School of Applied Linguistics

Digital Discourse Lab

Korpuszentrierte Diskursanalyse: Diskursmodellierung und quantitative Analyseansätze

Philipp Dreesen & Julia Krasselt

Zürcher Hochschule für Angewandte Wissenschaften

Methodologien der Quantitativen Sozialforschung, 17.02.2022

Outline

1. Theorierahmen und methodologische Ausrichtung
2. Formen der Quantifizierung
3. Diskussion

1. Theorierahmen und methodologische Ausrichtung

School of Applied Linguistics

Digital Discourse Lab

Ansatz *Diskurslinguistik in Anwendung* (DIA):

A. *in Anwendung*:
- Community of Practice aus Forschenden und Praxispartner:innen im Sinne von transformativem Forschungsverständnis zu gesellschaftlich relevanten real world problems
- Überführung von Praxisproblem in Forschungsfrage
- Beispiel:
Kuratorium für Verkehrssicherheit ist unsicher, wie Österreich über automatisiertes Fahren denkt.
> FF: Wie wird in öffentlichen Diskursen Österreichs der Gegenstand automatisiertes Fahren sprachlich konstruiert und kontextualisiert?

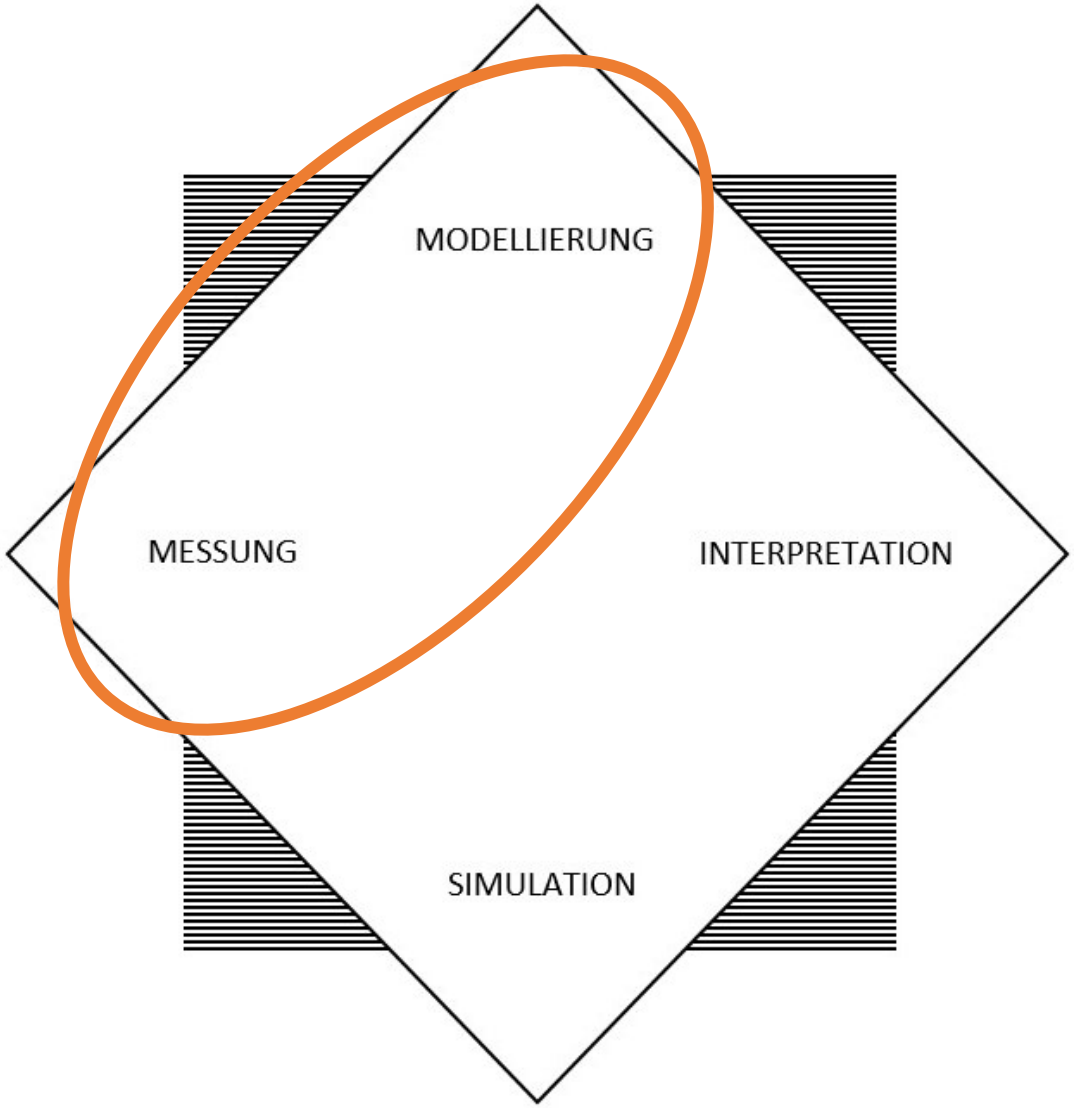
B. *Diskurslinguistik*:
- Diskursanalyse (sensu F. Foucault, R. Keller) als
- Diskurslinguistik (sensu I.H. Warnke, N. Bubenhofer)
> Soziale Sinnherstellung in Wissens- und Sagbarkeitsordnungen durch Sprachgebrauchsmuster

KFV (Kuratorium für Verkehrssicherheit) (2021): Automatisiertes Fahren in Österreichs medialem Diskurs. Wien. (KFV – Sicher Leben 28).
Dreesen, Philipp/Stücheli-Herlach, Peter (2019): Diskurslinguistik in Anwendung. Ein transdisziplinäres Forschungsdesign für korpuszentrierte Analysen zu öffentlicher Kommunikation. In: Zeitschrift für Diskursforschung 7(2), S. 123–162.

