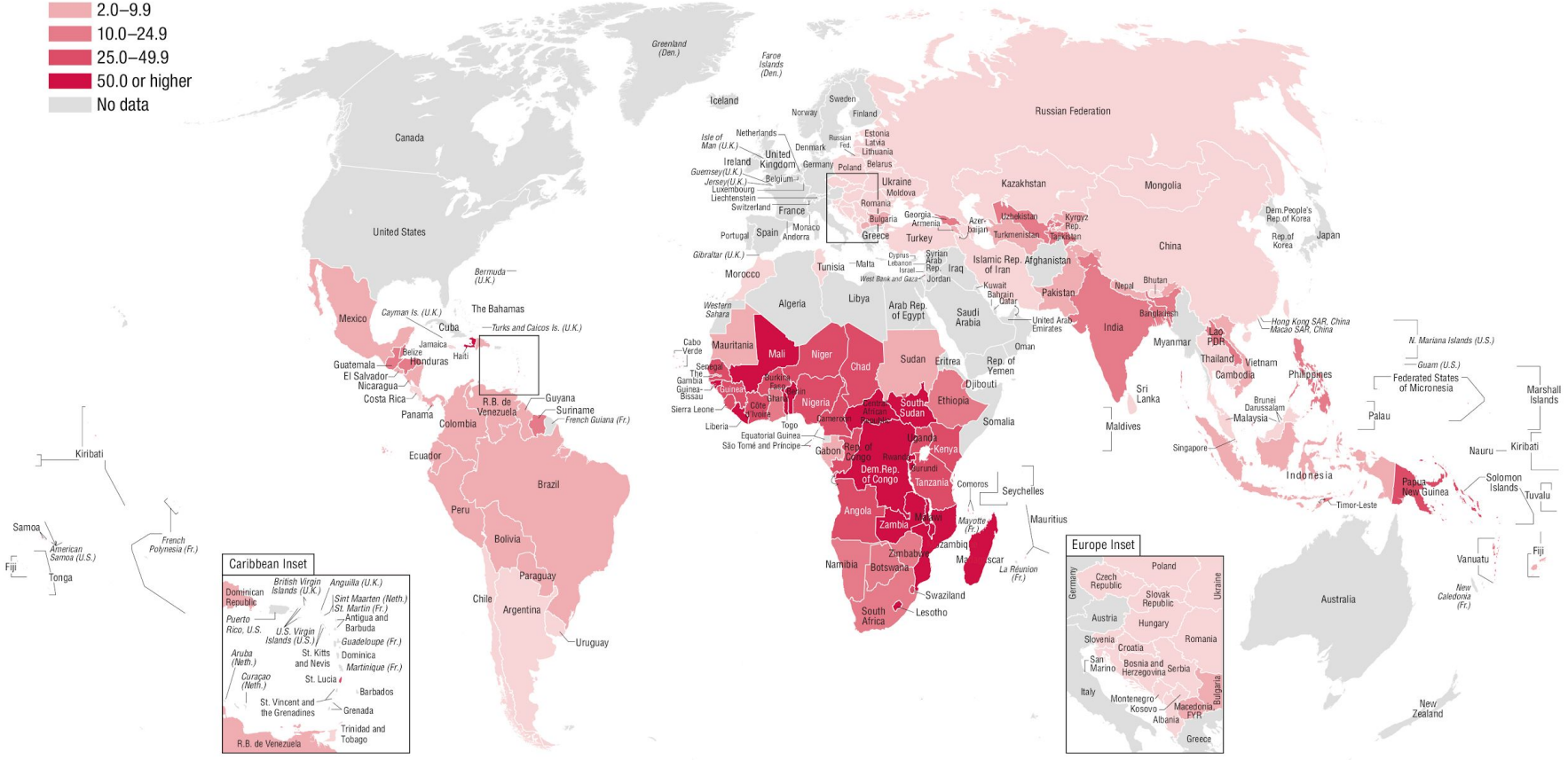
Texte als Daten für die Soziologie: Einblicke in globale Herausforderungen

Sophie Mützel & Alex Flückiger
Universität Luzern
DGS, Göttingen
27. September 2018

Poverty

Share of population living on less than 2011 PPP $1.90 a day, 2013 (%)

- Less than 2.0
- 2.0–9.9
- 10.0–24.9
- 25.0–49.9
- 50.0 or higher
- No data



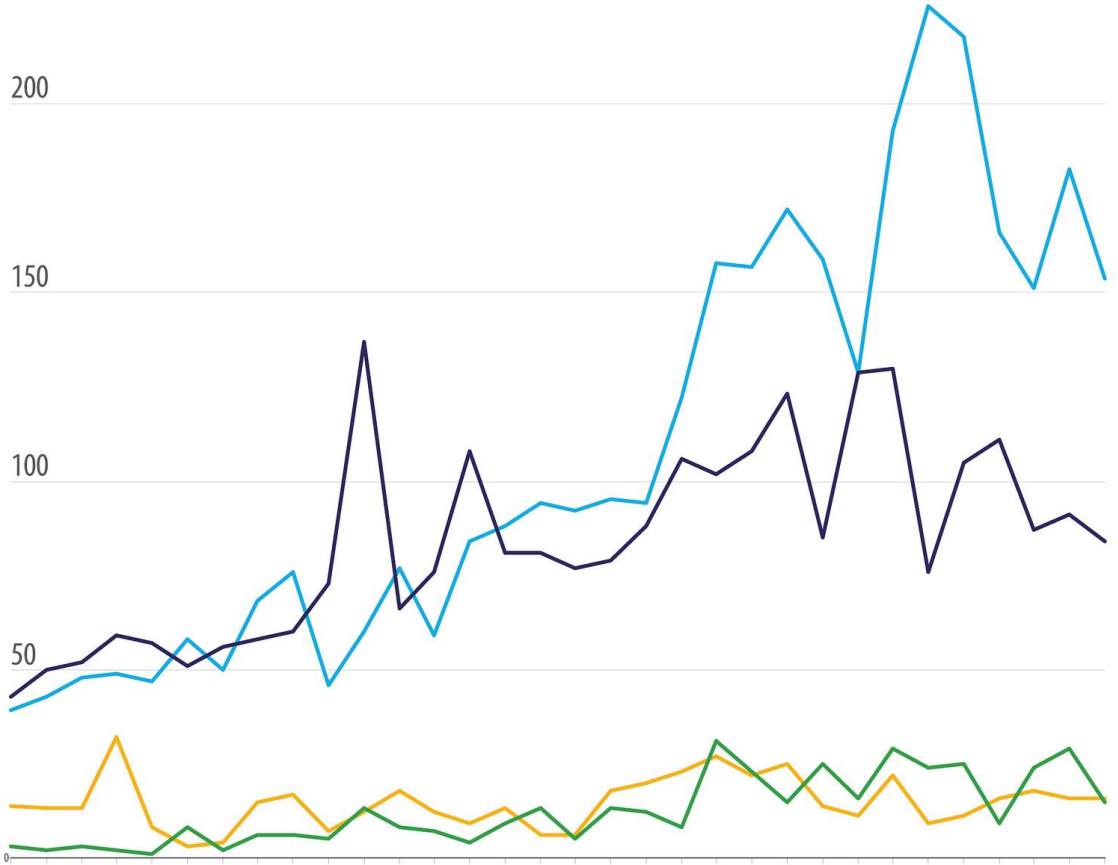
Number of Climate-related Disasters Around the World (1980-2011)

