Visualisierung sprachlicher Muster

—

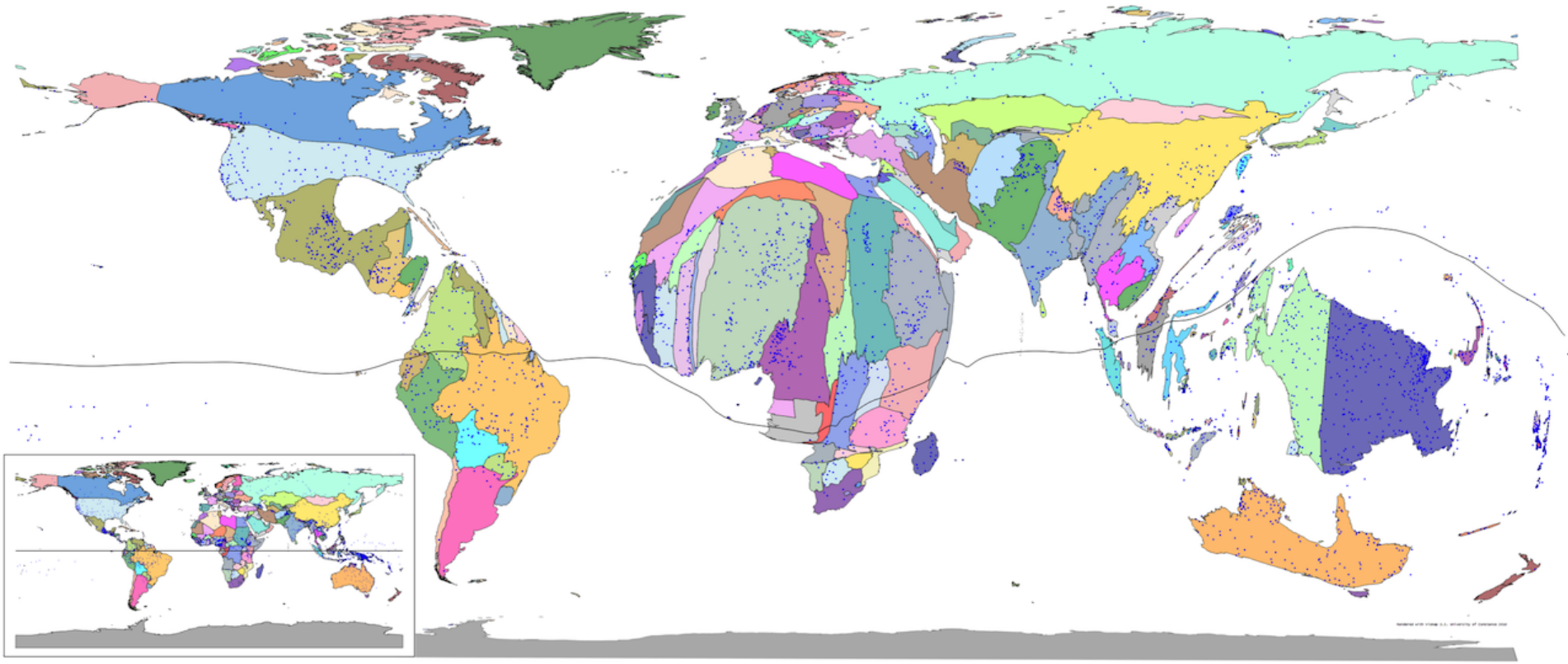
Den Wald UND die Bäume sehen

Annette Hautli-Janisz[1] Christian Rohrdantz[2]