1. Theorierahmen und methodologische Ausrichtung

School of Applied Linguistics

Digital Discourse Lab



Vortragsschwerpunkt auf quantitative Aspekte

Modul 1: Modellierung
Soziale Zeichenstrukturen werden als Diskurs modelliert, indem Texte relevanter Akteure nach Parametern (Gegenstand, Zeit, Sprache, Medien etc.) zusammengestellt werden. > Korpuskompilierung

Modul 2: Messung
Korpuszentrierte Analyse heisst, grosse Textvolumen insbesondere datengetrieben zu analysieren. > z.B. Analysen von Frequenzen und Kontexten im Sinne von Prämissen

a) der Diskurstheorie:
   Serien verweisen auf Sagbarkeiten
b) distributionellen Semantik:
   Kontexte prägen Bedeutung

2. Formen der Quantifizierung

Datengetriebene Analyse von Sprachgebrauchsmustern

1) Kollokation & Word Embedding
2) Keywords
3) Themen und Akteure

Daten

Schweizerischer Nationalfonds

zh aw School of Applied Linguistics

Digital Discourse Lab

- SNF-Projekt «COVIDisc: A Focus on Vector Populations»
- linguistisch annotiertes Korpus zum COVID-19-Diskurs in der Schweiz
  - Sprachen: deutsch (und italienisch)
  - Zeitraum: 01/2020–10/2020
  - suchwortbasiert *(corona, covid, sars.*cov)*
  - Untersuchte Akteursgruppen:
    - Bund, Kantone, Behörden
    - Gesundheitswesen
    - Berufsverbände
    - Universitäten und Fachhochschulen
    - NGOs
    - Journalistische Medien
  - Grösse: 55k Texte, 29 Mio. Wörter (deutschsprachiges Korpus)
- Computerlinguistische Verabeitungspipeline (Datenbeschaffung und -vorverarbeitung, Mehrebenenannotation, Metadatenerkennung)

Kollokationen & Word Embeddings

School of Applied Linguistics

Digital Discourse Lab

Forschungsteilfrage

Wie wird die Gruppe der 15- bis 34jährigen im Diskurs konstruiert?

Operationalisierung

- Datengeleitete Identifikation von Versprachlichungen für die Vektorpopulation
- Identifikation von typischen (signifikanten) Kontexten, in denen diese Versprachlichungen vorkommen
- Distribution dieser Versprachlichungen über Metadaten (z.B. Zeit & Akteursgruppen)

Kollokationen

School of Applied Linguistics

Digital Discourse Lab

text_source	text_year_month	left	node	right
blick	2020_10	eigentlich recht . »Sie selbst sei eher schüchtern gewesen als	Jugendliche	und hätte wohl kaum an solch einem Anlass wie hier
srf	2020_08	Das heisst , erstmals seit März gehen dann Kinder und	Jugendliche	wieder zur Schule . «Auch das wird für das ganze
grenchnertagblatt	2020_05	. 43. Emerson (Foul) . Zumindest Kinder und	Jugendliche	schienen bislang von schweren Erkrankungen durch das Coronavirus verschont zu
srf	2020_03	Staat sicher aufgehoben seien . Zürich , 27.04.20 – Zürcher	Jugendliche	erhalten mehr Zeit bei der Lehrstellensuche : Der Kanton Zürich
srf	2020_03	Zeitpunkt in früheren Jahren . Dank der verlängerten Frist können	Jugendliche	auch nach dem 17. August einen Lehrvertrag unterschreiben . Dies
srf	2020_03	Berufsfachschulen an einem Konzept , wie dies möglich ist .	Jugendliche	, die auf ihren Lehrantritt warten , sollen beispielsweise vorübergehend

	Schule	Komplikation	verschont	Lehrstelle	Frist	unterschreiben	Schule	...
Jugendliche	10	1	5	47	30	12	27	...

Evert (2008), Baker (2006)

Kollokationen

School of Applied Linguistics

Digital Discourse Lab

Kollokationen für *Jugendliche*; Log-Likelihood-Ratio > 3.84 ($p < 0.05$)

collocate	freq.collocate	freq.collocation	exp.freq	logLik	rankLL
Kind	17003	782	16.1891311775568	4589.11458620837	1
und	554484	1368	527.94296370384	952.226233405326	2
Erwachsene	1319	99	1.25586449586528	677.015876626306	3
Lehrstelle	485	50	0.461784898024763	374.726267630776	4
für	211293	446	201.179209192466	222.956408898361	5
Lehrstellensuche	70	21	0.0666493667252235	206.773885474772	6
Eltern	5847	53	5.56712638917688	144.46300506502	7
Schule	10142	65	9.65654110467452	137.603647464429	8
besonders	10320	63	9.82602092291866	128.148978962499	9
Alter	2454	34	2.33653637062426	119.202746260491	10
wählerisch	27	10	0.025707612879729	103.577821927054	11

