Linguistisch annotierte Korpora in der zeithistorischen Forschung: Methodik und Fallbeispiele

Thomas Werneke
ZZF Potsdam

Kay-Michael Würzner
BBAW

Anglistik-Fortbildung an der Staatsbibliothek zu Berlin

Teil 1: Methodik

- Textkorpora
- Annotation
  - Wort- und Satzgrenzenidentifikation
  - Grundformbildung
  - Wortartenbestimmung
  - Wortbeziehungen
  - Wortsemantik
- Werkzeuge
  - Suche
  - Wortverlaufskurven
  - Wortwolken
  - Wortprofile

Teil 2: Fallbeispiele

- Herkunft und Bedeutungswandel des Begriffes *Terrorismus*
- Begriffstransfer von West nach Ost am Beispiel *Bürgernähe*
- *Ost-* vs. *Westblock*: Verteilungsunterschiede von Eigen- und Fremdzuschreibung
- Rassismus in der Sprache? Bezeichnungswandel des Begriffes *Neger*

Textkorpora

Textkorpora

- Sammlungen von Texten
- *(Sprachwissenschaftliche) Referenzkorpora*
  - repräsentative Erfassung der Gesamtheit einer Sprache (bzw. eines Sprachstandes)
  - Englisch: *British National Corpus* *(Burnard, 1995)*
  - Deutsch:
    DWDS-Kernkorpus *(Geyken, 2007)*
    Deutsches Textarchiv *(Geyken und Klein, 2009)*
- *Spezialkorpora*
  - repräsentative Erfassung eines speziellen Ausschnitts einer Sprache
  - **medial**: Zeitungskorpora, Filmuntertitelkorpora, Internetkorpora
  - **inhaltlich**: *Berliner Wendekorpus*, *JuSpiL-Korpus* *(Dittmar, 2005)*
  - **forschungsbezogen**: *childLex* *(Schroeder et al., 2014)*

Textkorpora an der BBAW

Gegenwartssprachlich:

- DWDS-Kernkorpus, -Ergänzungskorpus
- Zeitungskorpora
  - Die ZEIT
  - Spiegel, Spiegel online
  - Bild & Welt (Auswahl)
  - Tagesspiegel, Potsdamer Neueste Nachrichten
- Filmuntertitelkorpus

Historisch:

- DTA-Kernkorpus
- DTA-Ergänzungen
  - Wikisource (Auswahl)
  - Gutenberg (Auswahl)
  - AEDit (Texte der frühen Neuzeit aus der HAB)
- Dinglers Polytechnisches Journal
- Leichenpredigten aus dem Bestand der ehemaligen Stadtbibliothek Breslau
- Neue Rheinische Zeitung

einheitliche Architektur für Annotation, Indizierung und Auswertung

DDR-Presseportal

- DFG-gefördertes Projekt an der Staatsbibliothek Berlin
- Digitalisierung und Volltexterschließung (OCR) von
  - Neues Deutschland (ND, 23. April 1946 - 3. Oktober 1990)
  - Berliner Zeitung (BZ, 21. Mai 1945 - 31. Dezember 1993)
  - Neue Zeit (22. Juli 1945 - 5. Juli 1994)
- verfügbar unter
  `http://zefys.staatsbibliothek-berlin.de/ddr-presse/`
- Recherche bisher auf einfache Volltextsuchen begrenzt
- ND und BZ: linguistische Annotation und Integration in CLARIN-D-Infrastruktur beantragt

Linguistische Annotation

Zum Begriff *Linguistische Annotation*

- Auszeichnung bestimmter linguistischer Eigenschaften
- Bezug zu einer *Einheit* des Textes:

  Wort Silbenstruktur, morphologische Zerlegung, lexikalische Semantik etc.

  Wortgruppe Mehrwortausdrücke, Namen, Kollokationen etc.

  Phrase syntaktische Kategorie, syntaktische Funktion

  Satz syntaktische Struktur, Satzsemantik, Funktion im Text (?)

