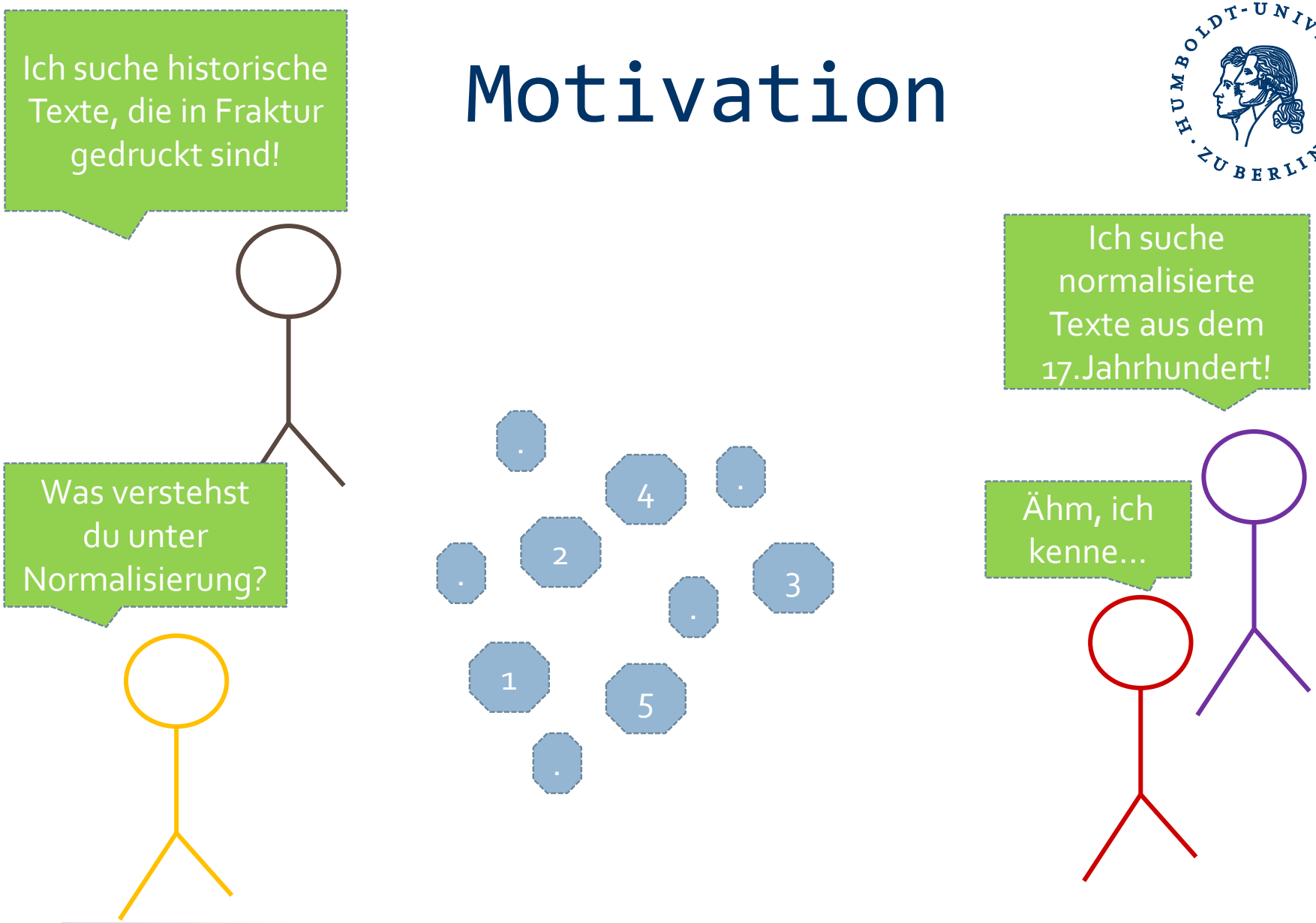
Carolin Odebrecht
Korpuslinguistik | Humboldt-Universität zu Berlin

