

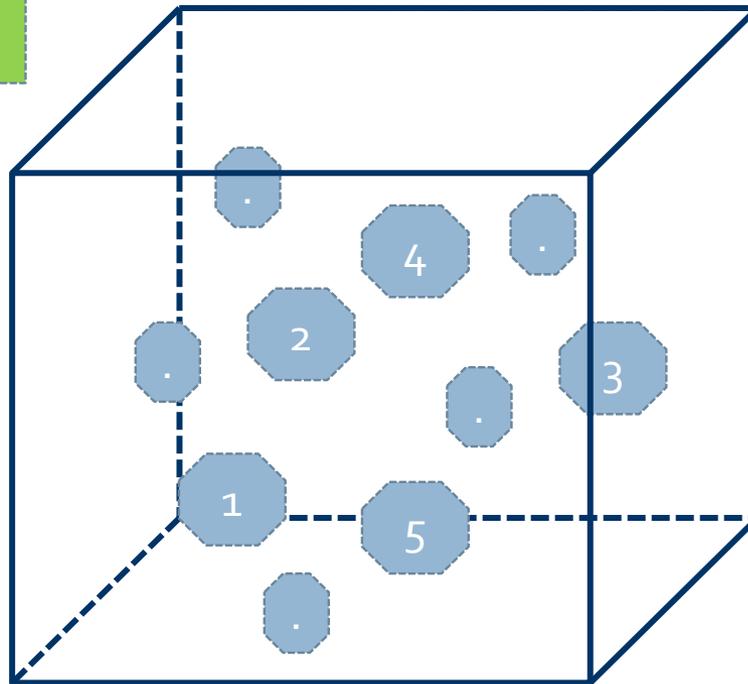
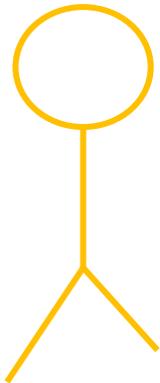
Carolin Odebrecht

Korpuslinguistik | Humboldt-Universität zu Berlin

INTERDISZIPLINÄRE NUTZUNG VON FORSCHUNGSDATEN MITHILFE EINER TECHNISCH-ABSTRAKTEN MODELLIERUNG

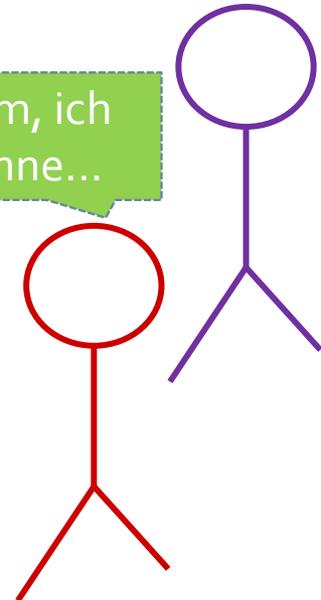
Aufgabe

Was ist eigentlich eine Normalisierung?



Ich brauche ein normalisiertes Korpus aus dem 16. Jahrhundert!

Ähm, ich kenne...

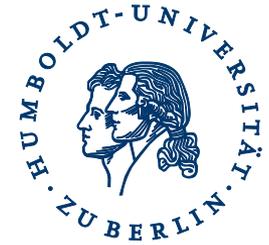


Fragestellung



- Wie können Metadaten eine Menge von (textbasierten) Korpusdaten beschreiben
 - um **Konzepte dritter** zu verstehen
 - um für die **eigene Forschung** eine **Auswahl** zu treffen
 - um Korpora **wiederverwenden**

Datenaustausch



- **Korpora und deren Wiederverwendung**

- ansehen, durchsuchen, analysieren
- vergrößern mit mehr Texten mit gleicher Annotation
- neu zusammenstellen mit gleicher, weniger oder anderer Annotation
- neue Annotationen hinzufügen
- ...

- Einführung in Korpora vgl.
Lemnitzer & Zinsmeister 2006, Lüdeling
2011
- Data Life Circle vgl. Rümpel 2011

- **Wissen über Korpora dazu erforderlich**

- Textgrundlagen/Textvorlagen
- Annotationsschemata, Annotationstools, Annotatoren
- Prüfverfahren, Versionen, Konvertierungen
- ...