- *manuelle* vs. *automatische* Annotation

Annotation von Korpora

Strukturierung und Anreicherung der Rohtexte zum Zwecke

- besserer Durchsuchbarkeit
- einfacherer Belegidentifikation
- moderner Korpuspräsentationsformen
- quantitativer Auswertungen

$\rightarrow$ Korpusumfang bedingt vollautomatische Analysekette

Annotation: Überblick

Standardanalysekette

- Wort- und Satzidentifikation (*Tokenisierung*)
- Wortartenauswahl und Grundformbildung (*morphologische Analyse*)
- Wortartenbestimmung im Kontext (*Part-of-Speech Tagging*)
- Wortbeziehungen (*Dependenz Parsing*)
- Wortbedeutung und semantische Beziehungen (*Thesaurus*, *Latent Semantic Indexing*)

$\rightarrow$ vollautomatisch mit akzeptabler Qualität

Annotation: Tokenisierung

- Unterteilung von Fließtext in **Wörter** (bzw. *Tokens*) und **Sätze**
- (Vor-)Klassifizierung der Tokens zur Beschleunigung der morphologischen Analyse
  - Abkürzungen
  - Zahlen
  - Sonderzeichen
  - Fremdalphabete
- Normalisierung der **Silbentrennung**
- Details: *Jurish und Würzner, 2013*

  → Multilingual approach; ships with an English model

Annotation: Tokenisierung – Beispiele

Problembereich *Satz*

Nach einer Schätzung des Industrieministeriums sind es mehr als 800.

„Österreich wurde alleingelassen in Europa“, beschwerte sich SPÖ-Zentralsekretär Josef Cap.

FR: Auf die Wahlerfolge der rechtsradikalen Parteien . . .

Problembereich *Token*

Kaiser's-Netz → Kaiser 's-Netz

mm. → mm. `[ORD]`

CDU/CSU → CDU / CSU

Jeanne d'Arc → Jeanne d' Arc

(Verwaltungs-)Personal → (Verwaltungs-) Personal

Annotation: Morphologie

- kontextfreie **Kategorisierung** von Wörtern bzgl. ihrer morpho-syntaktischen Merkmale (z.B. *Kasus*, *Numerus*, *Genus*)
- Abbildung einer Wortform auf eine Menge von Paaren aus **Grundform** und **Wortart**
- herausfordernd durch „unendliche“ Anzahl möglicher Wörter
  - umfassendes Lexikon (Wort, Kategorie, Flexionsklasse)
  - regelbasierte Umsetzung produktiver Wortbildungsprozesse
- Details: *Geyken und Hanneforth, 2006*

  → For English, use e.g. the **Porter stemmer** *(Porter, 1980)*

Annotation: Morphologie – Beispiele

[1]> apply "Wassern"

Wasser[NN SemClass=k_matstoff Gender=neut Number=pl Case=dat]

wasser/V~n[NN SemClass=none Gender=neut Number=sg Case=nom_acc_dat] <2>

wasser~n[VVINF] <2>

[1]> apply "Telekommunikation"

Telekommunikation[NN SemClass=abstr Gender=fem Number=sg Case=*]

tele/K+Kommunikation[NN SemClass=abstr Gender=fem Number=sg Case=*] <5>

Tele/N#Kommunikation[NN SemClass=abstr Gender=fem Number=sg Case=*] <10>

...

Telekom/PN#Muni/N#Kat/N#Ion[NN SemClass=k_g_dingnat Gender=neut Number=sg Case=nom_acc_dat] <30>

...

