

# Eine korpuslinguistische Analyse der Corona-Berichterstattung in der deutschen und französischen Presse

Sascha Diwersy<sup>1</sup>, Nathan Dykes<sup>2</sup>, Stephanie Evert<sup>2</sup>,  
Philipp Heinrich<sup>2</sup>, and Giancarlo Luxardo<sup>1</sup>

<sup>1</sup>*Praxiling*, Université Paul-Valéry (Montpellier 3)

<sup>2</sup>*Korpus- und Computerlinguistik*, Friedrich-Alexander-Universität Erlangen-Nürnberg

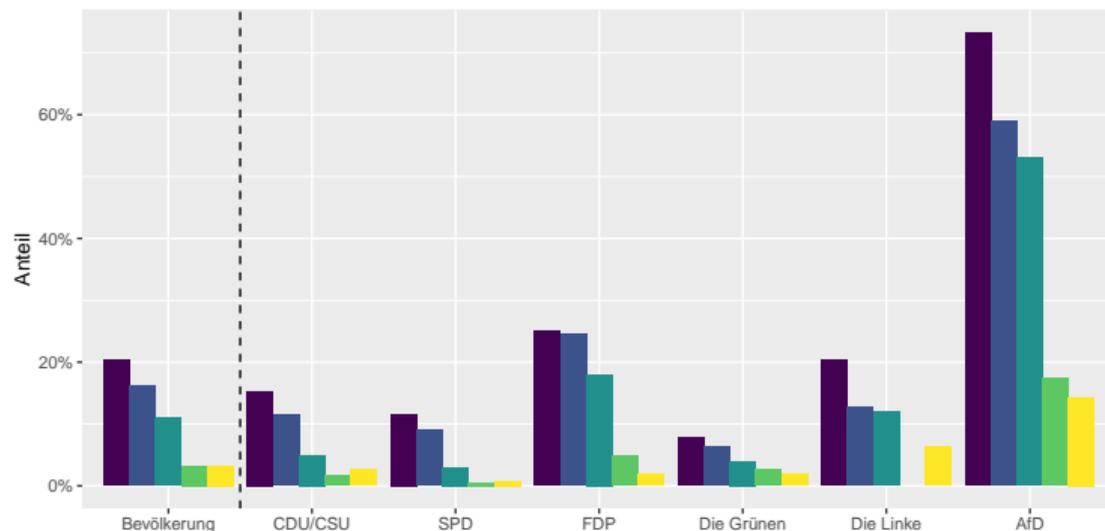


# Glaube an Verschwörungstheorien (→ Tracking the Infodemic)

## Zustimmung zu ausgewählten Verschwörungstheorien nach Parteianhängerschaft

Quelle: Allensbacher Archiv, IfD-Umfrage 12049 (2022)

- Viele Zahlen und Statistiken zu Corona sind gefälscht, um Angst zu verbreiten.
- Die Regierung nutzt die Krise als Vorwand, um dauerhaft die Grundrechte der Bürger einzuschränken.
- Die Corona-Impfung ist gefährlicher als die Krankheit selbst. Die Impfung verursacht schwere Schäden.
- Bill Gates ist der eigentlich Schuldige an der Corona-Krise. Er nutzt die Krise, um eine neue Weltordnung zu errichten.
- Ich fürchte, dass man allen Menschen Mikrochips einpflanzen will, um sie besser kontrollieren zu können.



# Corpus-assisted discourse studies (CADS)

- Daten: große, meist thematisch spezifizierte Korpora
- Ziel: Identifikation von **diskursiven Positionen**
- diskursive Position bzgl. Thema = Perspektiven, Standpunkte etc.
- Beispiel: *Covid-19 ist harmlos*
  
- Akteure können mittels Text diskursive Positionen
  - ▶ erschaffen und verbreiten,
  - ▶ akzeptieren oder ablehnen (→ Meinungsbildung)
  - ▶ positiv oder negativ evaluieren (→ semantische Prosodie)

