

# 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

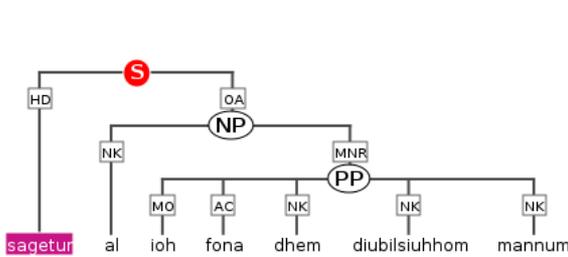
Base text Token Annotations Show Citation URL

1 / 394 > | > | Displaying Results 1 - 10 of 3933

1 Path: DDB.AHD > PAULA\_merge > Henchl

|   |               |             |     |      |             |                    |             |      |       |      |             |             |
|---|---------------|-------------|-----|------|-------------|--------------------|-------------|------|-------|------|-------------|-------------|
| # | sagetun       | al          | ioh | fona | dhem        | diubilsuihhom      | mannum      | enti | see   | saar | alle        | dhea        |
| # | sagen         | al          | io  | fona | ther        | tiuvalsioh         | man         | inti | sehan | sar  | al          | ther        |
| - | 3.Pl.Past.Ind | Acc.Sg.Neut | -   | -    | Dat.Pl.Masc | Pos.Dat.Pl.Masc.Wk | Dat.Pl.Masc | -    | 2.Sg  | -    | Nom.Pl.Masc | Nom.Pl.Masc |
| - | VVFIN         | PIS         | ADV | APPR | PDAT        | ADJA               | NN          | KON  | VVIMP | ADV  | PIAT        | PDAT        |

tree



1 Path: DDD-AD\_Tatian\_171012\_exmaralda > Tatian0

Prologus Bithiu uuanta manage zilotun ordinon

meta (grid)

|         |          |               |        |         |         |         |
|---------|----------|---------------|--------|---------|---------|---------|
| seite   | 25       |               |        |         |         |         |
| sprache | lat      | goh           | goh    | goh     | goh     |         |
| txt     | Prologus | Bithiu uuanta | manage | zilotun | ordinon |         |
| tok     | Prologus | Bithiu        | uuanta | manage  | zilotun | ordinon |

base (grid)

Wortarten  
Lemmatisierung  
Morphologie  
Syntax

.....  
Suche und Analyse der  
Daten in ANNIS

4 Path: Ridges\_Herbology\_Vers

nichts verhalten will , ob ichs

normalizations

|       |        |           |      |   |    |     |    |        |     |        |   |      |
|-------|--------|-----------|------|---|----|-----|----|--------|-----|--------|---|------|
| clean | nichts | verhalten | will |   |    |     |    |        |     |        |   |      |
| dipl  | nichts | verhalten | will |   |    |     |    |        |     |        |   |      |
| norm  | nichts | verhalten | will | , | ob | ich | es | gleich | tun | könnte | , | weil |

default\_ns (grid)

|                    |               |     |
|--------------------|---------------|-----|
| st                 |               |     |
| Pos_Masc_Pl_Nom_st | Ind_Past_Pl_3 |     |
| P                  | wk2           | wk2 |
| a,o                | wk2           | wk2 |
| Masc               |               |     |

|       |         |         |
|-------|---------|---------|
| om    |         |         |
|       | Ind     |         |
|       | Pl      |         |
|       | 3       |         |
| DJ    | VV      | VV      |
|       | Past    |         |
| anage | zilotun | ordinon |

# Linguistische Analyse historischer Korpora - Multikey-Annotationen

## Tatian DDD-AD

### Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsgegenstand

- Token und/oder Spannen-annotationen erhalten auf unterschiedlichen Ebenen Werte:

Numerus

Kasus

Genus

Modus

Part of Speech

1 Path: DDD-AD\_Tatian\_171012\_exmaralda > Tatian0

Prologus Bithiu uuanta manage zilotun ordinon  
meta (grid)

Diplomatische Transkription

|         |          |               |        |         |         |         |
|---------|----------|---------------|--------|---------|---------|---------|
| seite   | 25       |               |        |         |         |         |
| sprache | lat      | goh           | goh    | goh     | goh     |         |
| txt     | Prologus | Bithiu uuanta | manage | zilotun | ordinon |         |
| tok     | Prologus | Bithiu        | uuanta | manage  | zilotun | ordinon |

Annotations eines einzigen Tokens

|         |          |        |                    |               |         |         |
|---------|----------|--------|--------------------|---------------|---------|---------|
| dekl    |          |        | st                 |               |         |         |
| flexB   | Sg_Nom   |        | Pos_Masc_Pl_Nom_st | Ind_Past_Pl_3 |         |         |
| flexkB  | o        |        | P                  | wk2           | wk2     |         |
| flexkL  | o        |        | a,o                | wk2           | wk2     |         |
| gen     |          |        | Masc               |               |         |         |
| genB    | Masc     |        |                    |               |         |         |
| genL    | Masc     |        |                    |               |         |         |
| kasus   | Nom      |        | Nom                |               |         |         |
| modus   |          |        |                    |               | Ind     |         |
| numerus | Sg       |        | Pl                 | Pl            |         |         |
| person  |          |        |                    |               | 3       |         |
| posL    | NA       | KO     | ADJ                | VV            | VV      |         |
| tempus  |          |        |                    |               | Past    |         |
| tok     | Prologus | Bithiu | uuanta             | manage        | zilotun | ordinon |

# Linguistische Analyse historischer Korpora - Multikey-Annotationen

## Tatian DDD-AD

### Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsgegenstand

- Token und/oder Spannen-annotationen erhalten auf unterschiedlichen Ebenen Werte:

Numerus

Kasus

Genus

Modus

Part of Speech

1 Path: DDD-AD\_Tatian\_171012\_exmaralda > Tatian0

Prologus Bithiu uuanta manage zilotun ordinon  
meta (grid)