- Funktion von Metadaten vgl.
Odebrecht & Krause 2013

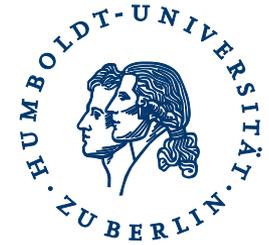
Problematik

- Ein Beispiel: ‚tok‘ enthält
 - primäre Textgrundlage
 - virtuelle Texteinheiten
 - normalisierte Texteinheiten
- Welche Annotation basiert auf welcher anderen Annotation?

Verbart				VV											VV	MV	
Verbform				Fin											Inf	Fin	
Satzglied				praed			subj	KOR							praed	praed	
Realisation_gram							Pron										
E						E											
dir						V											
tok		in	hesischer	Conterbution	waren	und	das	wir	also	das	Mal	vor	den	Schweden	bleiben	konten	.

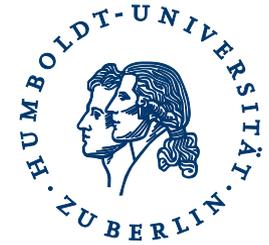
Kasseler Junktionskorpus Vilmos Ágel /Mathilde Hennig (Justus-Liebig-Universität Gießen), Bauernleben <http://hdl.handle.net/11022/0000-0000-2102-8>

Forschungsdaten

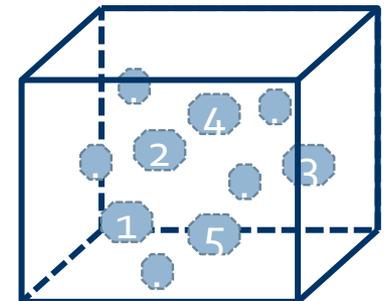


- notwendige Diversität in der Forschung
 - Kategorien
 - Wortart, Organisation, Komposition, Beziehung, Ellipsen, Benennung, Korrespondenz, Argumentation, Diskurs etc.
 - unterschiedliche Kategorisierungen nach Feinkörnigkeit, Ausprägung, Semantik
 - Definitionen, Skopus, Theorien (vgl. Lüdeling 2012)
 - Konsequenz I: keine einheitlichen Annotationen (und Formate)
 - Konsequenz II: theorieabhängige Tagsets
 - Konsequenz III: keine Vorhersage von Annotationen möglich (wenige/keine „Standards“)

Problematik



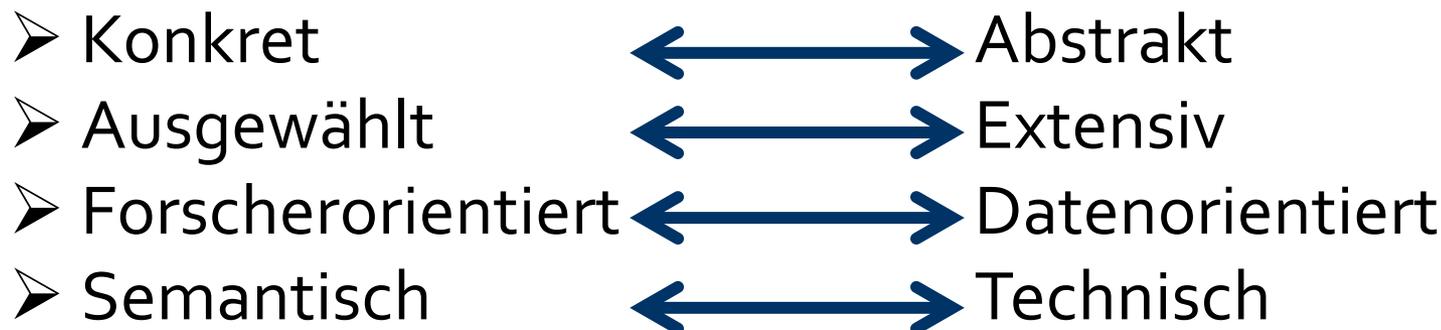
- In vielen Repositorien wissen wir nicht, was in den vorhandenen Korpora enthalten ist!
 - textbasiert
 - verschiedene Register, Sprachen, Sprachstufen etc.
 - annotiert
 - Token-, Spannenannotationen, Baumbanken, Inline-Annotationen etc.
 - Formate
 - vielfältig (vgl. Zipser 2014 für linguistische Formate)



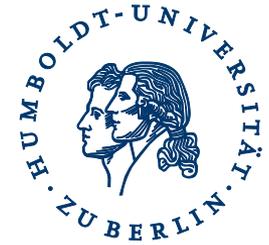
Metadaten

- Metadaten können im Repository zeigen, was in Korpora enthalten ist
 - Textgrundlagen/Textvorlagen
 - Annotationsschemata, Annotationstools, Annotatoren
 - Prüfverfahren, Versionen, Konvertierungen