https://swiss-al.linguistik.zhaw.ch/rsconnect/workbench/

Word Embeddings

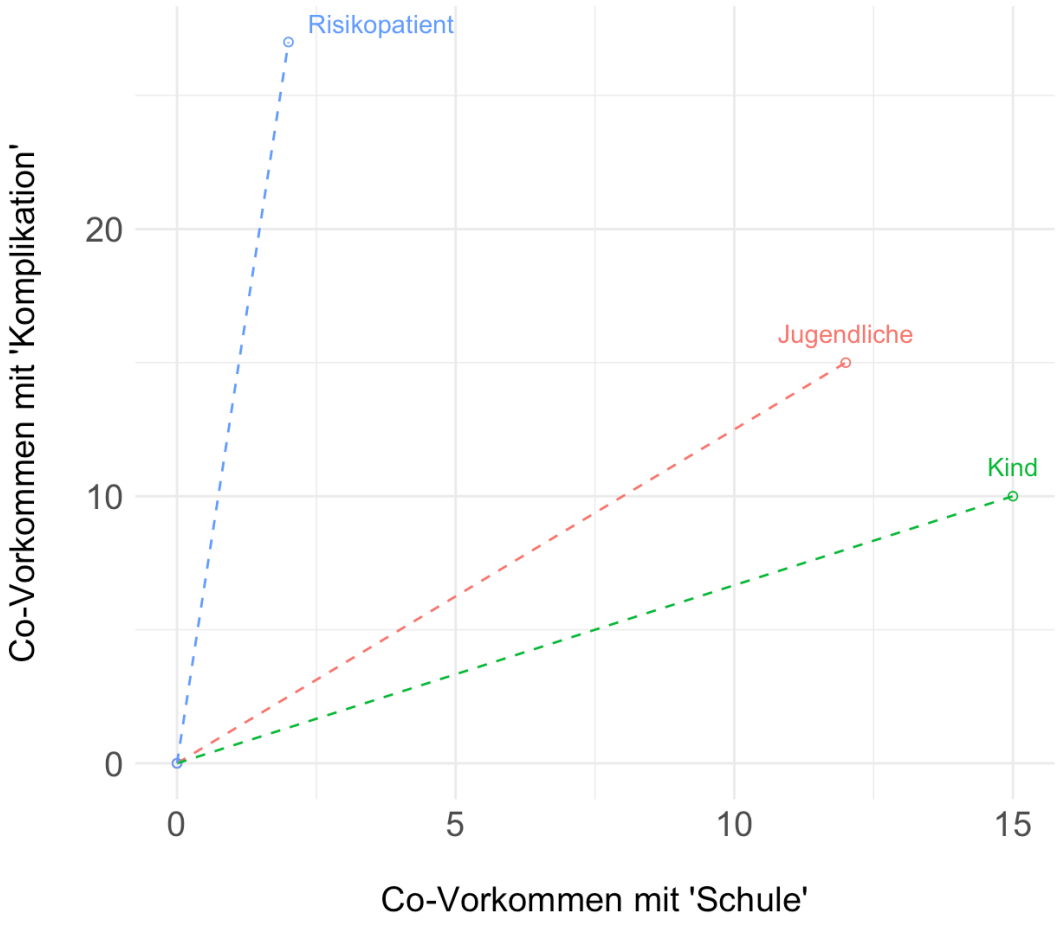
School of Applied Linguistics

Digital Discourse Lab

	Schule	Komplikation	...
Jugendliche	15	25	...
Kind	17	23	
Risikopatient	1	45	

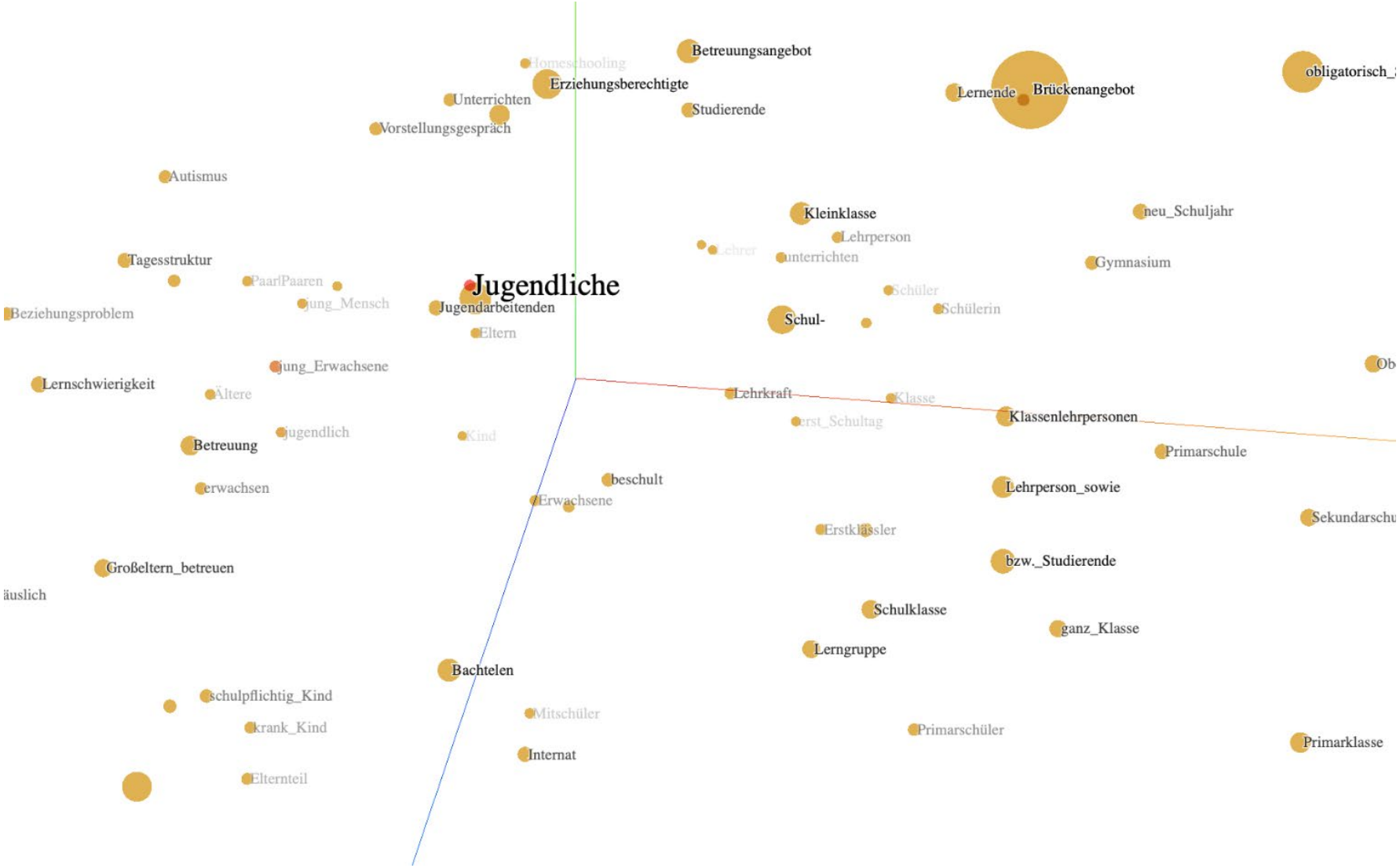
...

- Geometrische Repräsentation des Korpusvokabulars in einem n-dimensionalen Vektorraum
- Wörter mit ähnlichen distributionellen Vektoren (d.h. mit ähnlichen Kontexten) sind sich semantisch ähnlich



Lenci (2018), Mikolov et al. (2013), Bubenhofer (2020)

Word Embeddings



junge Erwachsene
jugendlich
jung Mensch
Kind
Ältere
Erwachsene
Eltern
Paar
Beziehungsproblem
erwachsen
Jugendarbeitende
häuslich
Berufslehre
Klasse
Lehrer
Schüler
...