Annotation: Wortartenbestimmung

- Auswahl der **wahrscheinlichsten** Wortart im konkreten Satzkontext aus der Menge der **möglichen** Wortarten eines Wortes → **PoS Tagging**
- statistischer Ansatz, trainiert auf **manuell kategorisierten** Daten
  - Modell über *Trigramme* aus Wörtern und Kategoriemengen (i.e. *Wortklasse*)
  - Bestimmung der wahrscheinlichsten Kategoriesequenz für einen Satz
  - heuristische Auswahl der „einfachsten“ **Grundform**
  - angepasste Modelle für historische Sprache, gesprochene Sprache, Kindersprache etc.
- Details: *Jurish, 2003*

  → Multilingual approach; but check also the **TreeTager**

(Schmid, 1994)

Annotation: Wortartenbestimmung – Beispiel

Input:	Linda	wird	die	Mannschaft	verstärken	.
Morphological Analysis:	NE.*first*, NE.*last*	VAFIN.3*rd*.*sg*.*pres*, VVFIN.3*rd*.*sg*.*pres*,	ART.*sg*.*nom*.*fem*, ⋮ PDS.*nom*.*sg*.*fem*, ⋮ PRELS.*acc*.*pl*	NN.*masc*.*sg*.*nom*, ⋮ NN.*fem*.*sg*.*	VVFIN.1*st*.*pl*.*pres*, ⋮ VVINF	{$.}
Tag Extraction:	{NE}	VVFIN, VAFIN	ART, PDS, PRELS	{NN}	VVFIN, VVINF	{$.}
Disambiguation:	NE	VAFIN	ART	NN	VVINF	$.

