Modellierung von Metadaten Technisch-abstrakte Perspektive

Korpuslinguistik Kolloquium | 10.12.2014 |
Carolin Odebrecht | carolin.odebrecht@hu-berlin.de
LAUDATIO-Repository | laudatio-repository.org

Themen

- Modellierung linguistischer Forschungsmetadaten
  - Gegenstand, Ziel, Anwendung
- Semantisch-eindeutiges Metadatenmodell
  - CLARIN, CMDI
- Technisch-abstraktes Metadatenmodell
  - eigener Ansatz

Modellierung linguistischer Forschungsmetadaten

- Gegenstand
  - diachrone, synchrone historische Korpora
  - verschiedene Register (Briefe, Kräuterbücher, Protokolle, Predigten etc.)
  - Sprachstufen und Sprachen (AHD, ModD, Jiddisch, etc.)
  - Token-, Spannenannotation, Baumannotationen, Dependenzen, Referenzlisten usw.

Modellierung linguistischer Forschungsmetadaten

- Ziel
  - Modell für Metadaten der Forschungsdaten
  - Technische Umsetzung

- Anwendung
  - LAUDATIO-Repository (Krause et al. 2014)
    - Speicherort für Korpora
    - Zugriff auf diese Korpora durch Metadaten

Modellierung linguistischer Forschungsmetadaten

- Kontext der Metadaten
  - Wiederverwendung von Korpora
    - ansehen, durchsuchen, analysieren
    - vergrößern mit mehr Texten mit gleicher Annotation
    - neu zusammenstellen mit gleicher, weniger oder anderer Annotation
    - neue Annotationen hinzufügen
    - ...
  - Wissen über Korpora dazu erforderlich
    - Textgrundlagen/Textvorlagen
    - Annotationsschemata, Annotationstools, Annotatoren
    - Prüfverfahren, Versionen, Konvertierungen
    - ...

(vgl. Odebrecht & Krause 2013)

Modellierung linguistischer Forschungsmetadaten

- Veranschaulichung
  - Forschergruppe baut ein historisches Briefe-Korpus
    - darin Untersuchung von Nebensätzen mit Hilfe von Syntaxbäumen
  - Suche nach einem weiteren, ähnlichem Korpus, um die Ergebnisse der eigenen Studie replizieren zu können
  - dazu umfangreiches Wissen über Korpora von Dritten notwendig
  - Metadaten liefern dieses Wissen über Korpora
- heute: Fokus auf Annotationen in einem Korpus

Modellierung linguistischer Forschungsmetadaten

- Diversität linguistischer Forschung
  - Wir forschen über Kategorien wie Wortart, Satz, Komposition, Wortform, Benennung, etc.
  - Je nach Forschungsfrage sind unterschiedliche Kategorisierungen (bspw. Feinkörnigkeit, Ausprägung, Semantik) gefragt!
    - Definitionen, Skopus, Theorien
  - Konsequenz I: keine einheitliche Annotation
  - Konsequenz II: theorieabhängige Tagsets
  - Konsequenz III: keine Vorhersage von Tagsets möglich (wenige/keine „Standards“)

Modellierung linguistischer Forschungsmetadaten

- Was machen die Metadaten (hier)?
  - beschreiben die Korpora und deren Annotationen
- Was beschreiben sie?
  - objektgebunden: das was als Korpus definiert wurde
  - zweckgebunden: alle Eigenschaften, die Korpora haben können, um sie für einen bestimmten Zweck zu beschreiben
- heute: Wie können Metadaten Korpora beschreiben?
  - semantisch-eindeutig vs. technisch-abstrakt

Semantisch-eindeutiges Metadatenmodell

- Bsp. CLARIN (EU, Research Infrastructure Project):
  - CMDI (Broeder et al. 2010): Component Metadata Infrastructure
    - Framework für Metadaten

  - "**No single** metadata scheme could **ever address** all the needs of the **heterogeneous community of humanities** and social sciences researchers[...]."

    (http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml)

Semantisch-eindeutiges Metadatenmodell

- "There is a **clear need for semantically explicit metadata descriptions**. **Ambiguity** could otherwise threaten the usefulness of metadata when many metadata descriptions, **coming from a multitude of sources**, are made searchable."