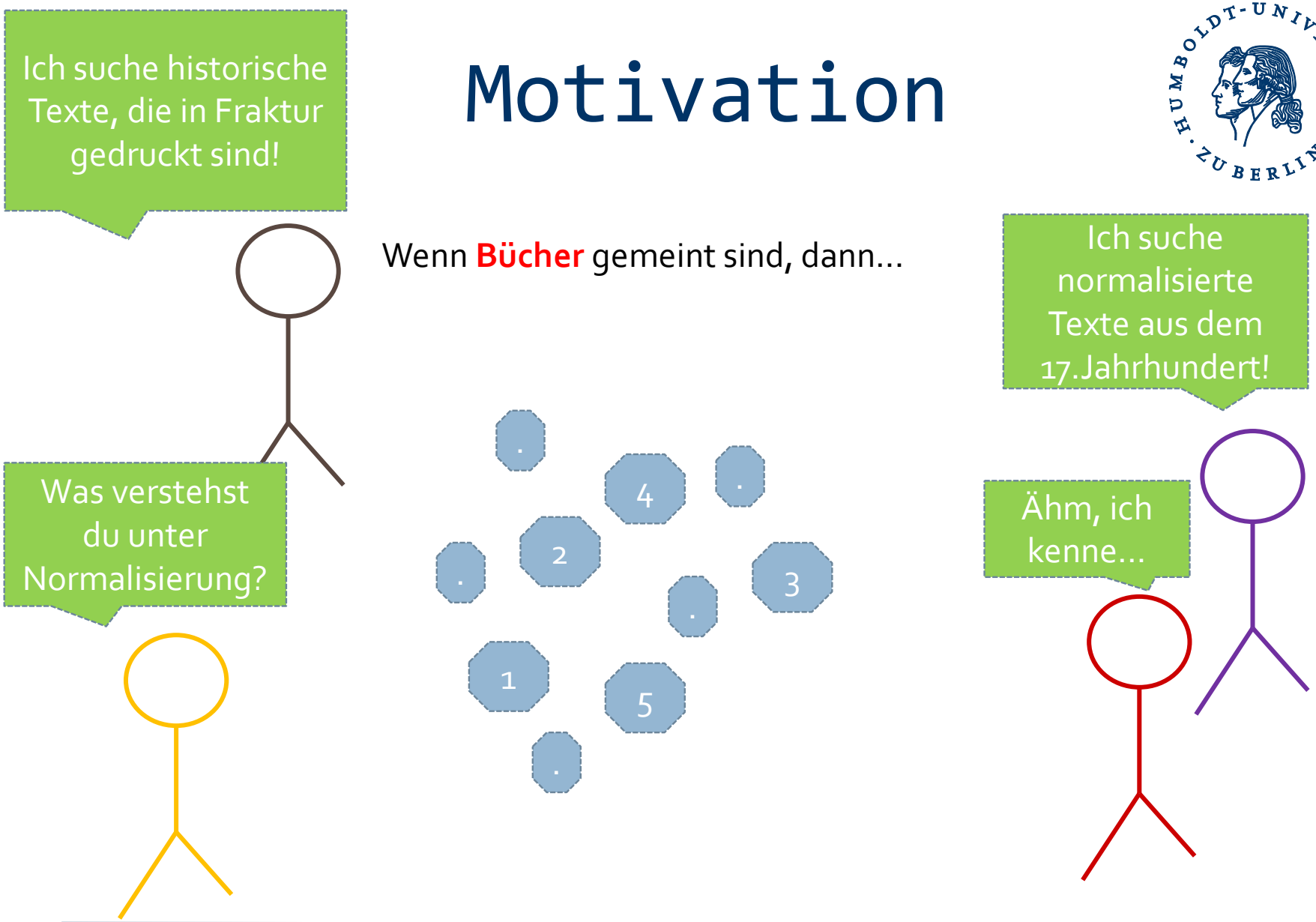
TECHNISCH-ABSTRAKTE METADATEN FÜR DIE SUCHE NACH HISTORISCHEN KORPORA

Motivation

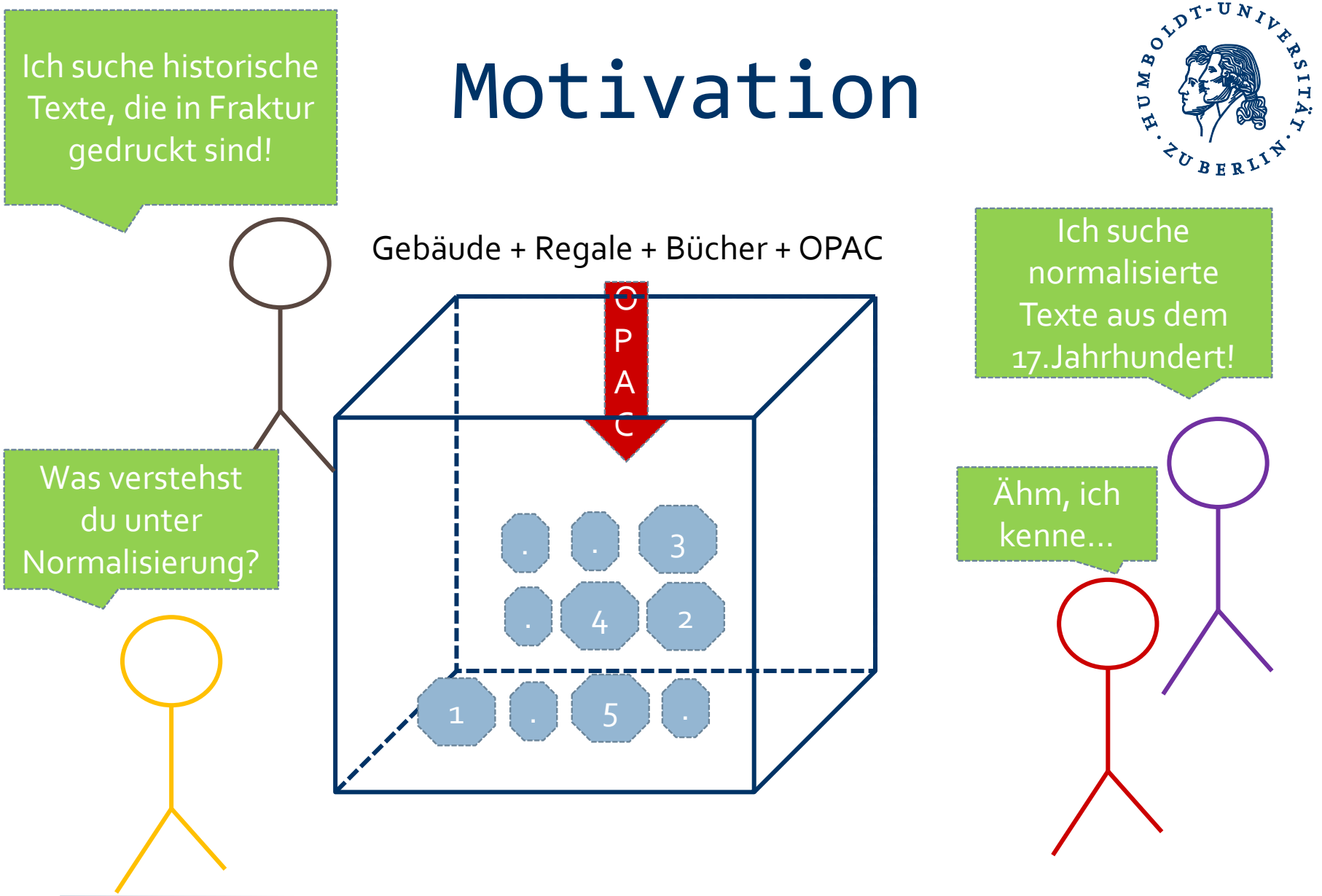

Ich suche historische Texte, die in Fraktur gedruckt sind!
Ich suche normalisierte Texte aus dem 17.Jahrhundert!
Was verstehst du unter Normalisierung?
Ähm, ich kenne…
4
2
3
1
5