Word Embeddings

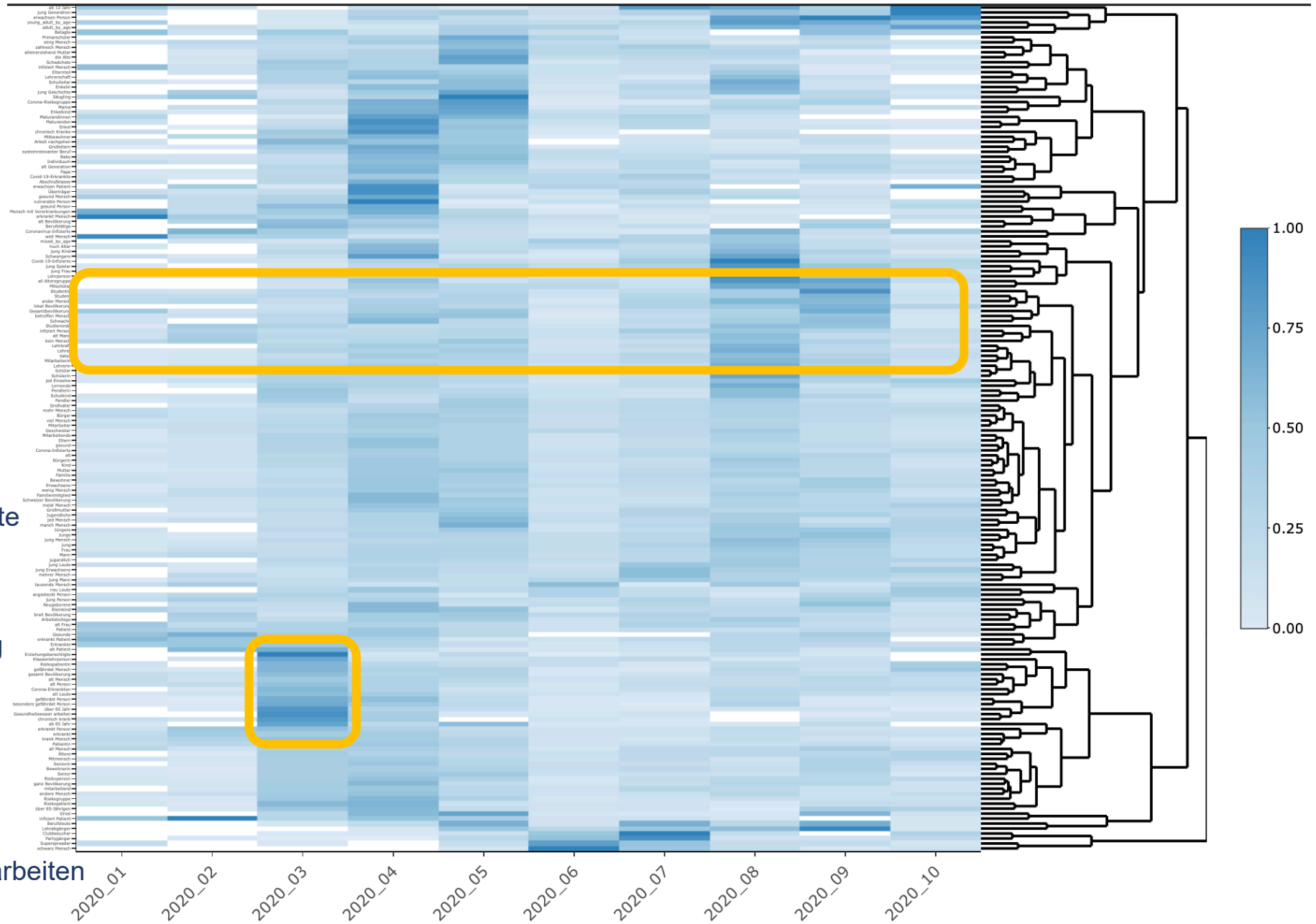
School of Applied Linguistics

Digital Discourse Lab

	2020_01	2020_02	2020_03	2020_04	2020_05	2020_06	2020_07	2020_08	2020_09	2020_10
Jugendliche	0.63	0.52	16.57	14.54	18.5	10.69	14.24	10.84	9.58	3.88
jugendlich			21.65	15.46	10.82	13.92	11.86	12.89	8.25	5.15
jung Mensch	0.93	1.32	19.87	11.13	11.13	12.85	12.05	14.83	11.92	3.97
jung Frau	0.38	4.75	11.79	10.84	9.89	10.65	16.92	17.49	14.83	2.47
jung Leute	1.86	1.06	18.83	12.2	9.28	10.34	21.75	13	8.49	3.18
jung Mann	0.2	3.79	15.97	12.38	7.58	19.56	19.56	10.18	7.39	3.39
jung Generation	1.4		8.39	13.29	13.99	11.19	13.29	13.99	9.79	14.69
jung Erwachsene		0.57	13.22	14.37	9.2	11.49	24.14	10.92	9.2	6.9
jung Spieler	0.55		8.84	9.94	12.15	13.81	15.47	22.1	12.71	4.42
jung Person	2.04	7.14	22.45	15.31	7.14	11.22	14.29	5.1	11.22	4.08

Lehrperson
Mitschüler
Student
Studentin
Studierende
Lehrkraft
Lehrer
Lehrerin

Alt Patient
Erziehungsberechtigte
Klassenlehrperson
Risikopatientin
Gefährdet Mensch
Gesamt Bevölkerung
Alt Mensch
Alt Person
Corona-Erkrankten
Alt Leute
Gefährdet Person
Über 65 Jahr
Gesundheitswesen arbeiten
Chronisch krank
Ab 65 Jahr



Keywords

- Forschungsteilfragen:

1) Welche datengeleitete Binnenperiodisierung für den Zeitraum 01/2020 bis 10/2020 ergibt sich hinsichtlich des Gebrauchs von Vokabular?
2) Welche Schlussfolgerungen lassen sich daraus in Bezug auf die Entwicklung eines Common Ground zu Covid-19 ziehen?

- Ermittlung signifikanter Lexik durch Vergleich zweier Korpora (Welche Wörter werden in Korpus A signifikant häufiger/seltener gebraucht als in Korpus B?)
- Korpus A und B stehen in einem für die jeweilige Fragestellung spezifischen Verhältnis zueinander (z.B. temporal, semantisch, akteursgruppenspezifisch, quellenspezifisch)

Gabrielatos (2018), Baker (2006)

Keywords

1. Keywords pro Woche
2. Ähnlichkeit zwischen Wochen basierend auf Keywords (via Jaccard Similarity Distance)
3. Hierarchisches, agglomeratives Clustering (Baayen 2008, Gries 2009, Manning et al. 2008)

keyword	DIFF	BIC	week
Sars	14475.6	110.6	3
Sars	5962.6	76.7	4
Sars	13379.8	959.2	3
Maskenpflicht	70.3	29.1	26
Maskenpflicht	[illegible]	[illegible]	[illegible]
…			

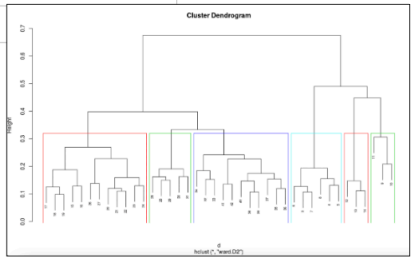
1) Calculation of keywords per week

	Sars	Maskenpflicht	…
3	1	0	
4	1	0	
…			
26	0	1	
…			

2) Matrix (rows = weeks, columns = keywords)