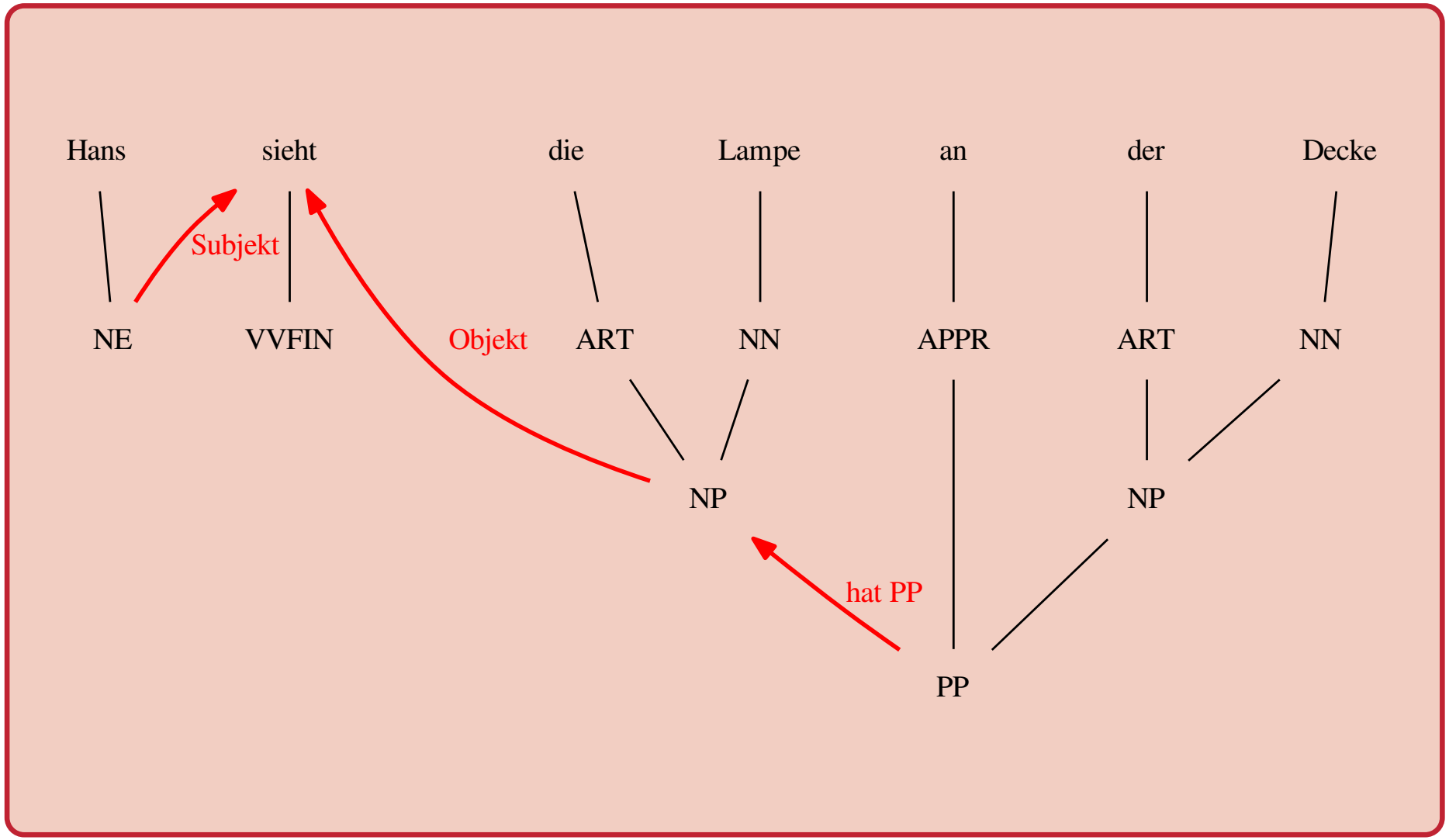
Annotation: Wortbeziehungen

- Bestimmung der **strukturellen** Beziehungen zwischen Wörtern im Satz
- **regelbasierter** Ansatz
  - handgeschriebene Grammatik
  - Grundform, Kategorie und morphosyntaktische Merkmale als Beschreibungseinheit
  - Implementierung mit Hilfe endlicher, gewichteter Automaten (**schnell**!)
- Details: *Didakowski, 2008*

  → For English, use e.g. the **Stanford parser**

(de Marneffe et al., 2006)

Annotation: Syntax – Beispiel



Annotation: Wortsemantik

- **Thesauri / Ontologien / Wortnetze**
  - Nachschlagen semantischer Kategorien sowie Beziehungen (Synonymie, Hyponymie *(„ist-ein“)*, Hyperonymie, . . .)
  - **GermaNet** *(Lemnitzer & Kunze, 2002)*
    * `http://www.sfs.uni-tuebingen.de/GermaNet`; z.B. ‚*Bank*‘
  - **OpenThesaurus** *(Naber, 2005)*
    * `http://www.openthesaurus.de`; z.B. ‚*Bank*‘
- **Distributionelle Semantik** *(Berry, Dumais, & O'Brien, 1995)*
  - berechnete, „latente“ Wortassoziierungen (gemeinsame Kontexte)
  - Terme ↔ ‚Dokumente‘ (Seiten) ↔ ‚Kategorien‘ (Bücher)
  - Beispiel: ‚*Wolke*‘

Werkzeuge

Werkzeuge: DDC-Korpussuchmaschine

(Sokirko, 2003; Jurish, Thomas, & Wiegand, 2013)

- Indizierung der Analyseergebnisse und der Metadaten
- **Wortattribute**

  → Token, Lemma, PoS, kanonische Schreibung, . . .
- **Metadatenattribute** → Autor, Titel, Datum, Textsorte, . . .
- **Abfragetypen**
  - Term-Expansion („conflation“) *Query Lizard:* `Bank`
  - Konjunktion, satzlokal `(Bank {Geld,Kredit}) #asc_date`
  - . . . oder termlokal `(Kohl with $p=NE) #dsc_date[1982,1999]`
  - Wildcards, Phrasen, RegEx `("anti* #2 Propaganda") #left[0]`
  - . . .