Motivation

Wenn **Bücher** gemeint sind, dann…

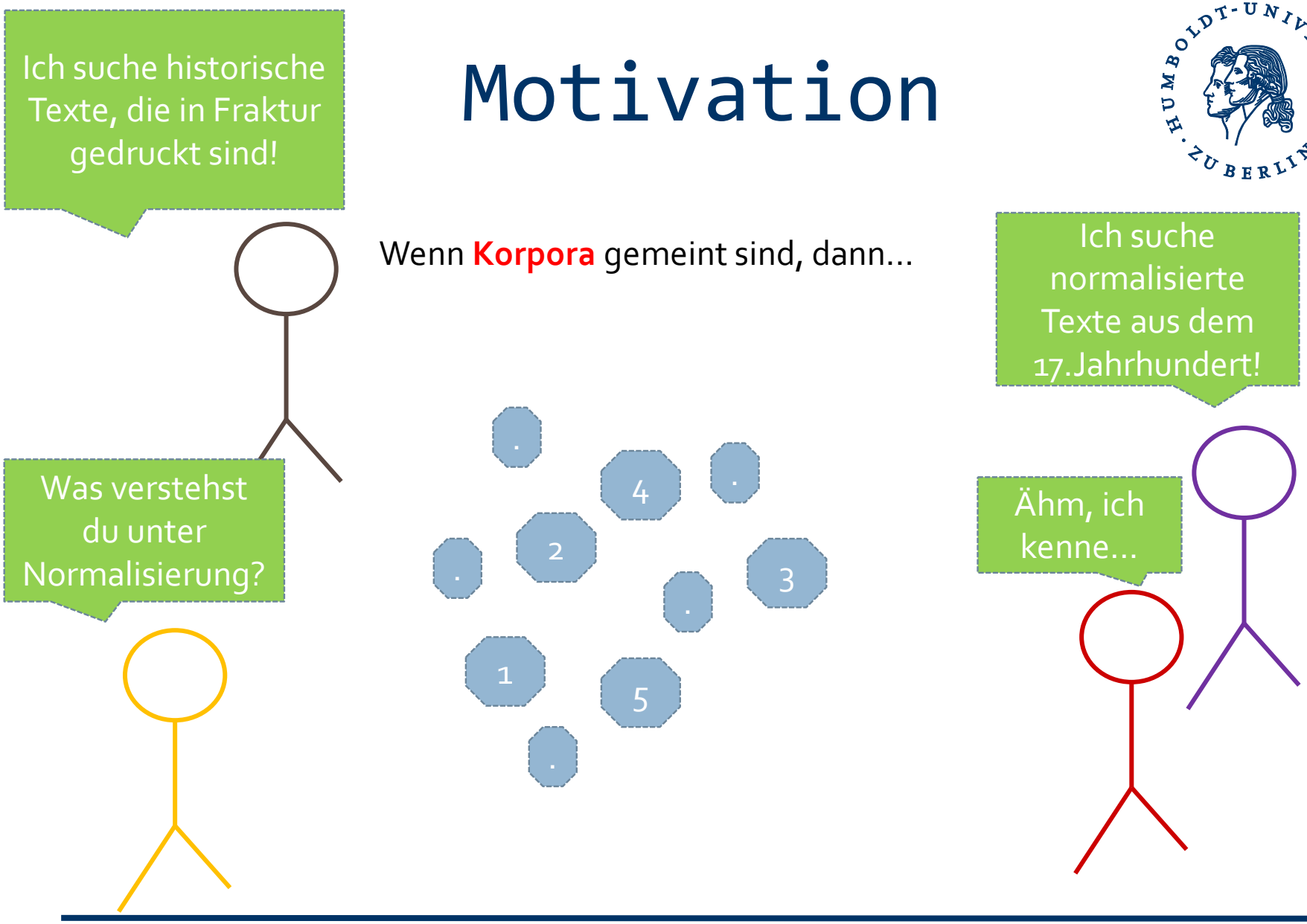


Motivation

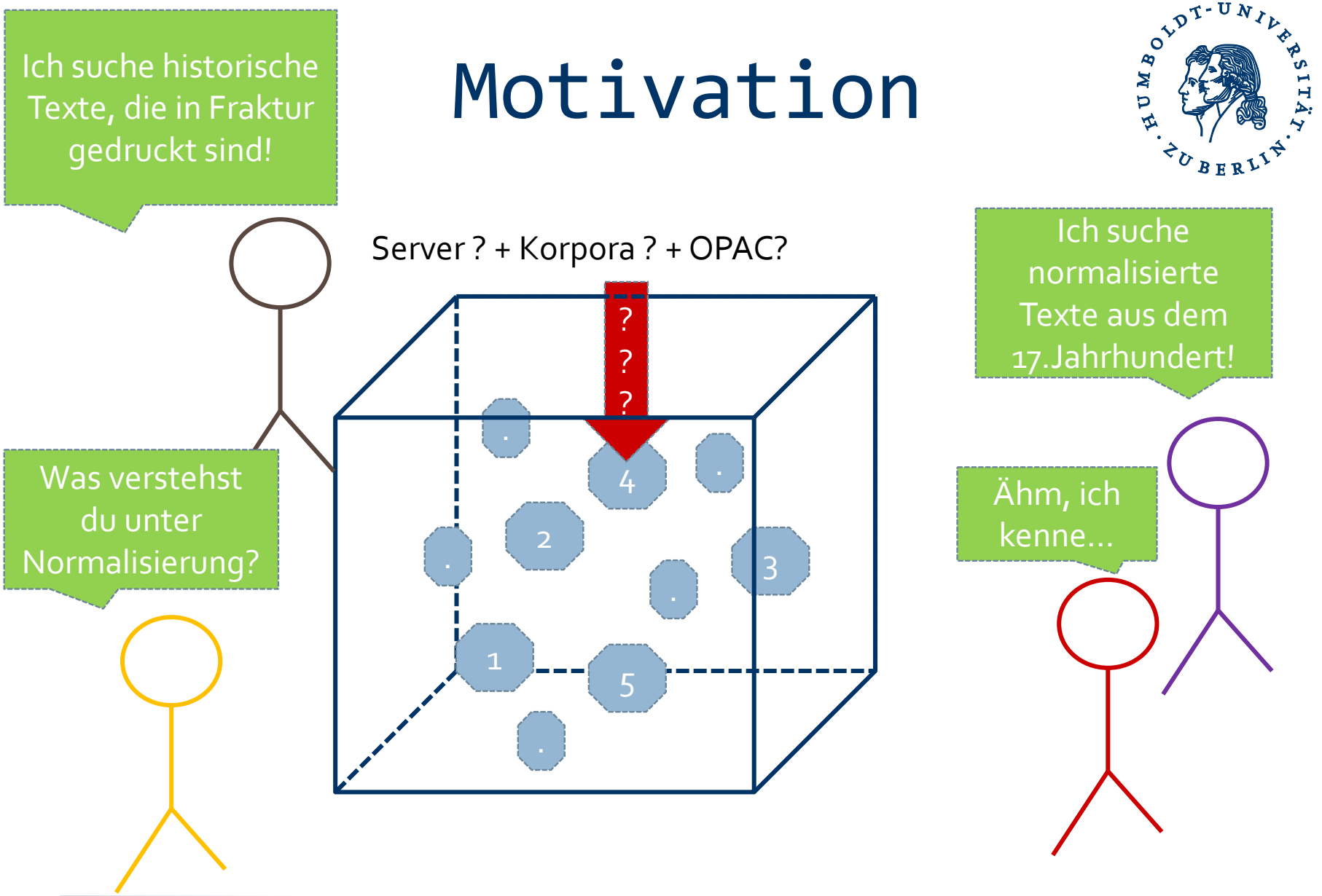

Ich suche historische Texte, die in Fraktur gedruckt sind!
Gebäude + Regale + Bücher + OPAC
OPAC
Ich suche normalisierte Texte aus dem 17.Jahrhundert!
Was verstehst du unter Normalisierung?
Ähm, ich kenne…

Motivation

Wenn **Korpora** gemeint sind, dann…



Motivation


Ich suche historische Texte, die in Fraktur gedruckt sind!
Server ? + Korpora ? + OPAC?
? ? ?
4
2
3
1
5
Ich suche normalisierte Texte aus dem 17.Jahrhundert!
Was verstehst du unter Normalisierung?
Ähm, ich kenne…

Beispiel Schönberg-Verein

Dateiexterne Informationen

text	Text

Wien, den 17. Jänner 1921.
Sehr geehrter <persName role="IAV" type="Präsident" ref="Personenliste.xml#P00001">Herr Schönberg</persName>!

<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.

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

Beispiel KAJUK

Dateiexterne Informationen

tok	Text

- Beispiel: ‚tok' enthält primäre, virtuelle und normalisierte Texteinheiten
- Annotation kennzeichnen Texteigenschaften

norm						dass										
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 Àgel /Mathilde Hennig (Justus-Liebig-Universität Gießen), Bauernleben 17.Jh
http://hdl.handle.net/11022/0000-0000-2102-8

Beispiel RIDGES

Dateiexterne Informationen

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, Anke Lüdeling, Carolin Odebrecht, Amir Zeldes (Humboldt-Universität zu Berlin Berlin) PflantzGart_1639, http://hdl.handle.net/11022/0000-0000-2D85-8

Herausforderung der Erschließung

- Korpora direkt über deren Daten erschließen
  - Tools
    - wie z.B. Oxygen (http://www.oxygenxml.com), ELAN (Slotjes & Wittenburg 2008), EXMARaLDA (Schmidt et al. 2011)
  - Analyse-Software
    - wie z.B. ANNIS (Krause & Zeldes 2014) , CQP (Hardie 2012)