  (http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml)

- indivuelle Kategorien, semantisch-eindeutig, keine Ambiguitäten
  - Bedeutungen von Kategorien werden in einer Datenbank, speziell dafür, gespeichert (ISO-Cat, Wright, Kemps-Snijders & Windhouwer 2007)
    - neue Bedeutungen hinzufügen, mehrfache Einträge möglich
    - Einträge werden referenziert, z.B. Type DC-3900

Semantisch-eindeutiges Metadatenmodell

- „[...] CMDI is not just another format. It is much more: as a **meta-model** it provides a well-defined framework **to define** and **use your own format**" (http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml)

- nicht zweckgebunden
- ALLES damit beschreibbar, nicht an ein Objekt gebunden
  - nicht nur Korpora, Annotationen, sondern auch Tools, Formate und Repositorien (u.v.m.)

- Es gibt natürlich Vorschläge von CLARIN zu Metadatenschema für Korpora und darin enthaltene Kategorien!

Semantisch-eindeutiges Metadatenmodell

- CMDI-Ansatz – Metadata Profile
  - genaue, umfangreiche Beschreibung einer Ressource (ist jede Ressource ein Korpus?), u.a.
    - Primärtext, Annotationen (written resource Type DC-3900)
    - Bedeutung der Annotationsebenen (annotation scheme DC-4085, tagset
    - und damit auch der Forschung (theoretic model DC-2501)
  - Profil-Struktur
    - Profile besitzen einzelne Komponenten, die wiederum Kategorien bündeln (Baukastenprinzip)
    - Struktur und Bezug zueinander frei
  - freie Anwendung dieses Profils

Semantisch-eindeutiges Metadatenmodell

- Vermeidung von semantischen Ambiguitäten in den Metadaten
- Anwendung des CMDI-Metadata Profile
  - jedes Korpus besitzt in irgendeiner Weise eine (primäre) „Text“-Ebene im Vergleich zu Annotationen (written resource Type DC-3900)
    - Metadaten zu Text-Ebenen in einem Korpus!
  - Beispiel: drei historische Korpora

Text-Ebenen in Korpora – KAJUK

text	Diplomatisches Transkript.
@norm	Normierte Schreibung.
E, @E	Ellipsen.

```
<lb n="8a,00,3003">
<J IR="kop"><KON>und</KON></J>
<J IR="kons" norm="dass" type="E"
dir="V"><SUB>das</SUB></J>
<subj real="Pron">wir</subj>
<KOR>also</KOR>
das Mal vor den Schweden <!--hier line-->
<praed><V ID="Inf"><VV>bleiben</VV></V></praed>
<praed><V ID="Fin"><MV>konten.</MV></V></praed></lb>
<line n="13"/>
```

KAJUK, Bauernleben

Text-Ebenen in Korpora - RIDGES

dipl	Die Transkription von Faksimiles stellt für die korpuslinguistische Aufbereitung zumeist die grundlegende, diplomatische Ebene (dipl). [...]
clean	Die clean-Ebene enthält erste vollautomatisch erstellte Normalisierungen hinsichtlich Sonderzeichen und grafischer Strukturierungen. [...].
norm	Die norm-Ebene stellt einen weiteren Normalisierungsschritt dar, indem hier die Tokenisierung und die Orthografie einheitlich nach modernen Orthografieregeln (vgl. Duden) angepasst werden, wobei die Flexion, wie z.B. Kasuszuweisungen, nicht berücksichtigt wird.[...].

dipl	von	Geiſ	ſen	vnnd	Haſen	zuverſtehen		/
clean	von	Geissen		vnnd	Hasen	zuverstehen		/
norm	von	Geißen		und	Hasen	zu	verstehen	/

RIDGES 4,1, PflantzGart_1639

Text-Ebenen in Korpora – HSJ

text	Diplomatischer Text. Zusammengeschriebene Wörter und Abkürzungen so dargestellt, wie sie vorgefunden wurden. Unklare Schreibung = "#".
aug-text	Diplomatischer Text. Der Text, der als Grundlage genutzt wird. Unklare Schreibung = "#".
mean	Hier wird die neuhochdeutsche Übersetzung/Entsprechung angeführt. Diese wurde dem Wörterbuch, aus dem das jeweilige Lemma stammt, entnommen.

#	máeše_	.	Alécsándér	Mukdun	,	der	war	ain	gėwaltigér	kinég
[141]	máeše_	.	Alécsándér	Mukdun	,	der	war	ain	gėwaltigér	kinég
	maese/geschihte					der/der	zayn/sîn	an/ein	gevaldik/gewaltic	kinig/künec

HSJ, BM_141

Semantisch-eindeutiges Metadatenmodell

- Beispiele KAJUK, RIDGES, HSJ zeigen
  - verschiedene Definitionen von Text-Ebenen
    - text (2), dipl, aug-text,
    - clean, norm (2), mean
    - Was ist mit Parallelkorpora (z.B. AHD-Latein?)