Fachbereich Sprachwissenschaft[1] Fachbereich Informatik[2]
Universität Konstanz

6. Arbeitstreffen "Internetlexikographie", IDS Mannheim
20./21. November 2013

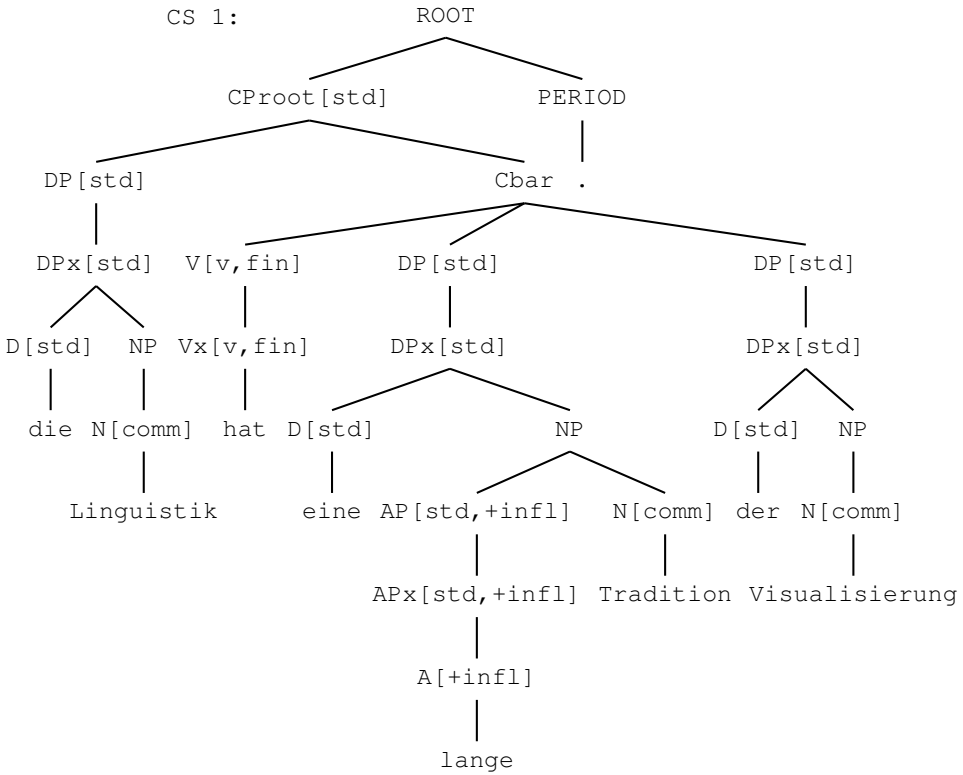
Motivation

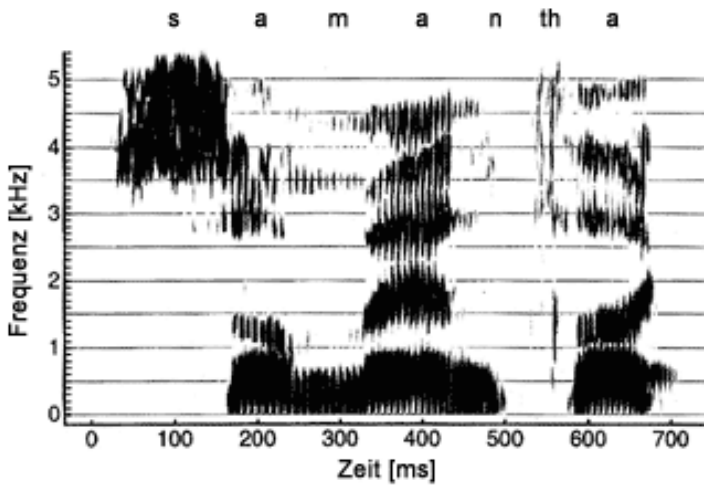


Rohrdantz et al. (2012) (`http://th-mayer.de/cartogram/`)

Motivation

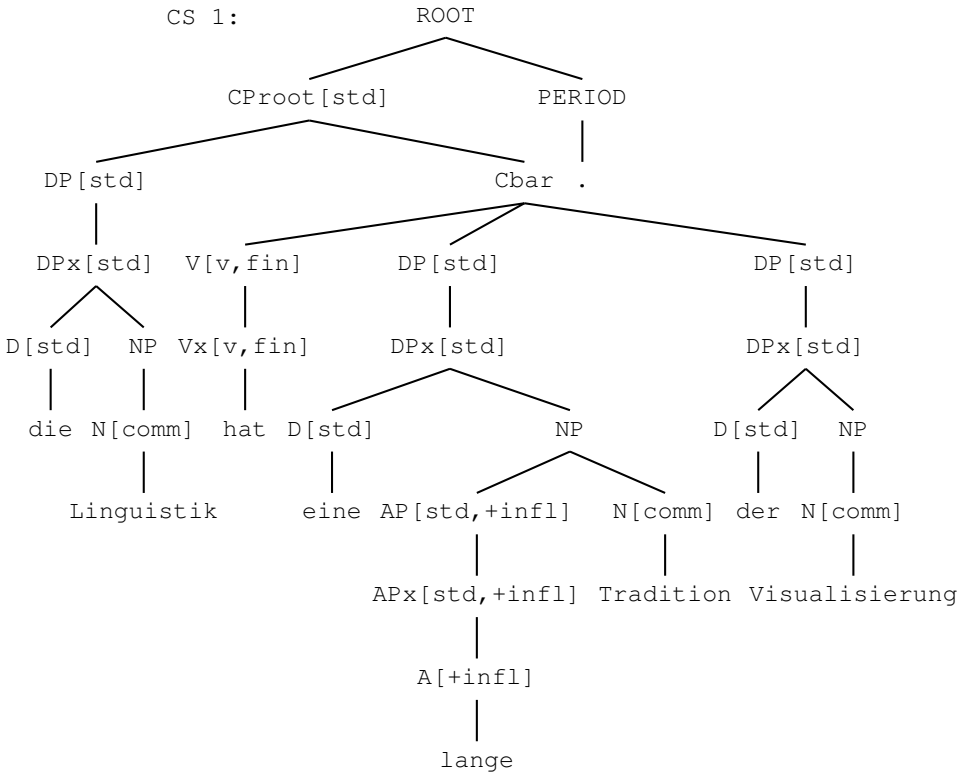
Die Linguistik hat eine lange Tradition der Visualisierung:

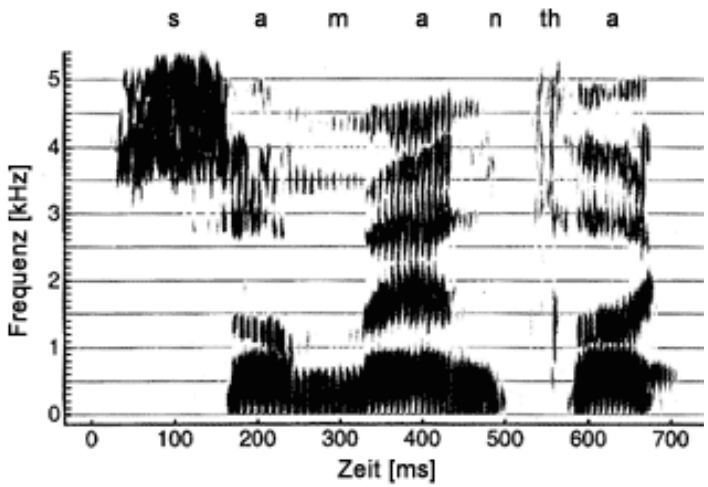




Motivation

Die Linguistik hat eine lange Tradition der Visualisierung:



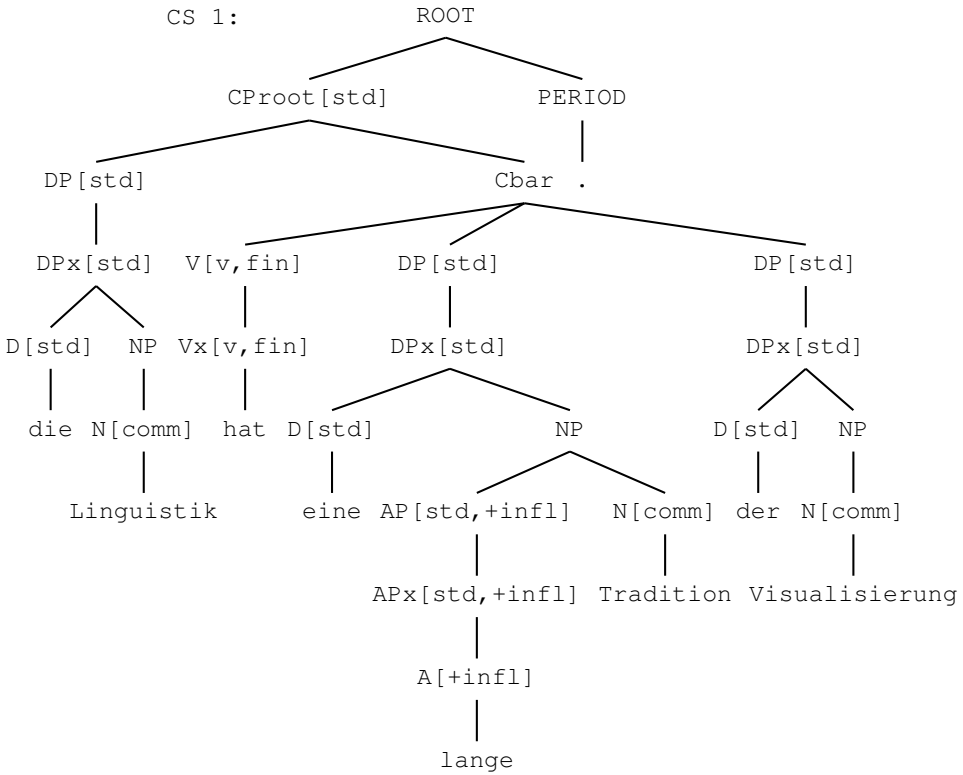


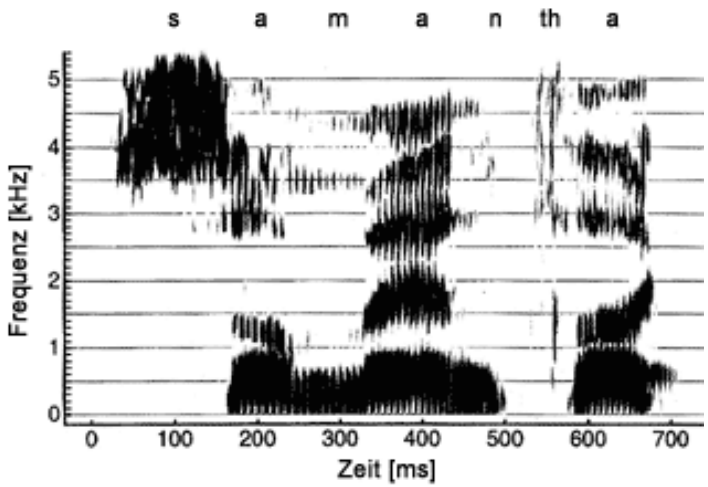
Anforderung im Digital Age:

"Data is the new oil." (Clive Humby, 2006)

Motivation

Die Linguistik hat eine lange Tradition der Visualisierung:





Anforderung im Digital Age:

"Data is the new soil." (David McCandless, 2010)

Motivation

Herausforderungen an die Visualisierung großer (linguistischer) Datenmengen:

- Brückenschlag zwischen theoretisch motivierter Fragestellung und adäquater maschineller/visueller Implementierung
- Kompletter Überblick über die Daten bei gleichzeitiger Möglichkeit der detaillierten Untersuchung einzelner Datenpunkte

Vorteil:

$\rightarrow$ Visualisierung als Schnittstelle zwischen Mensch und Computer löst vorbewusste Wahrnehmung aus.

Motivation

Vokalharmonie: "Standard" darstellung

Abweichung der Assoziierungsstärke zwischen Vokalen von der erwarteten Assoziierung.

	a	ı	u	o	ö	ü	i	e
a	0.266	0.427	-0.141	-0.060	0.019	-0.125	-0.261	-0.275
ı	0.162	0.292	-0.107	0.077	-0.010	-0.075	-0.190	-0.191
u	0.129	-0.143	0.464	0.017	-0.003	-0.051	-0.138	-0.140
o	0.066	-0.112	0.434	-0.015	0.006	-0.045	-0.104	-0.111
ö	-0.107	-0.092	-0.052	-0.026	0.006	0.366	-0.091	0.164
ü	-0.120	-0.114	-0.059	0.014	-0.006	0.507	-0.112	0.134
i	-0.201	-0.224	-0.118	0.071	-0.004	-0.087	0.319	0.211
e	-0.256	-0.251	-0.132	-0.062	-0.010	-0.097	0.400	0.276

Turkish

	a	i	o	e	u
a	-0.003	-0.075	0.094	-0.025	-0.018
i	-0.025	-0.004	0.064	-0.036	0.005
o	-0.028	-0.006	-0.075	0.098	0.026
e	-0,001	0.063	-0.073	0.016	0.021
u	0.077	0.038	-0.036	-0.057	-0.043

Spanish

	a	o	i	ü	ö	ä	u	e
a	0.019	0.009	-0.061	-0.034	-0.008	-0.025	0.018	0.035
o	-0.023	-0.004	-0.052	-0.013	-0.020	-0.013	-0.013	0.068
i	-0.069	-0.054	-0.050	-0.039	-0.036	-0.044	-0.003	0.133
ü	-0.067	-0.045	0.070	-0.028	-0.021	-0.033	-0.021	0.050
ö	-0.049	-0.032	0.049	-0.024	-0.013	-0.021	-0.013	0.036
ä	-0.067	-0.037	0.124	-0.033	-0.018	-0.028	-0.038	0.020
u	0.012	-0.018	-0.019	0.046	-0.002	-0.013	0.004	-0.001
e	0.108	0.084	0.026	0.069	0.063	0.096	0.021	-0.195