# Diskurseme

- Diskursem = Menge von *lexical items* mit bestimmter diskursiver Funktion
  - ▶ (POS-disambiguierte) Lexeme
  - ▶ zusätzlich: Named Entities, Zeitausdrücke, fixe Idiome, Phrasen, ...
  - ▶ Beispiele:
    - ★ *COVID-19, Corona, Covid*
    - ★ *Grippe, Erkältung, harmlos, ungefährlich*
- Diskursem  $\neq$  diskursive Position
  - ▶ Diskurseme fangen oft nur einen Teil-Aspekt der diskursiven Position ein
  - ▶ können Teil mehrerer diskursiver Positionen sein
  - ▶ *lexical items* können mehrdeutig sein und auch zu mehreren Diskursemen gehören
- Themendiskurseme
  - ▶ *topic node* in Kollokationsanalysen = Diskursem
  - ▶ a priori definiert

# Diskurseme

- Diskursem = Menge von *lexical items* mit bestimmter diskursiver Funktion
  - ▶ (POS-disambiguierte) Lexeme
  - ▶ zusätzlich: Named Entities, Zeitausdrücke, fixe Idiome, Phrasen, ...
  - ▶ Beispiele:
    - ★ *COVID-19, Corona, Covid*
    - ★ *Grippe, Erkältung, harmlos, ungefährlich*
- Diskursem  $\neq$  diskursive Position
  - ▶ Diskurseme fangen oft nur einen Teil-Aspekt der diskursiven Position ein
  - ▶ können Teil mehrerer diskursiver Positionen sein
  - ▶ *lexical items* können mehrdeutig sein und auch zu mehreren Diskursemen gehören
- Themendiskurseme
  - ▶ *topic node* in Kollokationsanalysen = Diskursem
  - ▶ a priori definiert

# Diskurseme

- Diskursem = Menge von *lexical items* mit bestimmter diskursiver Funktion
  - ▶ (POS-disambiguierte) Lexeme
  - ▶ zusätzlich: Named Entities, Zeitausdrücke, fixe Idiome, Phrasen, ...
  - ▶ Beispiele:
    - ★ *COVID-19, Corona, Covid*
    - ★ *Grippe, Erkältung, harmlos, ungefährlich*
- Diskursem  $\neq$  diskursive Position
  - ▶ Diskurseme fangen oft nur einen Teil-Aspekt der diskursiven Position ein
  - ▶ können Teil mehrerer diskursiver Positionen sein
  - ▶ *lexical items* können mehrdeutig sein und auch zu mehreren Diskursemen gehören
- Themendiskurseme
  - ▶ *topic node* in Kollokationsanalysen = Diskursem
  - ▶ a priori definiert

# Korpuszusammenstellung und -aufbereitung

## Samplingverfahren

- Pressearchive *Europresse* und *Factiva*
- Suchbegriffe:
  - ▶ dt.: Corona\* | Covid\* | Sars\* | Epidem\* | epidem\* | Pandem\* | pandem\*
  - ▶ frz.: Corona\* | corona\* | Covid\* | covid\* | Sars\* | sars\* | épidém\* | pandém\*
- nationale und regionale Tageszeitungen; nationale Wochenzeitungen
  - ▶ dt.: *Süddeutsche Zeitung*, *Frankfurter Rundschau*, *TAZ*, *Die Welt*, *Bild-Zeitung*, *Handelsblatt*; *Tagesspiegel*, *Hamburger Abendblatt*, *Kölner Stadt-Anzeiger*, usw.; *Die Zeit*, *Der Spiegel*, *Focus*
  - ▶ frz.: *Le Monde*, *Le Figaro*, *Libération*, *L'Humanité*, *La Croix*, *Les Echos*; *Le Parisien*, *La Provence*, *Le Progrès*, usw.; *L'Obs*, *L'Express*, *Le Point*, etc.
- Zeitraum: Dezember 2019 bis August 2021

# Korpuszusammenstellung und -aufbereitung

## Korpusverarbeitung

- Stanza:
  - ▶ Tokenisierung und Satzsegmentierung
  - ▶ PoS-Tagging (Universal Dependencies)
  - ▶ Lemmatisierung
  - ▶ Dependenz-Parsing (UD)
  - ▶ NER
- TreeTagger:
  - ▶ PoS-Tagging (STTS)
  - ▶ Lemmatisierung (frz.)
- SMOR:
  - ▶ Lemmatisierung (dt.)
- CWB:
  - ▶ Datenindexierung

