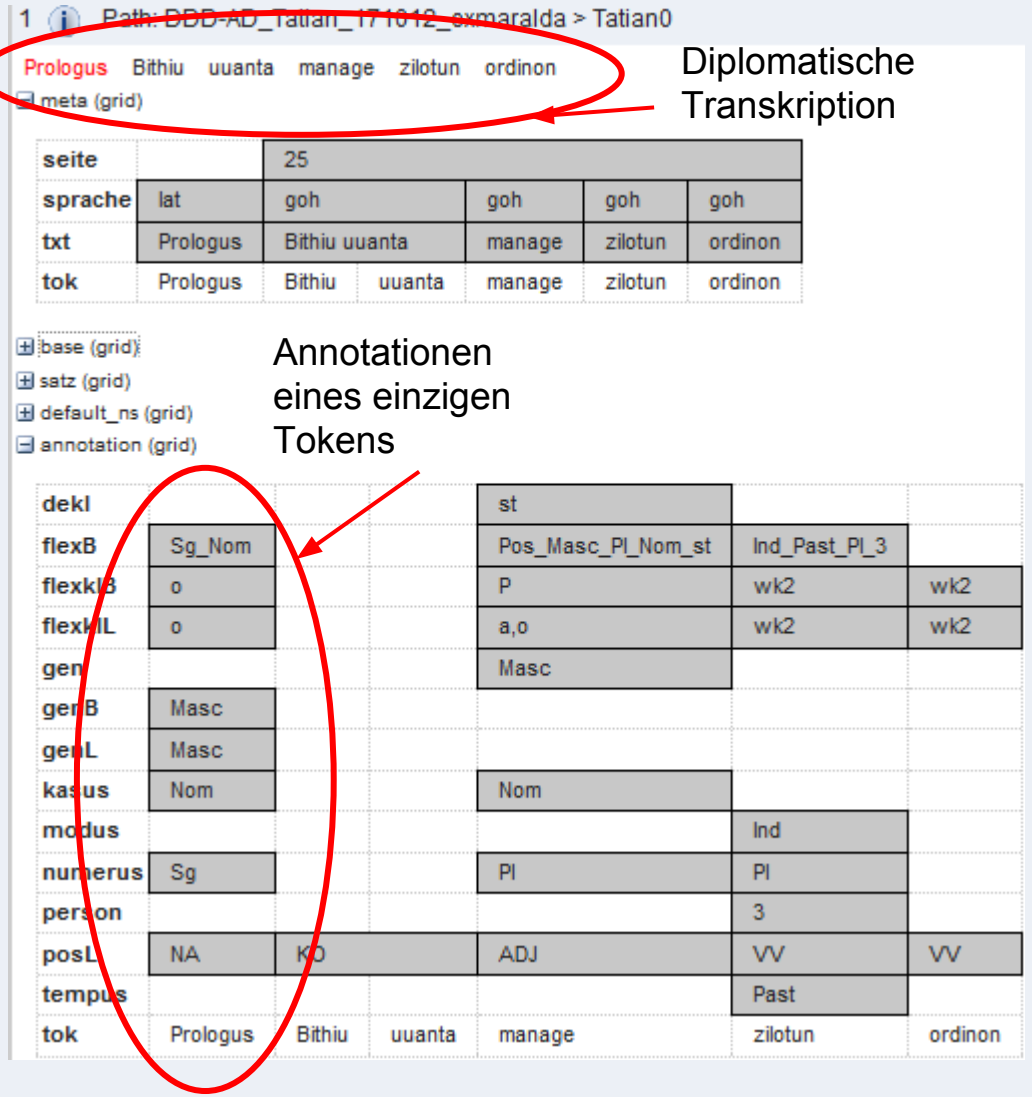


Linguistische Analyse historischer Korpora - Multikey-Annotationen

Tatian DDD-AD

Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsgegen-stand
- Token und/oder Spannen-annotationen erhalten auf unterschiedlichen Ebenen Werte:
  - Numerus
  - Kasus
  - Genus
  - Modus
  - Part of Speech

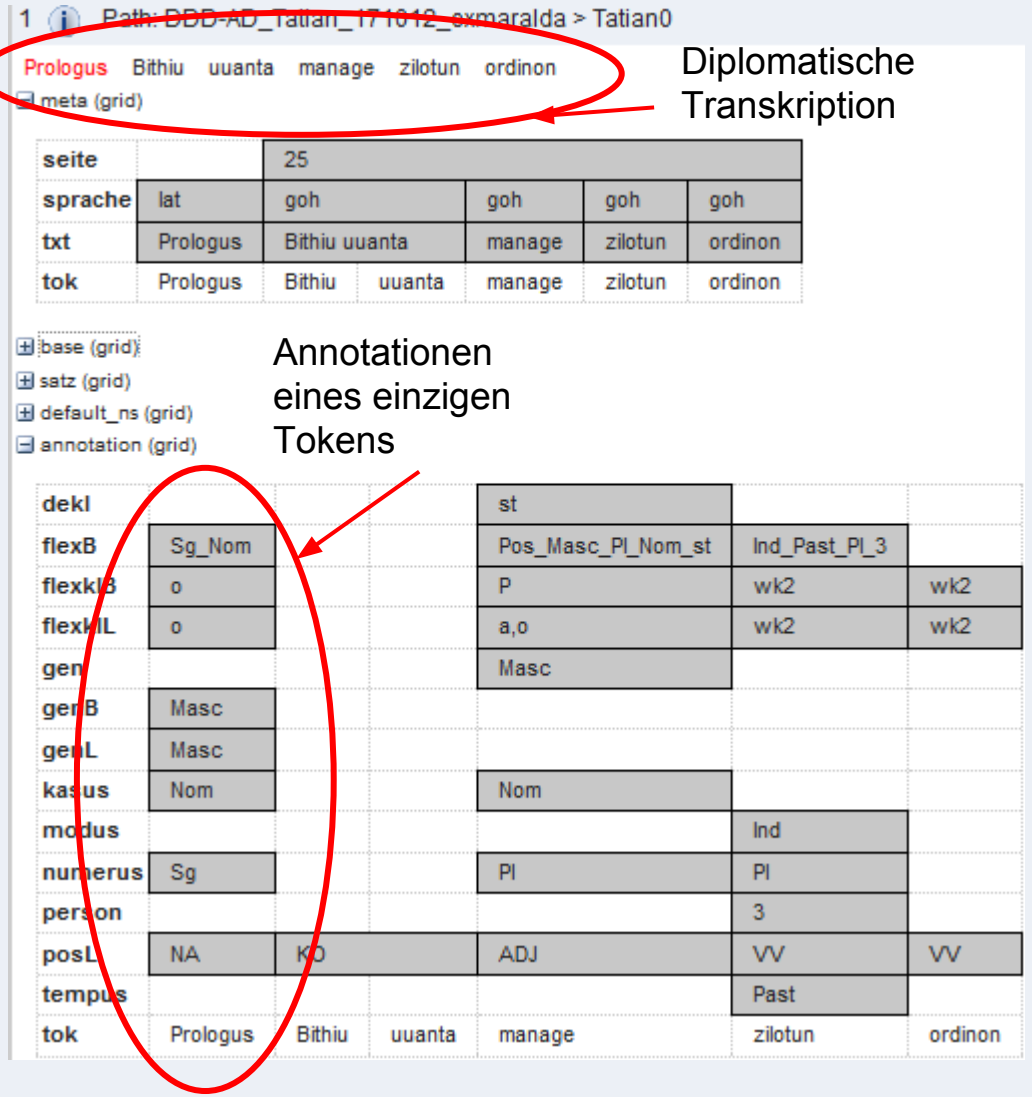


Linguistische Analyse historischer Korpora - Multikey-Annotationen

Tatian DDD-AD

Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsgegen-stand
- Token und/oder Spannen-annotationen erhalten auf unterschiedlichen Ebenen Werte:
  - Numerus
  - Kasus
  - Genus
  - Modus
  - Part of Speech



