

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 ³2015: 13)

Begriffserklärung – Annotation

- » „linguistische Anreicherung der Primärdaten“ (Lemnitzer/
Zinsmeister ³2015: 196)

- » Lemmatisierung
- » Part of Speech (POS)
- » Abhängigkeiten
- » 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

[Startseite](#) / [Erweiterte Suche](#)

Erweiterte Suche: News

[Link zu dieser Seite erstellen](#)

News

(schwul* or lesbisch* or lesbe* or homosexuell* or lgbt* or gleichgeschlechtlich* or queer* or schwuchtel* or tunte* or schwanzlutscher* or leckschwester* or homophob*) and length > 1000 and length < 1250



News | Wählen Sie einen anderen Inhaltstyp aus ▾

▼ Begriffe

Alle diese Begriffe

Geben Sie Begriffe ein, die Sie alle mit Ihrer Suche finden möchten. Diese Option funktioniert genauso wie das Einfügen von

Hinzufügen ↑

Einer oder mehrere dieser Begriffe

Geben Sie Begriffe ein, die Sie mindestens zum Teil mit Ihrer Suche finden möchten. Diese Option funktioniert genauso wie das Einfügen von

Hinzufügen ↑

Genau diese Wortgruppe

Geben Sie Begriffe ein, die Sie in Ihrer Suche in genau der eingegebenen Schreibweise verwenden möchten. Diese Option funktioniert genau

Hinzufügen ↑

Diese Begriffe ausschließen

Geben Sie einen Begriff ein, den Sie aus Ihrer Suche ausschließen möchten. Diese Option funktioniert genauso wie das Einfügen von

Hinzufügen ↑

Operatoren verwenden

- " " Genau Wortgruppe
- und 2 oder mehr Wörter an einer beliebigen Stelle im Dokument (alternativ: &)
- oder Ein oder mehrere Wörter einschließen
- und nicht Hiermit schließen Sie Dokumente aus, die das Wort oder die Wortgruppe enthalten. Dies muss der letzte Operator sein; andernfalls werden möglicherweise unerwartete Ergebnisse generiert
- /n Das erste Wort innerhalb von "n" Wörtern zum zweiten Wort (alternativ: w/n oder near/n)
- ! Wortvarianten mit diesem Stamm (alternativ: *)

[Alle Operatoren und Befehle anzeigen](#) 

Segment Examples



Nexis Uni – Ergebnisdarstellung

Nexis Uni® Menu ▾

News; German ▾ (schwul* or lesbisch* or lesbe* or homosexuell* or lgbt* or gleichgesch) 🔍

DE ▾ Verlauf ▾ Hilfe Anmelden | Registrieren

[Startseite](#) / [Erweiterte Suche](#) / [Ergebnisse](#)

Ergebnisse für:(schwul* or lesbisch* or lesbe* or homosexuell* or lgbt* or gleichgeschlechtlich* or queer* or schw... | [Aktionen ▾](#) Sprache auswählen ▾ [Haftungsausschluss](#)

News	1,274
------	-------

Filtern nach

- German ✕
- Jun 01, 2017 bis Dez 31, 2017 ✕
- International ✕
- Europe ✕
- Germany, Federal Republic of ✕

[Löschen](#) ☆

Innerhalb der Ergebnisse suchen

 🔍

- Zeitachse
- Publikationsort
- Publikationsart

News (1 274)

Duplikate gruppieren: On Aus ☰ ☰



Sortieren nach: Relevanz ▾

1. [Die Ehe für alle bedeutet nicht das Ende der Diskriminierung](#) Vorschau

News | Germany, Federal Republic of | 1165 Wörter | 01 Jul 2017 | Von Marco Krefting, dpa | dpa Infoline Politik und Wirtschaft etc.