Diplomatische Transkription

|         |          |               |        |         |         |         |
|---------|----------|---------------|--------|---------|---------|---------|
| seite   | 25       |               |        |         |         |         |
| sprache | lat      | goh           | goh    | goh     | goh     |         |
| txt     | Prologus | Bithiu uuanta | manage | zilotun | ordinon |         |
| tok     | Prologus | Bithiu        | uuanta | manage  | zilotun | ordinon |

Annotations eines einzigen Tokens

|         |          |        |                    |               |         |         |
|---------|----------|--------|--------------------|---------------|---------|---------|
| dekl    |          |        | st                 |               |         |         |
| flexB   | Sg_Nom   |        | Pos_Masc_PI_Nom_st | Ind_Past_PI_3 |         |         |
| flexkB  | o        |        | P                  | wk2           | wk2     |         |
| flexkL  | o        |        | a,o                | wk2           | wk2     |         |
| gen     |          |        | Masc               |               |         |         |
| genB    | Masc     |        |                    |               |         |         |
| genL    | Masc     |        |                    |               |         |         |
| kasus   | Nom      |        | Nom                |               |         |         |
| modus   |          |        |                    |               | Ind     |         |
| numerus | Sg       |        | PI                 | PI            |         |         |
| person  |          |        |                    |               | 3       |         |
| posL    | NA       | KO     | ADJ                | VV            | VV      |         |
| tempus  |          |        |                    |               | Past    |         |
| tok     | Prologus | Bithiu | uuanta             | manage        | zilotun | ordinon |

# Linguistische Analyse historischer Korpora - Multikey-Annotationen

## Tatian DDD-AD

### Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsgegenstand

- Token und/oder Spannen-annotationen erhalten auf unterschiedlichen Ebenen Werte:

Numerus

Kasus

Genus

Modus

Part of Speech

1 Path: DDD-AD\_Tatian\_171012\_exmaralda > Tatian0

Prologus Bithiu uuanta manage zilotun ordinon  
meta (grid)

Diplomatische Transkription

|         |          |               |        |         |         |         |
|---------|----------|---------------|--------|---------|---------|---------|
| seite   | 25       |               |        |         |         |         |
| sprache | lat      | goh           | goh    | goh     | goh     |         |
| txt     | Prologus | Bithiu uuanta | manage | zilotun | ordinon |         |
| tok     | Prologus | Bithiu        | uuanta | manage  | zilotun | ordinon |

Annotations eines einzigen Tokens

|         |          |        |                    |               |         |         |
|---------|----------|--------|--------------------|---------------|---------|---------|
| dekl    |          |        | st                 |               |         |         |
| flexB   | Sg_Nom   |        | Pos_Masc_Pl_Nom_st | Ind_Past_Pl_3 |         |         |
| flexkB  | o        |        | P                  | wk2           | wk2     |         |
| flexkL  | o        |        | a,o                | wk2           | wk2     |         |
| gen     |          |        | Masc               |               |         |         |
| genB    | Masc     |        |                    |               |         |         |
| genL    | Masc     |        |                    |               |         |         |
| kasus   | Nom      |        | Nom                |               |         |         |
| modus   |          |        |                    |               | Ind     |         |
| numerus | Sg       |        | Pl                 | Pl            |         |         |
| person  |          |        |                    |               | 3       |         |
| posL    | NA       | KO     | ADJ                | VV            | VV      |         |
| tempus  |          |        |                    |               | Past    |         |
| tok     | Prologus | Bithiu | uuanta             | manage        | zilotun | ordinon |

# Linguistische Analyse historischer Korpora - Multikey-Annotationen

## Tatian DDD-AD

### Morphologie:

- Annotationskey und Werte als direkte Analysehilfe für den jeweils konkreten Untersuchungsstand
- Token und/oder Spannen-annotationen erhalten auf unterschiedlichen Ebenen

Numerus

Kasus

Genus

Modus

Part of Speech

1 Path: DDD-AD\_Tatian\_171012\_exmaralda > Tatian0

Prologus Bithiu uuanta manage zilotun ordinon

meta (grid)

Diplomatische Transkription

|         |          |               |        |         |         |         |
|---------|----------|---------------|--------|---------|---------|---------|
| seite   | 25       |               |        |         |         |         |
| sprache | lat      | goh           | goh    | goh     | goh     |         |
| txt     | Prologus | Bithiu uuanta | manage | zilotun | ordinon |         |
| tok     | Prologus | Bithiu        | uuanta | manage  | zilotun | ordinon |

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

|         |          |        |        |        |         |         |
|---------|----------|--------|--------|--------|---------|---------|
| gerB    | Masc     |        |        |        |         |         |
| gerL    | Masc     |        |        |        |         |         |
| kasus   | Nom      |        | Nom    |        |         |         |
| modus   |          |        |        | Ind    |         |         |
| numerus | Sg       |        | Pl     | Pl     |         |         |
| person  |          |        |        | 3      |         |         |
| posL    | NA       | KO     | ADJ    | VV     | VV      |         |
| tempus  |          |        |        | Past   |         |         |
| tok     | Prologus | Bithiu | uuanta | manage | zilotun | ordinon |

# Linguistische Analyse historischer Korpora - Bäume und Konstituenten

## Deutsche Diachrone Baubank

- 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

The screenshot displays a linguistic analysis interface. At the top, there are navigation buttons and a search bar. Below that, a table lists tokens with their corresponding grammatical annotations. A red circle highlights the first row of the table, which corresponds to the first tree diagram. The tree diagram shows a hierarchical structure of constituents, with a red circle highlighting the root node 'S' and its children 'HD' and 'OA'. The second tree diagram shows a similar structure for the second row of the table.