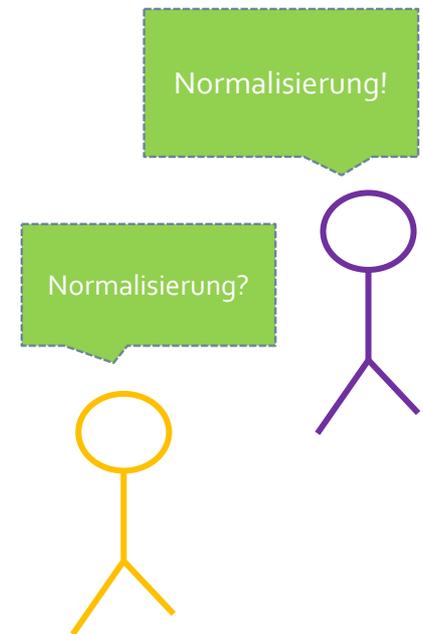
- Funktion von Metadaten vgl. Miller 2011, NISO 2004



Metadaten

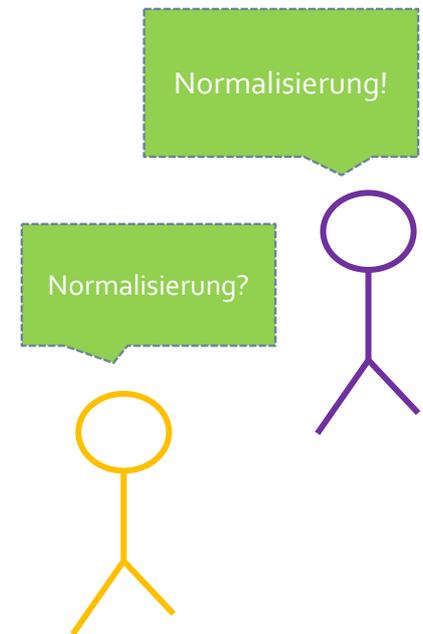


- Für „Text“ leichter?
 - primär, sekundär
 - transkribieren, normalisieren
 - Wortformen, Buchstaben
 - ...
- Normalisierung
 - nach moderner Grammatik?
 - nach Zeichen?
 - nach Lexemen?
 - nach ...
- Was wird eigentlich normalisiert?
 - Primärtext
 - Transkript
 - ...

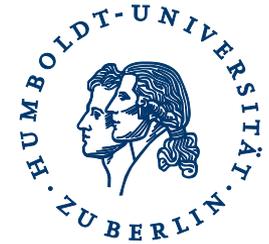


Metadaten

- Text-Ebenen für jedes Korpus
 - in Abgrenzung zu Annotationen
- Annahme:
In welchem Repository wir auch suchen, kaum eine Text-Definition kann auf mehrere Korpora angewendet werden.
- **Es muss mehrere Text-Definitionen geben können.**
- zurück zum Beispiel



Kasseler Junktionskorpus



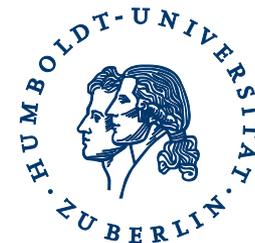
text	Diplomatisches Transkript.
@norm	Normierte Schreibung.
E, @E	Ellipsen.

```
<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"/>
```

Kasseler Junktionskorpus Vilmos Ágel /Mathilde Hennig (Justus-Liebig-Universität Gießen), Bauernleben <http://hdl.handle.net/11022/0000-0000-2102-8>

RIDGES

Kräuterkundekorpus

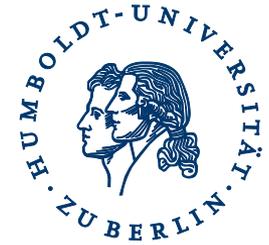


dipl	Die Transkription von Faksimiles stellt für die korpuslinguistische Aufbereitung zumeist die grundlegende, diplomatische Ebene (dipl). [...]
clean	Die clean-Ebene enthält erste vollautomatisch erstellte Normalisierungen hinsichtlich Sonderzeichen und grafischer Strukturierungen. [...].
norm	Die 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	Geiß	fen	vnnd	Hasen	zuverftehen	/
clean	von	Geissen		vnnd	Hasen	zuverstehen	/
norm	von	Geißen		und	Hasen	zu verstehen	/

