Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

30.11.2019

Wie erstelle und annotiere ich ein Korpus?

Möglichkeiten der digitalen Textanalyse

Lea Röseler

Gliederung

1) Begriffsklärung – Korpus und Annotation
2) Ausgangspunkt – Fragestellung
3) Korpuserstellung mit Nexis Uni
4) Annotation mit CATMA
5) Auswertung und Ergebnisse
6) Weitere Korpora und Tools
7) Literatur

Begriffserklärung – Korpus

das Korpus – Pl. die **Korpora**

„Ein Korpus ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus bestehen aus den Daten selber sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind."

(Lemnitzer/Zinsmeister ^{3}2015: 13)

Begriffserklärung – Annotation

» „linguistische Anreicherung der Primärdaten“ (Lemnitzer/ Zinsmeister ^{3}2015: 196)

» Lemmatisierung

» Part of Speech (POS)

» Dependenzen

» Koreferenz

» …

Ausgangspunkt – Fragestellung

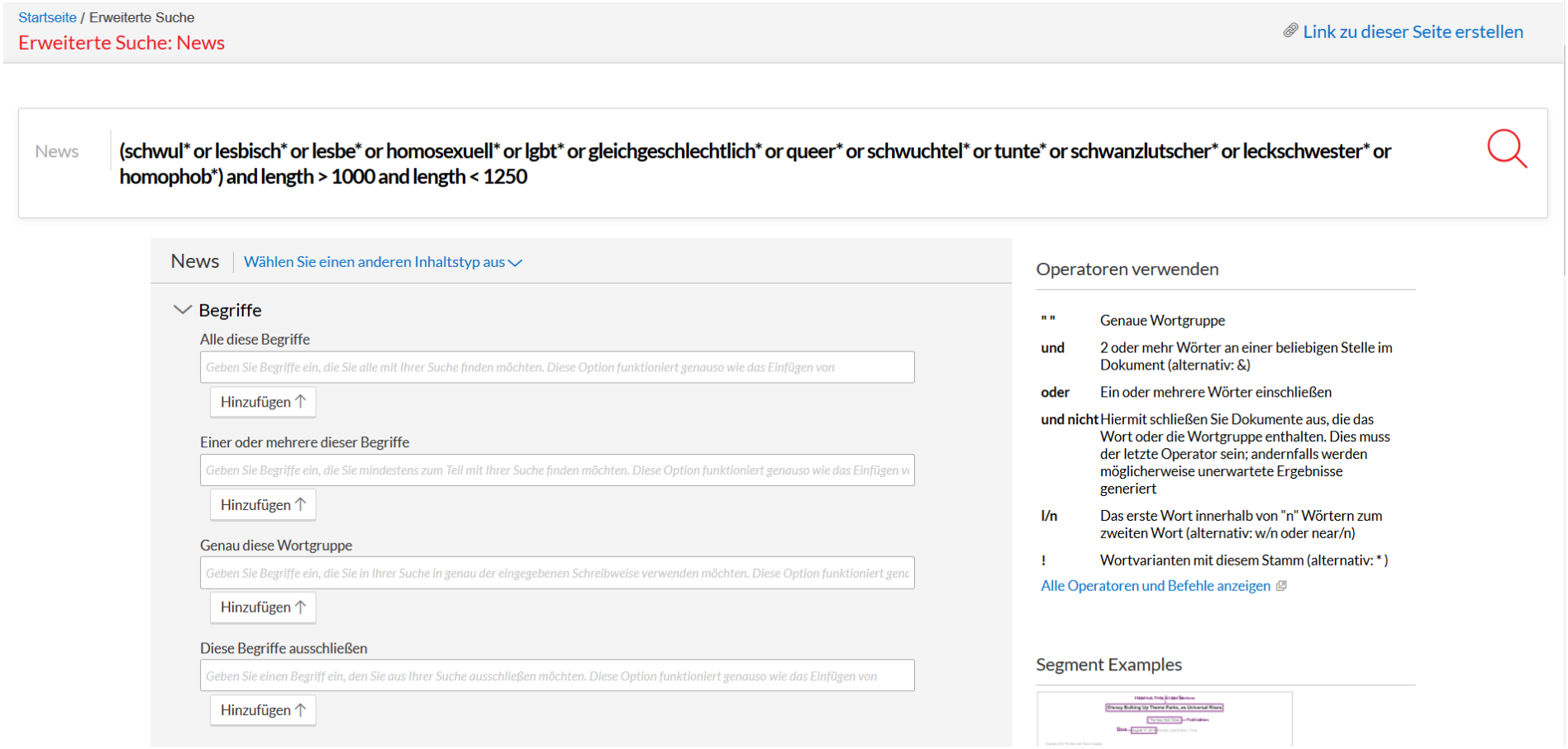
- globale Struktur von Zeitungstexten
  - These: Texte, die ein positives Ereignis als Ausgangspunkt der Berichterstattung haben, sind anders aufgebaut als Texte, die sich mit einem negativen Ereignis befassen.