	1	2	3	…	42
1	0.0	0.11	0.19	…	0.27
2	0.11	0.0	0.16	…	0.33
3	0.19	0.16	0.0	…	0.46
…	…	…	…	…	…
42	0.27	0.34	0.46		

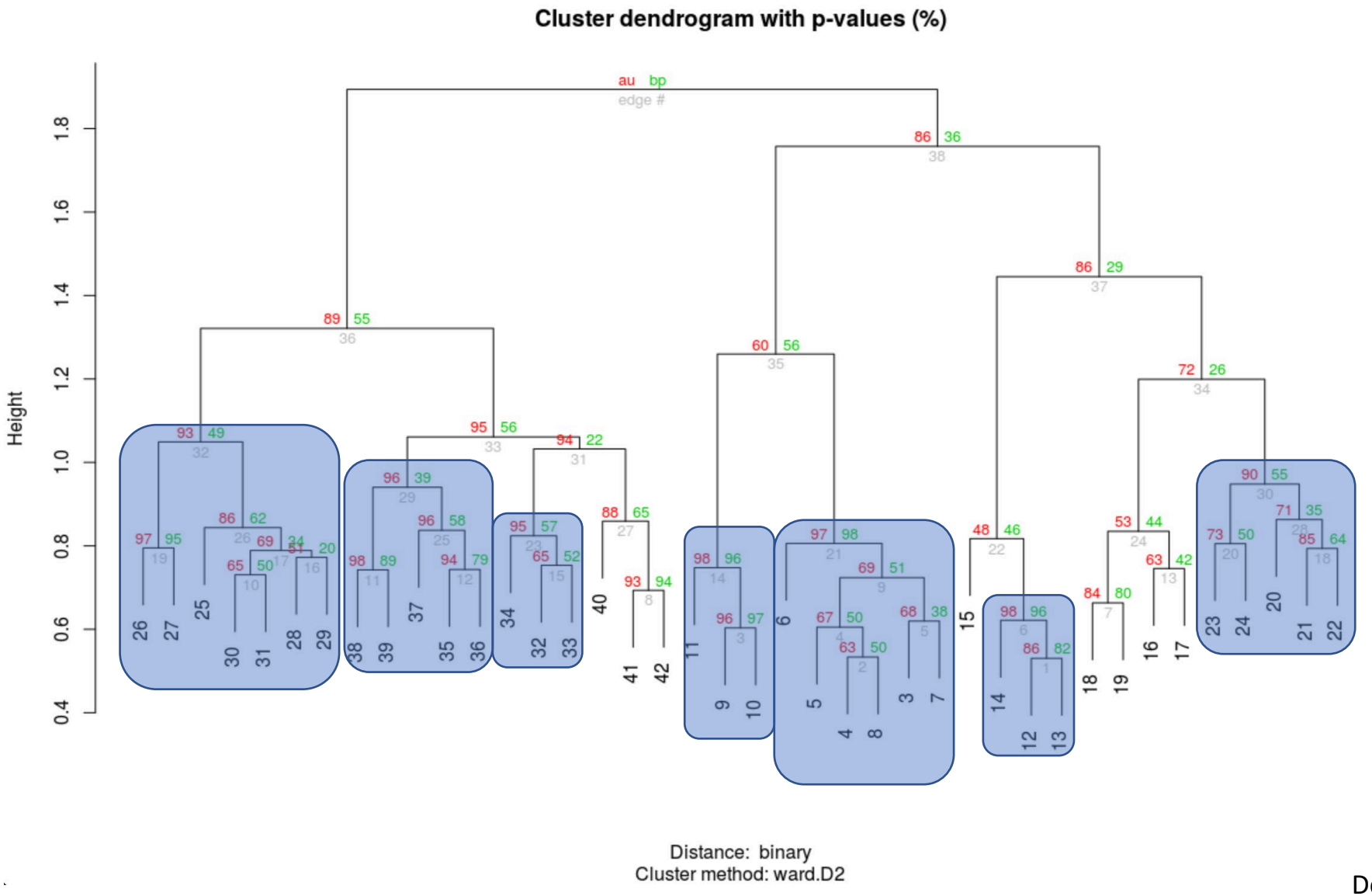
3) Similarity matrix via simple matching coefficient



4) Agglomerative Clustering (ward D2)

School of Applied Linguistics

Digital Discourse Lab



- 7 Cluster mit AU > 90

1. Woche 3-8 (Cluster 21)
2. Woche 9-11 (Cluster 14)
3. Woche 12-14 (Cluster 6)
4. Woche 21-24 (Cluster 30)
5. Woche 25-31 (Cluster 32)
6. Woche 32-34 (Cluster 23)
7. Woche 35-38 (Cluster 29)

Desagulier (2014), Divjak & Fieller (2014)