RIDGES 4,1, Anke Lüdeling, Carolin Odebrecht, Amir Zeldes (Humboldt-Universität zu Berlin Berlin) PflanzGart_1639,
<http://hdl.handle.net/11022/0000-0000-2D85-8>

Musiksoziologie Vereinsgeschichte



text	Text
------	------

<p>Wien, den 17. Jänner 1921.</p>

<p>Sehr geehrter <persName role="IAV" type="Präsident" ref="Personenliste.xml#P00001">Herr Schönberg</persName>!</p>

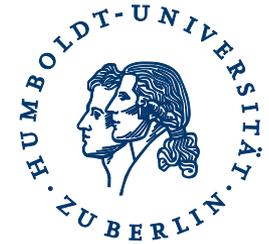
<p>

<persName role="IAV" type="Sonstige" ref="Personenliste.xml#P0007">Herr Berg</persName>

sagte mir, dass Sie einen Bericht über den Verkauf der Mitteilungen wünschen. Infolge des von Mittwoch bis gestern dauernden Poststreiks war ich bisher nicht in der Lage, Ihnen die Aufstellung zu übersenden. Jetzt, da er beendet ist, beeile ich mich, Ihrem Wunsche nachzukommen.</p>

Klarfeld_AS_1921001^7_5672, Katrin Bicher (Humboldt-Universität zu Berlin) https://www.muwi.hu-berlin.de/soziologie/mitarbeiter_soz/katrin-bicher

Koptisch - Sahidisch Shenoute

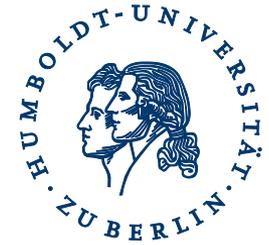


tok	Characters, Morphemes
dipl	Coptic Scriptorium preserves spelling, punctuation, line breaks, column breaks, page breaks.
dipl_word	Diplomatic word form.
norm	Normalization of word forms.

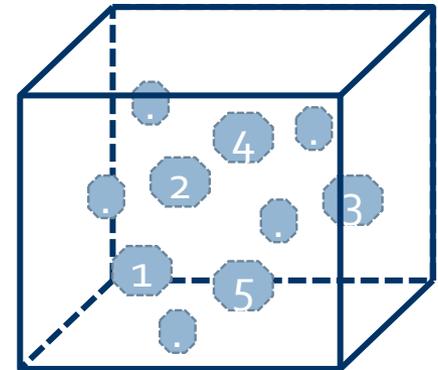
cb	cb					
dipl	ε	τρε	π	νοϣ[τε]		ϸϩⲚⲧ
dipl_word	ετρεπνοϣ[τε]					ϸϩⲚⲧ
hi_rend				superscript		
lb	lb				lb	
norm	ε	τρε	π	νοϣτε		ϸϩⲚⲧ
note				o sits directly above the κϣ		manuscript
p	p					
pb_xml_id	YA422					
pos	PREP	ACAUS	ART	N		N
tok	ε	τρε	π	κ	ο	ϣ [τε] ϸϩⲚⲧ

Shenoute A 22, Amir Zeldes (Georgetown University) Caroline T. Schroeder (University of the Pacific)
<http://hdl.handle.net/11022/0000-0000-4680-0>