Linguistische Analyse historischer Korpora - Multikey-Annotationen

Tatian DDD-AD

Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsg stand
- Token und/oder Spann annotationen erhalten auf unterschiedlichen Ebenen

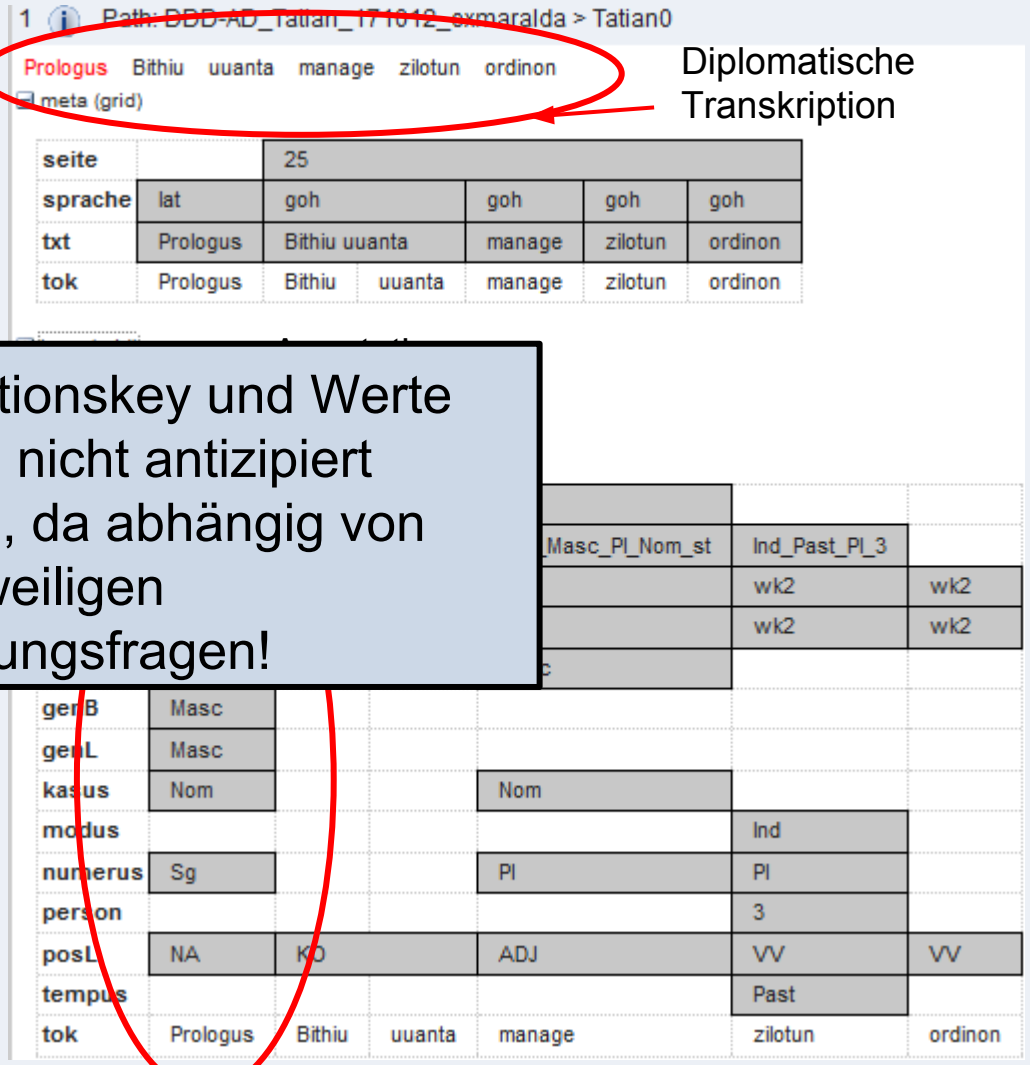
  Numerus

  Kasus

  Genus

  Modus

  Part of Speech



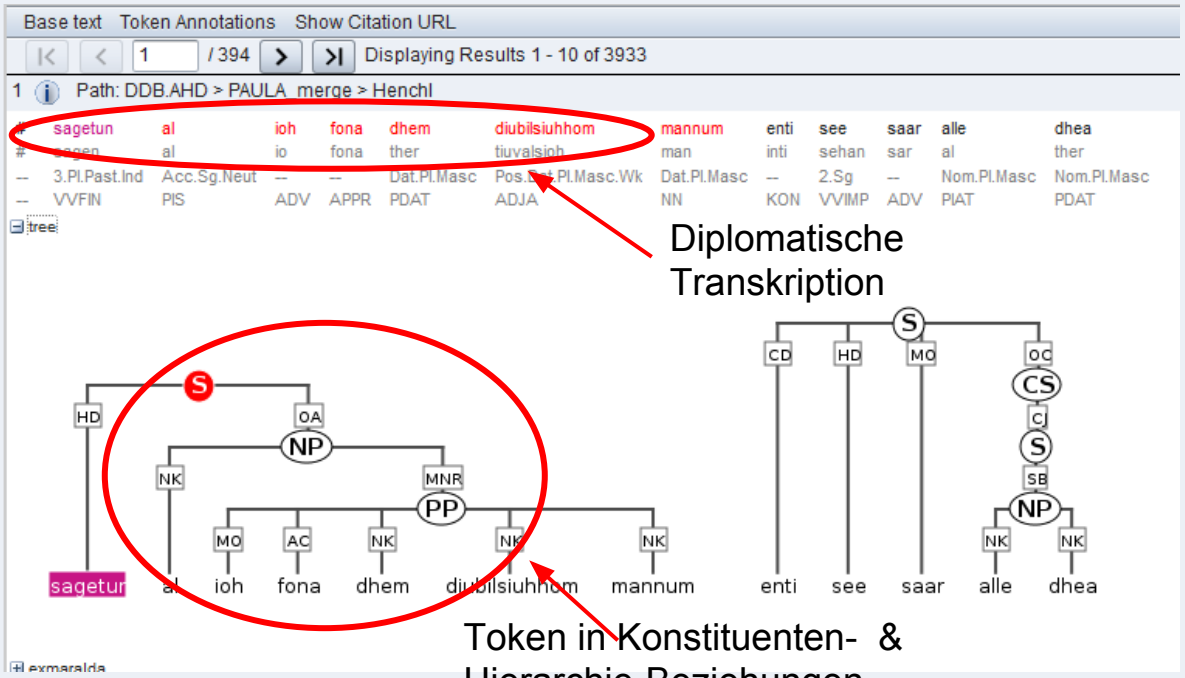
Annotationskey und Werte können nicht antizipiert werden, da abhängig von den jeweiligen Forschungsfragen!