Werkzeuge: Wortverläufe / Histogramme

- Visualisierung relativer Frequenzen nach Dekade, Textsorte
- beliebige DDC-Abfragen
- parametrisierte Aggregation (`slice`), Glättung (`window`)
- automatische Erkennung von Ausreissern (`prune`)
- Ausgangspunkt für (Sub-)Korpusvergleiche *(Rayson & Garside, 2000)*
- Beispiele (Dingler): *`Pferd`* vs. *`Motor`*
- Beispiele (DTA): *Latein* *`($p=FM.la)`* vs. *Englisch* *`($p=FM.en)`*

Werkzeuge: Wortwolken / Distributionelle Semantik

- korpusspezifisches, semantisches Modell (z.Z. nur DTA)
- 3-Wege-Relation zwischen:
  - „Terme“ ∼ Substantiv Grundformen
  - „Dokumente“ ∼ Buchseiten
  - „Kategorien“ ∼ Bücher
- Beispiel: Bank in GermaNet Geldinstitute im DTA
- Beispiel: Kant & Hegel
  - Wolken: *Terme (Kant)*, *Bücher (Kant|Hegel)*, *Seiten (Hegel)*
  - DDC: Kant-typische Terme bei Hegel
    `$1='cat=kant@50'|sem #has[author,Hegel*]`
  - Histogramm, Kant-Hegel'sches Vokabular
    `$1='cat=kant|hegel@50'|sem !#has[author,/Kant|Hegel/]`

Werkzeuge: Wortprofile

- Darstellung signifikanter **Kookkurrenten** eines Wortes
- untergliedert nach festgelegten syntaktischen Relationen
  - *flache* syntaktische Analyse *(Didakowski, 2008)*
  - Bewertung der statistischen Signifikanz der extrahierten Relationen *(Geyken et al., 2009)*
- Reliabilität durch sehr große zugrunde liegende Korpora (ca. 1,7 Milliarden Tokens)
- Filterung nach Teilkorpus möglich
- Beispiele

  Propaganda, *Bank*, *Pferd*

Fallbeispiele

Fallbeispiele

- **Explorativ**
  - Wordstatistiken und Häufigkeitsverläufe
  - signifikante Veränderungen im Inhaltswortbereich
  - Hinweise auf Bedeutungsverschiebung bzw. Bezeichnungswandel
- **Hypothesengeleitet**
  - spezifische Korpusrecherchen zur effizienten Belegauswahl
  - Belegauswertung
  - quantitatives „Untermauern“

Fallbeispiel 1: *Terror und Terrorismus*

Stand der Forschung

- Übernahme Ende 18. Jh. aus dem Französischen mit Bezug auf „la grande terreur“
- Rückbezug auf frühere Schreckensherrschaft
- Terrorismus als Kommunikationspraxis *(Hoffman, 2003)*
- Terrorismus als Teil des Politischen *(Münkler, 2004)*

terreur → Terrorismus → Terror

Hypothesenüberprüfung anhand von Korpusrecherche

- Begriffsherkunft: **Deutsches Textarchiv**
- Bedeutungswandel: **Die ZEIT**

Fallbeispiel 1: *Terror und Terrorismus*

Ergebnisse

- frühe lateinische Belege für *terror* im Sinne von „Schrecken“
- Laukhard *Leben und Schicksale*, Leipzig 1797
  - Terrorismus *als „Schreckenssystem in Frankreich“*
  - keine Belege für *Terror* und *terreur*
- Mommsen *Römische Geschichte*, Band 2, Berlin 1855
  - Terrorismus *als Handlung im Bürgerkrieg (Untermauerung von Herrschaftsansprüchen durch die* Hochverratskommission*)*
  - keine Belege für *Terror* und *terreur*
  - terrorisieren sowohl Schrecken als auch (Staats-)Gewalt

Fallbeispiel 1: *Terror und Terrorismus*

Ergebnisse II

- Verwendungshäufigkeit in der *ZEIT*
  - 1977: 2. Generation RAF
  - 1986: Anschlag auf Diskothek *La Belle*
  - 2001: Anschlag auf das *World Trade Center* und das *Pentagon*
- Terror vs. Terrorismus anhand häufiger *„Begleiter“*
  - Terror als Gewalt von oben (z.B. „stalinistisch“, „nationalsozialistisch“)
  - staatlicher Terror in national begrenztem Einflussgebiet
  - Terrorismus als Gewalt von unten (Partisanen, Guerillas, Widerstandskämpfer → Terroristen)
  - *entgrenzter* Terrorismus als (illegitime) Widerstandsform