- je mehr Korpora, desto mehr Definitionen
- flexible Attribut-Wert-Paare
- Was ist jetzt die semantisch-eindeutige Definition von Text-Ebene?

Semantisch-eindeutiges Metadatenmodell

- Was ist für alle Korpora die semantisch-eindeutige primäre Text-Ebene, was die Normalisierungsebene?
  - Primäre Text-Ebene, die konzeptionell am nächsten am Original ist? (Was heißt am nächsten?)
  - Primäre Text-Ebene, auf der die meisten Annotationen basieren?
  - Zählen „einzelne Annotationen“, die eigentlich konzeptionell Normalisierungen sind, zu Normalisierungstextebenen?
  - Mehrere Normalisierungen?
  - Parallelkorpora?
- Alles „Text-Ebenen“? Vorteil einer solchen Kategorisierung?

- Erwartung an eine semantische Beschreibung des suchenden Forschers, der möglicherweise seine eigene Definition von „Text“ besitzt?

Semantisch-eindeutiges Metadatenmodell

- Aufgabe: Beschreibung von Text-Ebenen in Korpora
  - Anwendung von CMDI-Profilen im Virtual Language Observatory (VLO)
    - Annotation, Normalisierung, Transkription, Text
- Semantisches Mapping der verschiedenen Kategorien (resource type) ist schwer zu bewerten:
  - Text (186)
  - text (81)
  - textAnnotation(32111)
  - Corpus (55)
  - corpus (43)
  - Dataset (19)
  - Anwendung von CMDI nicht in gleicher Form und gleichem Umfang
  - keine genaue Vorhersage, „wo“ in der Struktur der Metadaten das Schlagwort gefunden wurde

Diskussion des CMDI-Ansatzes

- Anwendung des Metadatenschemas für EINE Ressource
  - semantisch-eindeutiger Ansatz
  - eigenständige und spezielle Dokumentation von Daten UND Forschung
  - Semantik muss für jeden Fall beschrieben werden
    - Nutzung der ISO-Cat Datenbank
    - Verstehen der Forschung

Diskussion des CMDI-Ansatzes

- Anwendung des Metadatenschemas für MEHRERE Ressourcen
  - Vereinheitlichung dieser Metadaten schwierig, siehe VLO
  - semantische Eindeutigkeit gelingt durch ISO-Cat
    - „mein“ Nomen hat folgende Eigenschaften
    - „meine“ Textebene hat folgende Eigenschaften
  - viele Mengen, von dem was die Metadaten beschreiben
    - Dokumente, Tools, Tagsets, Autorenschaft, Forschung etc.
  - Einfluss des Zwecks der Beschreibung, Zielgruppen, eigene Kategorien
    - nicht alle nutzen CMDI gleich

- Wenn man keine festen Semantiken vorgeben kann, dann kann man den verschiedenen Forschungsrichtungen nur gerecht werden, wenn man diese nicht in das Metadatenmodell mit aufnimmt!

Technisch-abstraktes Metadatenmodell

- Warum technisch-abstrakte Metadaten?
  - alle Korpora unabhängig von der Forschung zusammen abzubilden (objekt- und zweckorientiert!)
  - Gemeinsamkeiten der Korpora
    - technische Realität
    - abstrakte Konzepte (idealerweise semantik-/theorie-neutral)
  - dafür ein allgemeines Format nutzen
    - Text Encoding Initiative (Burnard & Bauman 2008)
    - Guidelines nicht nur von einer Forschungsrichtung geprägt
    - **sehr allgemeine Semantiken, die durch Kontext spezifisch werden – neue Interpretation möglich**