- Verteilung von Begriffen des LGBTQ-Diskurses
  - These: Es lassen sich Unterschiede in der Verwendung von Begriffen des LGBTQ-Diskurses in den verschiedenen Textteilen feststellen.

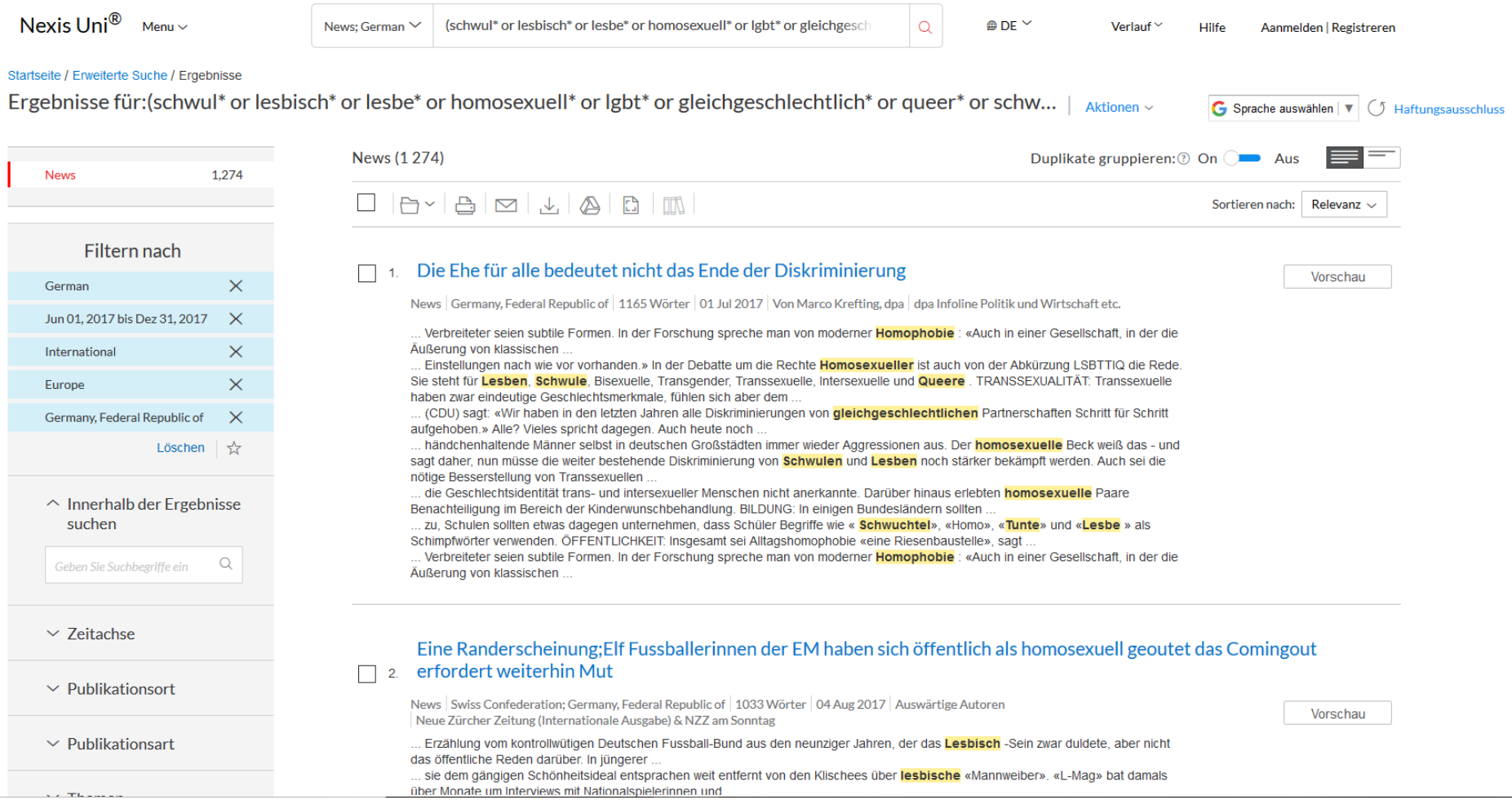
Nexis Uni

- keine öffentliche Datenbank
  - Zugang über Bibliotheken
- Volltexte von Zeitungsartikeln
- erweiterte Suche möglich
  - reguläre Ausdrücke
  - verschiedene Filter
    - Publikationsart
    - Sprache
    - Zeitraum
    - Textlänge
- Download der Ergebnisse in verschiedenen Formaten

Nexis Uni – Erweiterte Suche



Nexis Uni – Ergebnisdarstellung



Nexis Uni – Download der Ergebnisse

- Titelliste
  - bis zu 250 Treffer
  - Format: PDF, Word, Excel, Rich Text
- Volltexte
  - Texte müssen zunächst ausgewählt werden
  - bis zu 100 Treffer
  - Format: PDF, Word, Rich Text