3455 FLOODS

2689 STORMS

470 DROUGHTS

395 EXTREME TEMPS



	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
FLOOD	39	43	48	49	47	58	50	68	76	46	60	77	59	84	88	94	92	95	94	122	158	157	172	159	129	193	226	218	166	151	183	154
STORM	43	50	52	59	57	51	56	58	60	73	137	66	76	108	81	81	77	79	88	106	102	108	123	85	129	130	76	105	111	87	91	84
DROUGHT	14	13	13	32	8	3	4	15	17	7	12	18	12	9	13	6	6	18	20	23	27	22	25	14	11	22	9	11	16	18	16	16
EXTREME TEMPERATURE	3	2	3	2	1	8	2	6	6	5	13	8	7	4	9	13	5	13	12	8	31	23	15	25	16	29	24	25	9	24	29	15

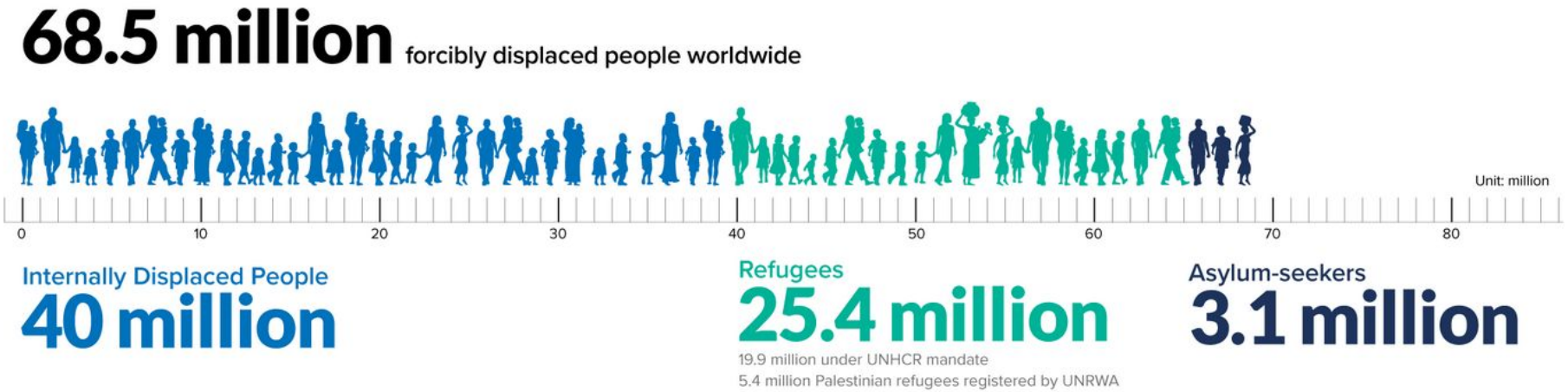
UNISDR
The United Nations Office for Disaster Risk Reduction
http://www.unisdr.org

Version: 13 June 2012

DATA SOURCES

EM-DAT - http://www.emdat.be/ - The OFDA/CRED International Disaster Database; Data version: 13 June 2012 - v12.07

Humanitarian Symbol Set (2008): http://www.ungiwg.org/map/guideline.php



UNHCR 2018

“We are now witnessing the highest levels of displacement on record.”

Forschungsfragen

1. Wie sind die Diskussionen um **globale Herausforderungen** in einem Forum der world polity **strukturiert**?
2. Wie werden diese Herausforderungen auf globaler Ebene **thematisiert**?
3. Wie hat sich das globale Verständnis von **Migration** über Jahrzehnte verändert?

Forschungsstand

UN-Generaldebatte (Baturo et al. 2017; Pomeroy et al. 2018a, 2018b; Hecht 2016)

Konzepte und world polity aus globalem Blick (Kategorien: Bennani 2017, Heintz 2015; organisationale Netzwerke: z.B. Beckfield 2010)

globale Migration (z.B. SI *Social Networks* zu Migration 2018)

große Textdatenmengen, um Einblick in semantische Strukturen und Muster zu erhalten

- **historische Soziologie und Kultursoziologie** (z.B. Bail 2014; Bearman 2015; DiMaggio et al. 2013; Edelmann/Mohr 2018; Hoffman et al. 2018; Kozlowski et al. 2018; Mohr/Bogdanov 2013; Mützel 2015; Rule et al. 2015)
- **Computational social science** und **politische Textkorpora** (z.B. Alvarez 2016; Blätte et al. 2018; Grimmer/Stewart 2013; Lazer et al. 2009; Merz et al. 2016; McFarland et al. 2016)

Thesen

In den Texten dieses globalen Forums lassen sich relationale und semantische Strukturen aufzeigen.

In den Texten dieses globalen Forums gibt es inhaltliche Brüche, Verschiebungen von Themen und Begriffswandel.

Daten und Methoden

- Alle Reden von UN Mitgliedsstaaten an der UN-Generaldebatte, gehalten 1970-2017 (öffentlich zugänglicher maschinenlesbarer Korpus, Mikhaylov et al. 2017)

United Nations A/69/PV.6

General Assembly

Sixty-ninth session

Official Records

6th plenary meeting
Wednesday, 24 September 2014, 9 a.m.
New York

President Obama: We come together at a crossroads between war and peace, between disorder and integration, between fear and hope. Around the globe, there are signposts of progress. The shadow of the Second World War that existed at the founding of this institution has been lifted, and the prospect of war between major Powers reduced. The ranks of Member States have more than tripled, and more people live under Governments that they elected. Hundreds of millions of human beings have been freed from the prison of poverty, with the proportion of those living in extreme poverty cut in half. And the world economy continues to strengthen after the worst financial crisis of our lives.