  - Tagsets abgebildet in dem jeweiligen Format
    - Übersicht zu linguistischen Formaten (vgl. Zipser 2014)

- Projekt-Homepage
  - nicht immer aussagekräftig
  - eigenständig aufgebaut
  - konzentriert auf das Forschungsziel

Erschließung von Interpretationen

- notwendige Diversität in der Forschung
  - theoretische Kategorien
  - unterschiedliche Kategorisierungen nach Feinkörnigkeit, Ausprägung, Semantik

  - ➢ keine einheitlichen Annotationen und Formate, kaum exhaustive Auflistung möglich (wenige/keine „Standards")
  - ➢ keine genaue Vorhersage über deren Kombinationen
  - ➢ keine genaue Vorhersage von Innovationen möglich

  - ➢ **mehr Komplexität durch überfachlichen Austausch von Korpora**

Forschungsfrage

- 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 **wiederzuverwenden?**

➢ wichtig dabei: Funktionen von Metadaten
  - Erschließung
  - einheitliche, gleichzeitige Suche
  - Referenzierung

- Funktion von Metadaten vgl. NISO (2004), Haynes (2004), Miller (2011), Odebrecht & Krause (2013)

Forschungsfrage Kontext

- **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 Korpora vgl. Lemnitzer & Zinsmeister (2006), Lüdeling (2011), Data Life Circle vgl. Rümpel (2011)

- ➢ **Wissen über Korpora dazu erforderlich**
  - Textgrundlagen/Textvorlagen
  - Korpusarchitekturen, Formate
  - Annotationsschemata, Annotationstools, Annotatoren
  - Prüfverfahren, Versionen, Konvertierungen
  - …

Forschungsfrage Eingrenzung

- 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 **wiederzuverwenden?**

- **unter zwei festen Bezügen:**
  - Objekt →historische textbasierte Korpora
  - Zweck →Erschließung von Korpora
  - ➢ Wiederverwendung/Referenzierung/Dokumentation

➢ **Erschließung von Korpora über deren Metadaten**
  - nicht(!) über deren Erstellung (Annotationstool)
  - nicht(!) über deren Analyse (Analysetool)