Gemeinsamkeiten

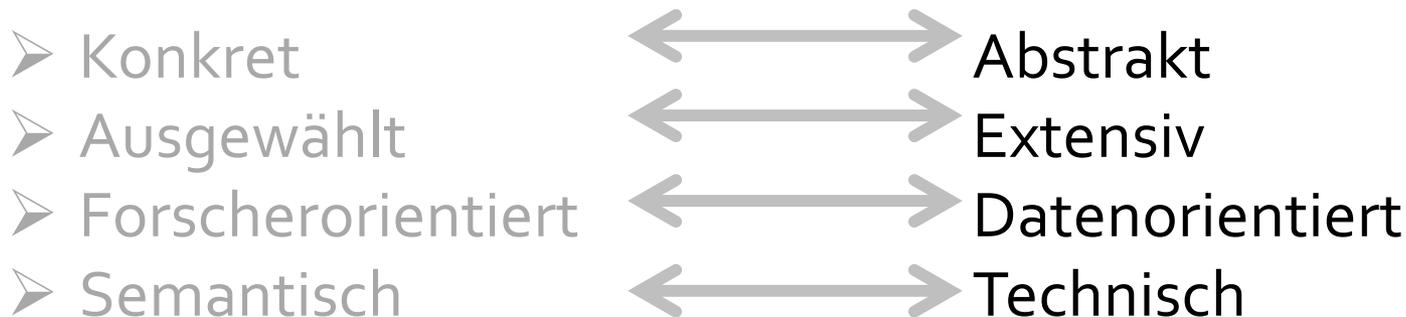


- verschiedene Definitionen von „Text“
 - zwischen Korpora
 - innerhalb eines Korpus
 - abhängig von der Forschung
 - text (3), dipl, aug-text, dipl_word, tok
 - clean, norm (3), mean
 - ...
- Normalisierungen
 - mit verschiedenen Skopi
 - für unterschiedliche Zwecke
 - ohne scharfe Abgrenzung zu anderen Annotationen wie bspw. Beziehungstyp, Wortart oder Edition



Metadaten

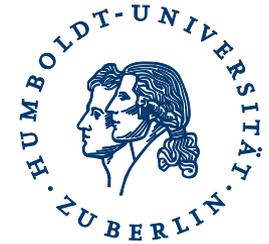
- Metadaten können im Repository zeigen, was in Korpora enthalten ist
 - Textgrundlagen/Textvorlagen
 - Annotationsschemata, Annotationstools, Annotatoren
 - Prüfverfahren, Versionen, Konvertierungen
- um dem gerecht zu werden, Argumentation für



Metadaten

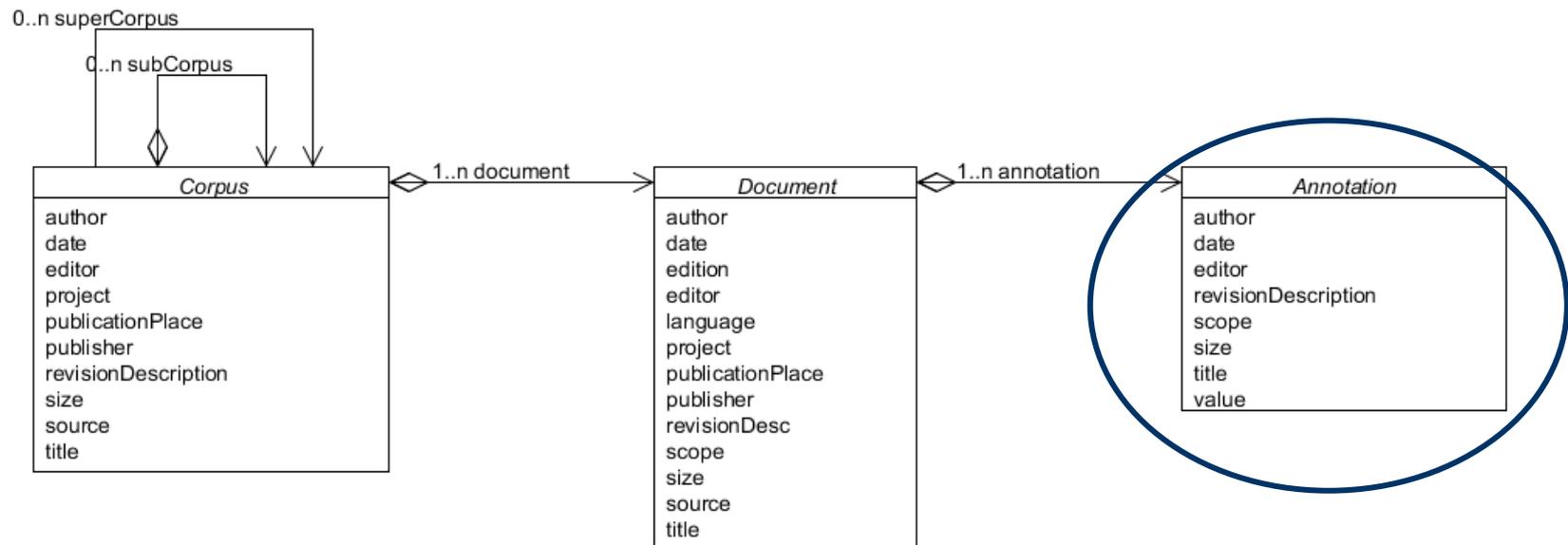
- Vermeidung von semantischen Konzepten im Modell
 - **Beispiele zeigen: Keine Vorhersage möglich, welche Konzepte noch erfasst werden müssen!**
- jedes Korpus besitzt eine Art „Text“-Ebene
 - aber keine bestimmte Text-Ebene
 - mehr als eine Text-Ebene