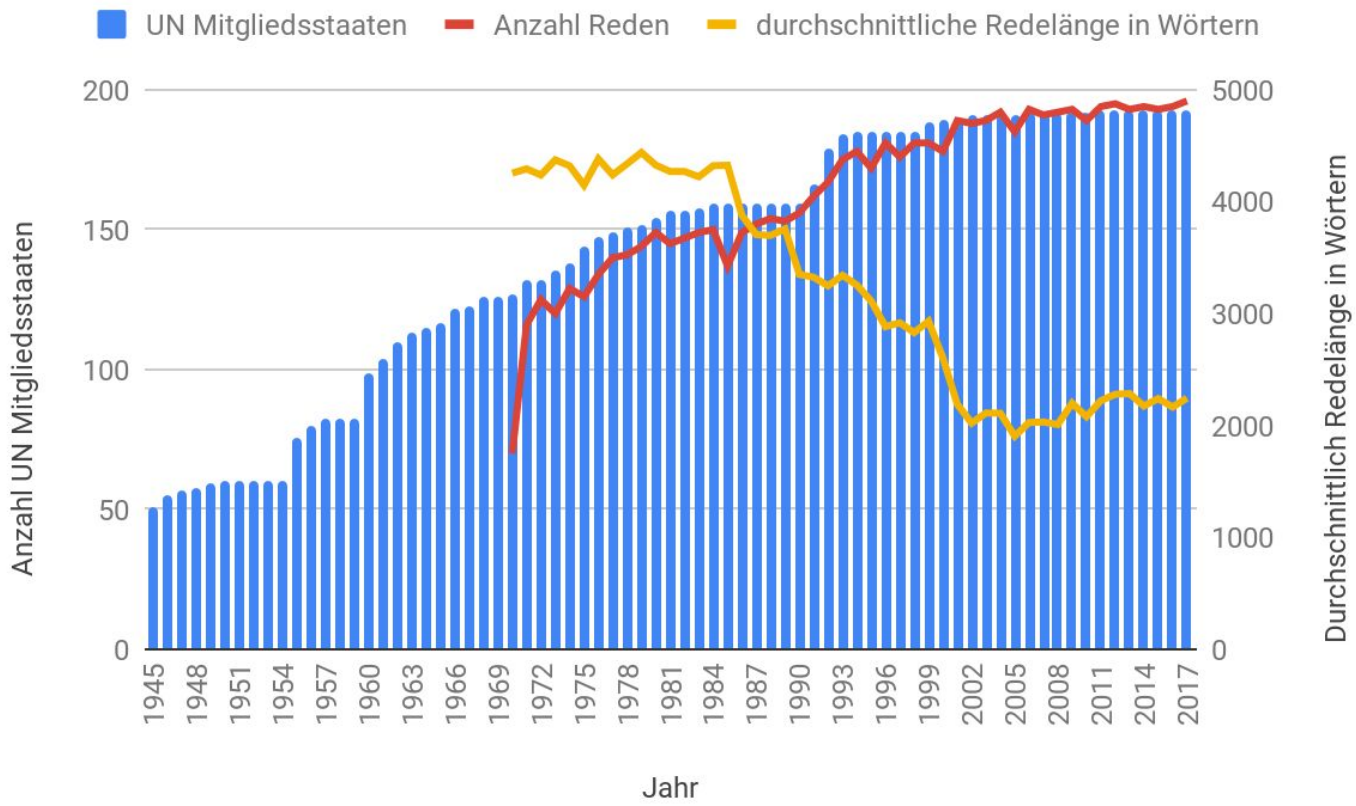
Recently, Russia's actions in Ukraine have challenged that post-war order. Here are the facts. After the people of Ukraine mobilized popular protests and calls for reform, their corrupt President fled. Against the will of the Government in Kyiv, Crimea was annexed by Russia. Russia poured arms into eastern Ukraine, fuelling violent separatists and a conflict that has killed thousands. When a civilian airliner was shot down from areas that those proxy forces controlled, those forces refused to allow access to the crash site for days. When Ukraine started to reassert control over its territory, Russia gave up the pretence of merely supporting the separatists and moved troops across the border.

Daten und Methoden

- Alle Reden von UN Mitgliedsstaaten an der UN-Generaldebatte, gehalten 1970-2017 (öffentlich zugänglicher maschinenlesbarer Korpus, Mikhaylov et al. 2017)
- Alle Reden: N=7897
- Über 23 Millionen Wörter, 827’595 Sätze

- Natural Language Processing (NLP) und netzwerkanalytische Verfahren

Beschreibung Daten



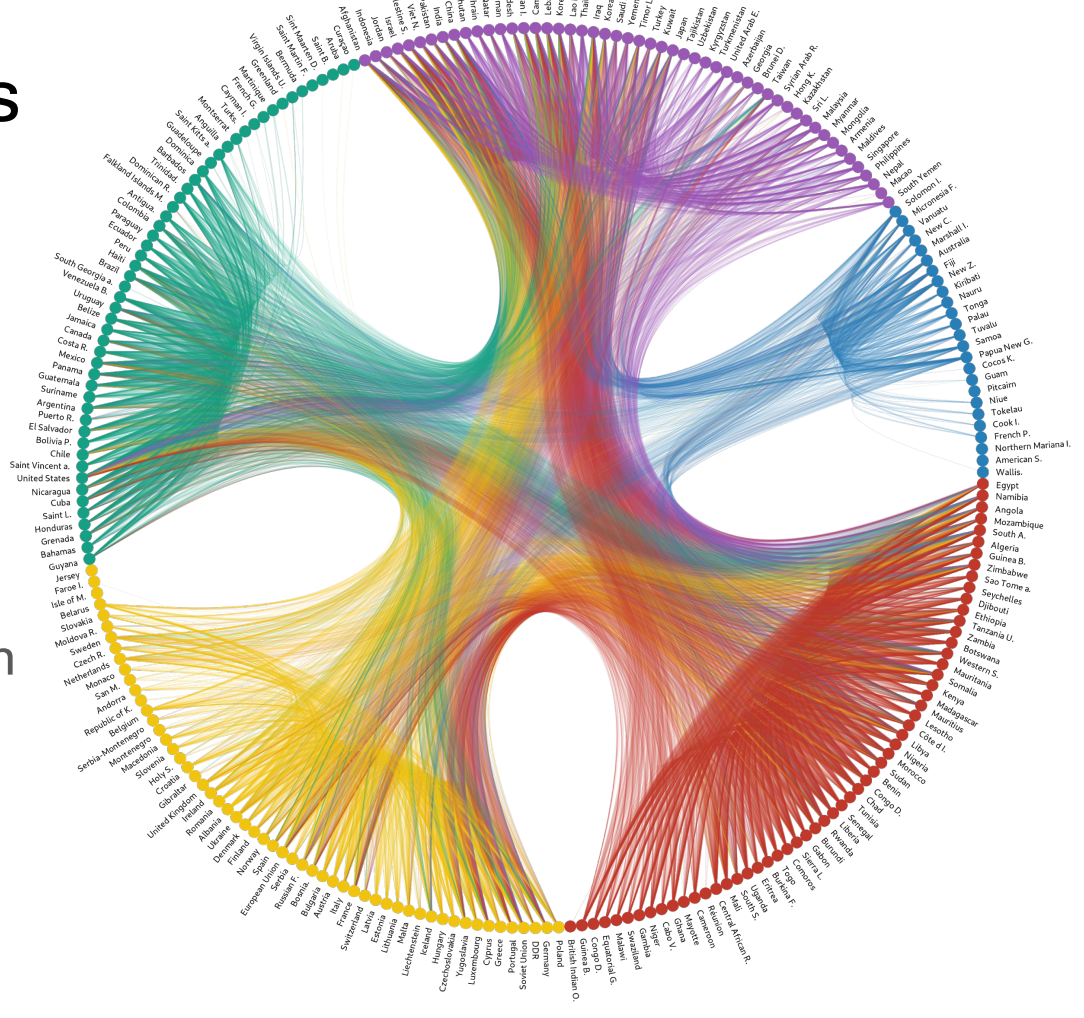
Ergebnisse

1. **Struktur** der Generaldebatten (Netzwerkanalysen)
2. **Themen** der Generaldebatten (LDA)
3. **Fokus**: Migration als Kategorie (word2vec)

Struktur: Mentionings

Wer erwähnt wen?