Linguistische Analyse historischer Korpora - Bäume und Konstituenten

Deutsche Diachrone Baumbank

- Syntax-Annotationen ermöglichen eine Suche und Darstellung von komplexen Hierarchie-Beziehung

- Suche nach syntaktisch abhängigen Strukturen:



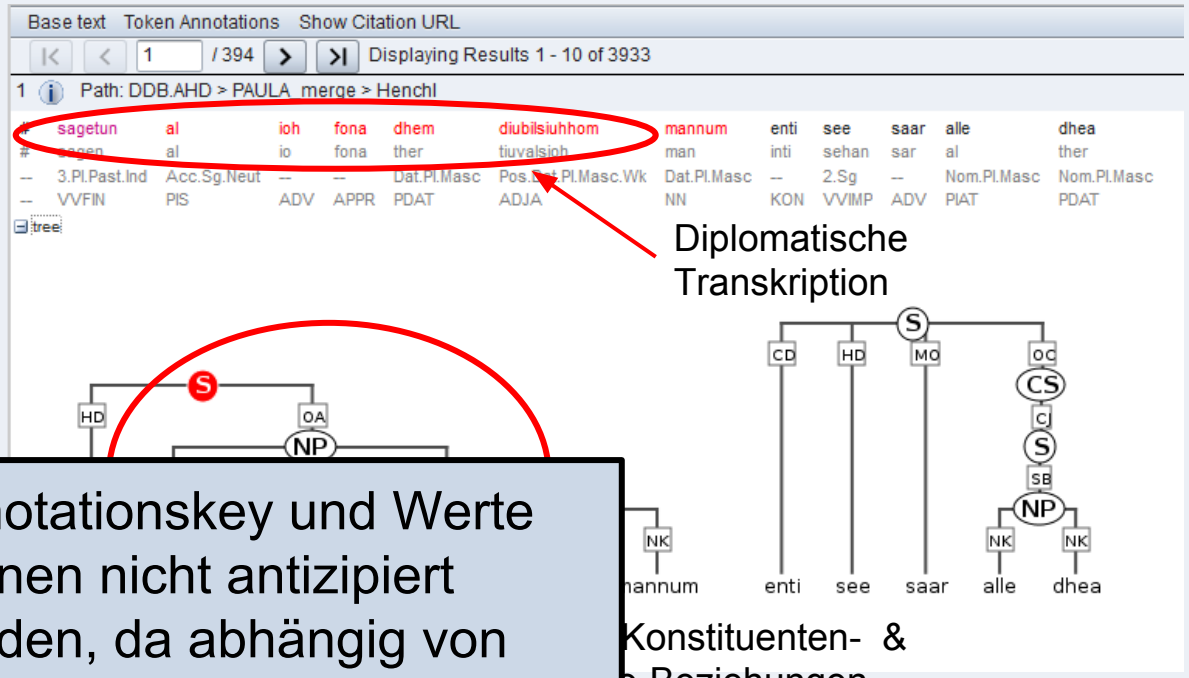
bspw. Nominalphrase, die mit einer bestimmten grammatischen Funktion, wie Subjekt, gelabelt sind

→ theoretische Schlüsse über die Syntax so möglich

Linguistische Analyse historischer Korpora - Bäume und Konstituenten

Deutsche Diachrone Baumbank

- Syntax-Annotationen ermöglichen eine Suche und Darstellung von komplexen Hierarchie-Beziehung



- Suche nach syntakt Konstituenten- & abhängigen Strukturen: e-Beziehungen

Annotationskey und Werte können nicht antizipiert werden, da abhängig von den jeweiligen Forschungsfragen!

bspw. Nominalphrase, die mit einer bestimmten grammatischen Funktion, wie Subjekt, gelabelt sind

→ theoretische Schlüsse über die Syntax so möglich

Linguistische Analyse historischer Korpora - Multiple Tokenisierung

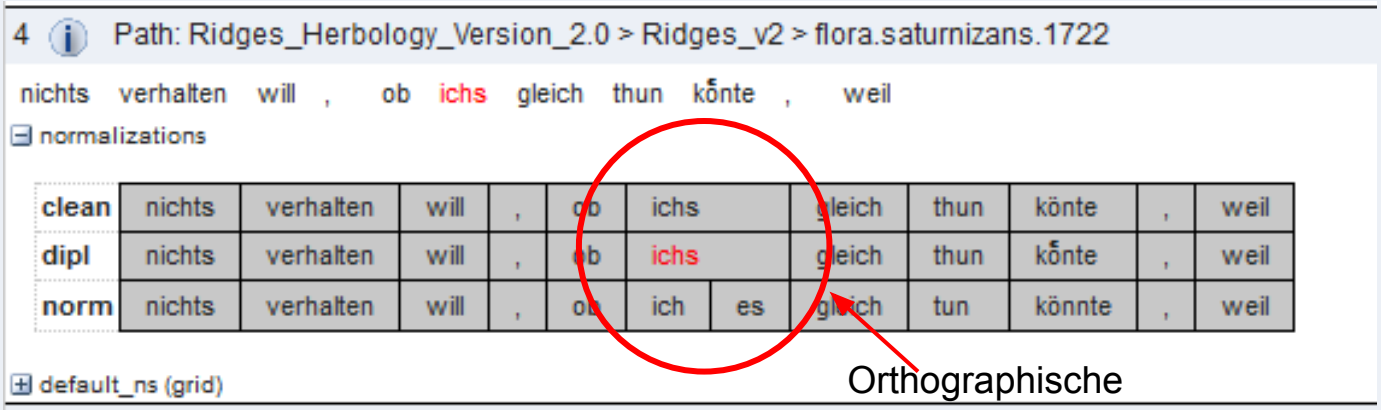
RIDGES Herbology

- Erhalt der diplomatischen Transkription
- Normalisierungen ermöglichen eine verbesserte Mustersuche/Suche nach Typen sowie bessere Performanz von Tagging- bzw. Parsingtools
- morphologisch wie syntaktisch interessante Phänomene so erst suchbar:

  bspw. Partikelverb-Konstruktionen, Klitika, Komposition

→ Mehrebenen-Normalisierung durch multiple Tokenisierung

→ linguistischer Vorteil: Varianz der Orthographie noch nachvollziehbar



Orthographische Realisierung von Klitika + Normalisierung

Linguistische Analyse historischer Korpora - Multiple Tokenisierung

RIDGES Herbology

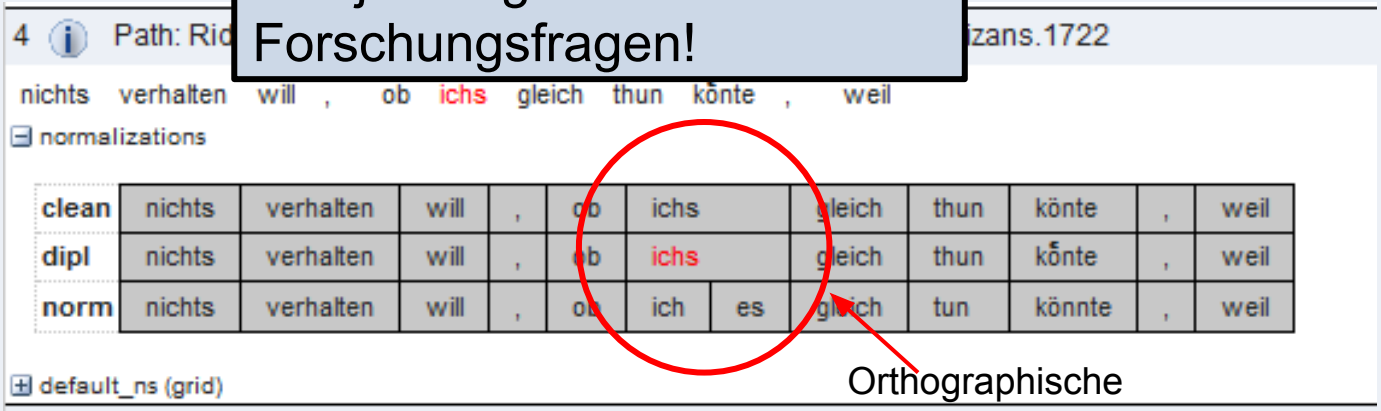
- Erhalt der diplomatischen Transkription
- Normalisierungen ermöglichen eine verbesserte Mustersuche/Suche nach Typen sowie bessere Performanz von Tagging- bzw. Parsingtools
- morphologisch wie syntaktisch interessante Phänomene so erst suchbar:

bspw. Partikelve…sition

→ Mehrebenen-Nor…sierung

→ linguistischer Vor…ch nachvollziehbar

Annotationskey und Werte können nicht antizipiert werden, da abhängig von den jeweiligen Forschungsfragen!