| Token         | Annotation  | Token | Annotation | Token       | Annotation        | Token       | Annotation | Token | Annotation | Token       | Annotation  |
|---------------|-------------|-------|------------|-------------|-------------------|-------------|------------|-------|------------|-------------|-------------|
| sagetun       | al          | ioh   | fona       | dhem        | diubilsiuhhom     | mannum      | enti       | see   | saar       | alle        | dhea        |
| 3.Pl.Past.Ind | Acc.Sg.Neut | --    | --         | Dat.Pl.Masc | Pos.Pl.Pl.Masc.Wk | Dat.Pl.Masc | --         | 2.Sg  | --         | Nom.Pl.Masc | Nom.Pl.Masc |
| VVFIN         | PIS         | ADV   | APPR       | PDAT        | ADJA              | NN          | KON        | VVIMP | ADV        | PIAT        | PDAT        |

Path: DDB.AHD > PAULA merge > Henchl

Displaying Results 1 - 10 of 3933

Diplomatische Transkription

Token in Konstituenten- & Hierarchie-Beziehungen

# Linguistische Analyse historischer Korpora - Bäume und Konstituenten

## Deutsche Diachrone Baubank

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

- Suche nach syntaktisch abhängigen Strukturen:

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

The screenshot shows a web interface for a linguistic corpus. At the top, there are tabs for 'Base text', 'Token Annotations', and 'Show Citation URL'. Below that, navigation controls and a search bar are visible. The main content area displays a text snippet with various tokens and their grammatical annotations. A red circle highlights a specific phrase: 'sagetun al ioh fona dhem diubilsiuhhom'. Below the text, a constituent tree diagram is shown, with a red arrow pointing from the highlighted phrase to a node in the tree. The tree structure includes nodes for 'HD', 'OA', 'NP', 'S', 'MO', 'OC', 'CS', 'CJ', 'SB', 'NP', 'NK', and 'N'. The text 'Diplomatische Transkription' is written next to the tree. Below the tree, the text 'Konstituenten- & e-Beziehungen' is visible.

| Token         | Annotation  | Token | Annotation | Token       | Annotation        | Token       | Annotation | Token | Annotation | Token       | Annotation  |
|---------------|-------------|-------|------------|-------------|-------------------|-------------|------------|-------|------------|-------------|-------------|
| sagetun       | al          | ioh   | fona       | dhem        | diubilsiuhhom     | mannum      | enti       | see   | saar       | alle        | dhea        |
| 3.Pl.Past.Ind | Acc.Sg.Neut | --    | --         | Dat.Pl.Masc | Pos.Sg.Pl.Masc.Wk | Dat.Pl.Masc | --         | 2.Sg  | --         | Nom.Pl.Masc | Nom.Pl.Masc |
| VVFIN         | PIS         | ADV   | APPR       | PDAT        | ADJA              | NN          | KON        | VVIMP | ADV        | PIAT        | PDAT        |

# 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

4 ⓘ Path: Ridges\_Herbology\_Version\_2.0 > Ridges\_v2 > flora.saturnizans.1722

nichts verhalten will , ob **ichs** gleich thun könnte , weil

normalizations

|       |        |           |      |   |    |             |    |        |      |        |   |      |
|-------|--------|-----------|------|---|----|-------------|----|--------|------|--------|---|------|
| clean | nichts | verhalten | will | , | ob | ichs        |    | gleich | thun | könte  | , | weil |
| dipl  | nichts | verhalten | will | , | ob | <b>ichs</b> |    | gleich | thun | könte  | , | weil |
| norm  | nichts | verhalten | will | , | ob | ich         | es | gleich | tun  | könnte | , | weil |

default\_ns (grid)

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. Partikelver-
- Mehrebenen-Nor- können nicht antizipiert werden, da abhängig von den jeweiligen
- linguistischer Vor- Forschungsfra-

4 ⓘ Path: Rid ... zans.1722

nichts verhalten will , ob **ichs** gleich thun könnte , weil

normalizations

|       |        |           |      |   |    |             |        |        |       |        |      |      |
|-------|--------|-----------|------|---|----|-------------|--------|--------|-------|--------|------|------|
| clean | nichts | verhalten | will | , | ob | ichs        | gleich | thun   | könte | ,      | weil |      |
| dipl  | nichts | verhalten | will | , | ob | <b>ichs</b> | gleich | thun   | könte | ,      | weil |      |
| norm  | nichts | verhalten | will | , | ob | ich         | es     | gleich | tun   | könnte | ,    | weil |

default\_ns (grid)