- Geopolitische Allianzen
- Rolle der USA
- Regionale Aufmerksamkeit
- (Ritualisierte Danksagungen und Begrüßungen)



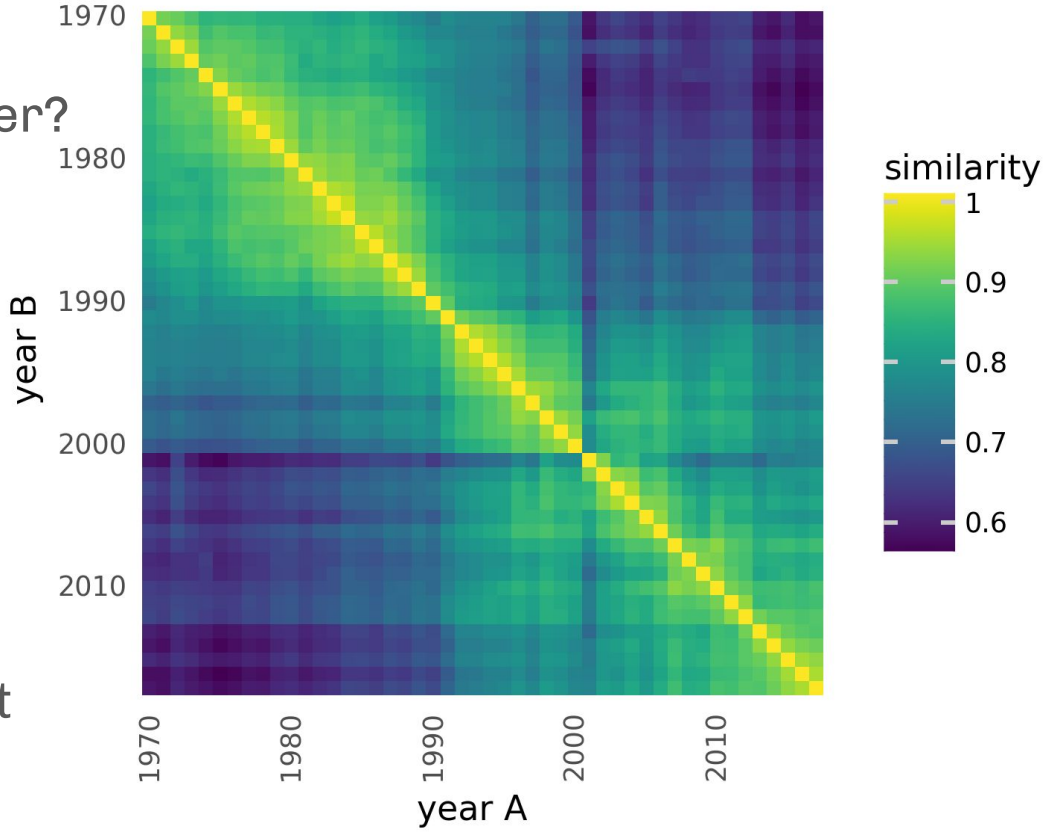
Struktur: Ko-Okkurrenzen

(Rule et al. 2015)

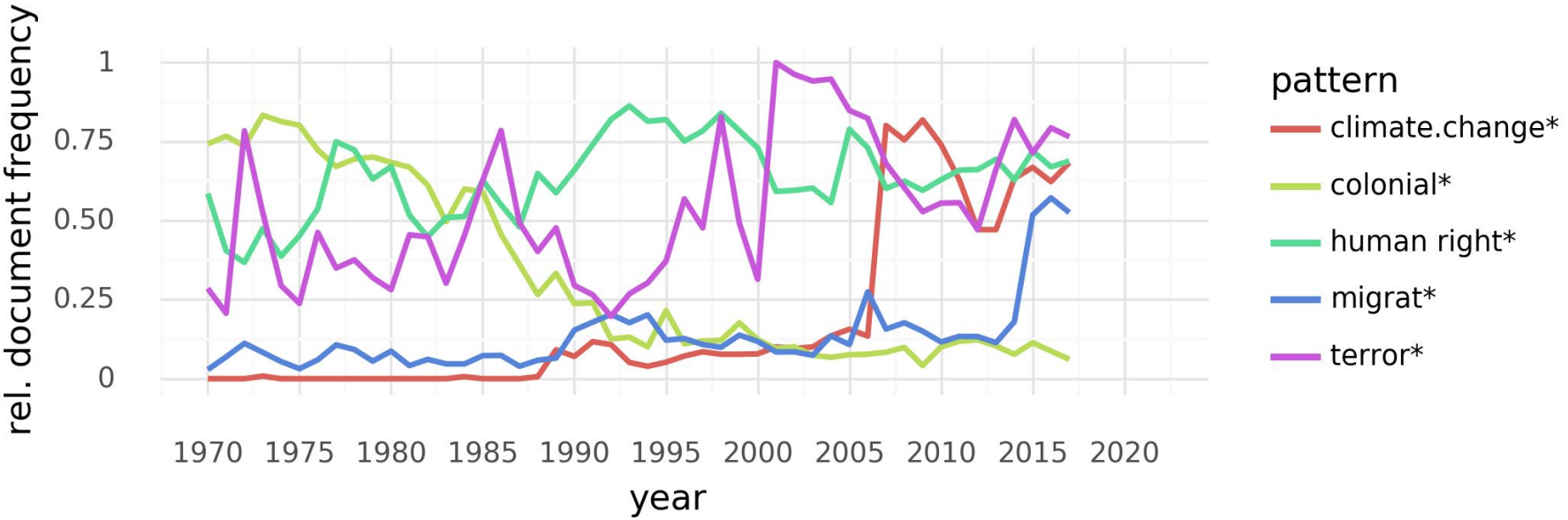
Wie ähnlich bzw. unähnlich sind die jährlichen Debatten zueinander?