Technisch-abstraktes Metadatenmodell

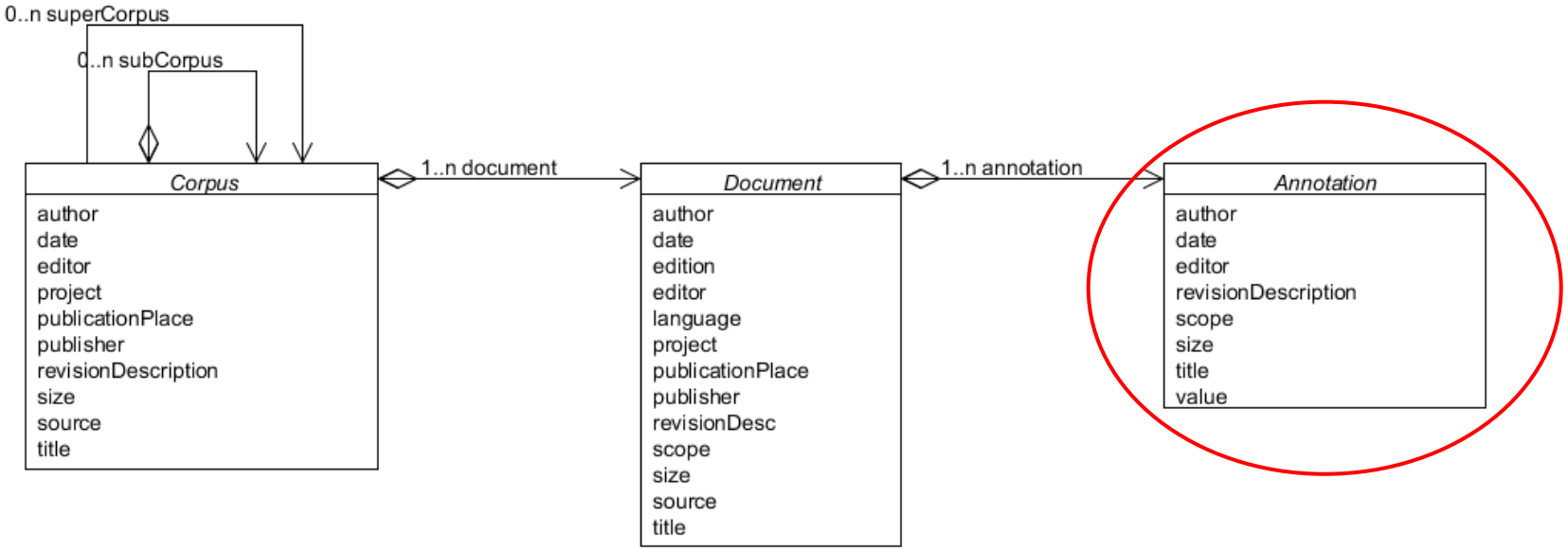
- Vermeidung von semantischen Konzepten in den Metadaten
  - eine Menge an Korpora soll durch Metadaten beschrieben werden
  - jedes Korpus besitzt in irgendeiner Weise „Text“-Ebenen
    - Die Frage nach einer bestimmten Text-Ebene wird nicht gestellt!
    - Es muss dennoch möglich sein, eine eindeutige, konkrete Beschreibung von Text-Ebenen ohne semantische Kategorien vorzunehmen!

Technisch-abstraktes Metadatenmodell

- Alle Text-Ebenen sind Annotationen!
  - Annotationen sind immer Interpretationen nach einem Annotationsschema
  - dipl, text, clean, aug-text sind ebenfalls Interpretationen nach einem Annotationsschema
- Damit sagt das Modell nicht, dass ein Korpus eine Textebene besitzen muss!
  - ein Korpus ist immer noch ein Korpus, wenn es mehrere Text-Ebenen besitzt
  - ein Korpus ist immer noch ein Korpus, wenn es eine völlig neue, nie gekannte, fachfremde (idiosynkratische, besondere ...) Text-Ebene besitzt
- Ein Korpus ist ein Korpus, wenn es Annotationen hat!

Metamodell linguistischer Korpora

Wir gehen davon aus, dass ...