Erstellung des LGBTQ-Korpus

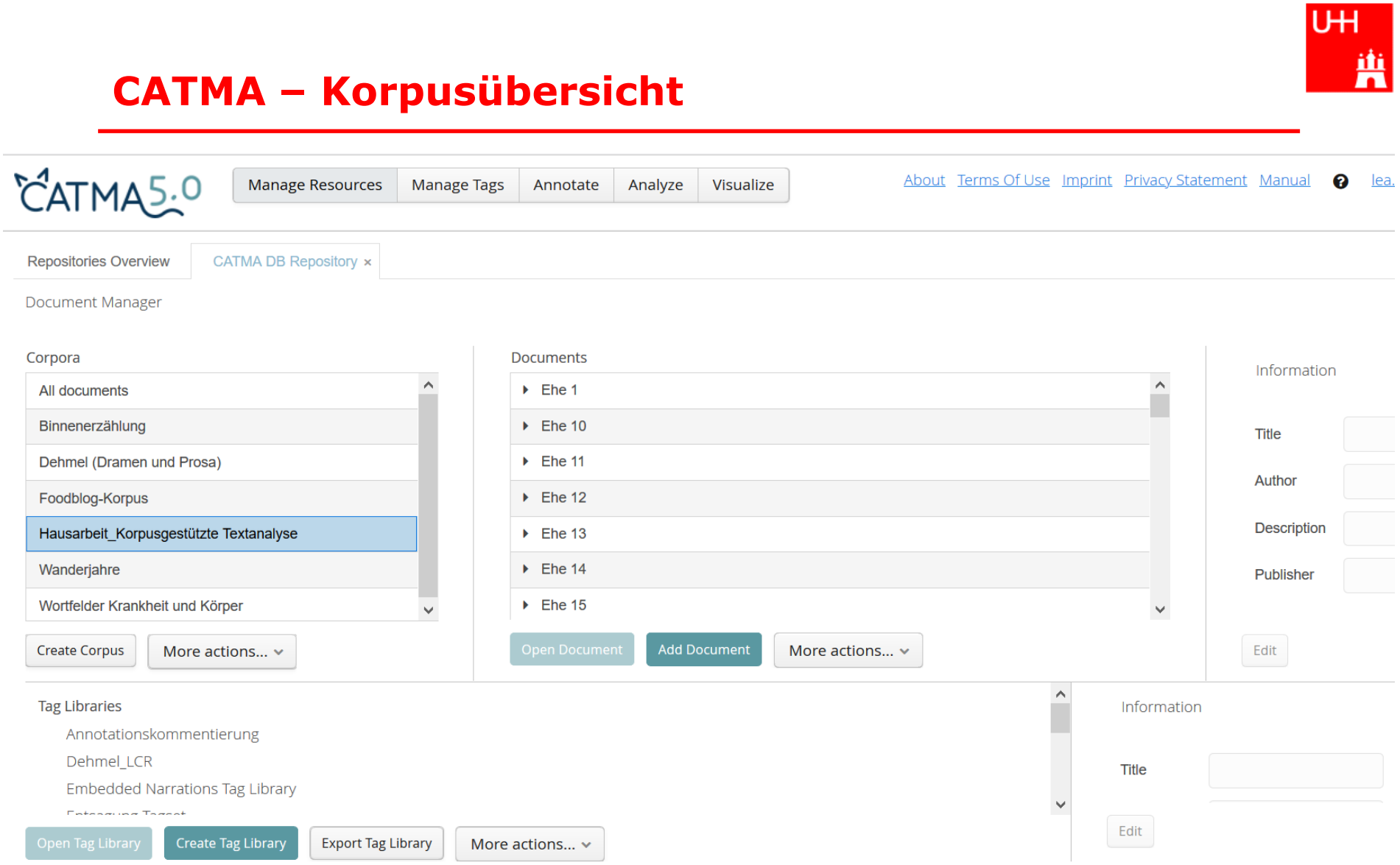
- zwei verschiedene Suchanfragen
  - zwei Zeiträume
    - 12.06.2016–31.12.2016
    - 01.06.2017–31.12.2017
  - mit „orlando“ bzw. „ehe“ spezifiziert
- insgesamt 273 Treffer
- Download der Volltexte als Word-Datei
  - Umwandlung in txt-Format
- zufällige Stichprobe von 30 Texten annotiert