Endogene Periodisierung
1975-1989
1990-2000
2001-2017

Umbruch 1990
2001 markant anders als andere Jahre, jedoch ähnlich wie Zukunft

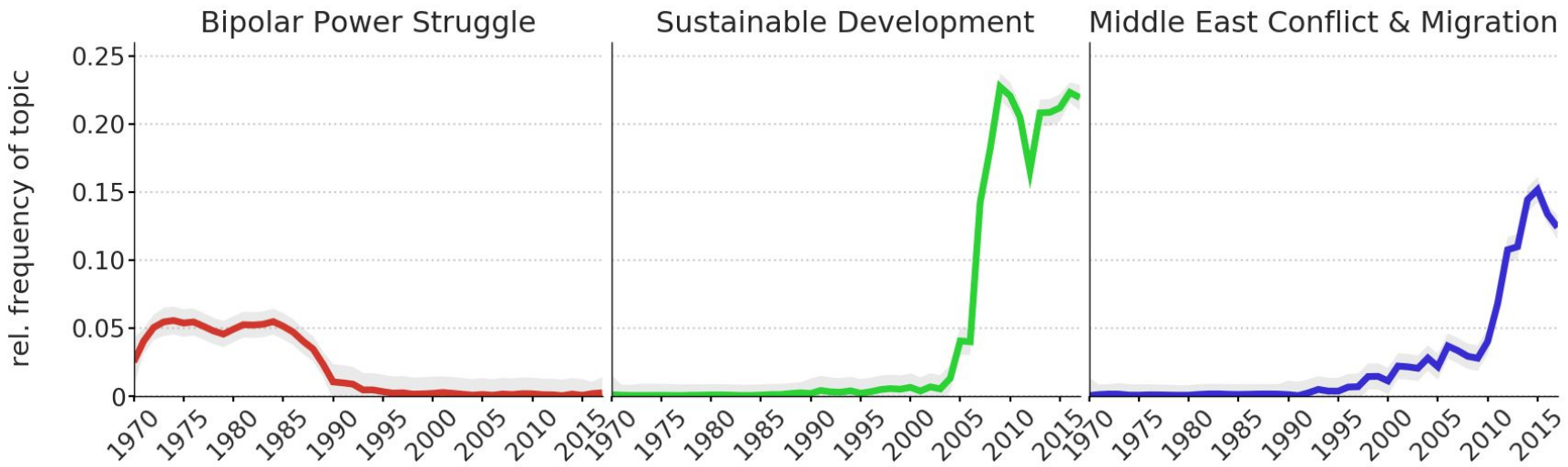


Themen: globale Herausforderungen



Themen: Topic Model

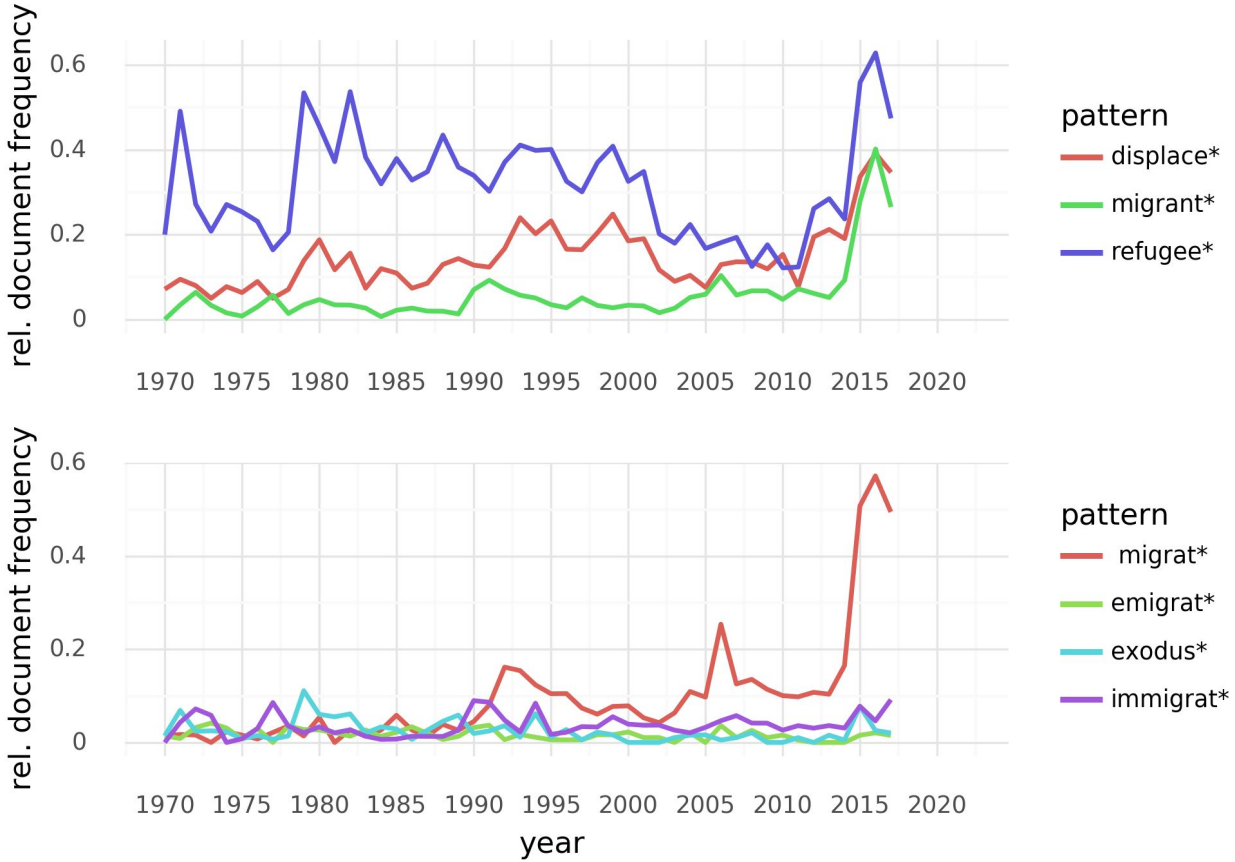
(LDA, Blei 2012)



