Modellierung linguistischer Forschungsdaten
Was weißt du alles über (d)ein Korpus?

Doktorandentag 2014 Institut für deutsche Sprache und Linguistik 06.10.2014
Carolin Odebrecht Betreuer: Prof. Dr. Anke Lüdeling, Dr. Laurent Romary

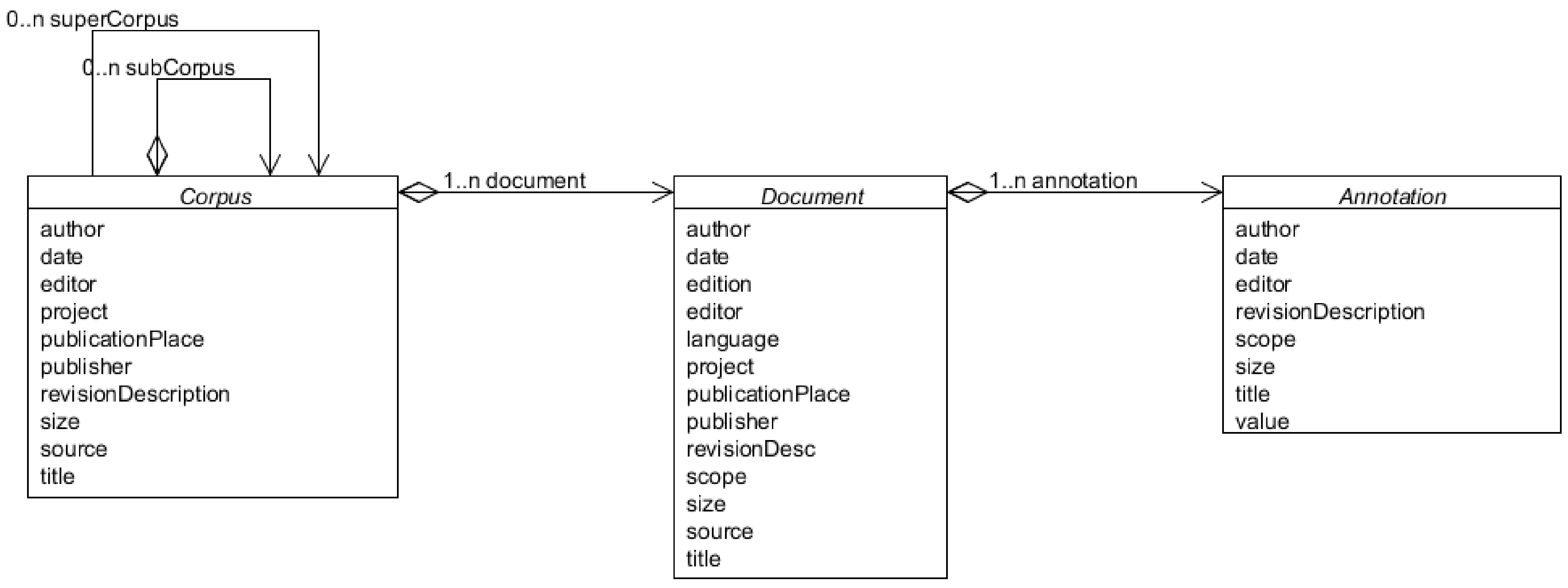
HUMBOLDT-UNIVERSITÄT ZU BERLIN

1. Linguistische Forschungsdaten, ihre Metadaten und Modellierung

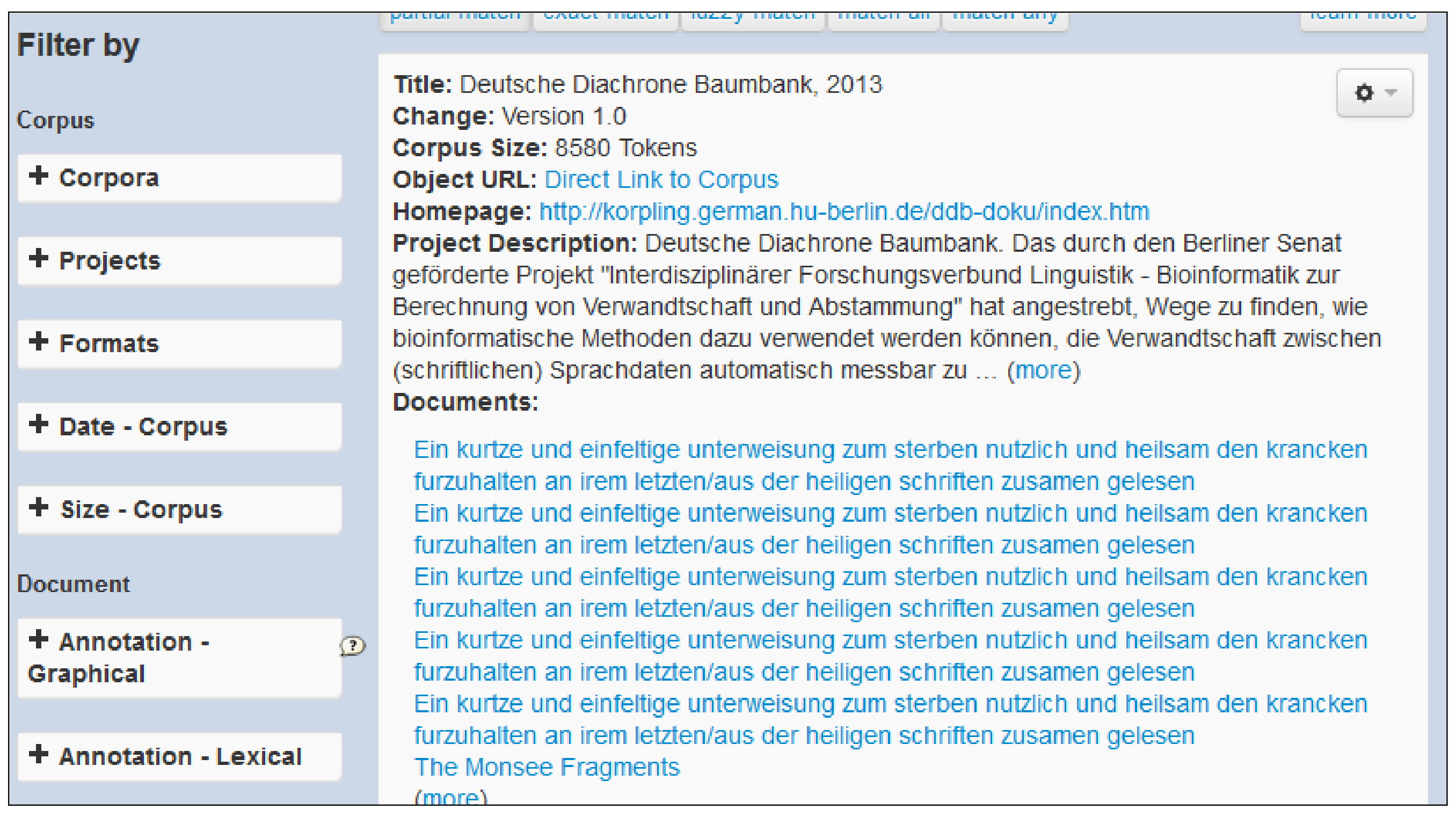
- **Was haben diese Korpora gemeinsam?**
  Eingrenzung auf historische Korpora aus der Linguistik (Lemnitzer & Zinsmeister 2008, Claridge 2008)
  - **Einheiten - Token**
  - Annotation
  - historische Vorlage für die Digitalisierung

Metadaten linguistischer Forschungsdaten (Odebrecht 2014)
- Daten zum Beschreiben der Daten (Metadaten der Korpora)
- Wofür sollen die Metadaten Korpora beschreiben?
  - Finden von Korpora für eigenen Forschung in einem Repositorium (Odebrecht et al. 2014)
    - bibliographische Eigenschaften der Texte (Titel, Jahr, Autor, Veröffentlichungsort, Genre etc.)
    - korpuslinguistische Eigenschaften (Format, Projekt, Annotatoren, Annotationsrichtlinien, etc.)
- Wie sollen die Metadaten beschreiben?
  - **einheitliche, strukturierte, umfangreiche Metadaten** historischer Korpora
    - unabhängig von Format
    - unabhängig von Annotation
    - unabhängig von der linguistischen Fragestellung

- **Worin unterscheiden sich Korpora?**
  - Art, Werte, Richtlinien, Tools für Annotationen
  - Texte (u.a. Register, Genre, Jahr, Ort)
  - Formate (u.a. TIGER XML, EXMARaLDA, Excel, PlainText)
  - Segmentierung (einfache oder multiple Segmentierung, vgl. Krause et al. 2012)
  - Projekt und Fragestellung (z.B. DDB, HSJ, KAJUK, RIDGES)