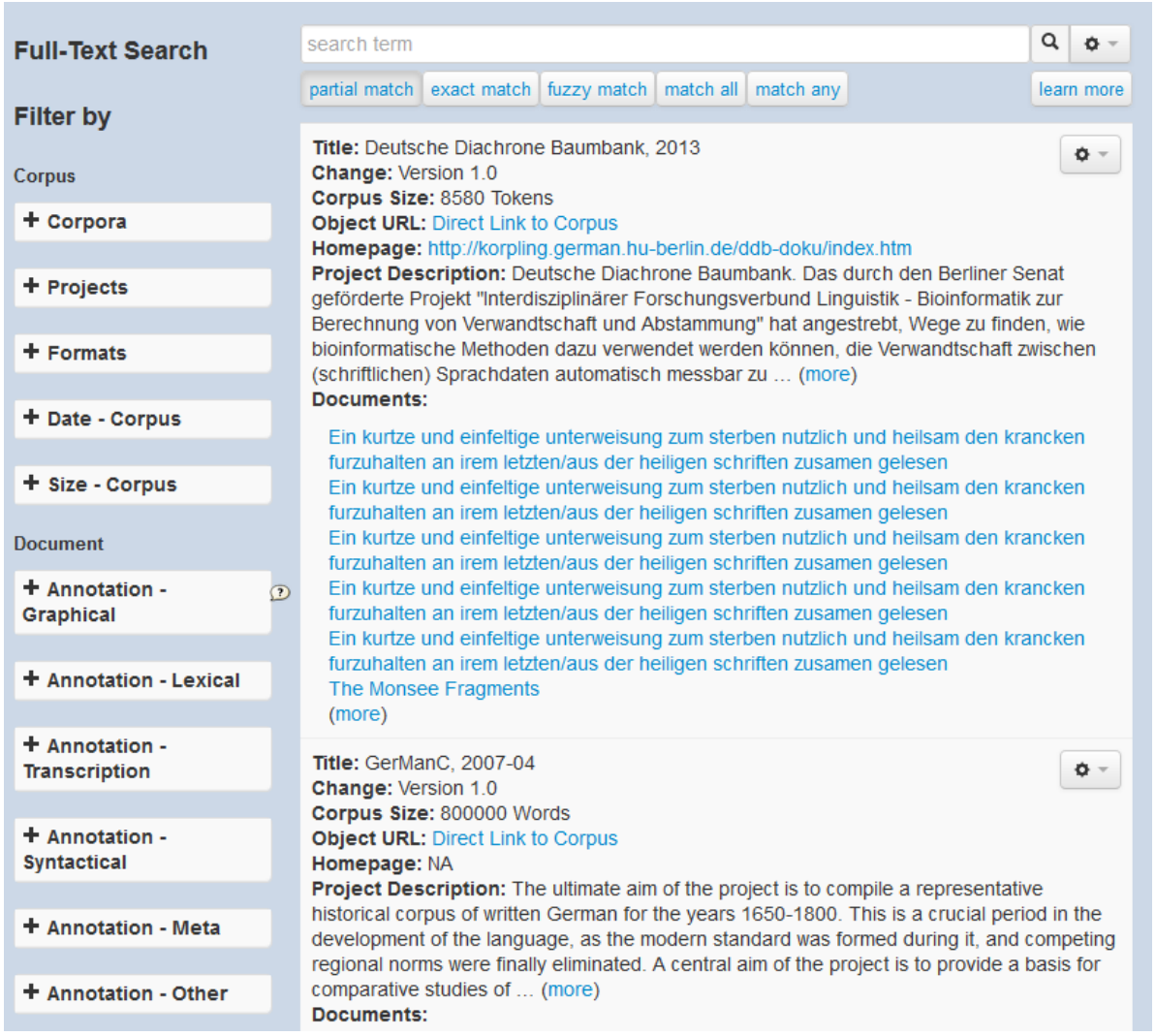
Technisch-abstraktes Metadatenmodell

- Aufgabe: Beschreibung von Text-Ebenen in Korpora
  - Annotationslisten gleich für alle Korpora
    - KAJUK: @norm, text, E
    - RIDGES: dipl, clean, norm
    - HSJ: text, aug-text, mean
  - Segmentierung (Tokenisierung) + String-Wert = Kandidaten für Text-Ebenen in Korpora
    - String = nicht feste (freie) Menge an Annotationswerten (z.B. Wortformen)
    - andere Annotation sind auf dieser Ebene (un-)mittelbar annotiert (in den Metadaten ist die Segmentierung angegeben)
    - Transkriptionsrichtlinie als Tagset für solche Ebenen (TEI tagsDecl)

Technisch-abstraktes Metadatenmodell

- Aufgabe: Beschreibung von Text-Ebenen in Korpora
  - Anwendung im LAUDATIO-Repository
    - KAJUK: @norm, text, E, J, lb, @n, GF [...]
    - RIDGES: dipl, clean, norm, pos, lb, pb, term [...]
    - HSJ: text, aug-text, mean, pos, morph [...]
  - Gruppierung dieser Annotationen für den (bekannten) Nutzer
    - alle Annotation sind in grobe Kategorien für graphischen Oberfläche eingeteilt (Lexikalische, Transkriptionen, Syntaktische)
    - kein Bestandteil des Metadatenmodells
    - Anpassung für Anwendungen und Anwender