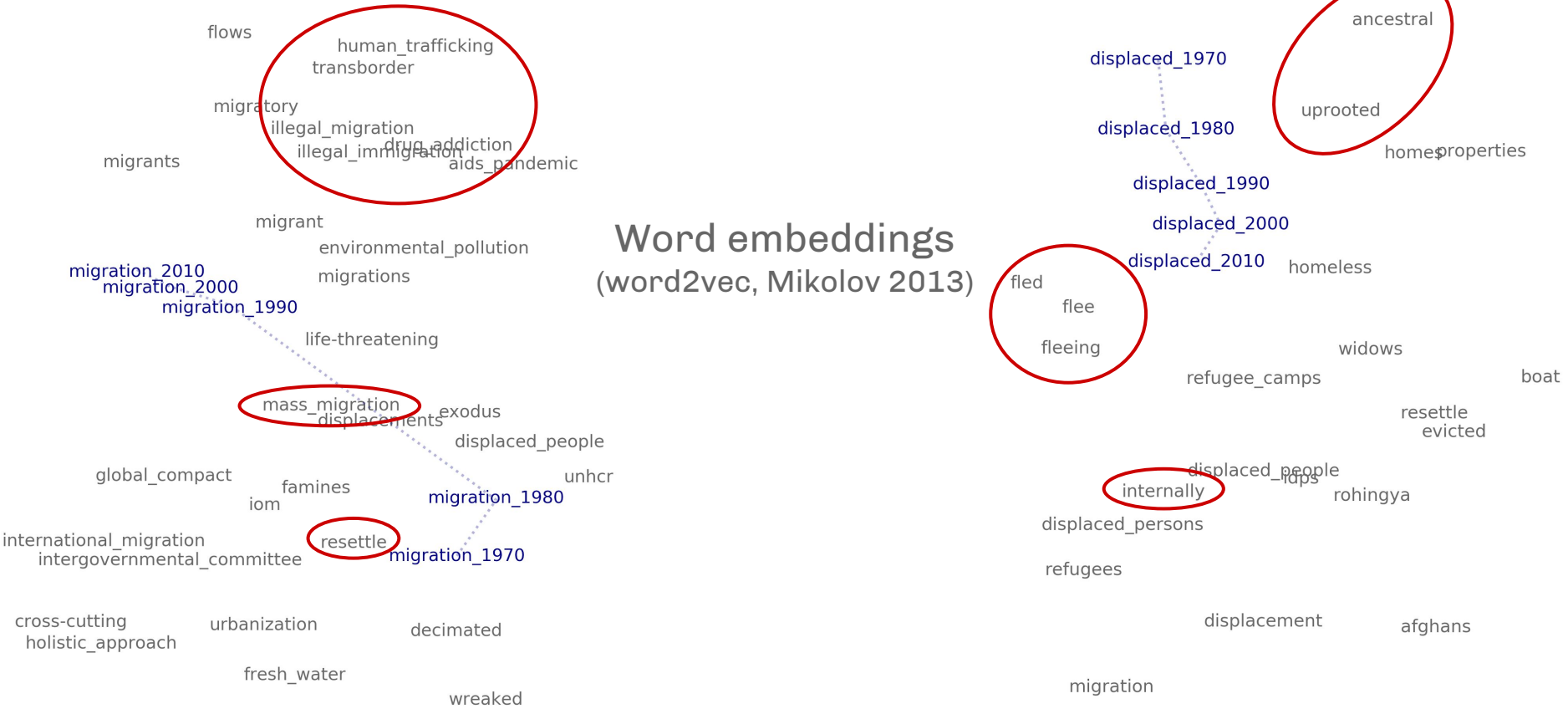
Top 5 Terme

Bipolar Power Struggle	Sustainable Development	Middle East Conflict & Migration
soviet_union	climate_change	syria
peoples	global	conflict
disarmament	agenda	terrorism
europe	sustainable_development	violence
policy	address	responsibility

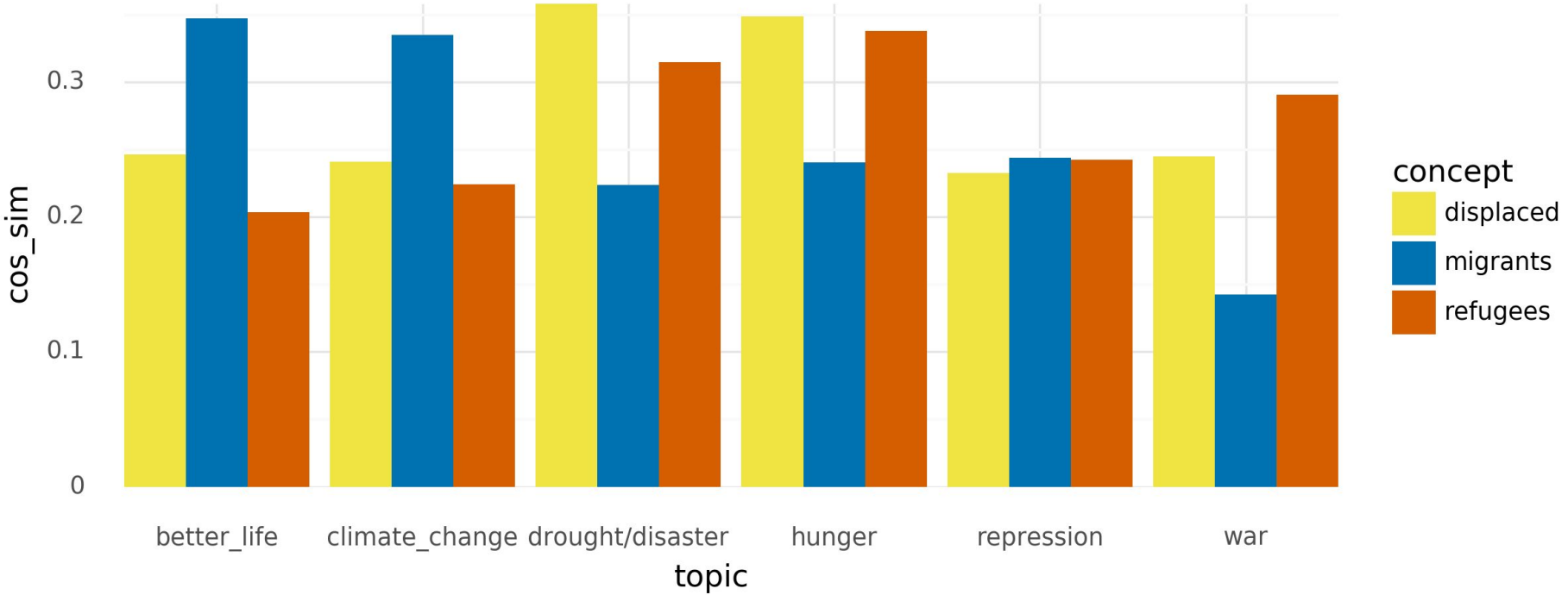
Fokus: Begriffskonjunkturen zu Migration



Fokus: Bedeutungswandel migration/displaced



Fokus: Kontextuelle Verwendung von Begriffen



Diskussion und Beitrag: Was konnten wir zeigen?

Einblicke in semantische Strukturen dieses globalen Forums

➔ Makroperspektive aufs Globale

➔ Globale Konstruktion von Konzepten

➔ Kategoriale und konnotative Verschiebungen von “Migration”

Methodisch

➔ Kombination von netzwerk- und textanalytischen Methoden

➔ Word Embeddings als wirkungsvolles Werkzeug für soziologische Analysen

Grenzen: neben Makroperspektive, Verknüpfung mit qualitativen Analysen notwendig

Vielen Dank!

sophie.muetzel@unilu.ch
alex.flueckiger@stud.unilu.ch

Präsentation und Daten auf:
https://github.com/aflueckiger/ungdc

Bibliographie

Alvarez, R. Michael (Hrsg.) (2016): Computational Social Science. Cambridge: Cambridge UP.
Baturo, Alexander, Niheer Dasandi und Slava J Mikhaylov (2017). "Understanding state preferences with text as data: introducing the UN General Debate Corpus." Research & Politics 4:1-9.
Bearman, Peter S. (2015): Big Data and historical social science. In: Big Data & Society, (2), http://bds.sagepub.com/content/2/2/2053951715612497.full.
Beckfield, Jason. 2010. "The Social Structure of the World Polity." American Journal of Sociology 115:1018-1068.
Bennani, Hannah. 2017. Die Einheit der Vielfalt: Zur Institutionalisierung der globalen Kategorie" indigene Völker". Frankfurt: Campus Verlag.
Blätte, Andreas/Joachim Behnke/Kai-Uwe Schnapp/Claudius Wagemann (Hrsg.). 2018. Computational Social Science. Die Analyse von Big Data. Baden-Baden: Nomos.
Blei, David M. 2012. Probabilistic Topic Models. In: Communications of the ACM, 55: 77-84.
DiMaggio, Paul/Nag, Manish/Blei, David (2013): Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. In: Poetics, 41: 570-606.
Grimmer, Justin/Stewart, Brandon M. (2013): Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. In: Political Analysis, 21: 267-297.
Hecht, Catherine. 2016. "The shifting salience of democratic governance: Evidence from the United Nations General Assembly General Debates." Review of International Studies 42:915-938.
Heintz, Bettina und Britta Leisering (Hrsg.). 2015. Menschenrechte in der Weltgesellschaft: Deutungswandel und Wirkungsweise eines globalen Leitwerts. Frankfurt: Campus Verlag.
Hoffman, Mark Anthony/Jean-Philippe Cointet/Philipp Brandt/Newton Key/Peter Bearman. 2018. "The (Protestant) Bible, the (printed) sermon, and the word(s): The semantic structure of the Conformist and Dissenting Bible, 1660–1780", in: Poetics 68: 89-103.