... Verbreiteter seien subtile Formen. In der Forschung spreche man von moderner **Homophobie**: «Auch in einer Gesellschaft, in der die Äußerung von klassischen ...
 ... Einstellungen nach wie vor vorhanden.» In der Debatte um die Rechte **Homosexueller** ist auch von der Abkürzung LSBTTIQ die Rede. Sie steht für **Lesben, Schwule**, Bisexuelle, Transgender, Transsexuelle, Intersexuelle und **Queere**. TRANSEXUALITÄT: Transsexuelle haben zwar eindeutige Geschlechtsmerkmale, fühlen sich aber dem ...
 ... (CDU) sagt: «Wir haben in den letzten Jahren alle Diskriminierungen von **gleichgeschlechtlichen** Partnerschaften Schritt für Schritt aufgehoben.» Alle? Vieles spricht dagegen. Auch heute noch ...
 ... händchenhaltende Männer selbst in deutschen Großstädten immer wieder Aggressionen aus. Der **homosexuelle** Beck weiß das - und sagt daher, nun müsse die weiter bestehende Diskriminierung von **Schwulen** und **Lesben** noch stärker bekämpft werden. Auch sei die nötige Besserstellung von Transsexuellen ...
 ... die Geschlechtsidentität trans- und intersexueller Menschen nicht anerkannte. Darüber hinaus erlebten **homosexuelle** Paare Benachteiligung im Bereich der Kinderwunschbehandlung. BILDUNG: In einigen Bundesländern sollten ...
 ... zu, Schulen sollten etwas dagegen unternehmen, dass Schüler Begriffe wie «**Schwuchtel**», «Homo», «**Tunte**» und «**Lesbe**» als Schimpfwörter verwenden. ÖFFENTLICHKEIT: Insgesamt sei Alltagshomophobie «eine Riesenbaustelle», sagt ...
 ... Verbreiteter seien subtile Formen. In der Forschung spreche man von moderner **Homophobie**: «Auch in einer Gesellschaft, in der die Äußerung von klassischen ...

2. [Eine Randerscheinung;Elf Fussballerinnen der EM haben sich öffentlich als homosexuell geoutet das Comingout erfordert weiterhin Mut](#) Vorschau

News | Swiss Confederation; Germany, Federal Republic of | 1033 Wörter | 04 Aug 2017 | Auswärtige Autoren | Neue Zürcher Zeitung (Internationale Ausgabe) & NZZ am Sonntag

... Erzählung vom kontrollwütigen Deutschen Fußball-Bund aus den neunziger Jahren, der das **Lesbisch** -Sein zwar duldete, aber nicht das öffentliche Reden darüber. In jüngerer ...
 ... sie dem gängigen Schönheitsideal entsprachen weit entfernt von den Klischees über **lesbische** «Mannweiber». «L-Mag» bat damals über Monate um Interviews mit Nationalspielerinnen und

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



[Manage Resources](#)
[Manage Tags](#)
[Annotate](#)
[Analyze](#)
[Visualize](#)

[About](#)
[Terms Of Use](#)
[Imprint](#)
[Privacy Statement](#)
[Manual](#)
[?](#)
[lea](#)

[Repositories Overview](#)
[CATMA DB Repository](#) ×

Document Manager

Corpora

- All documents
- Binnenerzählung
- Dehmel (Dramen und Prosa)
- Foodblog-Korpus
- Hausarbeit_Korpusgestützte Textanalyse
- Wanderjahre
- Wortfelder Krankheit und Körper

[Create Corpus](#)
[More actions...](#)

Documents

- ▶ Ehe 1
- ▶ Ehe 10
- ▶ Ehe 11
- ▶ Ehe 12
- ▶ Ehe 13
- ▶ Ehe 14
- ▶ Ehe 15

[Open Document](#)
[Add Document](#)
[More actions...](#)

Information

Title
 Author
 Description
 Publisher

[Edit](#)

Tag Libraries

- Annotationskommentierung
- Dehmel_LCR
- Embedded Narrations Tag Library
- Erzählung_Taxo

[Open Tag Library](#)
[Create Tag Library](#)
[Export Tag Library](#)
[More actions...](#)



Information

Title

[Edit](#)

CATMA – Annotation

Ehe 10 x

Leihmutterchaften; "Ich wollte uns das Dach über dem Kopf retten"

WELT ONLINE (Deutsch)

Montag 2. Oktober 2017 3:41 PM GMT+1

Copyright 2017 Axel Springer Alle Rechte vorbehalten

Section: PANORAMA; Panorama

Length: 1017 words

Byline: Tobias Käufer

Highlight: Seit die Industrieländer das Adoptionsrecht für Homosexuelle liberalisieren, boomt in armen Ländern wie Kolumbien die Leihmutterchaft. Für viele Frauen ist es der letzte Weg aus einer finanziellen Not.

Body

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 Leihmutterchaft konnte sie zumindest einen Teil der Schulden begleichen. Ein Paar bekam im Gegenzug endlich das ersehnte Baby.

Reiche Paare, arme Mütter

Die Geschichte ist kein Einzelfall. Kolumbien gehört zu den wachsenden "Märkten" für Leihmutterchaft. Kritiker sehen in der Entwicklung in dem

Active Tagsets
Active Annotations

Open Tagset
?