Keywords

Woche 3-8 (Cluster 21)	Woche 9-11 (Cluster 14)	Woche 12-14 (Cluster 6)	Woche 21-24 (Cluster 30)
Wuhan	,	wir	Rösti
China	zu	Sie|sie	SBB
Lungenkrankheit	:	Corona-Krise	Moderna
Mensch	können	unser	Gesetz
Patient	(	Klühr	Floyd
Fall	)	ich	Schwarze
Erreger	Kanton	Netflix	Protest
WHO	oder	ihr	Überbrückungsrente
Virus	sollen	Wohnung	Briefwahl
chinesisch	Person	Zeit	Polizeigewalt
neuartig	-	nicht	Kuster
Südkorea	März	Leben	Plünderung
zentralchinesisch	bleiben	Moment	Rassismus
Erkrankung	ab	Woche	schwarz
Sars	gegen	Balkon	Beirut
Thailand	dürfen	Distancing	SVP
Infektion	Bundesrat	wenn	Auto
bestätigen	jetzt	!	Libanon
sterben	April	Coronavirus-Pandemie	Biden
Weltgesundheitsorganisation	@ord@	Krise	Zuwanderung
Chinese	gelten	man	Nadal
Japan	Kind	?	Vorlage
Reisende	Bund	Miete	Jahr
Ausbruch	Schule	mein	Sion
Verdachtsfall	Haus	sie	Referendum
Gesundheitskommission	halten	Abstand	Amerika
Krankheit	Situation	Boeing	Stimme|Stimmen
Sars-Virus	Franke	täglich	3,3
neu	Mai	Ausgangsbeschränkung	Federer

Themen & Netzwerke

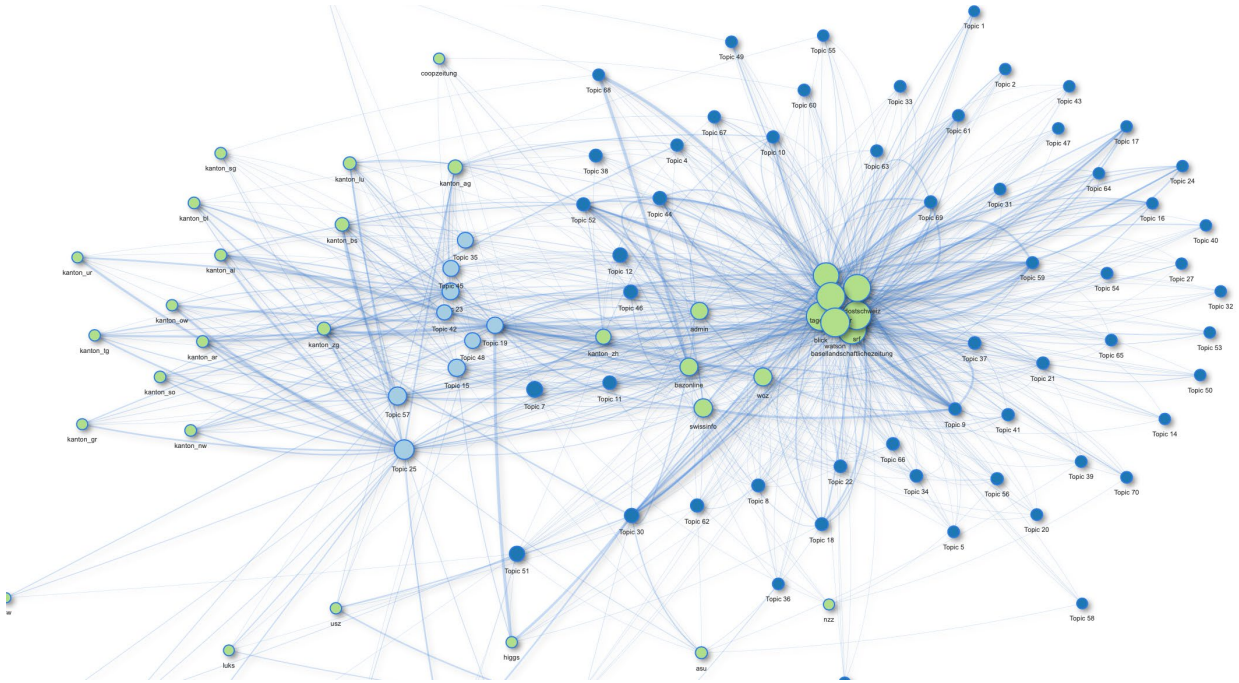
- Forschungsteilfragen: Welche Themen können im Diskurs identifiziert werden? In welcher Frequenz werden diese Themen von einzelnen Akteuren/Akteursgruppen verwendet?