German

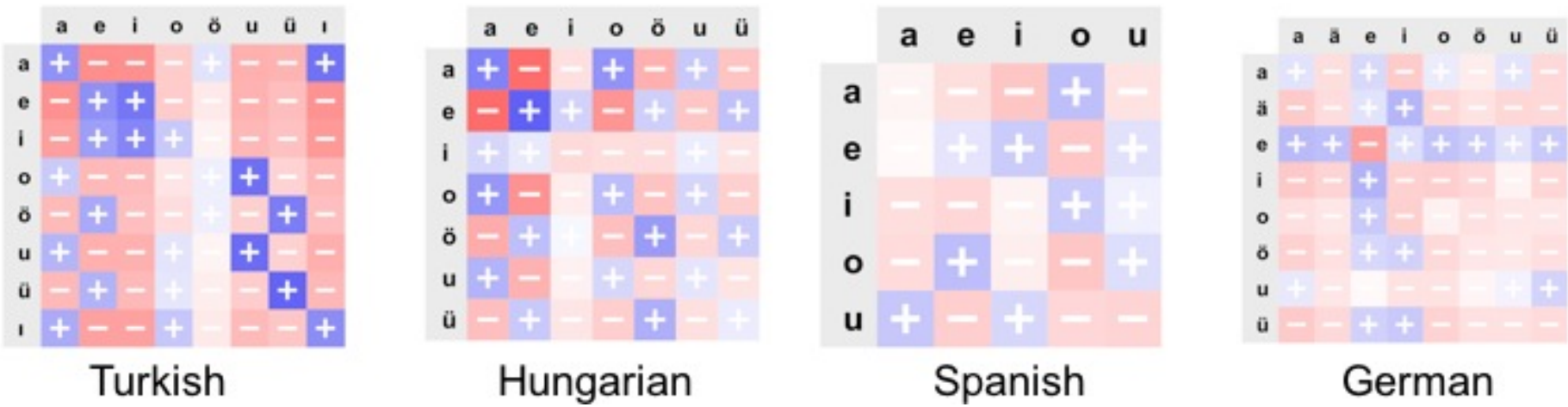
	a	o	u	i	ü	ö	e
a	0.339	0.263	0.070	-0.022	-0.081	-0.136	-0.431
o	0.239	0.099	0.041	-0.007	-0.052	-0.083	-0.253
u	0.132	0.038	0.015	-0.004	-0.017	-0.040	-0.131
i	0.037	-0.026	0.008	-0.030	-0.017	-0.027	0.011
ü	-0.093	-0.056	-0.022	-0.014	0.008	0.148	0.071
ö	-0.152	-0.093	-0.037	0.001	0.065	0.229	0.097
e	-0.435	-0.241	-0.076	0.048	0.091	0.054	0.531

Hungarian

Motivation

Eine erste Visualisierung

Vokale sind alphabetisch geordnet

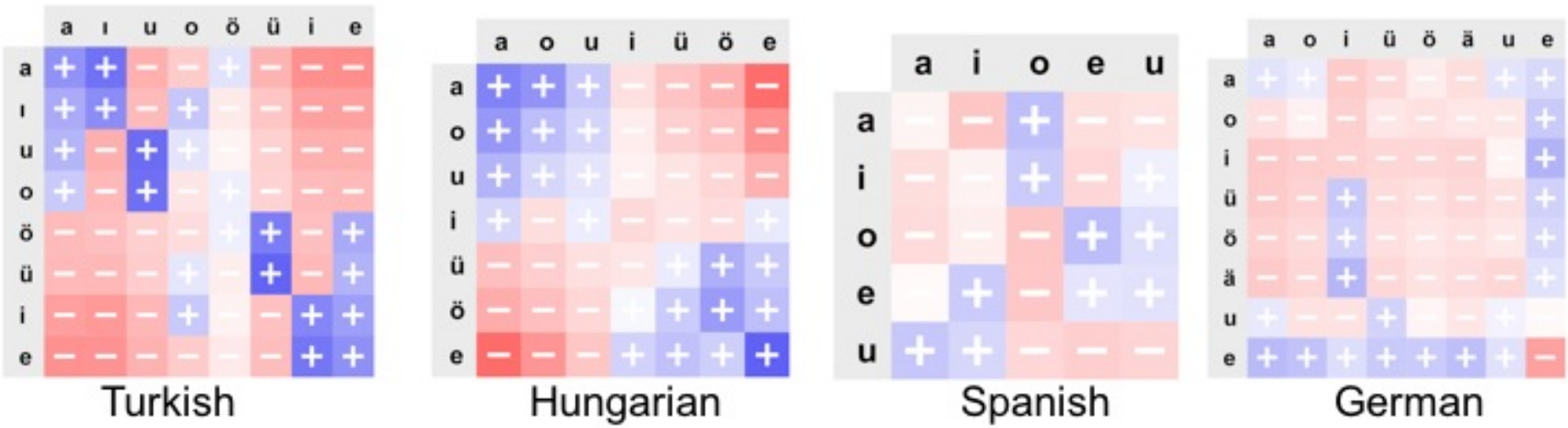


- Je gesättigter die Farbe, desto größer die Abweichung

Motivation

Eine sortierte Visualisierung

Können Sie jetzt ein Muster erkennen?



- Vokale sind nach Ähnlichkeit geordnet (Rohrdantz et al. 2010)
- Darstellung der Typen der Vokalharmonie

Visualisierung und Linguistik in Konstanz

Hauptverantwortlich: Miriam Butt und Daniel Keim

- Forschungsinitiative "Computational Analysis of Linguistic Development" (CALD) (2008-2010), Exzellenzinitiative Universität Konstanz
- eHumanities Projekt "Wie und wann überzeugen Argumente – Analyse und Visualisierung politischer Verhandlungen" (VisArgue) (2012-2015), gefördert vom BMBF
- "Visual Analysis of Language Change and Use Patterns" (2013-2014), gefördert durch die DFG
- Forschungsinitiative "Visual Analysis of Language Change and Use Patterns" (LingVisAnn) (2013-2015), Exzellenzinitiative Universität Konstanz

Agenda

1. Visualisierung von Bedeutungswandel
2. Clustervisualisierung
3. Fazit

Heute

1. Visualisierung von Bedeutungswandel
2. Clustervisualisierung
3. Fazit

Visualisierung von semantischem Wandel

Semantischer Wandel

- Bedeutungsänderung eines Wortes über die Zeit
- Verschiedene Arten des semantischen Wandels, z.B.
  - *Verengung* (Oberbegriff wird zum Unterbegriff), z.B. *skyline* im Englischen
  - *Erweiterung* (Begriff wird genereller), z.B. Englisch *horn*
- Wandel in den letzten 20 Jahren: Begriffe im Bereich des Internet

Ziel von Rohrdantz et al. (2011)