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

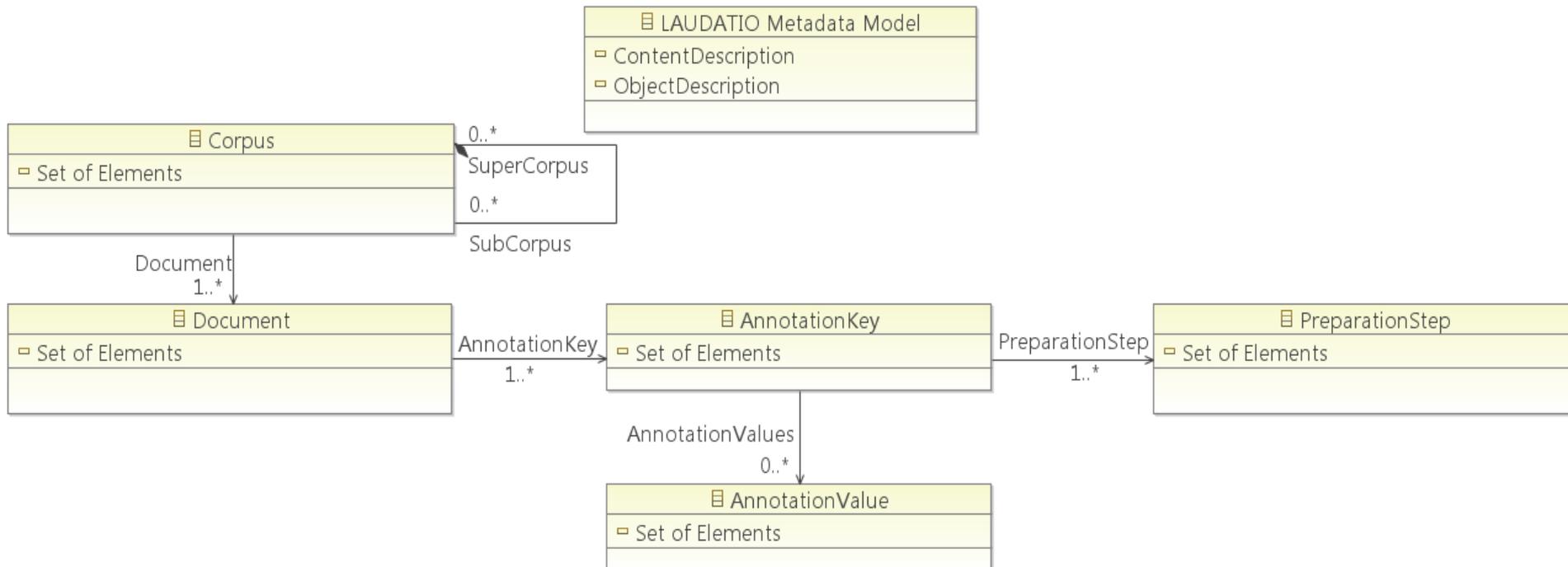
Neben klassischen Metadaten wollen wir (u.a.) wissen:

- Wie wurde transkribiert?
- Wie wurde annotiert?
- Mit welchem Tagset?
- Mit 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

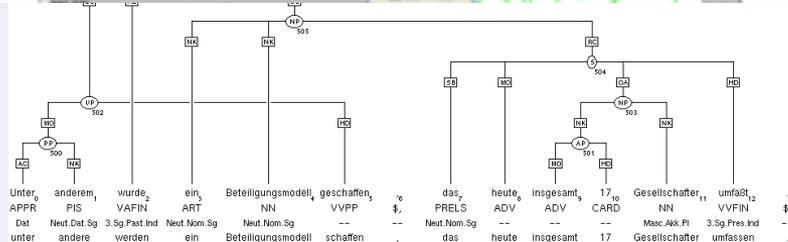
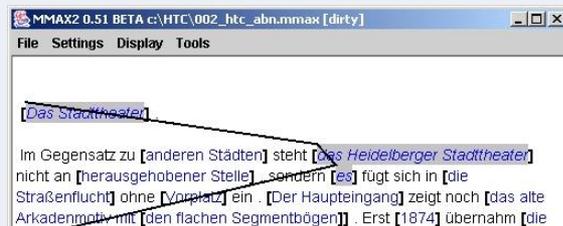
# Infrastruktur für historische Korpuslinguistik - Metadaten



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

# Infrastruktur für historische Korpuslinguistik - Daten annotieren

## /1/ Daten annotieren

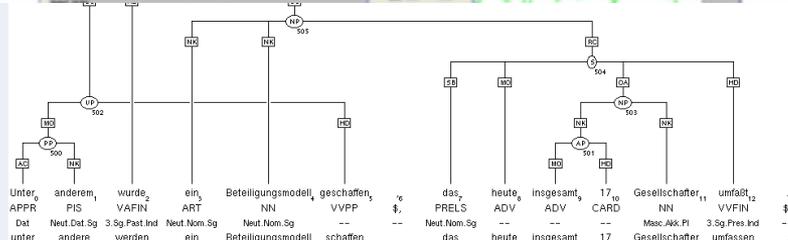


Viele spezifische Tools für unterschiedliche Annotationen:

- 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 :-)
- geringe Interoperabilität :-)
- manche Tools werden nicht mehr gepflegt :-)

MMAX2 0.51 BETA c:\HTC\002\_htc\_abn.mmax [dirty]

File Settings Display Tools

[Das Stadthaus]

Im Gegensatz zu [anderen Städten] steht [das nicht an [herausgehobener Stelle], sondern [e Straßenfucht] ohne [Vorplatz] ein . [Der Haupt Arkadenmotiv mit [den flachen Segmentbögen].

EXN  
Transkription, Ann

TEI  
TEXT ENCODING INITIATIVE

Diagram showing a tree structure for text encoding with nodes like NP, VP, PP, etc.

|                    |                      |                    |                  |                                 |                         |    |                  |                    |                         |                  |                              |                      |    |
|--------------------|----------------------|--------------------|------------------|---------------------------------|-------------------------|----|------------------|--------------------|-------------------------|------------------|------------------------------|----------------------|----|
| Unter <sub>6</sub> | anderem <sub>1</sub> | wurde <sub>2</sub> | ein <sub>3</sub> | Beteiligungsmodell <sub>4</sub> | geschaffen <sub>5</sub> | %  | das <sub>8</sub> | heute <sub>9</sub> | insgesamt <sub>10</sub> | 17 <sub>10</sub> | Gesellschafter <sub>11</sub> | umfaßt <sub>12</sub> | \$ |
| APPR               | PIS                  | VAFIN              | ART              | NN                              | VVPP                    | -- | PRELS            | ADV                | ADV                     | CARD             | NN                           | VVFIN                | -- |
| Dat                | Neut.Dat.Sg          | 3.Sg.Past.Ind      | Neut.Nom.Sg      | Neut.Nom.Sg                     | --                      | -- | Neut.Nom.Sg      | --                 | --                      | --               | Masc.Akk.Pl                  | 3.Sg.Pres.Ind        | -- |
| unter              | andere               | werden             | ein              | Beteiligungsmodell              | schaffen                | ,  | das              | heute              | insgesamt               | 17               | Gesellschafter               | umfassen             | .  |