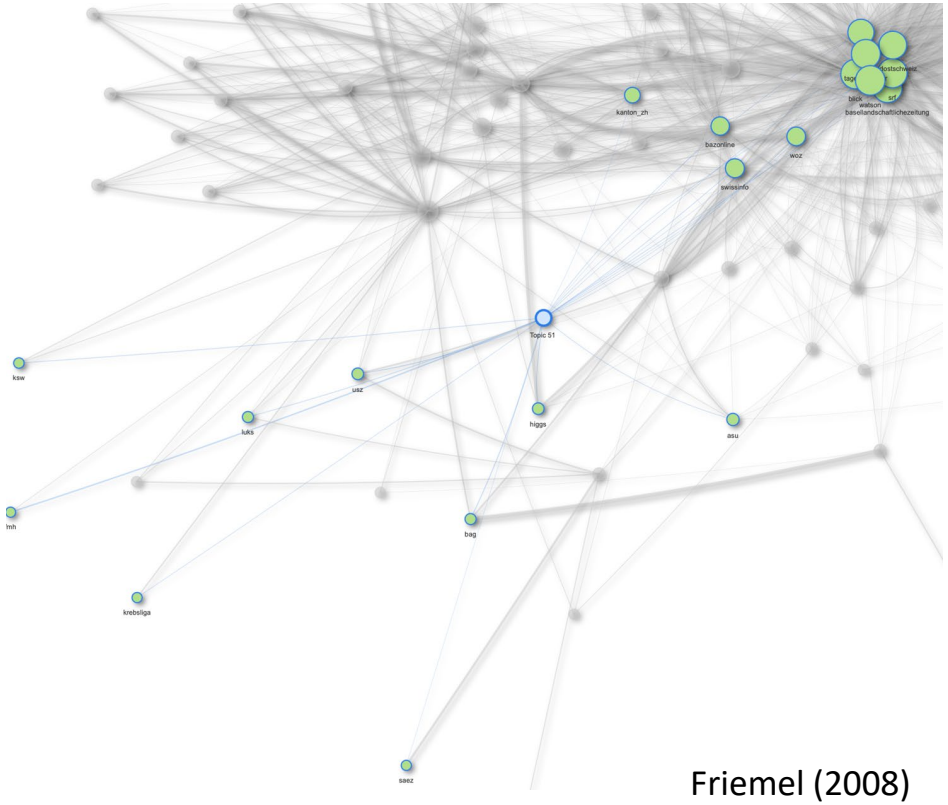
- Topic Modeling:
  - Identifikation latenter thematischer Strukturen in großen Textkorpora (Blei 2012)
  - Topic: Wahrscheinlichkeitsverteilung über Wörter innerhalb des Vokabulars eines Korpus
  - Topics weisen eine thematische Kohärenz auf

Topic 35	schule, kind, schüler, schülerin, eltern, lehrperson, unterricht, jugendliche, fernunterricht, lehrer, klasse, prüfung, finden, volksschule, präsenzunterricht, digital, lehrerin, schuljahr, lernende, gymnasium, sommerferien, lernen, unterrichten, statt, kindergarten
Topic 21	impfstoff, gegen, medikament, test, sollen, schweiz, erst, covid, impfung, million, roche, firma, unternehmen, klinisch, bereits, entwicklung, coronavirus, weltweit, usa, schweizer, produktion, studie, moderna, entwickeln, land
Topic 51	patient, spital, arzt, covid, schwer, behandlung, intensivstation, medizinisch, behandeln, krankheit, erkrankung, patientin, bett, klinik, medikament, ärztin, leiden, personal, betten, risiko, fieber, sterben, lunge, therapie, patienten

Themen & Netzwerke

Repräsentierte Akteure
Topic-Cluster 1
Topic-Cluster 2





Friemel (2008)

3. Diskussion

- Modellierung von Sprache und Diskurs ohne 'Noise'? Datafizierung von Sprache als 'natürlichem', 'fuzzy' Objekt
- Quantitative Datenanalyse vs. qualitative Einheiten

a) Sprache oder andere Datentypen als reine Zeichenfolge in der KI > z.B. im Sinne einer Digitalität als eigenes System/Mustererkennung zur Komplexitätsreduktion (A. Nassehi)

b) Auflösung von "Prä-Ideen" (L. Fleck) wie Text, Satz, Wort, Wortart?

Literatur

Baayen, H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baker, P. (2006). *Using corpora in discourse analysis*. Continuum

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Bubenhofer, N. (2020). Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings. *Zeitschrift Für Germanistische Linguistik*, *48*(3), 562–589. https://doi.org/10.1515/zgl-2020-2014

Bubenhofer, N. (2009). *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. De Gruyter.

Desagulier, G. (2014). Visualizing distances in a set of near-synonyms: *Rather, quite, fairly* , and *pretty*. In D. Glynn & J. A. Robinson (Eds.), *Human Cognitive Processing* (Vol. 43, pp. 145–178). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.06des

Divjak, D., & Fieller, N. (2014). Cluster analysis: Finding structure in linguistic data. In D. Glynn & J. A. Robinson (Eds.), *Human Cognitive Processing* (Vol. 43, pp. 405–441). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.43.16div

Dreesen, Ph./Krasselt, J. (2021). Exploring and analyzing linguistic environments. In F. Cooren/P. Stücheli-Herlach (Eds.), *Handbook of Management Communication* (pp. 389–408). De Gruyter.

Friemel, T. N. (2008). Netzwerkanalytische Methoden zur Identifizierung von Kommunikationsrollen. In C. Stegbauer (Ed.), *Netzwerkanalyse und Netzwerktheorie. Ein neues Paradigma in den Sozialwissenschaften* (pp. 179–190). Springer.

Gabrielatos, C. (2018). Keyness Analysis: Nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.), *Corpus Approaches to Discourse: A critical review*. Routledge.

Gries, S. Th. (2009). *Quantitative Corpus Linguistics with R. A Practical Introduction*. Routledge.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1212–1248). De Gruyter.

Keller, R. (2011). *Wissenssoziologische Diskursanalyse: Grundlegung eines Forschungsprogramms*. 3rd ed. VS Verlag für Sozialwissenschaften.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-030514-125254

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *ArXiv:1310.4546 [Cs, Stat]*. http://arxiv.org/abs/1310.4546

Warnke, I.H. (ed.) (2018). *Handbuch Diskurs*. De Gruyter. https://doi.org/10.1515/9783110296075.