CATMA

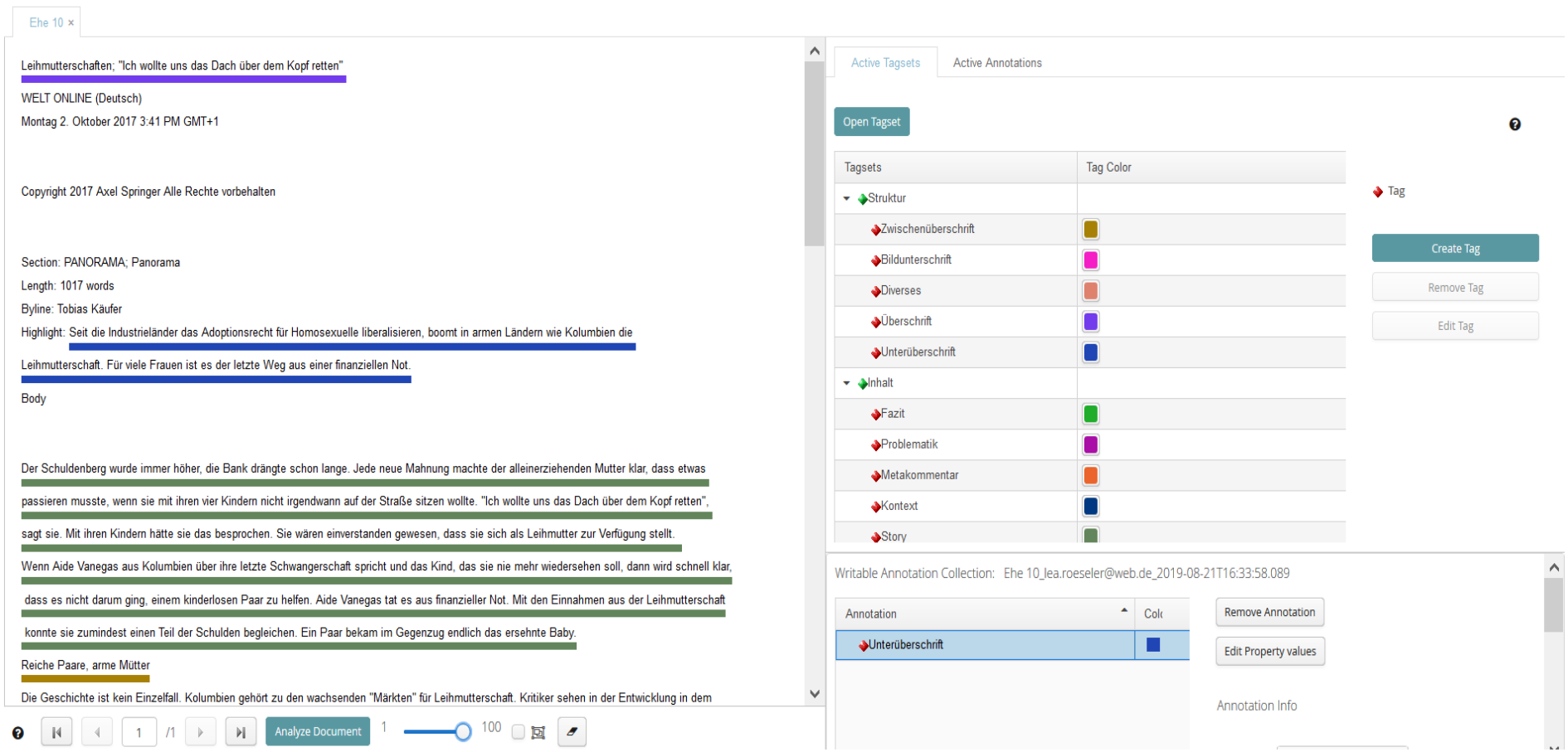
- entwickelt an der Uni Hamburg
- literaturwissenschaftliches Annotationstool
  - längere Textstellen
  - Erstellen eigener Tags und Tagsets
  - kollaboratives Arbeiten möglich
    - Teilen von Dokumenten und Tagsets
  - Analyse der Annotationen
  - Export der Annotationen

- https://catma.de/

CATMA – Korpusübersicht



CATMA – Annotation



CATMA - Tagset

Tagsets	Tag Color
Struktur	
Zwischenüberschrift	
Bildunterschrift	
Diverses	
Überschrift	
Unterüberschrift	
Inhalt	
Fazit	
Problematik	
Metakommentar	
Kontext	
Story	