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

↕ Tag
Create Tag
Remove Tag
Edit Tag

Writable Annotation Collection: Ehe_10_lea.roeseler@web.de_2019-08-21T16:33:58.089






Annotation	Colr	
↕ Unterüberschrift		

Remove Annotation
Edit Property values

Annotation Info

?
⏪
⏩
1 / 1
Analyze Document
1
100
🔍
✍️

CATMA - Tagset

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

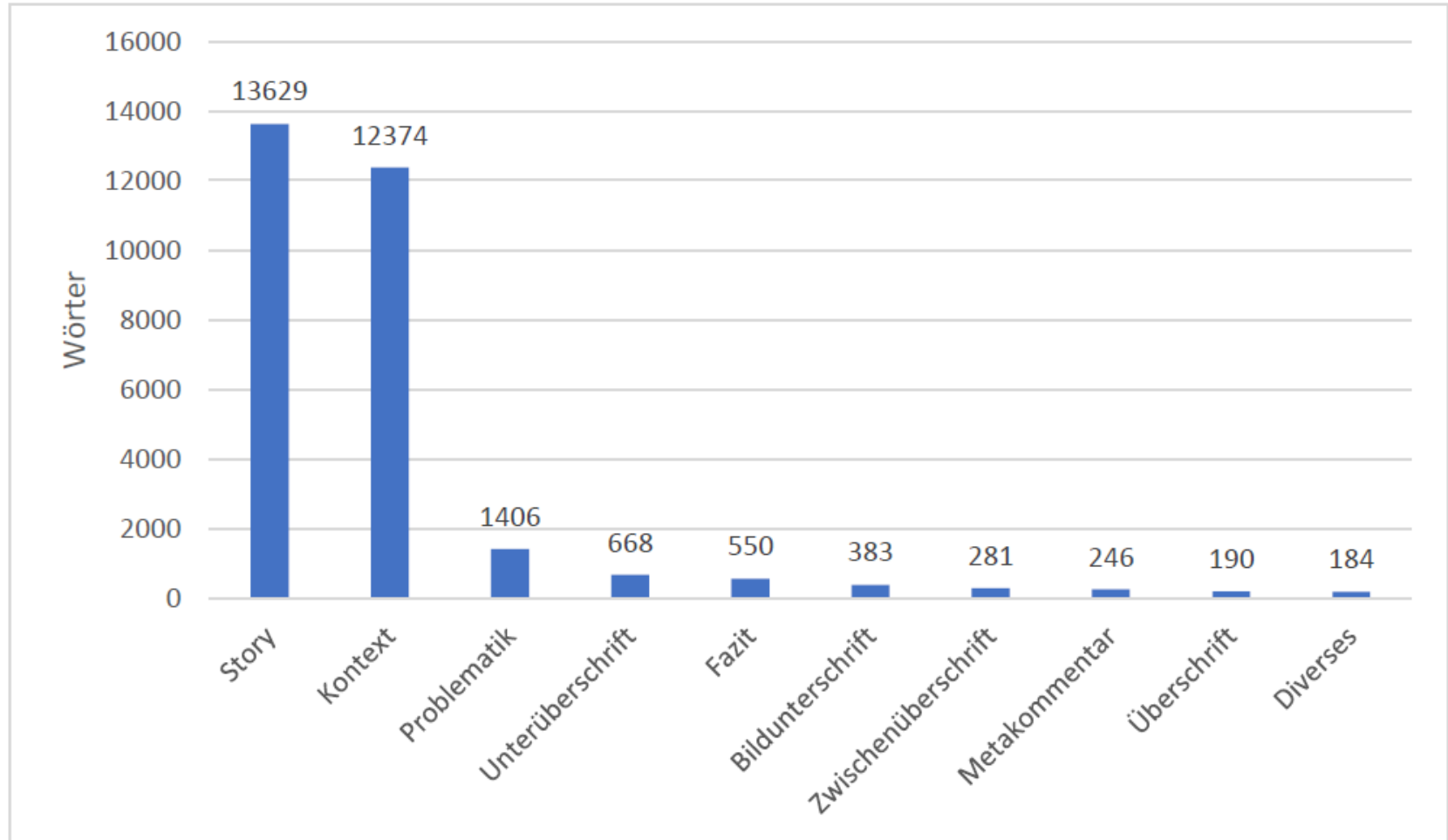
CATMA – Auswertung

- » 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
- ▶ Abhängigkeiten
- ▶ <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 (³2015): *Korpuslinguistik. Eine Einführung*, Tübingen: Narr Francke Attempto.