Verwandte Ansätze

- Objekte in Repositorien/Datenbanken
  a) Textgrid, DARIAH
     - engerer Objektbezug: TEI XML und Bilder
  b) Deutsches Textarchiv, CLARIN
     - engerer Objektbezug: TEI XML
  c) Collection Registry, DARIAH
     - engerer Objektbezug: Sammlungsbeschreibung
  d) Freebank
     - engerer Objektbezug: TEI XML, französisch

  e) Virtual Language Observatory, CLARIN
     - weiterer Objektbezug: Korpora, Metadatensätze, Tools, Tagsets
  f) The Language Archive, CLARIN
     - weiterer Objektbezug: Korpora, Metadaten, Tools
  g) Hamburger Zentrum für Sprachkorpora
     - weiterer/engerer Objektbezug: Tools Korpora (speziell gesprochene Sprache)

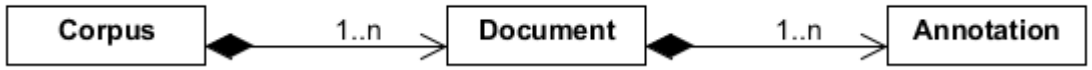
Zugänge zu Repositorien, siehe Anhang

Forschungsfrage Relevanz?

- „[…]Trotzdem ist zu bedenken, dass die Auswahl und Aufbereitung von Daten in den Geisteswissenschaften häufig so sehr an eine bestimmte Fragestellung und eine bestimmte theoretische sowie methodische Perspektive gebunden sind, **dass eine unmittelbare Übernahme fremder Grunddaten nicht ohne Weiteres möglich sein wird**. Deshalb ist auch die genaue Dokumentation dieser Daten von hoher Bedeutung. (Sahle & Kronenwett 2013, 82)."

Forschungsansatz

Metamodell (verkürzte Darstellung)



Modellierung ‚Buch' und dessen (relevante) Eigenschaften

Metadatenmodell (verkürzte Darstellung)