- Verfolgung des semantischen Wandels über die Zeit mit automatischer Annäherung
- Visuelle Darstellung der Entwicklung

Visualisierung von semantischem Wandel

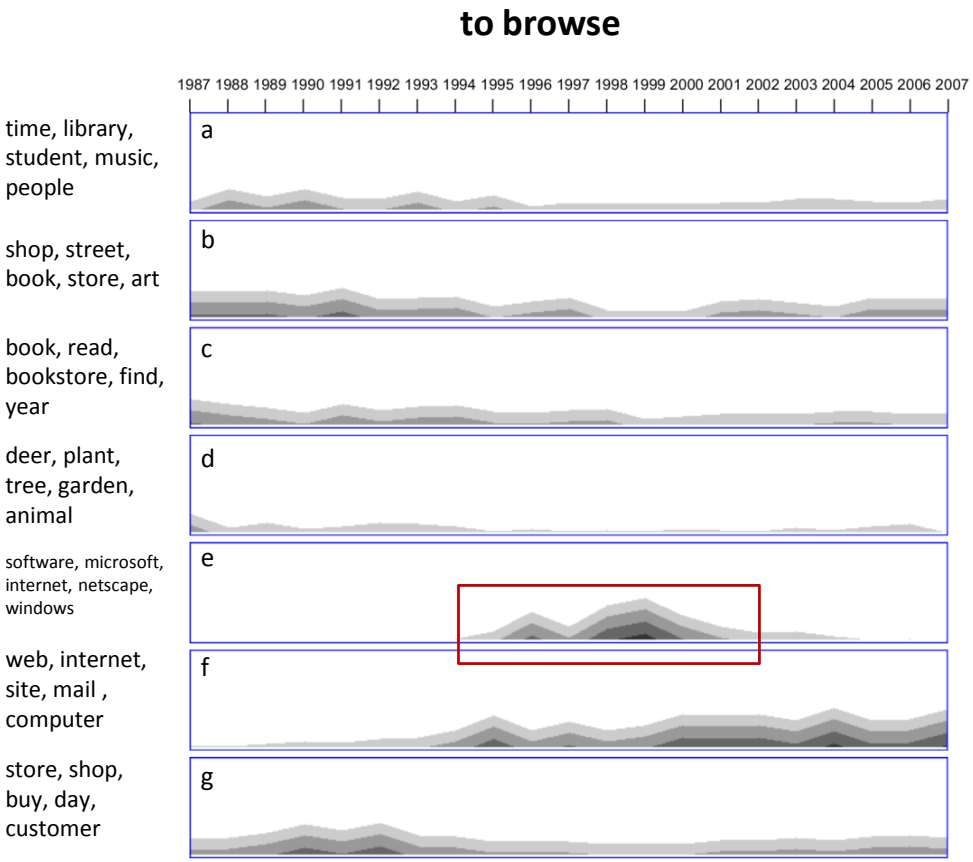
Vorgehensweise

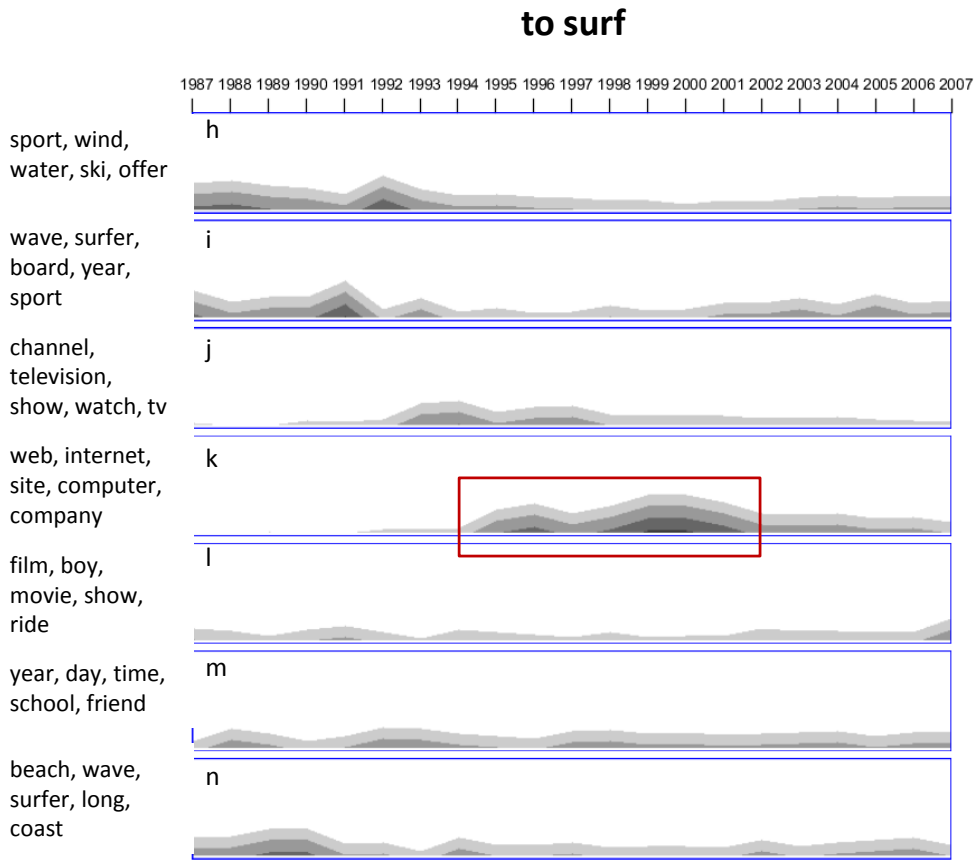
- Untersuchung auf Basis des New York Times Korpus
  - 1.8 Millionen Zeitungsartikel von 1987 bis 2007
  - Jeder Artikel ist kodiert mit einem Zeitstempel
- Extrahierung eines Kontextes von 25 Wörtern vor und nach dem Begriff
- Topic modeling zur Annäherung der Wortbedeutung aufgrund des Kontextes (LDA)
  - Jeder Kontext ist ein "Dokument"
  - Vorgegeben Anzahl an Topics, jeder Kontext wird einem Topic zugeordnet.
- Visualisierung interpretiert die Information aus der statistischen Analyse