Orthographische Realisierung von Klitika + Normalisierung

Infrastruktur für historische Korpuslinguistik - Annotationen

- Annotationen **als Grundlage** für Analyse
  - Forschungsfrage, immer theoretisch motiviert
  - Auswahl der Daten
  - unterschiedliche Tag-Sets
  - Datenaufbereitung

- Bildung von Kategorien, Zuweisung dieser (Annotation), Suche/Analyse dieser über große Menge von Daten (Korpus)
- viele Gemeinsamkeiten mit modernen Korpora (Aufbereitung und Analyse)

→ Aufgabe: Dokumentation dieser Informationen

Infrastruktur für historische Korpuslinguistik - Metadaten

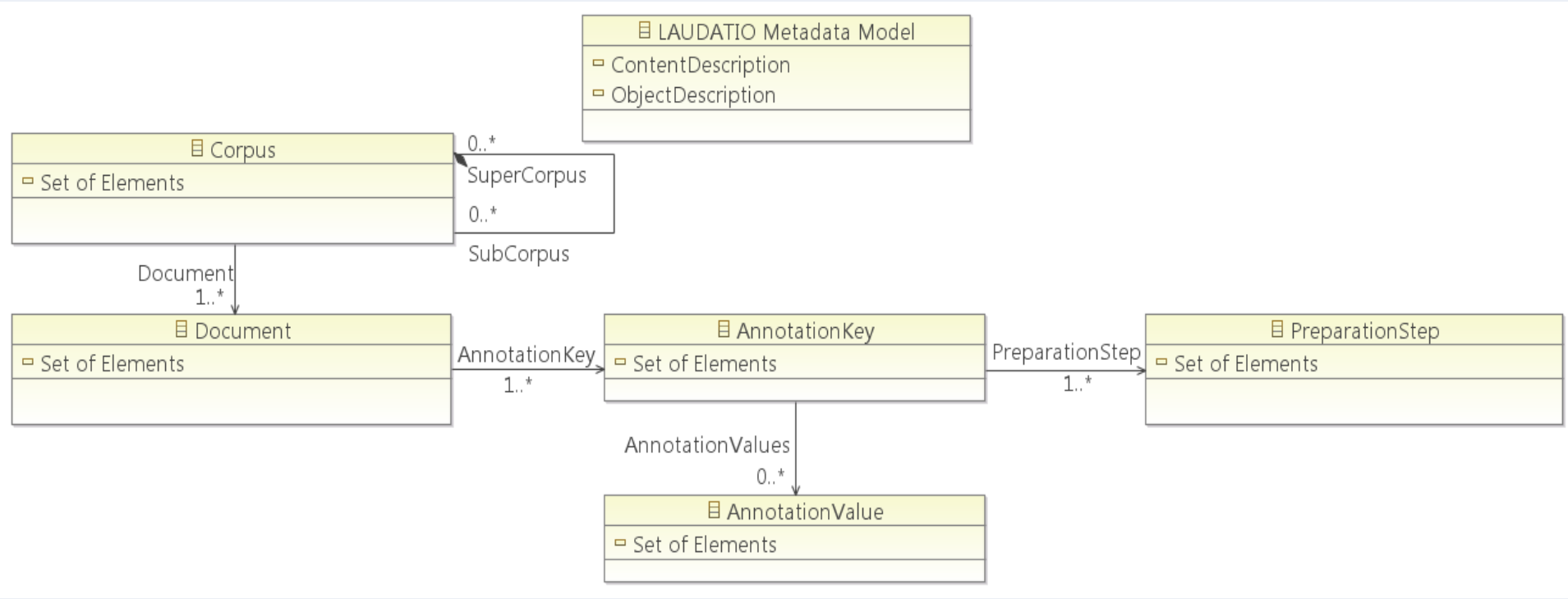
Neben klassischen Metadaten wollen wir (u.a.) wissen:

- Wie wurde transkribiert?
- Wie wurde annotiert?
- Mit welchen Tagset?
- Wit welchen Tools?
- Welche Ebenen beziehen sich auf einander?
- Wurde die Qualität überprüft?

→ (Weiter-)Verwendung dieser Daten durch Dokumentation der Verarbeitungsschritte eines Korpus

→ einheitlich strukturierte Erfassung (Dokumentation) der heterogenen Datengrundlage durch standardisierte Metadaten

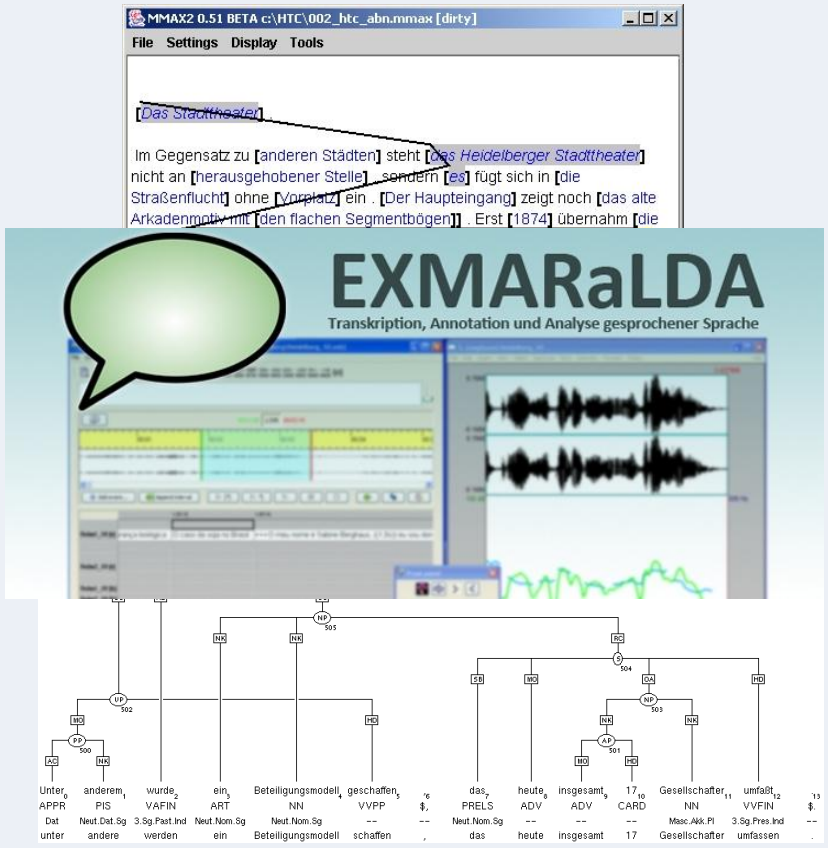
Infrastruktur für historische Korpuslinguistik - Metadaten



Modulares Metadatenmodell mit den Komponenten "Corpus", "Document", "AnnotationKey", "AnnotationValue" und "PreparationStep"