# Infrastruktur für historische Korpuslinguistik - TEI

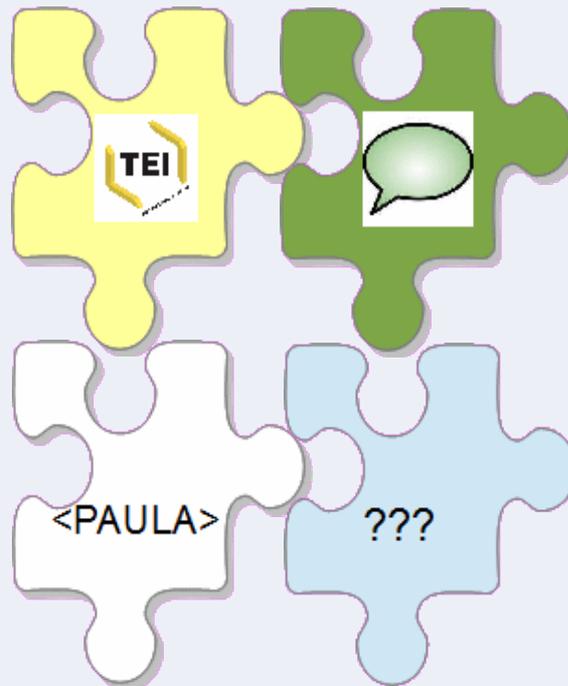
- Warum reicht uns TEI nicht?



- 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?

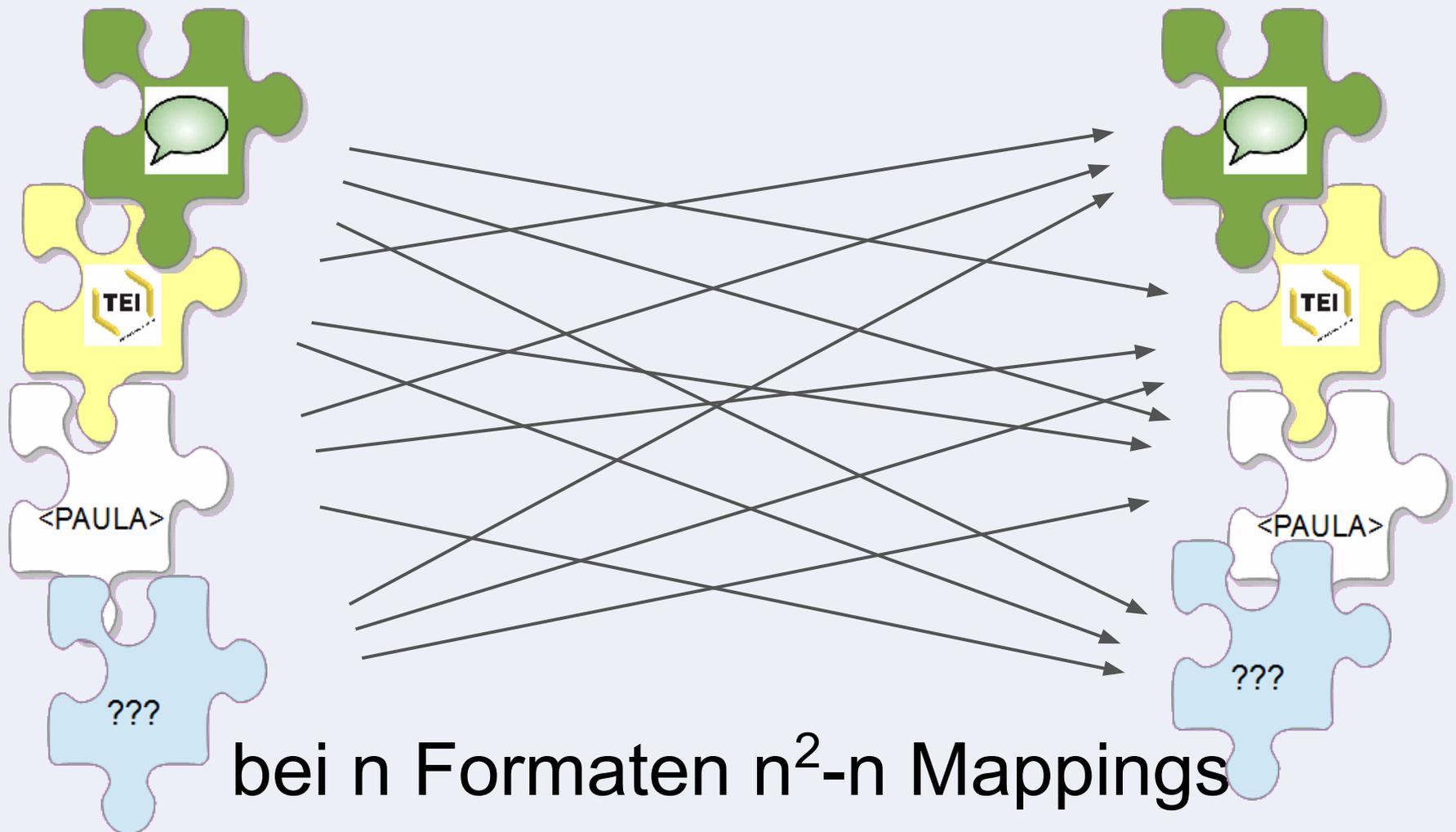
- erst Kombination der Formate ergibt möglichst umfangreiches Bild  
→ Formatpluriversum