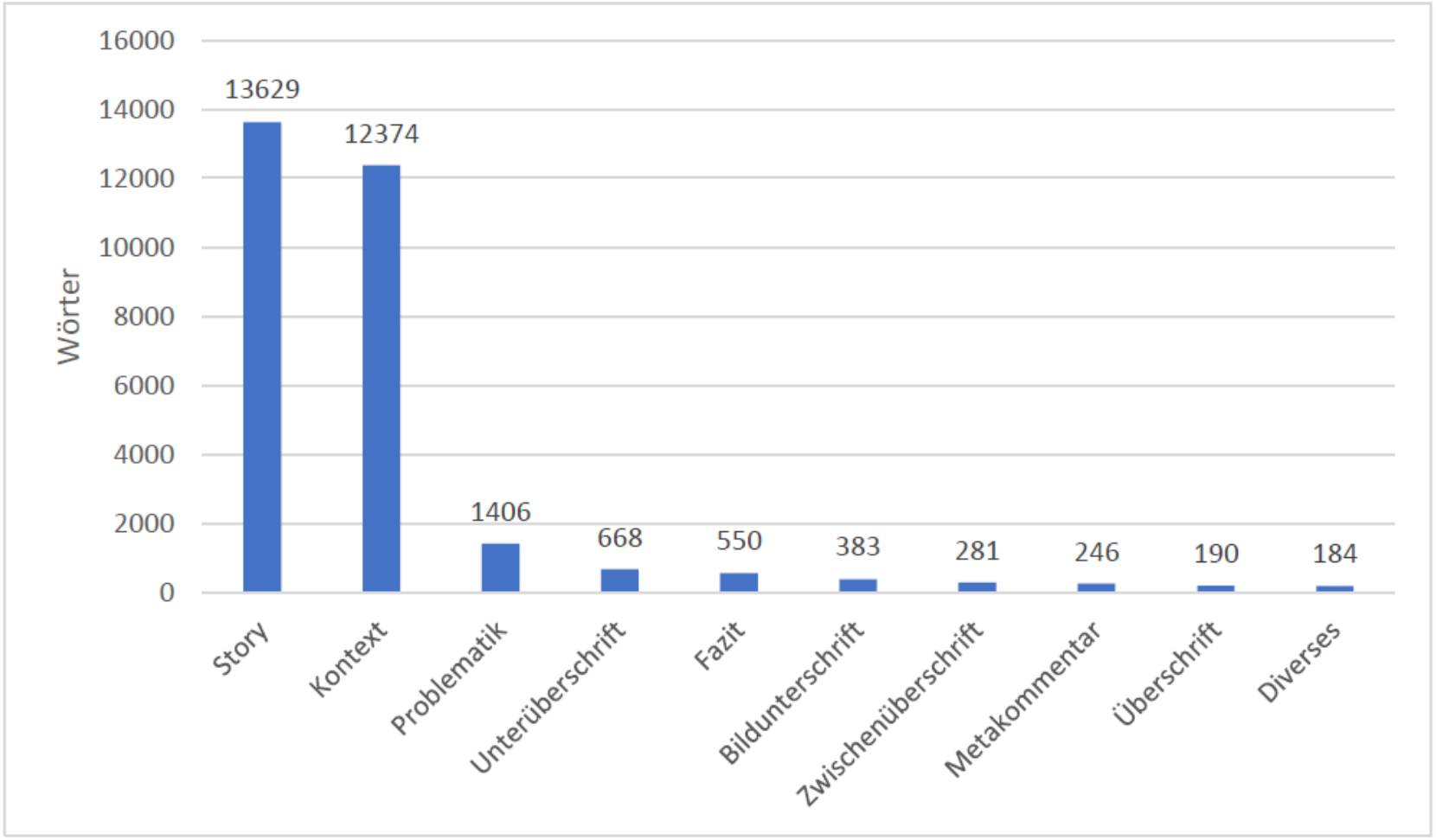
CATMA – Auswertung

- Korpus auswählen
  - Analyze Corpus
- Query: *tag*=“%”
- Download als csv-Datei
- Einlesen in Excel

CATMA – Export der Annotationen

	A	B	D	G	H
1	Tag	Text-ID	Textteil	Column7	Column8
2	/Story	Ehe 10	Aide Vanega ist vorerst froh, dass sie ihre Geldprobleme ein wenig lösen konnte. Was der Abschied von einem Kind, das sie ausgetragen hat und vermutlich nie wiedersehen wird, mit ihr macht, darüber lässt sich nur spekulieren.	1	FALSCH
3	/Story	Ehe 10	"Als der Moment der Geburt gekommen war, haben sie mir das Kind übergeben", berichtet Vanegas. "Als ich die Klinik verlassen habe, habe ich das Baby zum letzten Mal gesehen. Ich habe ‚Adios' gesagt, und das war es dann."	1	FALSCH
4	/Story	Ehe 10	Der Schuldenberg wurde immer höher, die Bank drängte schon lange. Jede neue Mahnung machte der alleinerziehenden Mutter klar, dass etwas passieren musste, wenn sie mit ihren vier Kindern nicht irgendwann auf der Straße sitzen wollte. "Ich wollte uns das Dach über dem Kopf retten", sagt sie. Mit ihren Kindern hätte sie das besprochen. Sie wären einverstanden gewesen, dass sie sich als Leihmutter zur Verfügung stellt. Wenn Aide Vanegas aus Kolumbien über ihre letzte Schwangerschaft spricht und das Kind, das sie nie mehr wiedersehen soll, dann wird schnell klar, dass es nicht darum ging, einem kinderlosen Paar zu helfen. Aide Vanegas tat es aus finanzieller Not. Mit den Einnahmen aus der Leihmutterschaft konnte sie zumindest einen Teil der Schulden begleichen. Ein Paar bekam im Gegenzug endlich das ersehnte Baby.	1	FALSCH
			Von solchen Beträgen kann Aide Vanegas nur träumen. Sie tut es trotzdem. Für den Kinderwunsch der gleichgeschlechtlichen Paare hat die Leihmutter Verständnis: "Sie haben		