# Datenmodell

## Textstruktureinheiten

- Einzeltext - Metadaten: Id, Titel, Datum (Jahr, Jahr+Monat, Tag), Autor, Quelle, Zeitungstyp (nationale vs regionale Presse, Tages- vs Wochenzeitung)
- Absatz - Metadaten: Typ (Titelzeile)
- Satz - Metadaten: Id, Satzart (Deklarativ-, Interrogativ-, Exklamativsatz)

## Linguistische Annotationen

- Wortart (UD + STTS)
- morphologische Merkmale (UD)
- Lemma
- Abhängigkeitsrelation + Satzfunktion (UD)
- NE

# Korpusumfang

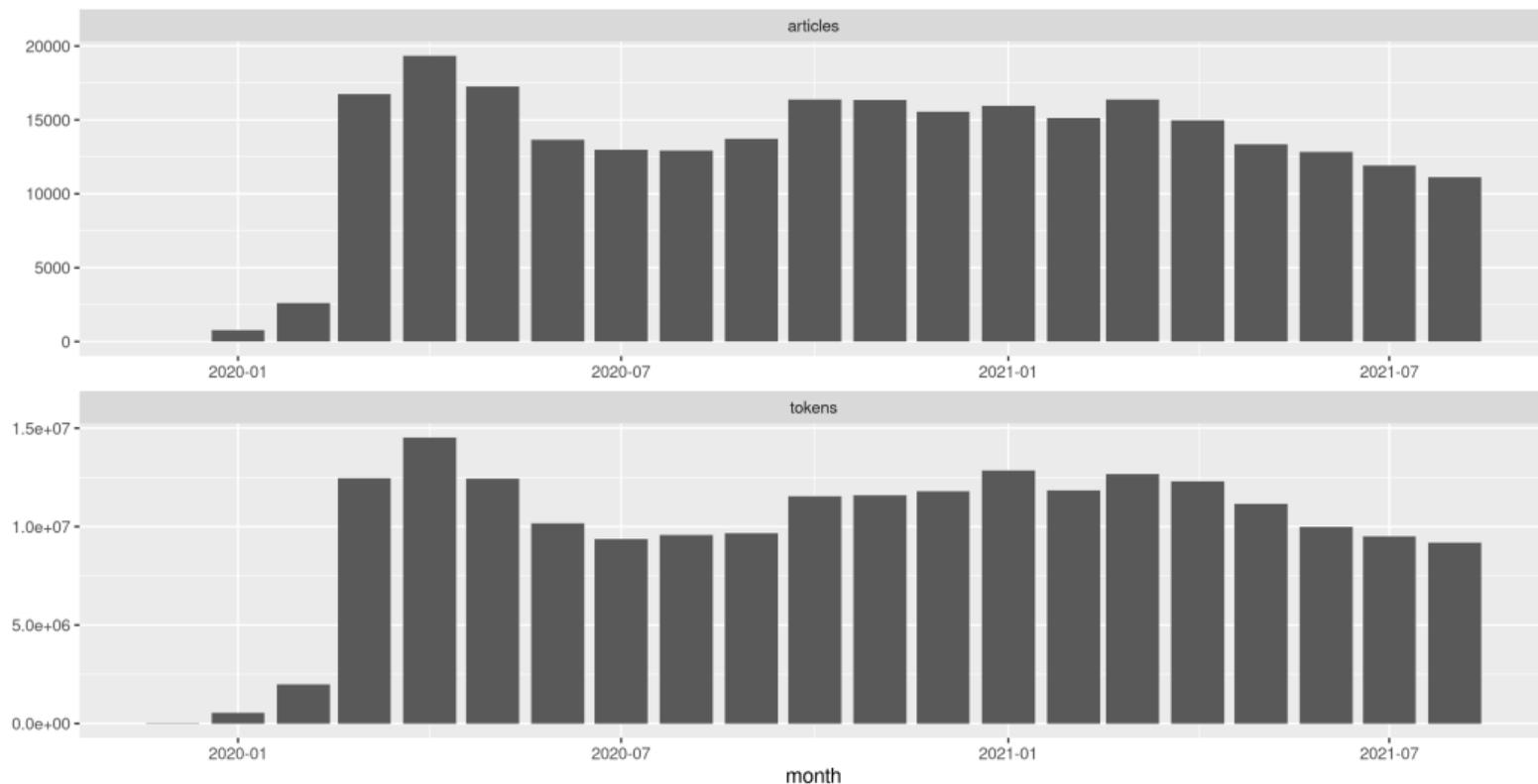
## deutsches Korpus

- 269,933 Artikel
- 12,897,608 Sätze
- 205,168,270 Token