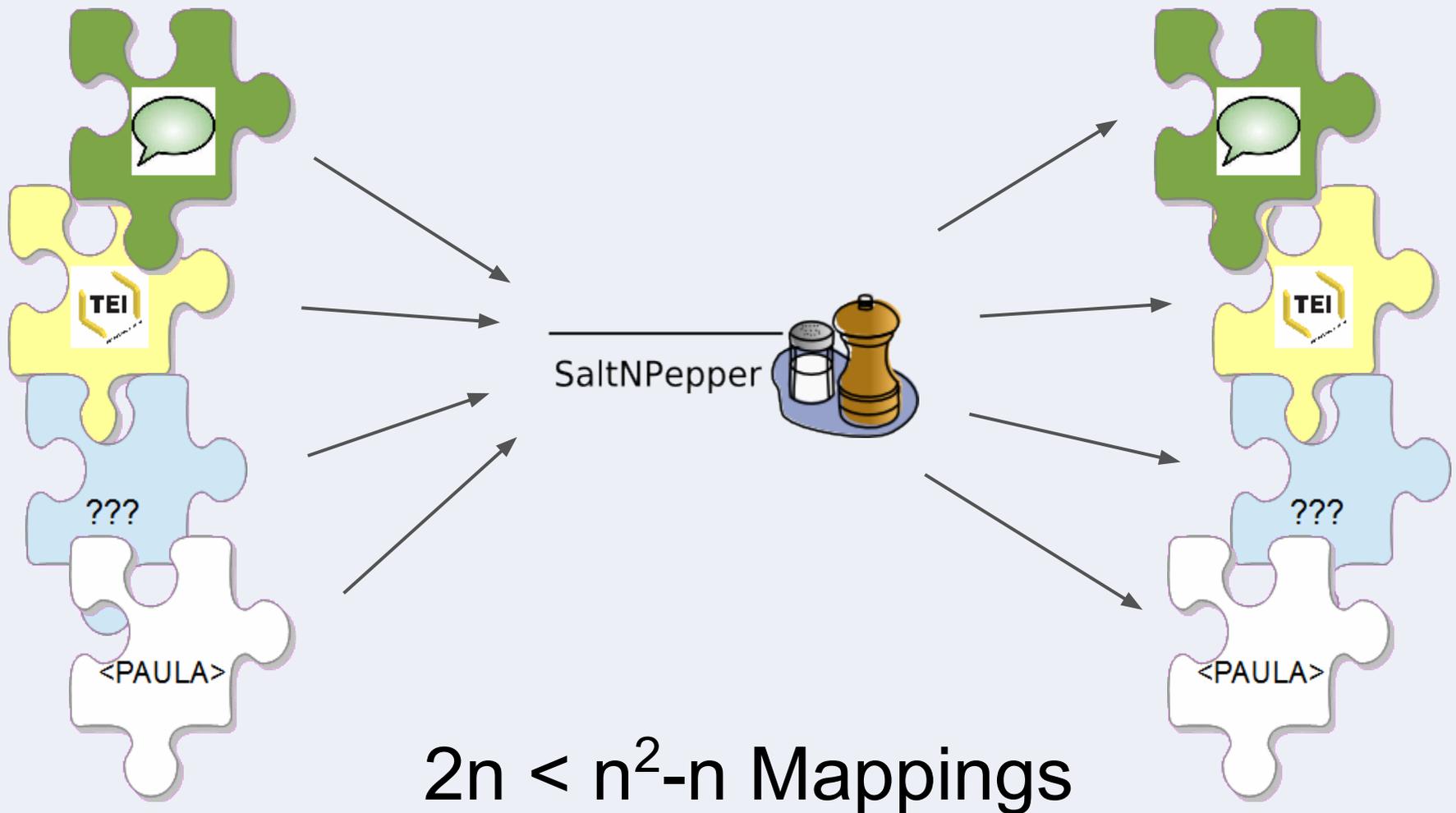
# Infrastruktur für historische Korpuslinguistik - Formatpluriversum

- aber wir wollen Interoperabilität zwischen Formaten  
soweit wie möglich
    - Digitaler Text,
    - Tokenisierung und
    - bestimmte Annotationin verschiedenen Tools nutzbar
  - Annotationen abbildbar auf Mehrebenenformate
- Überführen was zu überführen geht

# Infrastruktur für historische Korpuslinguistik - Formatpluriversum

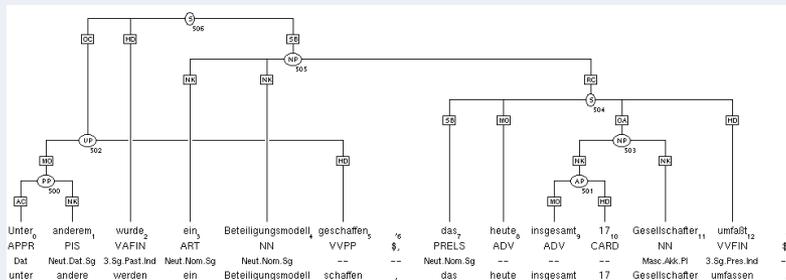


# Infrastruktur für historische Korpuslinguistik - Formatpluriversum



# Infrastruktur für historische Korpuslinguistik - Daten durchsuchen

## /3/ Daten durchsuchen



- auf einzelner Ebene

ANNSYS - Tutorial

Search Form

ANNSYSQL: [tok & tok & #1 ->dep [func="OA"] #2 & cat="S" & #2 \_ #1 & node & #3 ->secede #4 | correction="correcting" | correct]

Query Builder: Show >>

Result: 43

History: Query History

More Corpora

| Name                 | Texts | Tokens |
|----------------------|-------|--------|
| FalkEisner2V2_0      | 248   | 131511 |
| ONTONOTES_v1.5_small | 4     | 6450   |
| SMILTRON_Banana      | 2     | 2782   |
| TueSWS_no_cvc        | 2187  | 770949 |
| sgn_J                | 24    | 164    |
| b4.miller2.0         | 2031  | 11295  |
| pos-3                | 3     | 573    |
| pos2                 | 2     | 399    |
| tiger1_dep           | 1     | 929    |
| tiger2               | 1971  | 888578 |