Fallbeispiel 2: *Bürgernähe*

qualitativer Befund:

- Bürgernähe als Begriff in der Sprache des SED-Regimes
- **kein** Begriff der offiziellen Herrschaftssprache

quantitative Überprüfung des Befundes:

- ZEIT-Korpus (BBAW)
- DDR-Presseportal (SBB)

Fallbeispiel 2: *Bürgernähe*

Ergebnisse ZEIT:

- erster Beleg **1971**
- formulierter Anspruch an die (kommunale) Verwaltung
- Ende der 1970er Jahre als direkte Aufgabe der Politik formuliert
- etwa **gleichbleibende** Frequenz in 1970er und 1980er Jahren

Ergebnisse DDR-Presseportal:

- erster Beleg **1979** in *Berliner Zeitung*
- auch hier stark auf „Verwaltung“ bezogen
- Gegenbegriff „Bürokratismus“
- **starker Anstieg** der Frequenz in den Jahren 1987–1989

Fallbeispiel 2: *Bürgernähe*

Bedeutung für Erforschung des Sprachwandels:

- Begriffe mittleren Festlegungsgrades (in der Diktatur)
- subkutaner Sprachwandel politischer Kultur
- Identifizierung semantischer Netze und Wortfelder

Grenzen der Technik:

- Bezeichnungswandel (*Onomasiologische* Analyse)
- Qualitative Auswertung bleibt unverzichtbar

Fallbeispiel 3: *Ost- und Westblock*

Ausgangsbeobachtung

- „Ostblock“ als politisches Schlagwort des Kalten Krieges
- „Westblock“ als sprachliches Pendant für „Ostblock“
- Propagandabegriffe mit negativer Konnotation

Quantitative Überprüfung

- ZEIT-Korpus des DWDS
- DDR-Presse-Korpus der SBB
- „Ostblock“ und „Westblock“ jeweils in beiden Korpora untersucht

Fallbeispiel 3: *Ost- und Westblock*

„Westblock“ in der ZEIT:

- nur 60 Belege
- **immer** im Bedeutungszusammenhang mit *„Ostblock“*
- **kein** Selbstzuschreibungsbegriff
- für **1948** die meisten Belege

„Ostblock“ in der ZEIT:

- 4 424 Belege
- heiße Phasen des **Kalten Krieges** sind *„abbildbar“*
- Höhepunkt im Jahr 1961 (Mauerbau?)

Fallbeispiel 3: *Ost- und Westblock*

„Westblock“ im DDR-Presseportal:

- 738 Belege gesamt – 455 Belege davon auf der **Titelseite**
- offizielle Propagandasprache bis Mitte der 1950er, dann *„marginal“*
- für **1948** die meisten Belege (385 von 738 Belegen)
- „European Recovery Program“ (ERP) verabschiedet

„Ostblock“ im DDR-Presseportal:

- 1 280 Belege – 658 davon zwischen 1990–1994
- heiße Phasen des **Kalten Krieges** sind *„abbildbar“*
- Höhepunkt auch hier im Jahr 1961 (Mauerbau?)

Fallbeispiel 4: *Neger*

Ausgangsbeobachtung

- Das Verschwinden des „Negers“ aus der Sprache
- „Neger“ als Trägerbegriff von Rassismus

Quantitative Überprüfung

- DWDS-Zeitungskorpora
- Untersuchung bezeichnungsnaher Begriffe

Fallbeispiel 4: *Neger*

„Neger“ als Zuschreibungsbegriff

- gängige Bezeichnung – Pendant im Englischen ist der „negro“
- im Zusammenhang mit der amerikanischen Bürgerrechtsbewegung: **Hochphase** in den 1960ern
- Höhepunkt ZEIT 1964, Höhepunkt DDR-Presse 1963