Infrastruktur für historische Korpuslinguistik - Daten annotieren

/1/ Daten annotieren

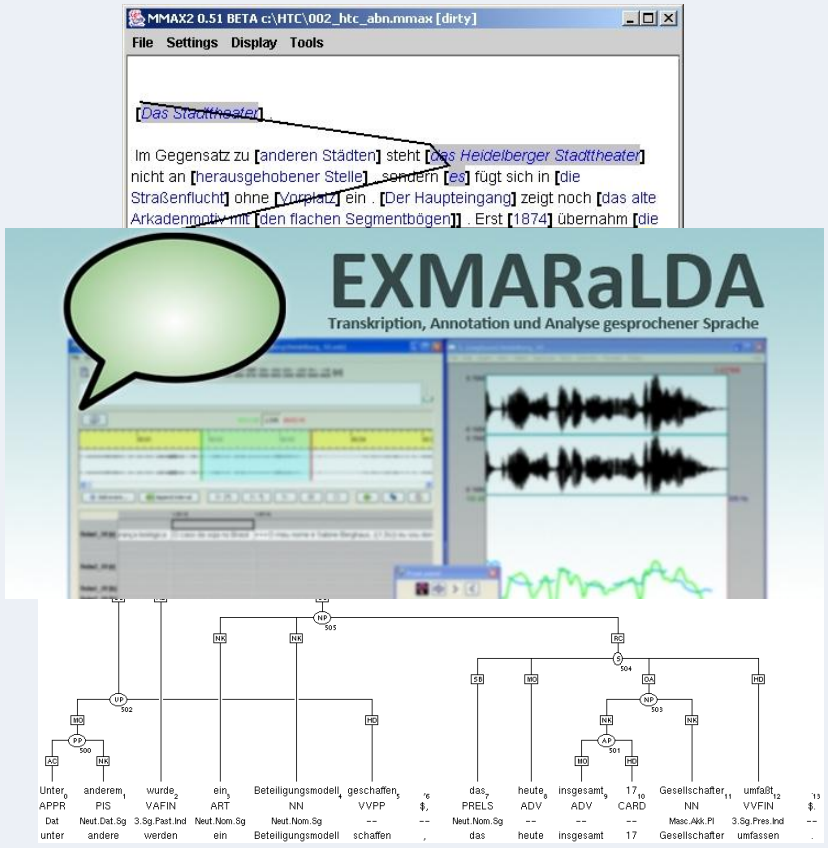


Viele spezifische Tools für unterschiedliche Annotationen:

- MMAX2, EXMARaLDA, TIGERSearch (Synpathie), RSTTool...

Infrastruktur für historische Korpuslinguistik - Daten speichern

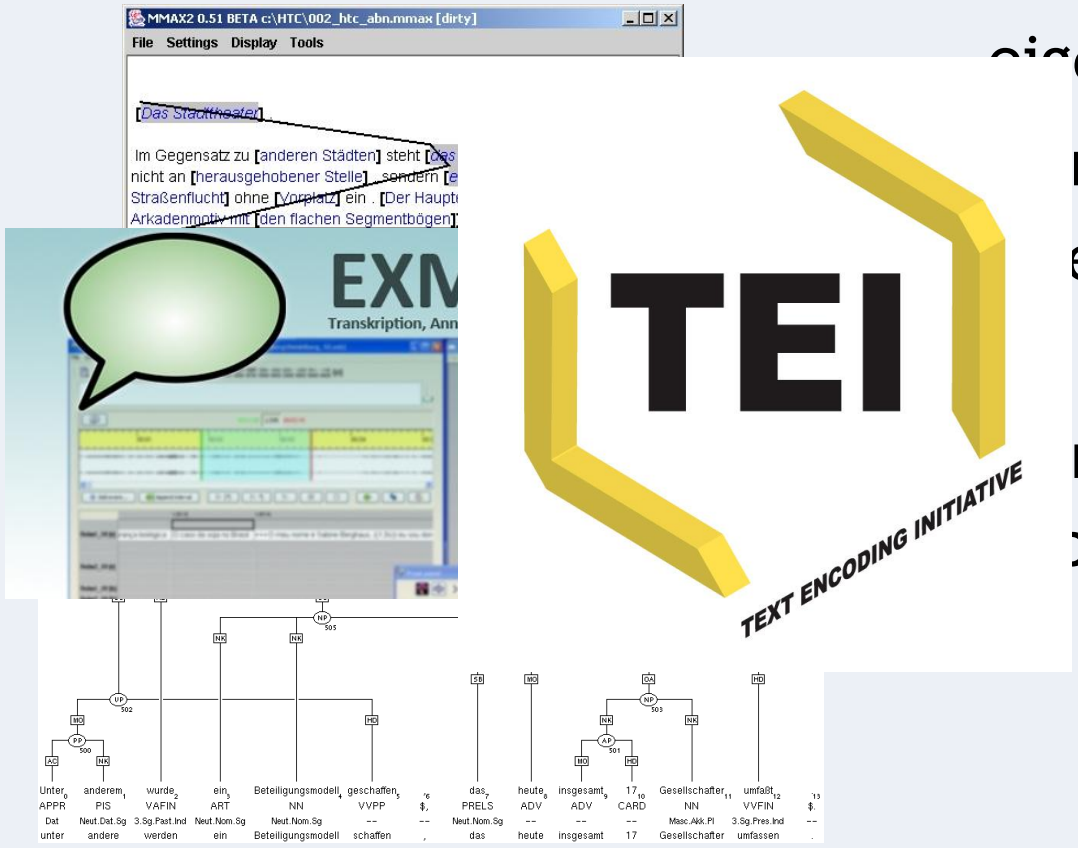
/2/ Daten speichern



- Jedes Tool hat ein eigenes Format :-(
- wenig Interoperabilität :-(
- manche Tools werden nicht mehr gepflegt :-(

Infrastruktur für historische Korpuslinguistik - Daten speichern

/2/ Daten speichern



- Jedes Tool hat ein
  eigenes Format :-(
  nig
  eroperabilität :-(
  nche Tools
  rden nicht mehr
  oflegt :-(

Infrastruktur für historische Korpuslinguistik - TEI

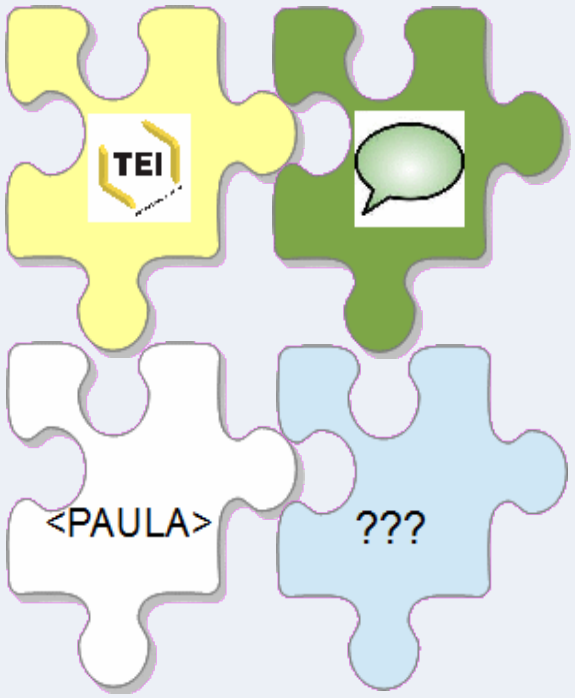
- Warum reicht uns TEI nicht?

- bereits existierende Korpora oft nicht in TEI
- viele bestehende Tools unterstützen kein TEI
- viele neue Kategorien pro Tag (individuelle Forschungsfrage)
  - Formate müssen flexibel/ generisch sein
- viele Annotationsarten nicht gut von TEI unterstützt
- Linguistik: nicht nur reine Digitalisierung des Originals

Infrastruktur für historische Korpuslinguistik - Was nun?