Search Export

Context Left: 0

Context Right: 0

Results per page: 10

Show Results

Search Result - tok & tok & #1 ->dep[func="OA"] #2 & cat="S" & #3 \_ #1 & node & #3 ->secede #4 (0, 0)

Page: 1 of 5

Token Annotations Show Citation URL

Displaying Results 1 - 10 of 43

während 78 Prozent sich für Bush und vier Prozent für Clinton aussprechen

während 78 Prozent sich für Bush und vier Prozent für Clinton aussprechen

KOLS\_CARD NN PRF APPR NE KON\_CARD NN APPR NE VVFIN

-- \*\*Neut 3.Acc.Pl -- Acc.Sa\* -- \*\*Neut -- Acc.Sg\* 3.Pf.Past.Ind

dependencies

constituents

Die Vase auf dem Tisch ist größer als die Vase

animacy (grd)

Select Displayed Annotation Levels

mmaxref\_type Baum

mmaxref\_type Baum

tok Die Vase auf dem Tisch ist größer als die Vase

conference (discourse)

Die Vase auf dem Tisch ist größer als die Vase auf der Fensterbank . ich fride . sie sieht nicht so gut aus . weil der Tisch zu klein ist

- parallel auf mehreren Ebenen

# Infrastruktur für historische Korpuslinguistik - Datenzugriff

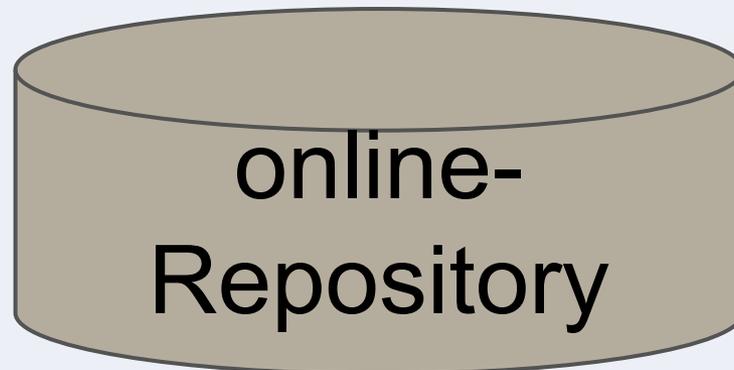
/4/ auf Daten zugreifen

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

# Infrastruktur für historische Korpuslinguistik - Datenzugriff

/4/ auf Daten zugreifen

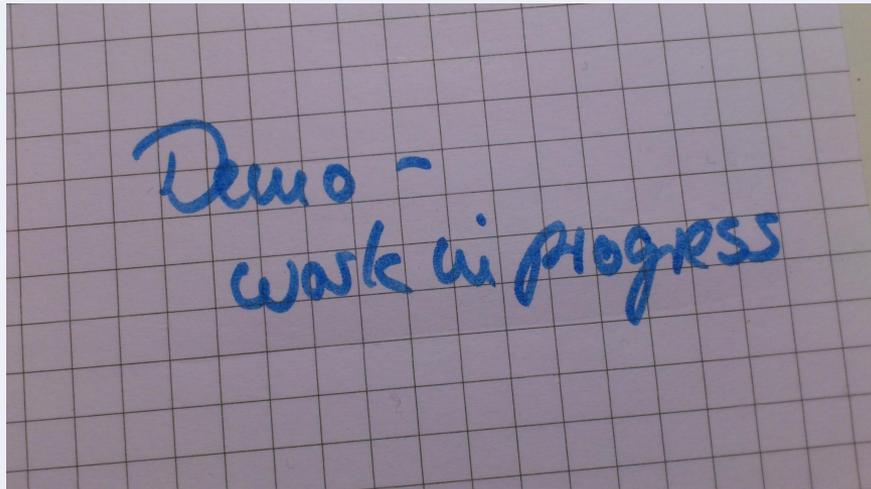
- immer und überall
- Upload/ Download
  - in verschiedenen Formaten- Facetten
- Versionierungssystem
- einheitliche Metadatensuche





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

# Referenzen

|                     |  |
|---------------------|--|
| RIDGES<br>Herbology | Krause, Thomas, Lüdeling, Anke, Odebrecht, Carolin & Amir Zeldes (2012) <b>Multiple Tokenization in a Diachronic Corpus</b> . <i>Exploring Ancient Languages through Corpora Conference (EALC), 14.-16.Juni 2012</i> .<br><a href="http://korpling.german.hu-berlin.de/ridges/documentation_en.html">http://korpling.german.hu-berlin.de/ridges/documentation_en.html</a><br><a href="https://korpling.german.hu-berlin.de/annis3/">https://korpling.german.hu-berlin.de/annis3/</a> |
| RSTTool             | O'Donnell, Michael (2000) "RSTTool 2.4 -- A Markup Tool for Rhetorical Structure Theory". <i>Proceedings of the International Natural Language Generation Conference (INLG'2000)</i> , 13-16 June 2000, Mitzpe Ramon, Israel. 253 - 256.   |
| SaltNPepper         | Zipser, Florian & Laurent Romary (2010) <b>A model oriented approach to the mapping of annotation formats using standards</b> In: <i>Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010</i> . Malta. URL: <a href="http://hal.archives-ouvertes.fr/inria-00527799/en/">http://hal.archives-ouvertes.fr/inria-00527799/en/</a>   |
| Tatian DDD          | Linda, Sonja, Unverzagt, Silke & Karin Donhauser (erscheint) <b>Old German Reference Corpus</b> . <i>32. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft</i> .<br><a href="https://korpling.german.hu-berlin.de/annis3/">https://korpling.german.hu-berlin.de/annis3/</a>   |
| TEI                 | Burnard, Lou & Syd Bauman (Eds.) (2008). <b>TEI P5: Guidelines for Electronic Text Encoding and Interchange</b> . Oxford.<br><a href="http://www.tei-c.org/Guidelines/P5/">http://www.tei-c.org/Guidelines/P5/</a> .   |
| TIGERSearch         | Lezius, Wolfgang (2002) <b>Ein Suchwerkzeug für syntaktisch annotierte Textkorpora</b> . <a href="http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/">http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/</a>   |
| Treetagger          | Schmid, Helmut (1994) <b>Probabilistic Part-of-Speech Tagging Using Decision Trees</b> . In: <i>Proceedings of International Conference on New Methods in Language Processing</i> .  |