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>Title</title>
    </titleStmt>
  </fileDesc>
</teiHeader>
```

für jedes Korpus

Burnard & Rahtz (2004)

Suchgrundlage für den OPAC einer Bibliothek

Korpusmetadaten

Ridges
Register in Diachronic German Science

KAJUK

u.v.m.

Metadaten zu den Büchern einer Bibliothek

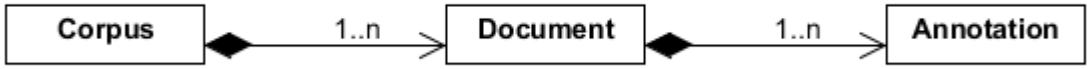
Forschungsansatz

- Interpretationen sind als Grundlage für die Erschließung eines Korpus ungeeignet.

- Datenerschließung als eine Voraussetzung für die Wiederverwendung!

- Gemeinsamkeiten von Korpora auf einer technisch-abstrakten Ebene identifizieren und als Grundlage für die Erschließung verwenden!
  - Identifikation und Klassifikation in einem Modell (Objektbezug)
  - einheitliche Beschreibung verschiedener Korpora
  - überfachlicher Zugang für Dokumentation, Referenzierung, Suche und Wiederverwendung (Zweckbezug)

Forschungsansatz



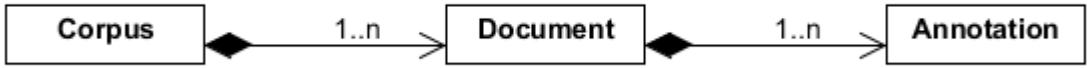
KAJUK

Satzglied				praed			subj	KOR		
Realisation_gram							Pron			
E						E				
dir						V				
tok	in	hesischer	Conterbution	waren	und	das	wir	also	das	Mal

Aufgabe

- Gesucht werden Texte aus dem 17.Jahrhundert mit Normalisierung, dazu Wissen über
  - Eigenschaften von Texten (*Document*)
    - Datum des Verfassens, des Veröffentlichens
  - Eigenschaften des „Text" (*Annotation*)
    - konkrete Werte der Annotation (Strings)
      - Wortformen, Buchstaben, Morpheme etc.
    - andere Annotationen basieren darauf
      - (alle) anderen Annotationen sind auf dieser Ebene (un-)mittelbar annotiert
  - Eigenschaften des Text-Korpus (*Corpus*)
    - Texte werden in einer Sammlungen veröffentlicht
    - Tagsets

Metamodell



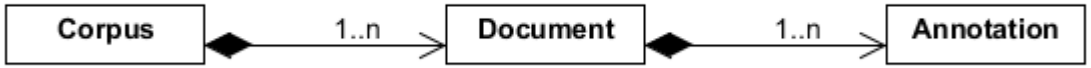
verkürzte Darstellung

- Ebenenname A - Werte (konkrete Werte der Ebene A:Wortformen, Buchstaben, Morpheme etc.)	- Datum des Verfassens, des Veröffentlichens - mindestens 1 Dokument im Korpus besitzt Ebene A	- Ebene A - Segmentierung - Abhängigkeiten von anderen Ebenen (B-Z)

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.

Aufgabe Text



- **„Text"** als Publikation beschreiben (Document)
  - u.a. Datum einer Veröffentlichung
- „**Text**" als Annotation beschreiben (Annotation)
  - abstrakt-technische Eigenschaften wie Tokenisierung, Relationen zwischen Annotationen

Metamodell ermöglicht Suche

- **Gesucht werden Texte mit Normalisierungen aus dem 17.Jh**
- **Kriterien für die Suche:**

- alle unabhängigen, eigenständig segmentierten Annotationen mit freien Annotationswerten
  - Annotationsrichtlinien dienen dem konkreten Verständnis
- gewünschte Eigenschaften der Normalisierung
  - durchgängig oder nicht annotiert
  - nach bestimmten Regeln (Tagset)
  - in bestimmter Form (Token-, Spannenannotation)
- Veröffentlichungsdatum des Dokumentes
  - Niederschrift, Handschrift, Edition

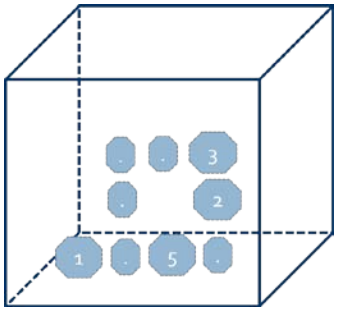
Forschungsansatz Vorteile

- flexibel auf neue Architekturen reagieren
  - ungesehene Korpora auch beschreibbar
- auf Modellebene keine stark interpretatorischen, fachspezifischen Konzepte
  - fördert den überfachlichen Austausch
- keine Auflösung konfligierender Annotationen oder Konzepte
  - keine Suche auf Konzepten, Konzepte sind die Treffer

- nur für historische Textkorpora
- letzter Schritt: in die Daten gucken
  - Einschränken der Menge, die angeschaut werden muss

Anwendung

- Metadatenmodell als Anzeige-und Suchgrundlage eingebunden in
  - Repositorien
    - zentraler Zugang zu Korpora
    - Dokumentation, Referenzierung von Korpora
  - Tools