- erst Kombination der Formate ergibt möglichst umfangreiches Bild

→ Formatpluriversum



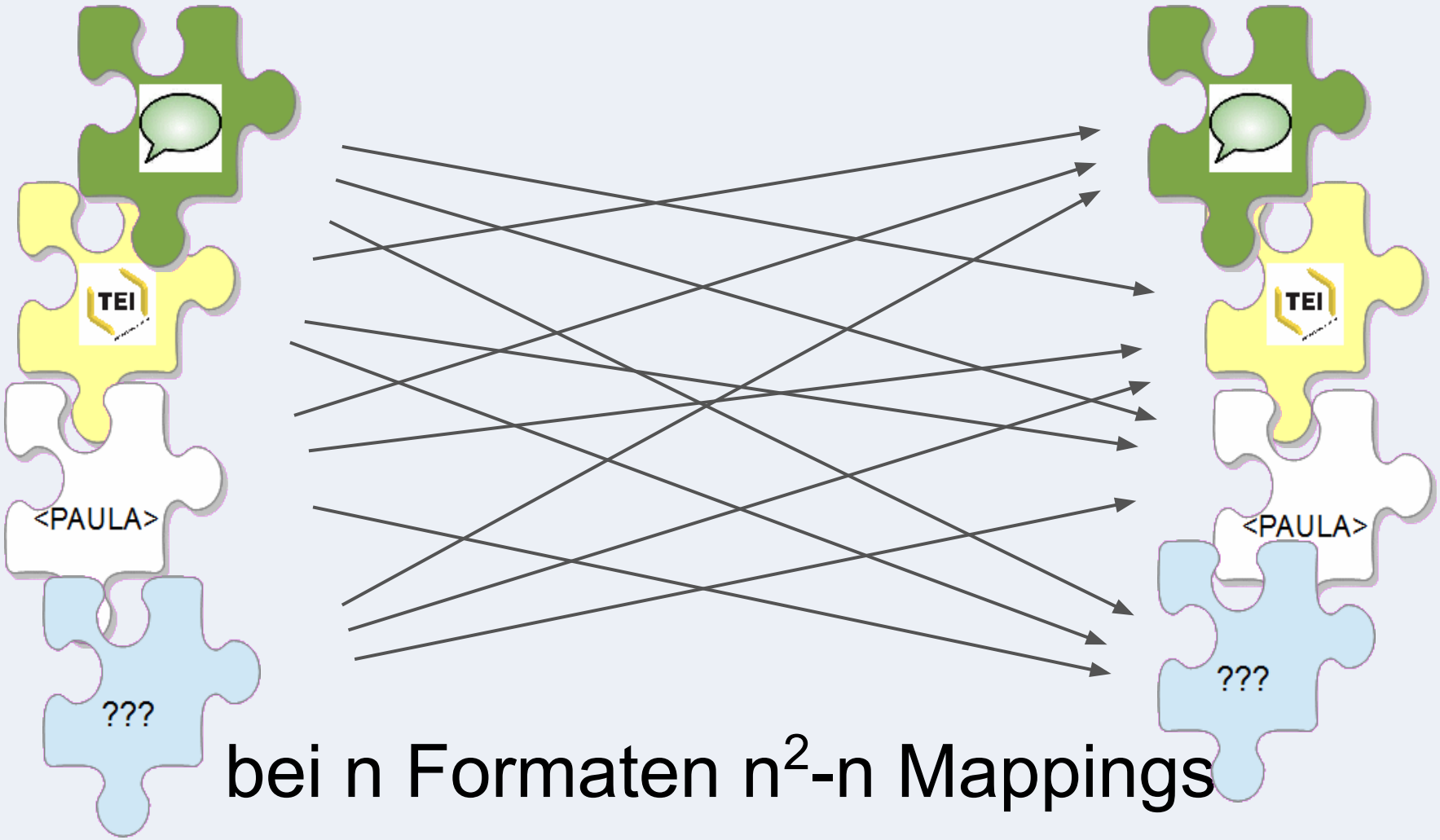
Infrastruktur für historische Korpuslinguistik - Formatpluriversum

- aber wir wollen Interoperabilität zwischen Formaten soweit wie möglich
  - Digitaler Text,
  - Tokenisierung und
  - bestimmte Annotation

  in verschiedenen Tools nutzbar
- Annotationen abbildbar auf Mehrebenenformate

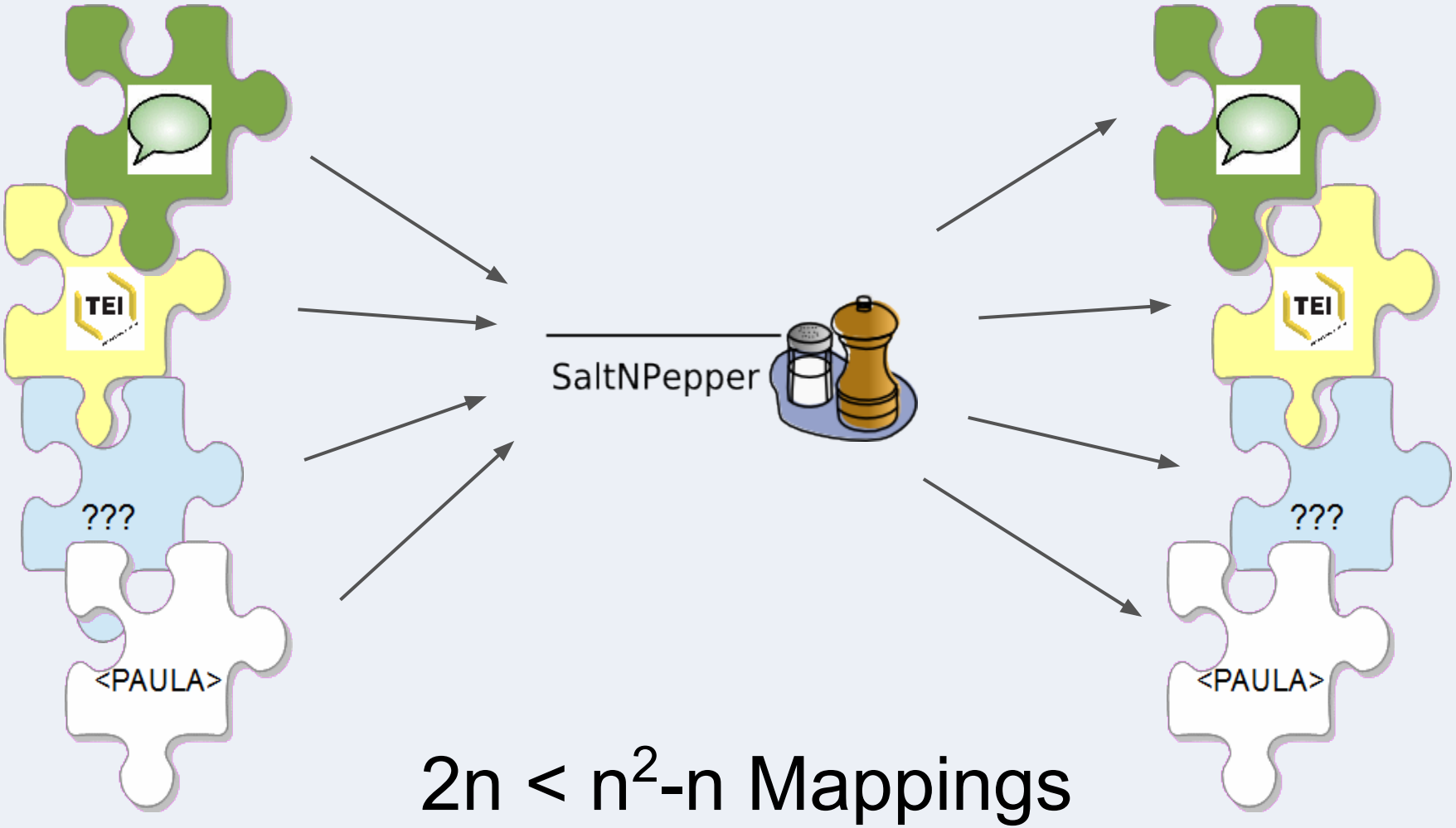
→ Überführen was zu überführen geht

Infrastruktur für historische Korpuslinguistik - Formatpluriversum



bei n Formaten n^2-n Mappings

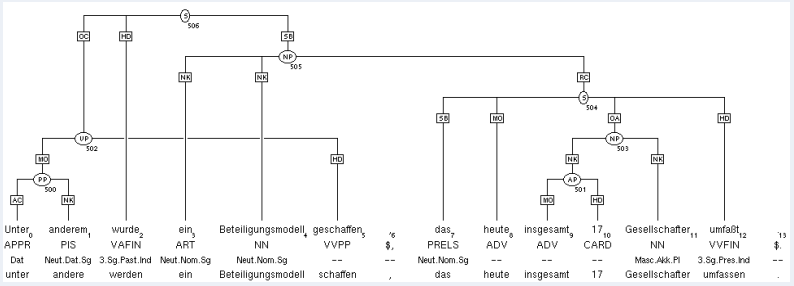
Infrastruktur für historische Korpuslinguistik - Formatpluriversum