2. Anwendung des Modells und technische Umsetzung



Title: Deutsche Diachrone Baumbank, 2013
Change: Version 1.0
Corpus Size: 8580 Tokens
Object URL: Direct Link to Corpus
Homepage: http://korpling.german.hu-berlin.de/ddb-doku/index.htm
Project Description: Deutsche Diachrone Baumbank. Das durch den Berliner Senat geförderte Projekt "Interdisziplinärer Forschungsverbund Linguistik - Bioinformatik zur Berechnung von Verwandtschaft und Abstammung" hat angestrebt, Wege zu finden, wie bioinformatische Methoden dazu verwendet werden können, die Verwandtschaft zwischen (schriftlichen) Sprachdaten automatisch messbar zu ... (more)
Documents:
Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen
Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen
Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen
Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen
Ein kurtze und einfeltige unterweisung zum sterben nutzlich und heilsam den krancken furzuhalten an irem letzten/aus der heiligen schriften zusamen gelesen
The Monsee Fragments
(more)

LAUDATIO-Repository
- Facettensuche über Metadaten
- Beschreibung von relevanten Klassen und deren Eigenschaften durch das Metamodell
- Suche und Auswahl eines Korpus nach diesen Eigenschaften
  - u.a. einzelne Facetten für solche Eigenschaften
    z.B. Annotationsebenen, Größe des Korpus, Projekte
    alle Eigenschaften sind über die Freitextsuche abrufbar

Umsetzung mit TEI XML
- Annahme: das muss man über sein Korpus wissen – drei Klassen mit ihren Attributen
- jede Klasse erhält einen teiHeader (Burnard & Baumann 2008)
  - Corpus – CorpusHeader
  - Document – DocumentHeader
  - Annotation – PreparationHeader
- teiHeader sind speziell dafür angepasst (ODD + Schema, vgl. Burnard & Rahtz 2004)
- TEI XML als Format, um diese Metadaten dem Repositorium zu geben

→ Anzeige der Metadaten für jedes Korpus: **einheitlich, strukturiert und umfassend**
→ Suche nach Korpora über diese Metadaten

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader type="CorpusHeader">
```

Liste aller PreparationHeader

Liste aller DocumentHeader

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader type=„DocumentHeader">
```