LAUDATIO-Repository



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)

Title: GerManC, 2007-04
Change: Version 1.0
Corpus Size: 800000 Words
Object URL: Direct Link to Corpus
Homepage: NA
Project Description: The ultimate aim of the project is to compile a representative historical corpus of written German for the years 1650-1800. This is a crucial period in the development of the language, as the modern standard was formed during it, and competing regional norms were finally eliminated. A central aim of the project is to provide a basis for comparative studies of … (more)
Documents:

Technisch-abstraktes Metadatenmodell

- Modell definiert, was beschrieben werden soll
  - kein Wissen über Bedeutungen von „Annotationsnamen“ benötigt
    - Bedeutung „text(1)“, „text (2)“, „norm“ etc. (deren Bedeutung wird natürlich in den Metadaten wiedergegeben)
    - wichtige strutkurelle Angabe der Segmentierung jeder Annotation
  - schwierigen Fragen nicht im Modell zu klären
  - Theorien als Interpretationen →Annotationen
    - Tagsets der Textebenen abgebildet, nicht Grundlage des Modells
  - Korpora können x Text-Ebenen/Annotationen besitzen
    - Innovation neuer Richtlinien zur Text-Ebenen-Erstellung
      - sowie wir es schon für Annotationen eigentlichen kennen
    - Uminterpretation des Primärdatums möglich

Zusammenfassung

- in diesem Modell:
  - keine Primärtexte
- ungleich: technische Modellierung von Korpusdaten
  - Format, atomare Token

- Modellierung der Metadaten
  - technisch-abstrakte Strukturierung
  - die Beziehung von Annotationen aufeinander beschrieben werden kann
  - nicht grundlegend auf uniforme Semantiken gebaut
- **Bis lang: Annotieren in Strukturen mit freien Semantiken**
- **Vorschlag: Dokumentieren in Strukturen mit freien Semantiken**

Ausblick

- weitere Korpora aus anderen (nicht-linguistischen) Fächern testen: Musikwissenschaft (Katrin Bicher, Musiksoziologie)
  - Annotieren mit Referenzlisten
    - Indexierung, Referenzierung
  - Annotieren mit kritischem Apparat
    - Kommentare (erklärende, kritische) „in“ den Texten
- Frage: Passt das ebenfalls in die technisch-abstrakte Modellierung?

Herzlichen Dank!

Vor allem an:

Anke, Florian, Laura, Laurent, Malte, Thomas, Vivian

Referenzen

- ISO Cat https://catalog.clarin.eu/isocat/interface/index.html
- TEI p5 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-name.html
- CLARIN clarin.eu
- VLO: http://catalog.clarin.eu/vlo/?0
- KAJUK: Kasseler Junktionskorpus http://hdl.handle.net/11022/0000-0000-2102-8
- RIDGES: Register in German Science, Herbology Corpus http://hdl.handle.net/11022/0000-0000-2D32-6
- HSJ: Historische Syntax des Jiddischen http://hdl.handle.net/11022/0000-0000-24F9-F
- CMDI http://www.clarin.eu/content/component-metadata

Referenzen

- **Broeder**, D., Kemps-Snijders, M., et al. (2010). A data category registry- and component-based metadata framework. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10) , Valletta, Malta. ELRA.
- **Burnard**, L., Bauman, S. (Ed.) (2008) *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford.*
- **Krause**, Th., Lüdeling, A., Odebrecht, C., Romary, L., Schirmbacher, P., Zielke, D. (2014) LAUDATIO-Repository: Accessing a heterogeneous field of linguistic corpora with the help of an open access repository. *Digital Humanities 2014 Conference. Poster Session. 8.7.-12.7.2014, Lausanne.* http://www.laudatio-repository.org/
- Odebrecht, Carolin (2014) Modeling Linguistic Research Data for a Repository for Historical Corpora. *Digital Humanities 2014 Conference. 8.7.-12.7.2014, Lausanne.*
- **Odebrecht**, Carolin, Krause, Thomas (2013) Metadata in an Infrastructure for Historical Corpora. *SFB 732 Incremental Specification in Context. Kolloquium. 20.06.2013, Stuttgart.*
- **Wright**, S.E., M. Kemps-Snijders, M., Windhouwer, M. (2007) ISO-Cats: The Revised and Future TC 37 Data Category Registry. Presentation at the *Pragmatic Applications for TC 37 Standards* (TC37 2007), Provo, UT USA, August 13, 2007.