Ergebnisse – Umfang der Textteile



Ergebnisse

- Leadtext und Fazit nicht obligatorisch
- Wechsel von Story- und Kontext-Teilen
  - Schwerpunkt variiert
  - Story-Teil nicht obligatorisch
- positiv: endet auf Story-Teil
- negativ: endet auf Kontext-Teil
- LGBTQ-Begriffe vermehrt in „Kontext“
- verhältnismäßig hohes Vorkommen in Leadtexten

Weitere Korpora

- Deutsches Referenz Korpus (DeReKo)
  - https://www1.ids-mannheim.de/kl/projekte/korpora/
  - geschriebene Gegenwartssprache
    - Zeitungstexte, Belletristik, (populär-)wissenschaftliche Texte etc.
  - 45,9 Milliarden Wörter
  - Suche in Subkorpora möglich
  - Anmeldung erforderlich
  - Cosmas II
  - Zugang: https://cosmas2.ids-mannheim.de/cosmas2-web/

Weitere Korpora

- Digitales Wörterbuch der Deutschen Sprache (DWDS)
  - 14 Milliarden Wörter
  - historische und gegenwartssprachliche Korpora
    - Zeitungstexte, Blogs, gesprochene Sprache etc.
  - zum Teil nur eingeschränkt durchsuchbar
  - Suche mit regulären Ausdrücken möglich
  - POS-annotiert
  - https://www.dwds.de/

Weitere Korpora

- Datenbank für Gesprochenes Deutsch (DGD)
  - 11 Millionen Wörter in 31 Korpora
  - gesprochene Sprache, größtenteils transkribiert
  - Lemma- und POS-annotiert
  - Anmeldung erforderlich
  - https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.welcome
- MoCoDa1 und MoCoDa2
  - SMS- und WhatsApp-Kommunikation
  - Anmeldung erforderlich
  - https://mocoda.spracheinteraktion.de/
  - https://db.mocoda2.de/#/c/home

Weitere Korpora

- Referenzkorpus Altdeutsch (ReA)
  - https://www.deutschdiachrondigital.de/
- Referenzkorpus Mittelhochdeutsch (ReM)
  - https://www.linguistics.rub.de/rem/access/search.html
- Referenzkorpus Mittelniederdeutsch/Niederrheinisch (ReN)
  - https://www.slm.uni-hamburg.de/ren/ueber-ren.html

- Lemma- und POS-annotiert
- durchsuchbar mit ANNIS

Weitere Annotationstools

- WebLicht
  - automatische Annotation eigener Daten
  - POS
  - Dependenzen
  - https://weblicht.sfs.uni-tuebingen.de/weblicht/
- WebAnno
  - eigene Annotation auf verschiedenen Ebenen
  - Korrektur bereits bestehender Annotationen
  - https://webanno.sfs.uni-tuebingen.de/

Literatur

» Andresen, Melanie/Zinsmeister, Heike (2019): *Korpuslinguistik*, Tübingen: Narr Francke Attempto.

» Lemnitzer, Lothar/Zinsmeister, Heike (^{3}2015): *Korpuslinguistik. Eine Einführung*, Tübingen: Narr Francke Attempto.