Metadatenmodell

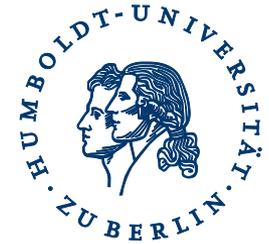


➤ Forschungsdaten beschreiben

- technisch-abstrakte, extensive Modellierung (Odebrecht 2014)



Metadatenmodell



- „**Text**“ als Annotation beschreiben und Annotation aufgrund ihrer **technische Gemeinsamkeiten (nicht Bedeutung)** zusammenfassen
 - technisch-abstrakt
 - **jede** Annotation beschreiben
 - extensiv
 - Annotation in ihrer **Realisierung** beschreiben
 - datenorientiert
- **zweckgebundenes und objektgebundenes technisch-abstraktes Metadatenmodell**

Metadatenmodell

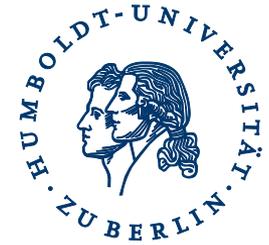


➤ Annotationen

- Damit sagt das Modell nicht, dass ein Korpus eine (bestimmte/primäre/etc.)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!

Technische Umsetzung



- allgemeines Format
 - Text Encoding Initiative (Burnard & Bauman 2008, Burnard & Rahtz 2004)
 - Guidelines nicht nur von einer Forschungsrichtung geprägt
 - **sehr allgemeine Semantiken, die durch Kontext spezifisch werden**
 - **neue Interpretation möglich**
- ja Klasse im Modell eine TEI Spezifikation (ODD)
 - unter CC_BY 3.0

<https://github.com/korpling/LAUDATIO-Metadata>

Aufgabe

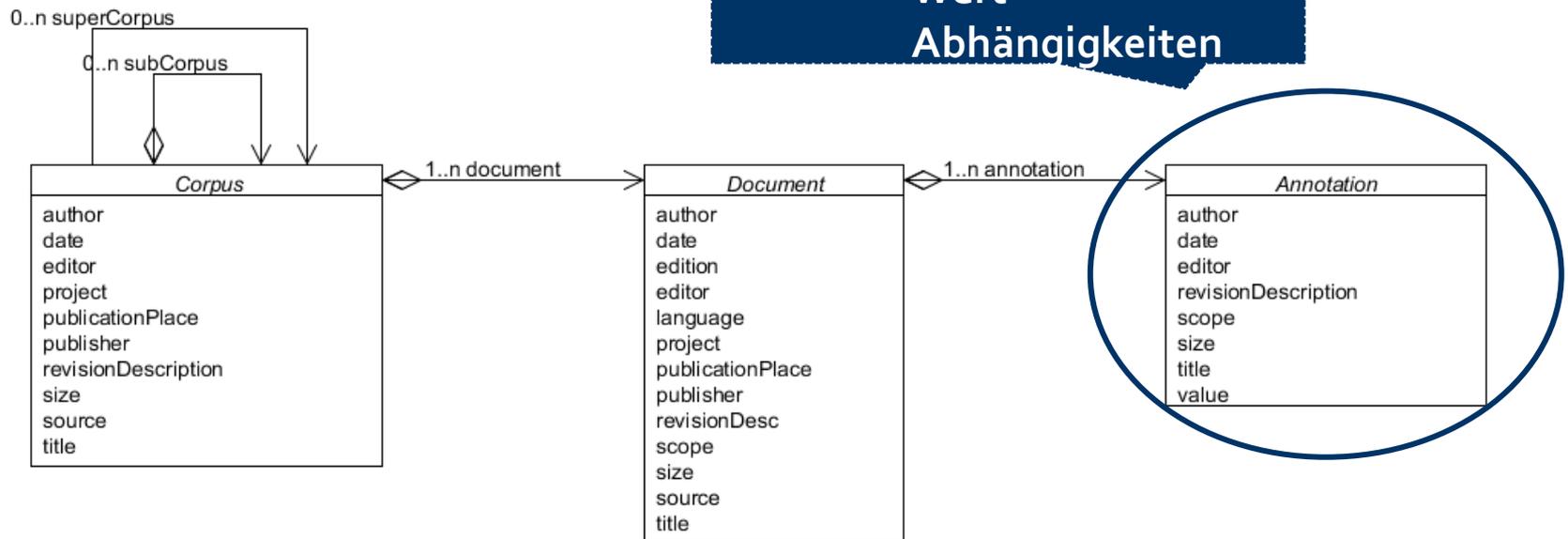
- Gesucht wird ein Korpus mit Normalisierung
 - Eigenschaften der Annotation „Text“
 - konkrete Werte der Annotation (Strings)
 - Wortformen, Buchstaben, Morpheme etc.
 - jedes Dokument im Korpus besitzt diese Annotation
 - andere Annotationen basieren darauf
 - (alle) anderen Annotationen sind auf dieser Ebene (un-)mittelbar annotiert
 - eigenständige Tokenisierung (Segmentierung, vgl. Krause et al. 2012)
- Das ist in dem Modell verankert:

Metadatenmodell

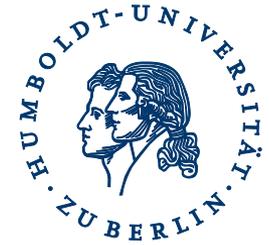
➤ Forschungsdaten

- technisch-abstrakte Modelle

Metadaten
 Token
 Segmentierung
 Name
 Wert
 Abhängigkeiten



Aufgabe Normalisierung

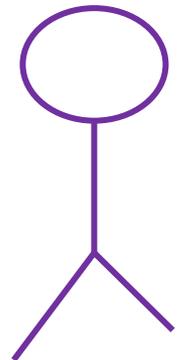
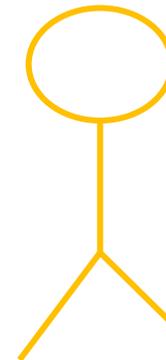


- **(technisch abstrakte) Text-Ebenen**
- finde alle unabhängigen, eigenständig tokenisierte Annotationen mit freien Annotationswerten
 - Annotationsrichtlinien dienen dem konkreten Verständnis

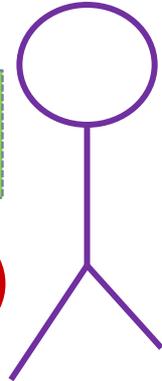
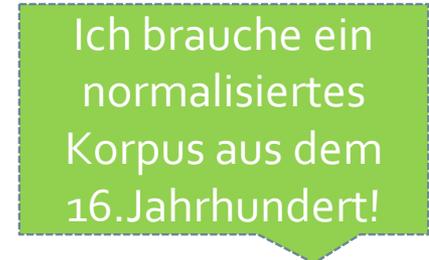
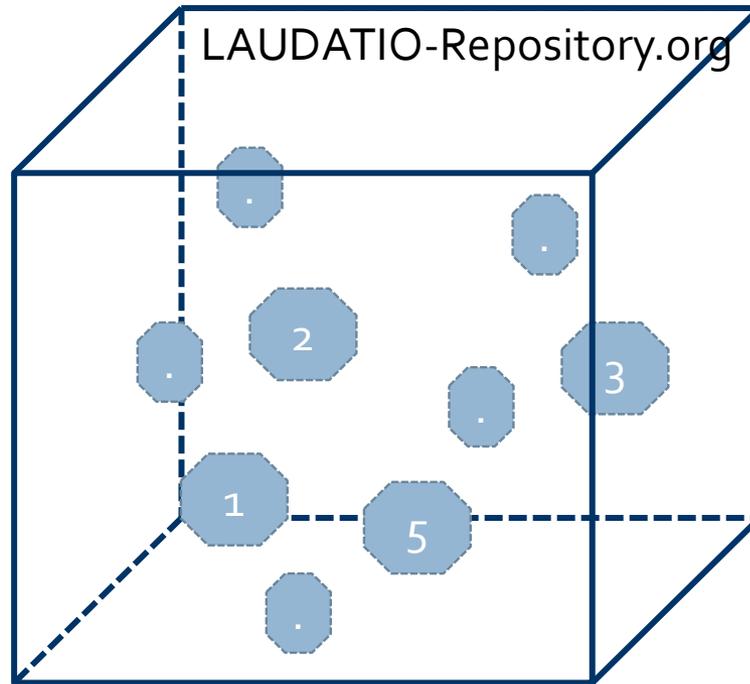
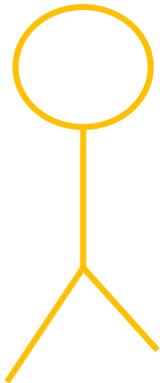
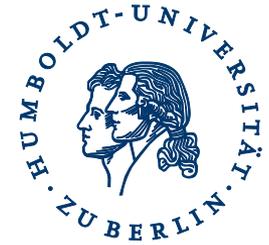
- gewünschte Eigenschaften der Normalisierung
 - durchgängig oder nicht
 - nach bestimmten Regeln
 - in bestimmter Form oder Format

Ich brauche ein
normalisiertes
Korpus aus dem
16. Jahrhundert!