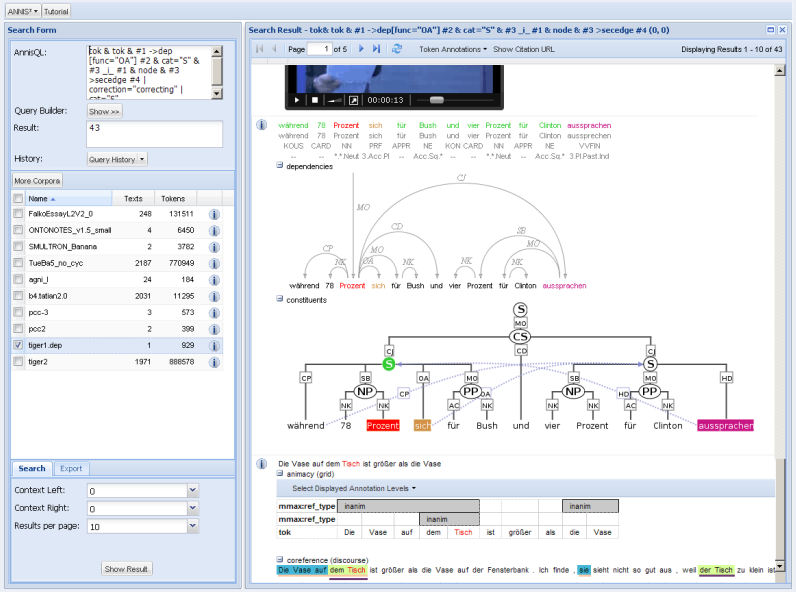
$2n < n^2-n$ Mappings

Infrastruktur für historische Korpuslinguistik - Daten durchsuchen

/3/ Daten durchsuchen



- auf einzelner Ebene



- parallel auf mehreren Ebenen

Infrastruktur für historische Korpuslinguistik - Datenzugriff

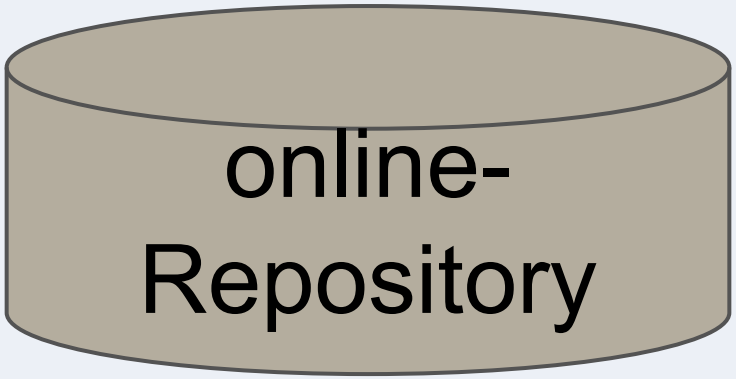
/4/ auf Daten zugreifen

- immer und überall
- Upload/ Download
  - in verschiedenen Formaten- Facetten
- Versionierungssystem
- einheitliche Metadatensuche

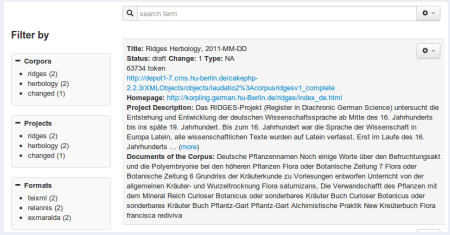
Infrastruktur für historische Korpuslinguistik - Datenzugriff

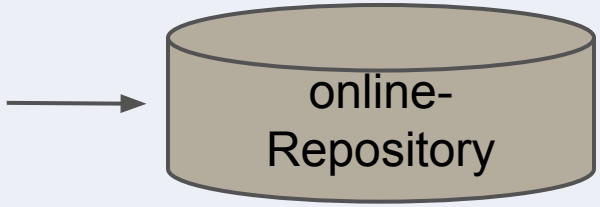
/4/ auf Daten zugreifen

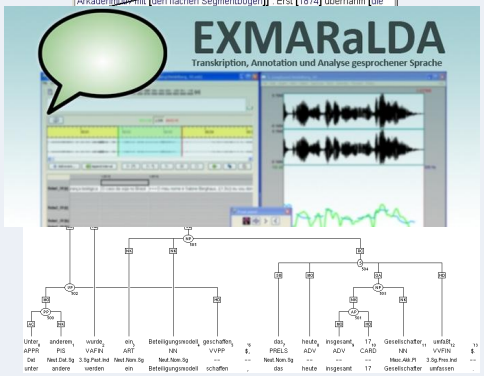
- immer und überall
- Upload/ Download
  - in verschiedenen Formaten- Facetten
- Versionierungssystem
- einheitliche Metadatensuche

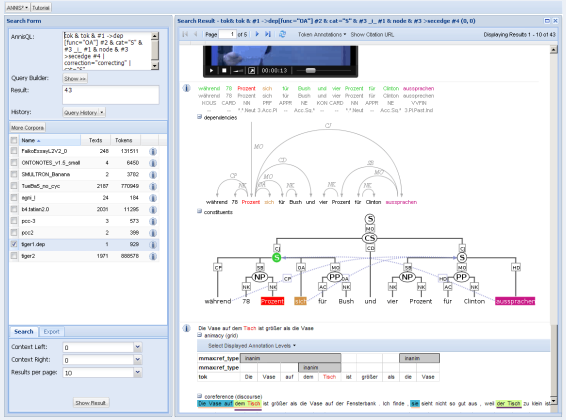


LAUDATIO - Long-term Access and Usage of Deeply Annotated Information



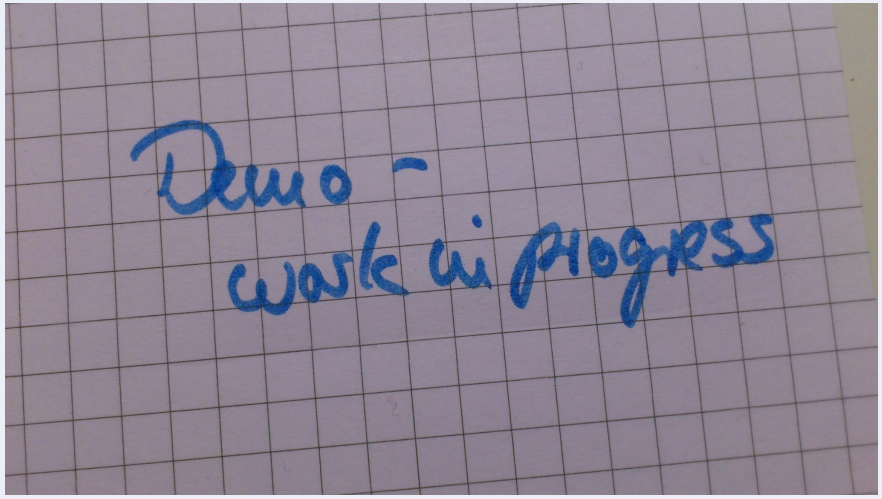






Vielen Dank für Ihre Aufmerksamkeit!

Demo von LAUDATIO
auf der DGfS-Tagung 2013 in Potsdam!