Liste aller PreparationHeader

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader type=„PreparationHeader">
```

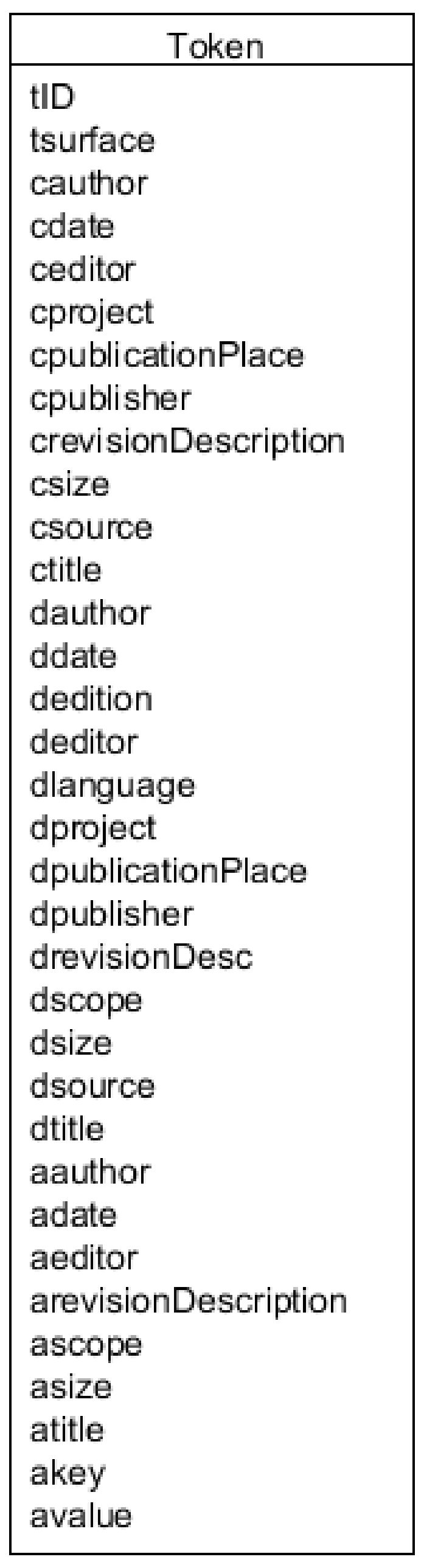
3. Neuer Ansatz für die Modellierung von Korpora

Neue Aufgabenstellung
- bislang noch keine Umsetzung von Superkorpus oder Subkorpus im Repositorium
- noch keine Annotationen, die zwischen Dokumenten bestehen, abgebildet (Anwendung vgl. Berti 2012),
- bislang nur ‚Annotation' innerhalb ‚Document'
- keine Einbindung von ‚Token'

Klasse ‚Token'
- **Mengen/Konzeptdefinitionen**
- Annahme: jedes zu beschreibende Korpus besitzt Token
- Token besitzen die Summe aller Attribute der verschiedenen Klassen
- diese Attribute sind auch so referenziert (a für Annotation / d für Document / c für Corpus)
- Token können also über ihre Attribute und deren Werte Gemeinsamkeiten aufweisen

- **Bestimmung der Zugehörigkeit zu einer Klasse: die Menge an gleichen Werten zu einer bestimmten Gruppe an Attributen**

→Relationen via Schnittmengen von Token mit bestimmten Attributen und deren Werten
→Zugehörigkeit zu einer Klasse über gemeinsame Attribute
→Zugehörigkeit zu Instanzen einer Klasse über gemeinsame Werte von Attributen



Welche Eigenschaften besitzt ein Token?
- referenzierbare, kleinste zu annotierende Einheit einer Segmentierung
- ohne Semantik, mit mindestens einer isolierten Eigenschaft
- ...
- ...

Beispiel ‚Token' und ‚Document' in RIDGES Herbology Corpus
- 154.266 Token in RIDGES Herbology 4.0
- um zu einer Klasse ‚Document' zu gehören, müssen Token gleiche Attribute besitzen:
  - dtitle, dauthor und ddate
- um zu einer Instanz der Klasse ‚Document' zu gehören, müssen Token diese Gruppe von Attributen mit gleichem Wert besitzen:
  - New Kreuterbuch, Fuchs, 1543

Beispiel ‚Token' und ‚Annotation' in RIDGES Herbology Corpus
- 154.266 Token in RIDGES Herbology
- um zu einer Klasse ‚Annotation' zu gehören, müssen Token folgende gleiche Attribute besitzen:
  - atitle, aauthor und adate
- um zu einer Instanz der Klasse ‚Annotation' zu gehören, müssen Token diese Gruppe von Attributen mit gleichem Wert besitzen:
  - pos, Perlitz, 2014-04-01

Nächste Schritte / Ausblick
- tokenbasiertes Modellieren erlaubt einfachere Mengendefinition
- Annotationen, die über die Grenze eines Dokumentes hinaus verweisen, anders zu formulieren: Token, die die Attribute und Werte zu einer Instanz der Klasse ‚Annotation' teilen, aber nicht die der Klasse ‚Document'
- Ausformulieren des Modells, technische Implementation sowie umfangreiches Testen

Referenzen Berti, M. (2012) Working with Text in a Digital Age An NEH Institute for Advanced Topics in the Digital Humanities Tufts University, 23.7.-10.8.2012. **Burnard**, L./Bauman, S. (Hg.) (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford. http://www.tei-c.org/Guidelines/P5/ **Burnard**, L./Rahtz, S. (2004) RelaxNG with Son of ODD. In: *Proceedings of Extreme Markup Languages* 2004. http://conferences.idealliance.org/extreme/html/2004/Burnard01/EML2004Burnard01.html **Claridge**, C. (2008) Historical Corpora. In: Lüdeling, A./Kytö, M. (Hg.) *Corpus Linguistics. An International Handbook*. Vol 1. Berlin: de Gruyter, 242–259. **Odebrecht**, C., Zielke, D., Krause, T., Weißenfels, B., Belz, M., Schernikau, T., Voigt, V. (2014) Wissenschaftliche Nutzung der korpuslinguistischen Infrastruktur LAUDATIO. *DGfS-CL Poster Session 2014. 36. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS). 6.3.2014, Marburg.* **Odebrecht**, C. (2014) Modeling Linguistic Research Data for a Repository for Historical Corpora. *Digital Humanities 2014 Conference. 8.7.-12.7.2014, Lausanne.* **Korpora** Deutsche Diachrone Baumbank DDB http://hdl.handle.net/11022/0000-0000-2107-3 , Historische Syntax des Jiddischen HSJ http://hdl.handle.net/11022/0000-0000-24F9-F, Kasseler Junktionskorpus KAJUK http://hdl.handle.net/11022/0000-0000-2102-8, Register in Diachronic German Science RIDGES Herbology http://hdl.handle.net/11022/0000-0000-2C89-5