    - Verarbeitungsmetadaten beim Erstellen



Forschungsdaten

- **Modellierung unter zwei festen Bezügen:**
  - Objekt →historische textbasierte Korpora
  - Zweck →Erschließung von Korpora über deren Metadaten

- **für**
  - überfachliche Identifikation von Forschungsdaten
  - überfachliche, einheitliche Beschreibung
  - Referenzierung und Dokumentation
  - einheitliche Suche und Erschließung
  - ➢ Wiederverwendung

- **nur möglich durch**
  - überfachliche Zusammenarbeit und Kooperationswille zum Austausch von Forschungsdaten

Weitere Themen

- Be- und Verarbeitung, Versionierung
  - prozedurale Metadaten
  - resultative Metadaten
- Referenzierung
  - bibliographische Metadaten: Datumsangaben (Corpus, Document, Annotation)
  - administrative Metadaten: Personen/Autoren/Annotatoren
- Forschung
  - Forschungspublikationen
  - Kontexte der Forschung

u.v.m.

Vielen Dank

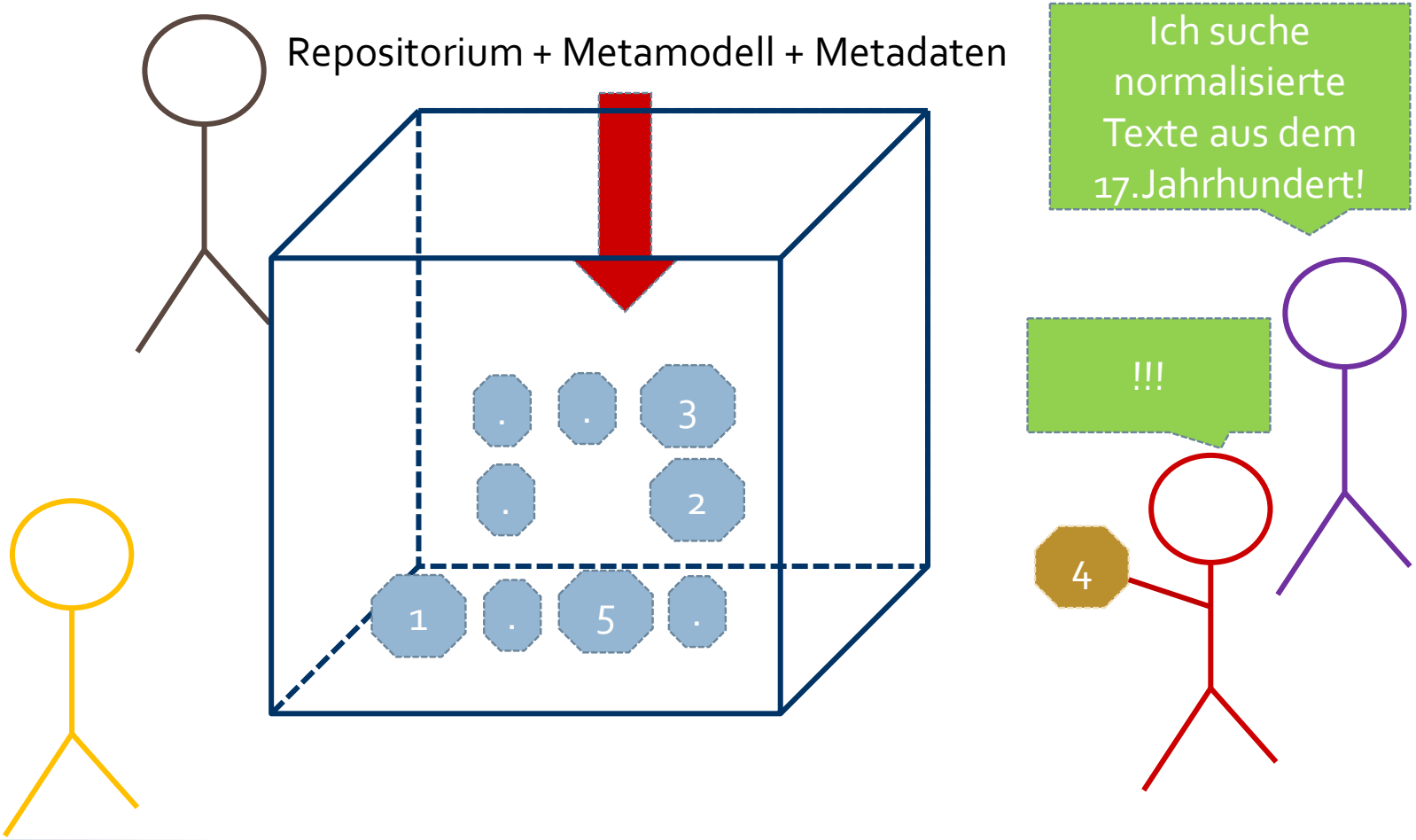
für die vielfältige Unterstützung an

Anke Lüdeling und Laurent Romary,

die korpling-Gruppe

und die Phonetiker!

Vielen Dank für Ihre Aufmerksamkeit



Referenzen

- **Burnard**, L., Rahtz, S. (2004). RelaxNG with Son of ODD. Extreme Markup Languages Proceedings 2004. Montréal, Québec.
- **Hardie** A. (2012). CQPweb - Combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics 2012;17(3):380-409.
- **Haynes**, D. (2004). Metadata for information management and retrieval. London: Facet publishing.
- **Krause**, T., Zeldes, A. (2014): ANNIS3: A new architecture for generic corpus query and visualization. in: Digital Scholarship in the Humanities 2014 http://dsh.oxfordjournals.org/cgi/content/abstract/fqu057?ijkey=GJBroLhNfKW1g8i&keytype=ref
- **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. (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**, S. 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., 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.

Referenzen

- **Sahle**, P., Kronenwett, S. (2013). Jenseits der Daten. Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner Data Center For The Humanities. In: Library Ideas (23), S. 76–96.
- **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>
- **Sloetjes**, H., Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In: Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).
- **Schmid**, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of International Conference on New Methods in Language Processing, 1994. Manchester.
- **Schmidt**, T., Wörner, K., Hedeland, H., Lehmberg T. (2011). New and future developments in EXMARaLDA In Thomas Schmidt, Kai Wörner, eds., Multilingual Resources and Multilingual Applications. Proceedings of GSCL Conference 2011 Hamburg.
- **Zipser**, F. (2014). SaltNPepper und das Formatpluriversum. LAUDATIO Workshop 2014. Berlin, 07.- 08.10.2014. http://dx.doi.org/10.5281/zenodo.17557

Ressourcen

- 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
- Verein Schönberg

  http://hdl.handle.net/11022/0000-0000-84DD-2
- TEI ODDs – für Metadatenschemata

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

Ressourcen

a) Textgrid https://textgridrep.de
b) Deutsches Textarchiv http://www.deutschestextarchiv.de/
c) Virtual Language Observatory https://vlo.clarin.eu/;jsessionid=CE89FC7E28DEF3932A2F14BC696F49DC?o
d) The Language Archive https://tla.mpi.nl/
e) Collection Registry http://colreg.de.dariah.eu/colreg/colreg/main?execution=e1s1
f) Freebank nria-00100194
g) Hamburger Zentrum für Sprachkorpora https://corpora.uni-hamburg.de/drupal/de/islandora/object/islandora%3Aroot