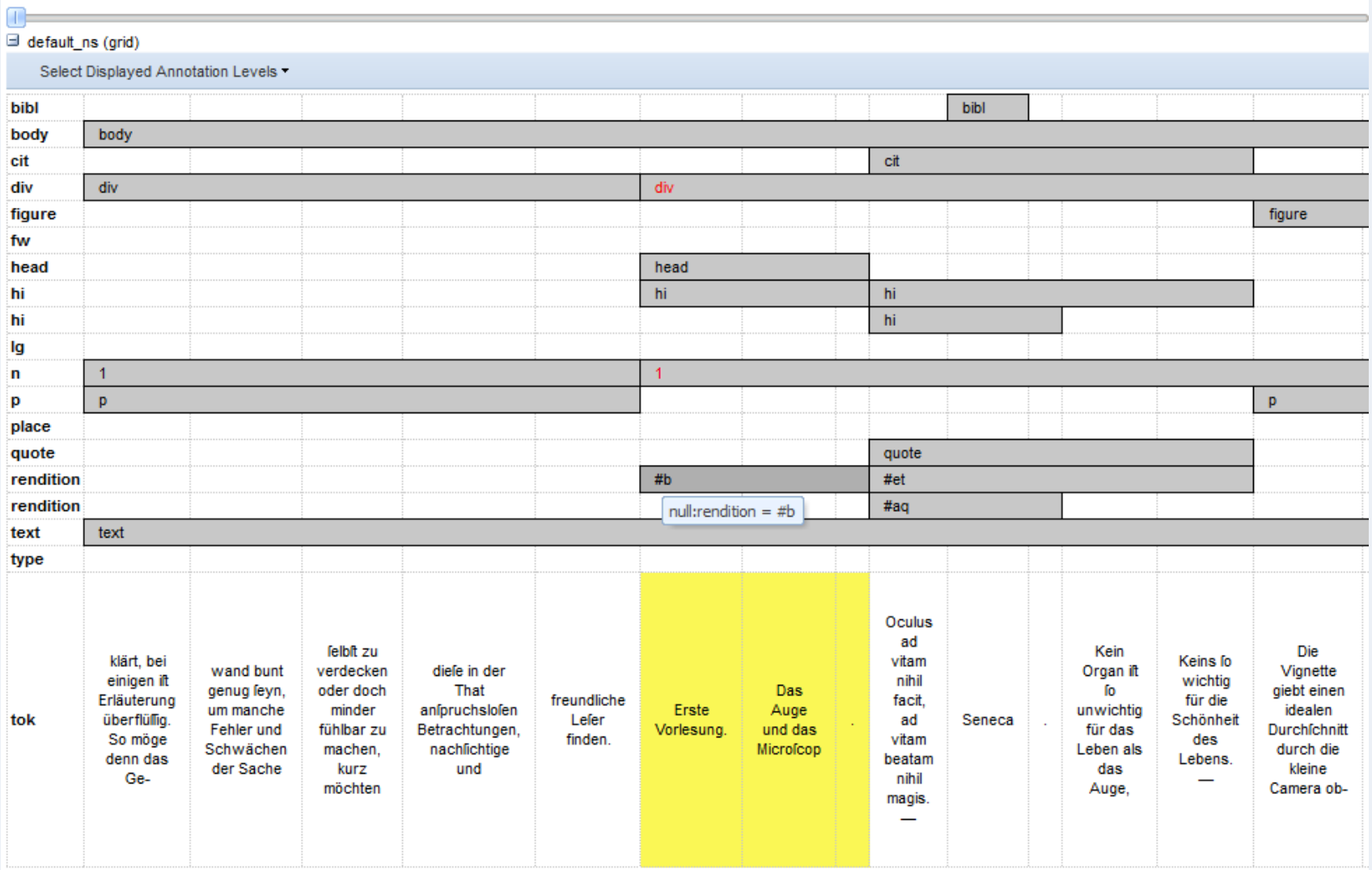
Referenzen

ANNIS	Zeldes, Amir, Ritz, Julia, Lüdeling, Anke & Christian Chiarcos (2009) **ANNIS: A Search Tool for Multi-Layer** Annotated Corpora. In: *Proceedings of Corpus Linguistics 2009*, Liverpool, July 20-23, 2009.
Deutsche Diachrone Baumbank	Hirschmann, Hagen & Sonja Linde (erscheint): **Annotationsbeschreibung für syntaktische Annotationen unterschiedlicher Sprachstufen des Deutschen**. Technical Report. Humboldt-Universität zu Berlin. https://korpling.german.hu-berlin.de/annis3/
EXMARaLDA	Schmidt, Thomas (2002) **EXMARaLDA - ein System zur Diskurstranskription auf dem Computer.** *Arbeiten zur Mehrsprachigkeit*, Folge B 34:1 ff. http://www.exmaralda.org/files/AZM.pdf.
LAUDATIO	Krause, Thomas, Odebrecht, Carolin & Dennis Zielke (erscheint) **Langfristiger Zugang und Nutzung von tief annotierten Korpora: LAUDATIO.** *32. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. http://www.laudatio-repository.org/
MMAX2	Müller, Christoph & Michael, Strube (2006): **Multi-Level Annotation of Linguistic Data with MMAX2**. In: Sabine Braun, Kurt Kohn & Joybrato Mukherjee (Eds.): *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, pp. 197-214. (English Corpus Linguistics, Vol.3).
PAULA	Dipper, Stefanie (2005) **XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation.** In: Eckstein R, Tolksdorf R (Eds.) *Berliner XML Tage*.

Referenzen

RIDGES Herbology	Krause, Thomas, Lüdeling, Anke, Odebrecht, Carolin & Amir Zeldes (2012) **Multiple Tokenization in a Diachronic Corpus**. *Exploring Ancient Languages through Corpora Conference (EALC), 14.-16.Juni 2012.* http://korpling.german.hu-berlin.de/ridges/documentation_en.html https://korpling.german.hu-berlin.de/annis3/
RSTTool	O'Donnell, Michael (2000) "**RSTTool 2.4 -- A Markup Tool for Rhetorical Structure Theory**". *Proceedings of the International Natural Language Generation Conference* (INLG'2000), 13-16 June 2000, Mitzpe Ramon, Israel. 253 - 256.
SaltNPepper	Zipser, Florian & Laurent 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. Malta.* URL: http://hal.archives-ouvertes.fr/inria-00527799/en/
Tatian DDD	Linda, Sonja, Unverzagt, Silke & Karin Donhauser (erscheint) **Old German Reference Corpus**. *32. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft*. https://korpling.german.hu-berlin.de/annis3/
TEI	Burnard, Lou & Syd Bauman (Eds.) (2008). **TEI P5: Guidelines for Electronic Text Encoding and Interchange.** Oxford. http://www.tei-c.org/Guidelines/P5/.
TIGERSearch	Lezius, Wolfgang (2002) **Ein Suchwerkzeug für syntaktisch annotierte Textkorpora.** http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/
Treetagger	Schmid, Helmut (1994) **Probabilistic Part-of-Speech Tagging Using Decision Trees**. In: *Proceedings of International Conference on New Methods in Language Processing*.

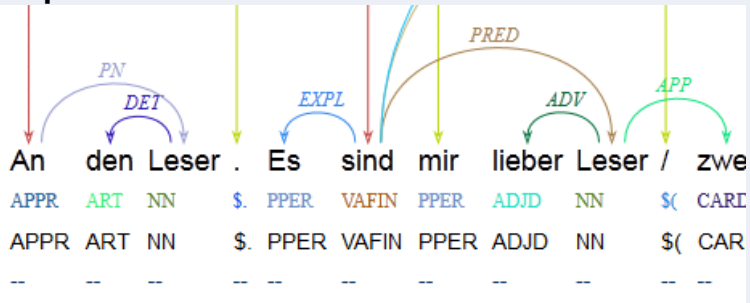
DTA Basis Format



Linguistische Analyse historischer Korpora - Bäume und Konstituenten

- weitere Möglichkeit der syntaktischen Annotation:

Dependenzen



Pointing Relation