Visualisierung von semantischem Wandel

Erste Ansicht

- Aggregierte, diachrone Sicht auf die Daten



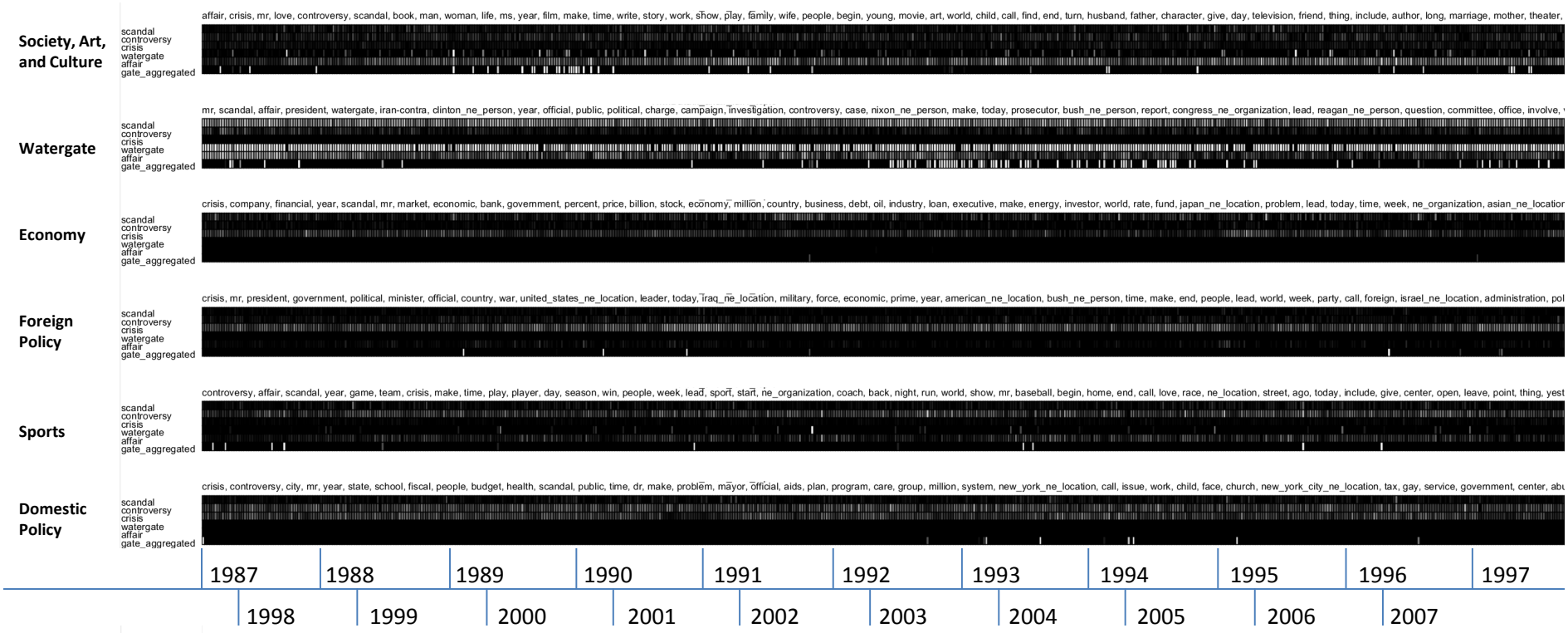


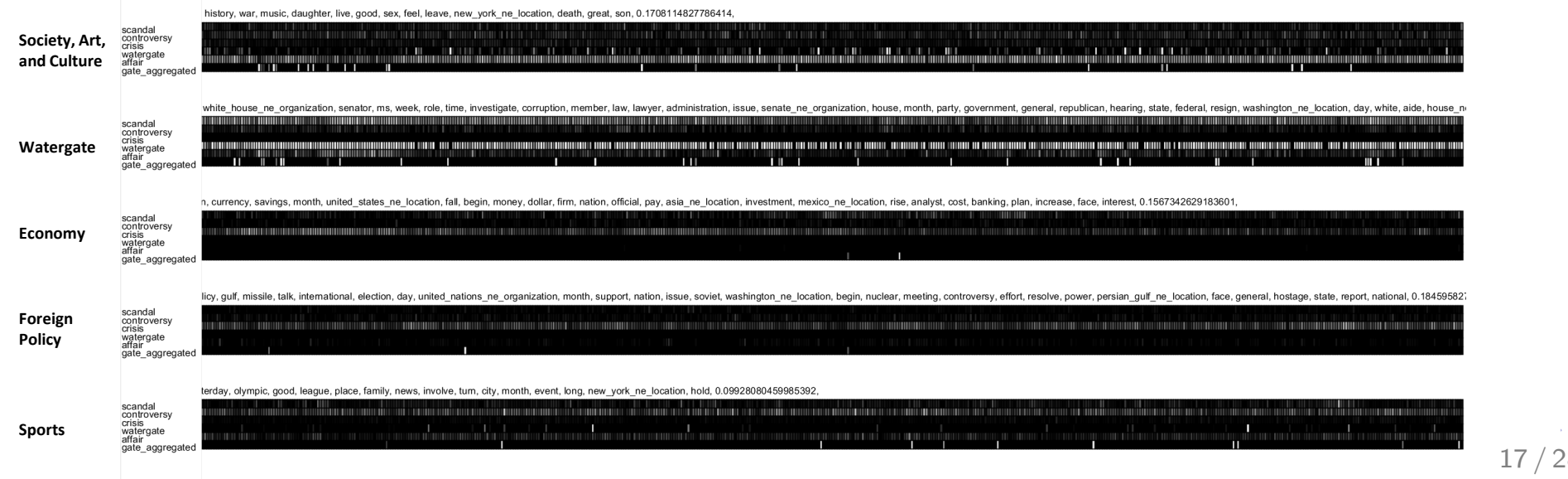
Visualisierung von semantischem Wandel

Bedeutungswandel des Suffixes *-gate* im Englischen (Rohrdantz et al. 2012)