Kozlowski, Austin C./Matt Taddy/James A. Evans. 2018. "The Geometry of Culture: Analyzing Meaning through Word Embeddings", in: arXiv:1803.09288 [cs].Lazer, David/Pentland, Alex/Adamic, Lada/Aral, Sinan/Barabasi, Albert-Laszlo/Brewer, Devon/Christakis, Nicholas/Contractor, Noshir/Fowler, James H./Gutmann, Myron/Jebara, Tony/King, Gary/Macy, Michael/Roy, Deb/Alstyne, Marshall Van. 2009. Computational social science. In: Science, 323: 721-723.
McFarland, Daniel A./Lewis, Kevin/Goldberg, Amir (2016): Sociology in the Era of Big Data: The Ascent of Forensic Social Science. In: The American Sociologist, 47: 12-35.
Merz, Nicolas, Sven Regel und Jirka Lewandowski. 2016. "The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis." Research & Politics 3:1-8.
Mikhaylov, Slava/Baturo, Alexander/Dasandi, Niheer. 2017, "United Nations General Debate Corpus", https://doi.org/10.7910/DVN/0TJX8Y, Harvard Dataverse, V4
Mikolov, Tomas/Sutskever, Ilya/Chen, Kai/Corrado, Greg S/Dean, Jeff (2013): Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.
Mohr, John W./Bogdanov, Petko. 2013. Introduction—Topic models: What they are and why they matter. In: Poetics, 41: 545-569.
Pomeroy, Caleb, Niheer Dasandi und Slava Mikhaylov. 2018a. "Disunited Nations? A Multiplex Network Approach to Detecting Preference Affinity Blocs using Texts and Votes." arXiv preprint arXiv:1802.00396.
Pomeroy, Caleb, Niheer Dasandi und Slava Mikhaylov. 2018b. "Multiplex Communities and the Emergence of International Conflict." arXiv preprint arXiv:1806.00615.
Rule, Alix/Cointet, Jean-Philippe/Bearman, Peter S. (2015): Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014. In: Proceedings of the National Academy of Sciences, 112: 10837-10844.

Appendix

Original Data Sources

- UN Corpus General Assembly
  https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/0TJX8Y
- Countries Dataset
  https://restcountries.eu/

Python Packages

- Natural Language Processing
  - spaCy: https://github.com/explosion/spaCy
  - Textacy: https://github.com/chartbeat-labs/textacy
- Topic Modeling
  - MALLET: https://github.com/mimno/Mallet
- word2vec
  - Gensim: https://github.com/RaRe-Technologies/gensim
- Network Analysis
  - Graph-Tool: https://github.com/antmd/graph-tool
  - Networkx: https://github.com/networkx/networkx
- Machine Learning (general)
  - Scikit-Learn: https://github.com/scikit-learn/scikit-learn
- Visualization (general)
  - Plotnine: https://github.com/has2k1/plotnine

Wieso computergestützte Textanalyse?

Formalisiert (explizite Annahmen) = reproduzierbar = kritisierbar
Effizienter & kollaborativer Analyseprozess (open-data, open source)
Endogenität von Analyse (data-driven)
Visualisierung sehr nützlich für Komplexitätsreduktion
Textmenge tlw. zu gross

Subtile Konzepte wie semantische Nähe/Konnotation greifbar machen (oft schlechte Intercoder-Reliabilität)

Semantische Verschiebungen über Zeit

	Wort	**Cosinus-Ähnlichkeit**
Stärkste Drifts (instabile Bedeutung)	islamic_republic_of_iran Iraq **climate** yugoslavia somalia taiwan **transparent** deliver **worldwide** ...11 Terme entfernt... agenda **migration**	0.324 0.354 **0.373** 0.392 0.402 0.406 **0.409** 0.413 **0.419** ... 0.471 **0.472**
Schwache Drifts (stabile Bedeutung)	has is have was warmest_congratulations gross_national_product	0.907 0.898 0.892 0.892 0.891 0.884