„Neger“ als Tabubegriff

- stark **fallende Frequenz** des Begriffes in den frühen 1970er Jahren in beiden Korpora
- Neger seit 1990er Jahren nahezu ausschließlich in *„Anführungszeichen“*
- 2013 lässt sich die „Neger“-Kinderbuch-Debatte nachweisen

Fallbeispiel 4: *Neger*

„Farbiger“ als Zuschreibungsbegriff

- weniger geläufig (ZEIT etwa 134 Treffer)
- Pendant im Englischen ist „colored“
- ebenfalls Hochphase in den 1960ern

„Schwarzafrikaner“ in *Koordination mit*

- **Häufig** in Koordination mit bestimmten Ethnien nichteuropäischer Herkunft (Araber, Inder, Marokkaner, Maghrebiner etc.)
- Psycholinguistik: unterbewusste **Kategorisierungsmechanismen**?
- Soziolinguistik: Versprachlichung sozialer (ethnischer) **Exklusion**?

Fallbeispiel 4: *Neger*

„Der Neger“ als Kollektivsingular

- nur **Einzelbeleg-Auswertung** möglich
- zuviele „false positives“ als Treffer

„Afroamerikaner“ als Zuschreibungsbegriff

- im ZEIT-Korpus *steigende* Frequenz ab 1989
- der Versuch einer **unbedenklichen** Zuschreibung
- umschließt nur amerikanische Schwarze

Resumé

- Korpusrecherche und Frequenzverläufe als Mittel zur Hypothesenüberprüfung im Bereich der neueren und neuesten Geschichte
- Grenzen der Adaption durch die Fachwissenschaften:
  - methodische Diskussion in den Fächern anregen
  - Verbreitung digitaler Analyseverfahren fördern
  - Digital Humanities als Teil des Propädeutikum in der Universitätsausbildung verankern
- Grenzen der Ressourcen und Werkzeuge:
  - Problem der vollständigen Digitalisierung historischer Dokumente
  - Historizität digitaler Korpora (besser) Rechnung tragen
  - technische Bereitstellung und dauerhafte Verfügbarkeit der Ressourcen sicherstellen – Aufgabe der Bibliotheken?

Vielen Dank für Ihre Aufmerksamkeit!

Literatur

Burnard, L. "Users' reference guide to the British National Corpus" Oxford University Press, Oxford, 1995. http://www.natcorp.ox.ac.uk/archive/worldURG/urg.pdf (Version von 2000)

Geyken, A. "The DWDS corpus: A reference corpus for the German language of the 20th century". In: C. Fellbaum (Ed.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects, pp. 23-41. Continuum Press, London. http://dwds.de/static/website/publications/text/DWDS-Corpus_Desc4_draft.pdf

Geyken, A. & W. Klein. "Deutsches Textarchiv". In: Jahrbuch der Berlin-Brandenburgischen Akademie der Wissenschaften 2009, pp. 320-324. Akademieverlag, Berlin, 2009. http://edoc.bbaw.de/volltexte/2010/1515/pdf/BBAW_Jahrbuch_2009.pdf

Dittmar, N. & N. Bahlo. "Jugendsprache". In: H. Anderlik and K. Kaiser (Eds.), Die Sprache Deutsch, pp. 264-268. Deutsches Historisches Museum, Dresden, 2008 http://www.geisteswissenschaften.fu-berlin.de/v/jugendsprache/ (HTML Draft)

Schroeder, S., K.-M. Würzner, J. Heister, A. Geyken & R. Kliegl. "childLex: a lexical database of German read by children". Behavior Research Methods DOI: 10.3758/s13428-014-0528-1, 2014. http://http://link.springer.com/article/10.3758/s13428-014-0528-1

Jurish, B., & K.-M. Würzner. "Word and Sentence Tokenization with Hidden Markov Models". Journal for Language Technology and Computational Linguistics, 28(2):61-83, 2013. http://www.jlcl.org/2013_Heft2/3Jurish.pdf