- Schöpfung: Water*gate* Affäre (1973-74)
- Seither: Gebrauch von *-gate* als Suffix (Iraq*gate*, Monica*gate*, Nipple*gate*, Ruby*gate*)
- Hypothese: Bedeutungsveränderungen des Suffixes
- Art der Messung: Vergleich der *-gate* Termini mit Wörtern wie *scandal, affair, crises, controversy*
  - Korpora: NYT Korpus und EMM Korpus
  - Extraktion der Kontexte aller Termini, Vergleich der Ähnlichkeit der Kontextvektoren

Visualisierung von semantischem Wandel



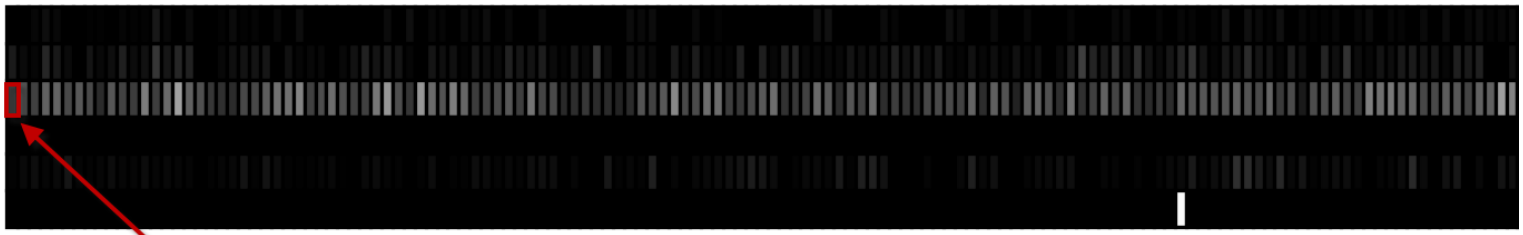


Visualisierung von semantischem Wandel

Foreign Policy

crisis, mr, president, government, political, minister, official, country, war, unite

scandal
controversy
crisis
watergate
affair
gate_aggregated

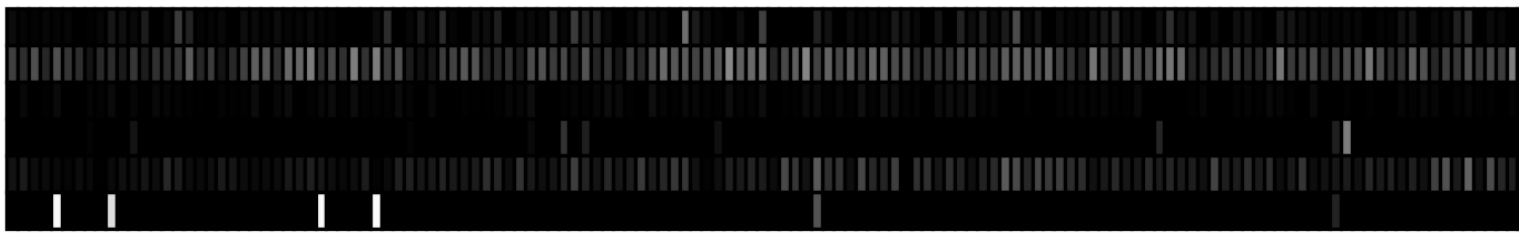


relative number of "crisis"-contexts of the first week that belong to the topic Foreign Policy

Sports

controversy, affair, scandal, year, game, team, crisis, make, time, play, player

scandal
controversy
crisis
watergate
affair
gate_aggregated



Domestic Policy

crisis, controversy, city, mr, year, state, school, fiscal, people, budget, health,

scandal
controversy
crisis
watergate
affair
gate_aggregated

Visualisierung von semantischem Wandel

Heylen et al. 2013

- Bedeutungswandel von englischen Adjektiven, z.B. *terrific*
- Hypothese: Zwischen 1860 und 2000, Wandel von *terrific* von negativ zu positiv konnotiertem Adjektiv.
- Methode:
  - Vergleich der Kontexte von *terrific* mit Kontexten von anderen positiven/negativen Adjektiven wie *terrible*, *horrible* and *magnificent*, *great*
  - Visualisierung mit den Google Vis Tools

[Demo]

Heute

Clustervisualisierung

Generelles Problem bei Methoden des maschinellen Lernens:

- Intransparenz der Ergebnisse
- (Oft) fehlender Einblick in die zugrundeliegenden Daten

$\rightarrow$ keine detaillierte (linguistische) Analyse möglich

Ansatz in Lamprecht et al. (2013):

- Visualisierung von Ergebnissen des automatischen Clusterings
- Clusterkohärenz und Zusammensetzung mithilfe eines interaktiven Visualisierungssystems