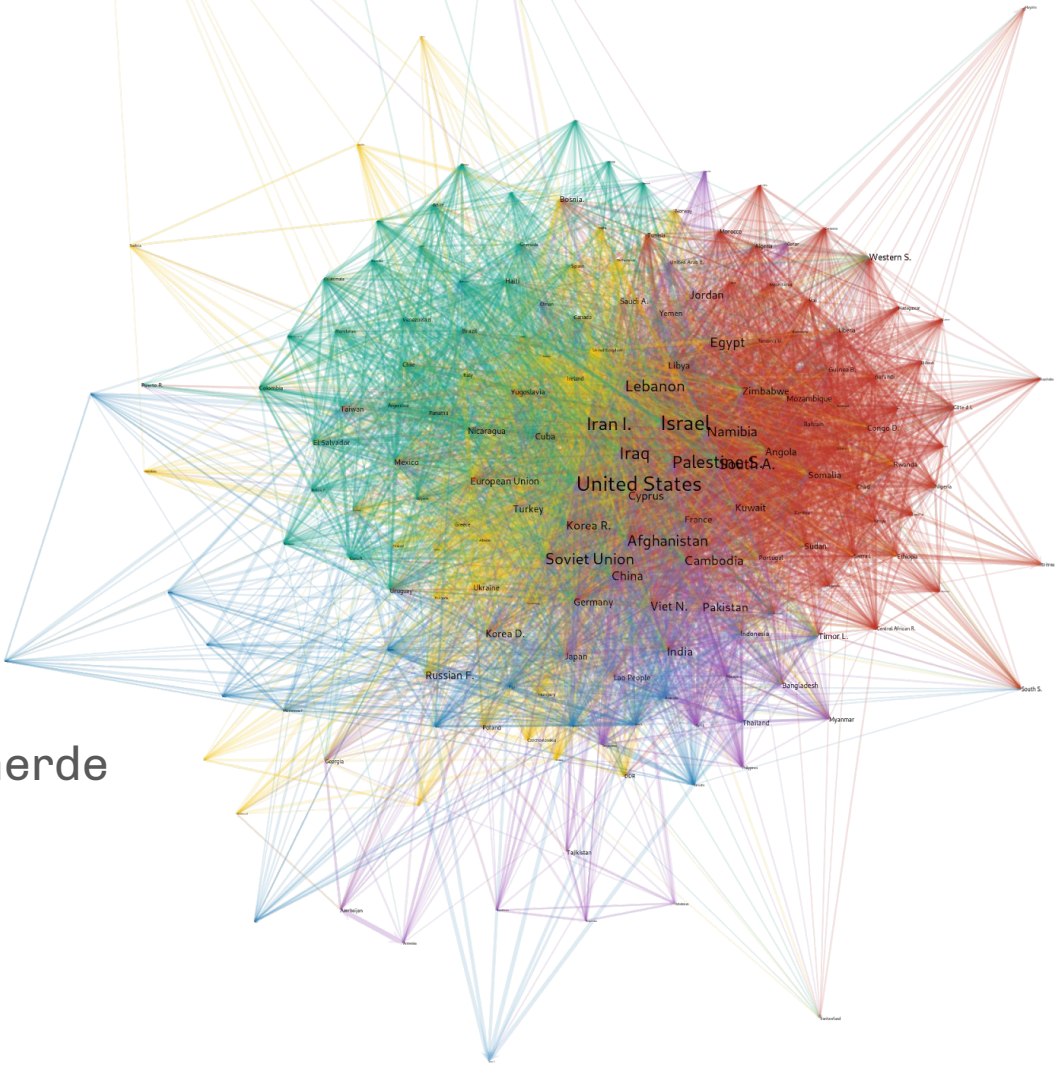
Fokus: Bedeutungswandel immigration

Word embeddings
(word2vec, Mikolov 2013)

Struktur: Mentionings

Wer erwähnt wen?

- Bipartiter Graph & Eigenvektorzentralität
- Geopolitische Allianzen
- Regionale Aufmerksamkeit
- Hegemonialstaaten & Konfliktherde



Preprocessing Korpus

OCR-Korrekturen
Phrasen zusammenfügen (nach Häufigkeit, POS- und NER-Tags)
Phrasenmatcher für Ländernamen (inkl. alternativen Schreibweisen)

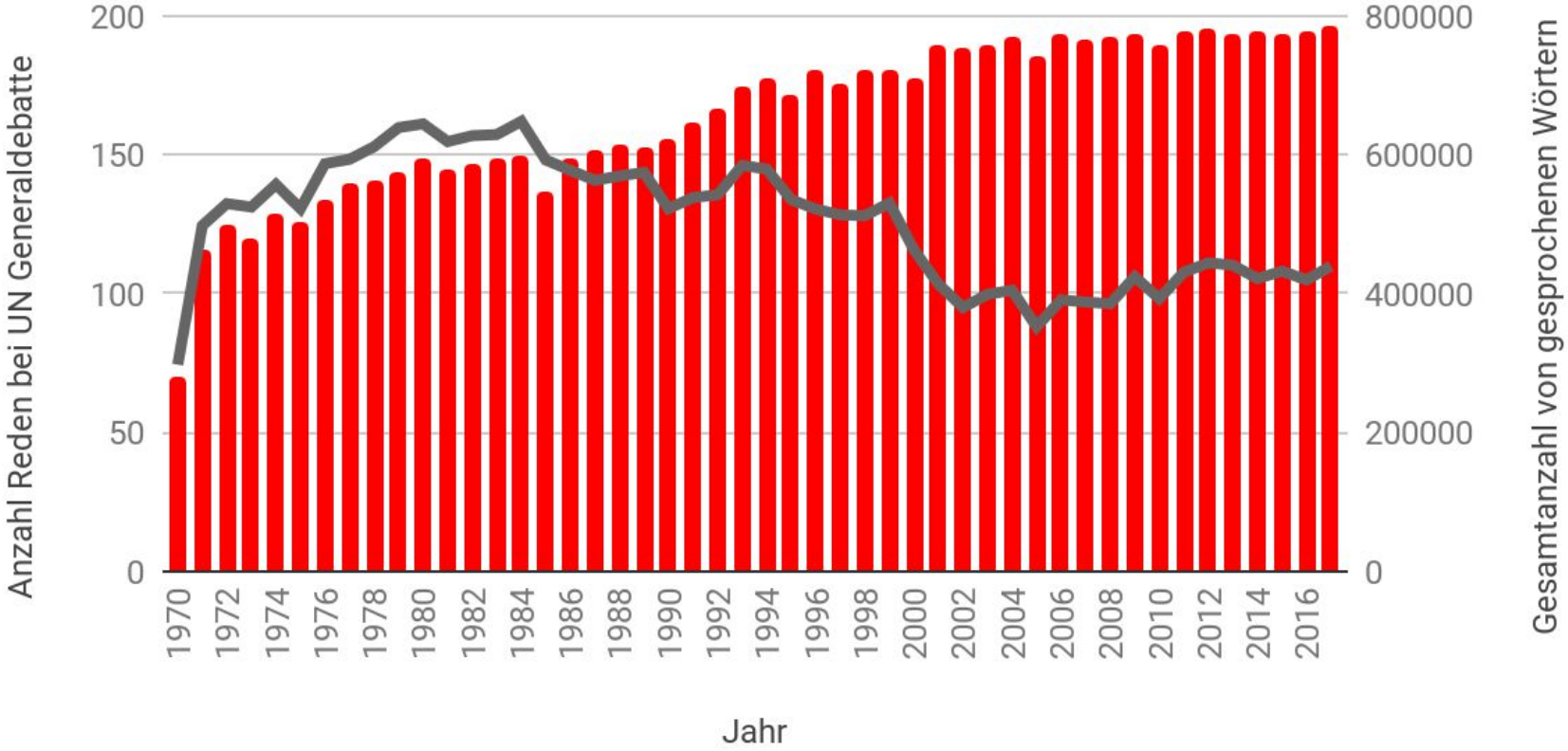
Modelle: Parameter & Evaluation

Year-to-Year Heatmap
Kleinschreibung, keine Stopwörter/Interpunktion/Zahlen, Filterung (min. DF=10, max. DF=0.5), lineare Textlängenormalisierung, kein IDF, Cosine-Similarity

Word2vec
Keine Interpunktion, Platzhalter für Zahlen, Filterung (min. DF=5), Modell-Alignierung über Procrustes-Transformation, Evaluation mithilfe von Analogie-Task, [algorithm=SGNS, size=200, negative=25, window=10], TSNE-Visualisierung

asymmetric LDA
Keine Interpunktion/Zahlen, Filterung (min. DF=10, max. DF=0.7),
Evaluation mithilfe von Kohärenz-Metrik C_V
[n_topics = 37, iterations=2000, optimize_interval = 50, alpha=50]

Anzahl der Reden bei UN Generaldebatte und Anzahl der Wörter aller Reden pro Jahr, 1970-2017



Häufigste Tokens (Top 20)

term	frequency
united nations	78935
world	62079
peace	54204
countries	44725
people	42652
development	40184
new	39923
states	38714
country	34875

term	frequency
security	32302
economic	30936
government	29119
organization	28974
efforts	27822
support	27502
international community	26488
international	25992
general assembly	25207