Geyken, A. & T. Hanneforth. "TAGH: A Complete Morphology for German based on Weighted Finite State Automata". In A. Yli-Jyrä, L. Karttunen, and J. Karhumäki, editors, Finite State Methods and Natural Language Processing, pp. 55–66. Springer, Berlin, Heidelberg, 2006. http://dwds.de/static/website/publications/text/Geyken_Hanneforth_fsmnlp.pdf

Porter, M.F. "An algorithm for suffix stripping". Program, 14(3):130-137, 1980. http://http://tartarus.org/~martin/PorterStemmer/

Jurish, B. "A hybrid approach to part-of-speech tagging." Final report, Project Kollokationen im Wörterbuch, Berlin-Brandenburgische Akademie der Wissenschaften, 2003. http://kaskade.dwds.de/~moocow/mirror/pubs/dwdst-report.pdf

Schmid, H. "Probabilistic Part-of-Speech Tagging Using Decision Trees". Proceedings of International Conference on New Methods in Language Processing, Manchester, 1994. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Didakowski, J. "Local Syntactic Tagging of Large Corpora Using Weighted Finite State Transducers". In A. Storrer et al. (Hrsg.), Text Resources and Lexical Knowledge Mouton de Gruyter, S. 65-78, 2008 http://dwds.de/dokumentation/syncop/

de Marneffe, M.-C., B. MacCartney & C.D. Manning. "Generating Typed Dependency Parses from Phrase Structure Parses". LREC, 2006. http://nlp.stanford.edu/software/lex-parser.shtml

Lemnitzer, L. & C. Kunze. Computerlexikographie: Eine Einführung. Narr, Tübingen, 2007. [Kap. 6.2: GermaNet] http://www.ssg-bildung.ub.uni-erlangen.de/computerlexikographie.pdf

Naber, D. "OpenThesaurus — Ein offenes deutsches Wortnetz", Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005 in Bonn (Frankfurt: Peter-Lang-Verlag), S. 422–433. http://www.danielnaber.de/publications/gldv-openthesaurus.pdf

Berry, M. W., S. T. Dumais & G. W. O'Brien. "Using Linear Algebra for Intelligent Information Retrieval", SIAM Review 37(4), 573–595. http://dx.doi.org/10.1137/1037127

Jurish, B., C. Thomas & F. Wiegand. "Querying the Deutsches Textarchiv." In: U. Kruschwitz, F. Hopfgartner, & C. Gurrin (Hg.): Proceedings of the Workshop MindTheGap 2014: Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities (co-located with iConference 2014, Berlin, 4. März, 2014), S. 25–30, 2014. http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf

Rayson, P. & R. Garside. "Comparing Corpora using Frequency Profiling." In: Proceedings of the workshop on Comparing Corpora, held in conjunction with ACL 2000. October 2000, Hong Kong, S. 1-6. http://eprints.lancs.ac.uk/11882/1/rg_acl2000.pdf

Didakowski, J. "SynCoP - Combining Syntactic Tagging with Chunking Using Weighted Finite State Transducers". In: T. Hanneforth and K.-M. Würzner (Eds.), Finite-State Methods and Natural language Processing, 6th International Workshop, FSMNLP 2007, pp. 107-118. Universitätsverlag Potsdam, 2007. http://opus.kobv.de/ubp/volltexte/2008/2381/pdf/fsmnlp07proc.pdf#page=123

Geyken, A., J. Didakowski & A. Siebert "Generation of word profiles for large German corpora". In: Y. Kawaguchi et al. (Eds.), Corpus Analysis and Variation in Linguistics, pp. 141-157, Tokyo University of Foreign Studies, Studies in Linguistics 1, John Benjamins Publishing Company, 2009. http://dwds.de/static/website/publications/text/Geyken_Didakowksi_Siebert_WordProfiles_Ms.pdf