Clustervisualisierung

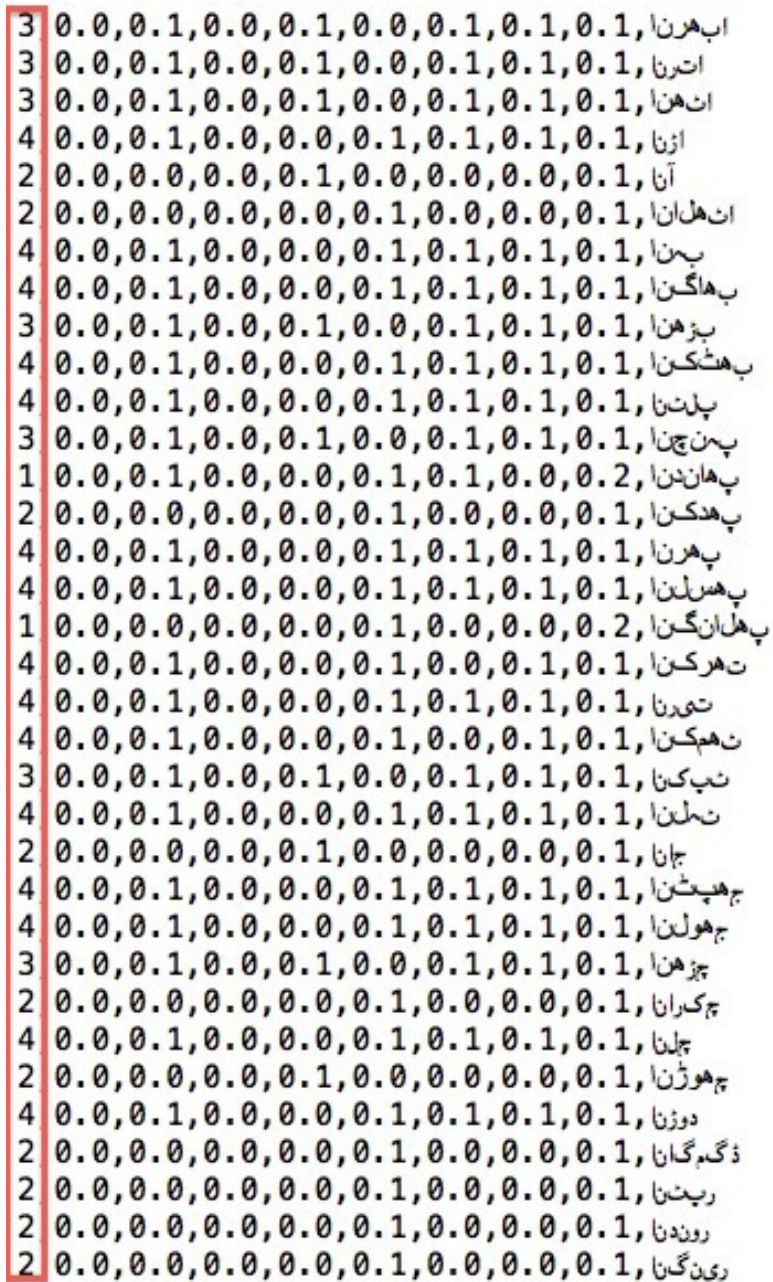
- Hier: Urdu Verben und ihre syntaktischen Merkmale
- Jeder Datenpunkt besteht aus einem Vektor an Attributen
- Ziel: Verbklassen als Grundlage für ein Urdu VerbNet

0.0,0.1,0.0,0.1,0.0,0.1,0.1,0.1,ابھرنا
0.0,0.1,0.0,0.1,0.0,0.1,0.1,0.1,اترنا
0.0,0.1,0.0,0.1,0.0,0.1,0.1,0.1,اٹھنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,اڑنا
0.0,0.0,0.0,0.1,0.0,0.0,0.0,0.1,آنا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.1,اٹھلانا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,بہنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,بھاگنا
0.0,0.1,0.0,0.1,0.0,0.1,0.1,0.1,بڑھنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,بھٹکنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,پلٹنا
0.0,0.1,0.0,0.1,0.0,0.1,0.1,0.1,پہنچنا
0.0,0.1,0.0,0.0,0.1,0.1,0.0,0.2,پھاندنا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.1,پھدکنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,پھرنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,پھسلنا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.2,پھلانگنا
0.0,0.1,0.0,0.0,0.1,0.0,0.1,0.1,تھرکنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,تیرنا
0.0,0.1,0.0,0.0,0.1,0.0,0.1,0.1,ٹھمکنا
0.0,0.1,0.0,0.1,0.0,0.1,0.1,0.1,ٹپکنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,ٹہلنا
0.0,0.0,0.0,0.1,0.0,0.0,0.0,0.1,چاہنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,چھپنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,چھولنا
0.0,0.1,0.0,0.1,0.0,0.1,0.1,0.1,چڑھنا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.1,چکرانا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,چلنا
0.0,0.0,0.0,0.1,0.0,0.0,0.0,0.1,چھوڑنا
0.0,0.1,0.0,0.0,0.1,0.1,0.1,0.1,دوڑنا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.1,ڈگمگانا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.1,رہنا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.1,روندنا
0.0,0.0,0.0,0.0,0.1,0.0,0.0,0.1,رینگنا

Clustervisualisierung

- Automatisches Clustering, z.B. k-means clustering
- Zuordnung jedes einzelnen Datenpunkts zu einem Cluster
- Fragen: Wie ähnlich sind sich Datenpunkte? Wie kohährent sind die Cluster?

[Demo]



Heute

1. Visualisierung von Bedeutungswandel
2. Clustervisualisierung
3. Fazit

Fazit

- Große digitale Datenmengen bergen großes Potential für die (theoretische) Linguistik
- Visualisierung unterstützt die Datenanalyse
  - Generierung und
  - Verifizierung von Hypothesen
- Bedingung: Darstellung der großen Muster sowie detaillierte Ansicht der zugrundeliegenden Daten

Fazit

- Große digitale Datenmengen bergen großes Potential für die (theoretische) Linguistik
- Visualisierung unterstützt die Datenanalyse
  - Generierung und
  - Verifizierung von Hypothesen
- Bedingung: Darstellung der großen Muster sowie detaillierte Ansicht der zugrundeliegenden Daten

Potentielle Anwendungsbereiche in der Lexikographie

- Semantische Ähnlichkeit von Begriffen
- Diachrone Muster

Vielen Dank!

References

K. Heylen, T. Wielfaert and D. Speelman. 2013. *Tracking change in word meaning. A dynamic visualization of a diachronic distributional semantic model*. DGfS Workshop on the Visualization of Linguistic Patterns, Potsdam, Germany.

A. Lamprecht, A. Hautli, C. Rohrdantz and T. Bgel. 2013. *A Visual Analytics System for Cluster Exploration*. In Proceedings of ACL Demo Session, pages 109-114, Sofia, Bulgaria.

C. Rohrdantz, M. Hund, T. Mayer, B. Wälchli and D. A. Keim. *The World's Languages Explorer: Visual Analysis of Language Features in Genealogical and Areal Contexts*. Computer Graphic Forum, 31(3):935-944, 2012.

C. Rohrdantz, A. Niekler, A. Hautli, M. Butt and D. A. Keim. 2012. *Lexical Semantics and Distribution of Suffixes — A Visual Analysis*. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH, pages 7-15, Avignon, France.

C. Rohrdantz, A. Hautli, T. Mayer, M. Butt, D. A. Keim and F. Plank (2011). *Towards Tracking Semantic Change By Visual Analytics*. In Proceedings of ACL, pages 305–310, Portland, USA.