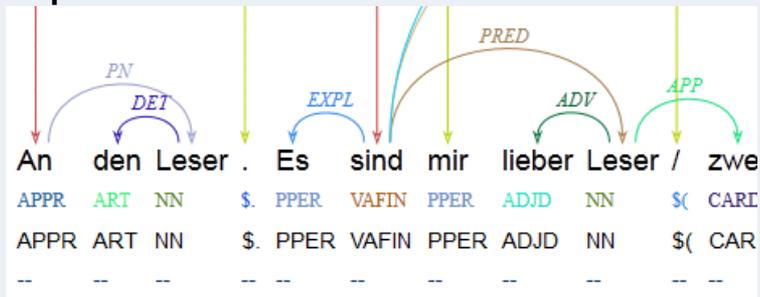
# DTA Basis Format

| default_ns (grid)                    |  |  |   |  |                           |                  |                            |   |        |   |  |   |
|--------------------------------------|--|--|---|--|---------------------------|------------------|----------------------------|---|--------|---|--|---|
| Select Displayed Annotation Levels ▾ |  |  |   |  |                           |                  |                            |   |        |   |  |   |
| bibl                                 |  |  |   |  |                           |                  |                            |   | bibl   |   |  |   |
| body                                 | body   |  |   |  |                           |                  |                            |   |        |   |  |   |
| cit                                  |  |  |   |  |                           |                  |                            |   | cit    |   |  |   |
| div                                  | div  |  |   |  |                           | div              |                            |   |        |   |  |   |
| figure                               |  |  |   |  |                           |                  |                            |   |        | figure  |  |   |
| fw                                   |  |  |   |  |                           |                  |                            |   |        |   |  |   |
| head                                 |  |  |   |  |                           | head             |                            |   |        |   |  |   |
| hi                                   |  |  |   |  |                           | hi               |                            | hi  |        |   |  |   |
| hi                                   |  |  |   |  |                           |                  |                            | hi  |        |   |  |   |
| lg                                   |  |  |   |  |                           |                  |                            |   |        |   |  |   |
| n                                    | 1  |  |   |  |                           | 1                |                            |   |        |   |  |   |
| p                                    | p  |  |   |  |                           |                  |                            |   |        | p   |  |   |
| place                                |  |  |   |  |                           |                  |                            |   |        |   |  |   |
| quote                                |  |  |   |  |                           |                  |                            |   | quote  |   |  |   |
| rendition                            |  |  |   |  |                           | #b               |                            | #et   |        |   |  |   |
| rendition                            |  |  |   |  |                           |                  |                            | #aq   |        |   |  |   |
| text                                 | text   |  |   |  |                           |                  |                            |   |        |   |  |   |
| type                                 |  |  |   |  |                           |                  |                            |   |        |   |  |   |
| tok                                  | klärt, bei einigen ist Erläuterung überflüssig. So möge denn das Ge- | wand bunt genug feyn, um manche Fehler und Schwächen der Sache | fehlt zu verdecken oder doch minder fühlbar zu machen, kurz möchten | diese in der That anspruchslosen Betrachtungen, nachrichtige und | freundliche Lefer finden. | Erste Vorlesung. | Das Auge und das Microscop | Oculus ad vitam nihil facit, ad vitam beatam nihil magis. — | Seneca | Kein Organ ist so unwichtig für das Leben als das Auge, | Keins so wichtig für die Schönheit des Lebens. — | Die Vignette giebt einen idealen Durchschnitt durch die kleine Camera ob- |

# Linguistische Analyse historischer Korpora - Bäume und Konstituenten

- weitere Möglichkeit der syntaktischen Annotation:

## Dependenzen



## Pointing Relation

|     |         |       |            |                |             |             |               |     |     |      |
|-----|---------|-------|------------|----------------|-------------|-------------|---------------|-----|-----|------|
| und | verzagt | wie   | die        | deutsche       | Abwehrreihe | der         | Fußballkicker | .   | Und | dann |
| und | verzagt | wie   | der        | deutsch        | Abwehrreihe | der         | Fußballkicker | .   | und | dann |
| --  | Pos     | --    | Nom.Sg.Fem | Pos.Nom.Sg.Fem | Nom.Sg.Fem  | Gen.Pl.Masc | Gen.Pl.Masc   | --  | --  | --   |
| KON | ADJD    | KOKOM | ART        | ADJA           | NN          | ART         | NN            | \$. | KON | ADV  |

mmax (discourse)

Steilpass Wunder gibt es immer wieder ! Erst spielen die Dallgower Gemeindevertreter so statisch und Tiefe solch ein fulminanter Steilpass , von dem man hofft , dass die Seeburger oder Groß-Glienicker , es vorerst keine Gefahr fürs Dallgower Tor gab . Die Seeburger und einige Groß-Glienicker haben de sie zeigen , wie sie die Chance verwerten . Eine Diskussion , wo künftig die Trainerkabine stehen si deutschen Grotten-Kickern gibt es immer noch . Auch wenn die Spieler aus den verschiedenen Verei einer Mannschaft " Döberitzer Heide " spielen . Und das heißt geme **Component: 5, Type: anaphor\_a** kleinsten Schubser gegenseitig zu zerfleischen , sind normalerweise überflüssig . Vorerst allerdings hi