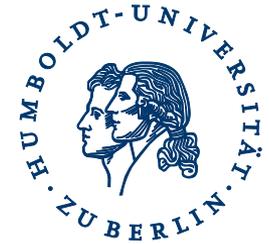
Was ist eigentlich
eine
Normalisierung?



Aufgabe gelöst!

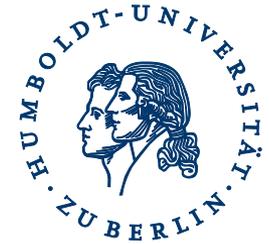


Referenzen



- Burnard, L., Bauman, S. (Ed.) (2008)** TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford.
- Burnard, L., Rahtz, S. (2004)** RelaxNG with Son of ODD. Extreme Markup Languages Proceedings 2004. Montréal, Québec. **Himmelman, N. P. (2012)** Linguistic Data Types and the Interface between Language Documentation and Description. In Language Documentation & Conservation 6. 187-207.
- 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/>
- Krause, T., Lüdeling, A., Odebrecht, C., Zeldes, A. (2012) Multiple Tokenization in a Diachronic Corpus. Exploring Ancient Languages through Corpora Conference (EALC), 14.-16.Juni 2012.
- Lemnitzer, L., Zinsmeister H. (2006)** Korpuslinguistik. Eine Einführung. Gunter Narr Verlag, Tübingen.
- Lüdeling, A. (2012)** A corpus-linguistics perspective on language documentation, data, and the challenge of small corpora. In Seifart, F., Haig, G., Himmelman, N. P., Jung, D.,; Margetts, A. & Trilsbeek, P. (Hg.) Potentials of Language Documentation: Methods, Analyses, and Utilization. Language Documentation & Conservation Special Publication No. 3 at the University of Hawai'i Press. 32-38.
- Lüdeling, A. (2011)** Corpora in Linguistics: Sampling and Annotation. In Grandin, K. (Hg.) Going Digital. Evolutionary and Revolutionary Aspects of Digitization. [Nobel Symposium 147]. Science History Publications/USA, New York. 220-243.
- Miller, Steven J. (2011)** Metadata for Digital Collections. A How-To-Do-It Manual. New York; London: Neal-Schuman Publishers (How-To-Do-It Manuals, 179).
- NISO (2004)** Understanding Metadatada. Hg. v. NISO. Bethesda. Online verfügbar unter <http://www.niso.org/publications/press/UnderstandingMetadata.pdf>, zuletzt geprüft am 13.02.2015.
- Odebrecht, C. (2014)** Modeling Linguistic Research Data for a Repository for Historical Corpora. Digital Humanities 2014 Conference. 8.7.-12.7.2014, Lausanne.
- Odebrecht, C., Krause, T. (2013)** Metadata in an Infrastructure for Historical Corpora. SFB 732 Incremental Specification in Context. Kolloquium. 20.06.2013, Stuttgart.
- Rümpel, St. (2011)** Der Lebenszyklus von Forschungsdaten. In Büttner, St., Hobohm, H. & Müller, L. (Hg.) Handbuch Forschungsdatenmanagement. Bock und Herchen Verlag. Bad Honnef. 25-31.
- Salmon-Alt, S., Romary, L., Pierrel, J. (2006)** Un modèle générique d'organisation de corpus en ligne : application à la FReeBank. Traitement Automatique des Langues, ATALA, 2006, 45, 145-169. <hal-00110970>
- Zipser, F. (2014)** SaltNPepper und das Formatpluriversum. LAUDATIO-Workshop 07.10.2014. Berlin.

Ressourcen



- TEI p5 <http://www.tei-c.org>
- Coptic Scriptorium Shenoute A 22
<http://hdl.handle.net/11022/0000-0000-4680-0>
- KAJUK: Kasseler Junktionskorpus
<http://hdl.handle.net/11022/0000-0000-2102-8>
- RIDGES: Register in German Science, Herbiology Corpus
<http://hdl.handle.net/11022/0000-0000-2D32-6>
- HSJ: Historische Syntax des Jiddischen
<http://hdl.handle.net/11022/0000-0000-24F9-F>
- TEI ODDs – für Metadatenschemata
<https://github.com/korpling/LAUDATIO-Metadata>