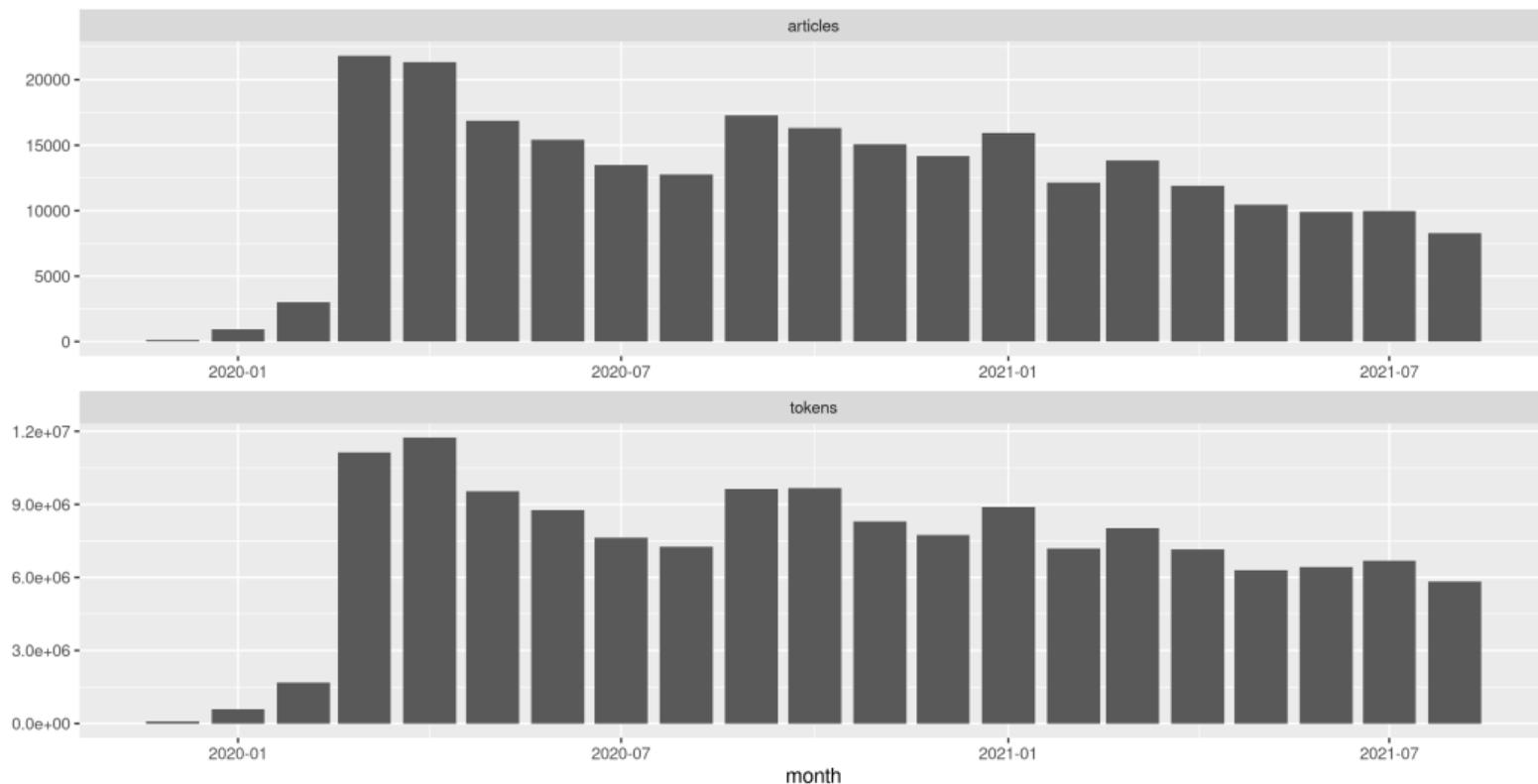
## französisches Korpus

- 261,026 Artikel
- 7,931,359 Sätze
- 150,223,266 Token

# Textmenge im zeitlichen Verlauf (deutsches Korpus)



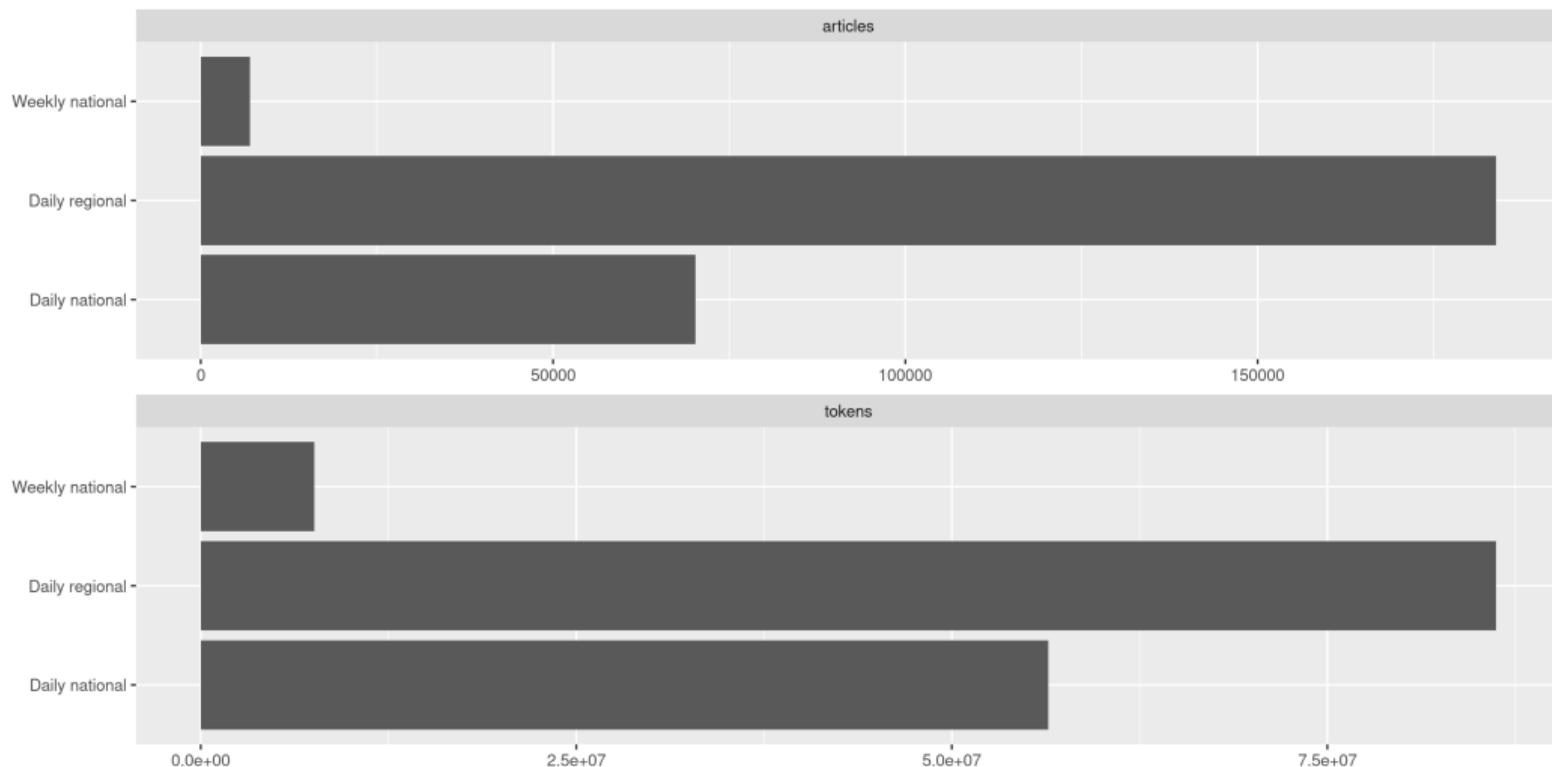
# Textmenge im zeitlichen Verlauf (französisches Korpus)



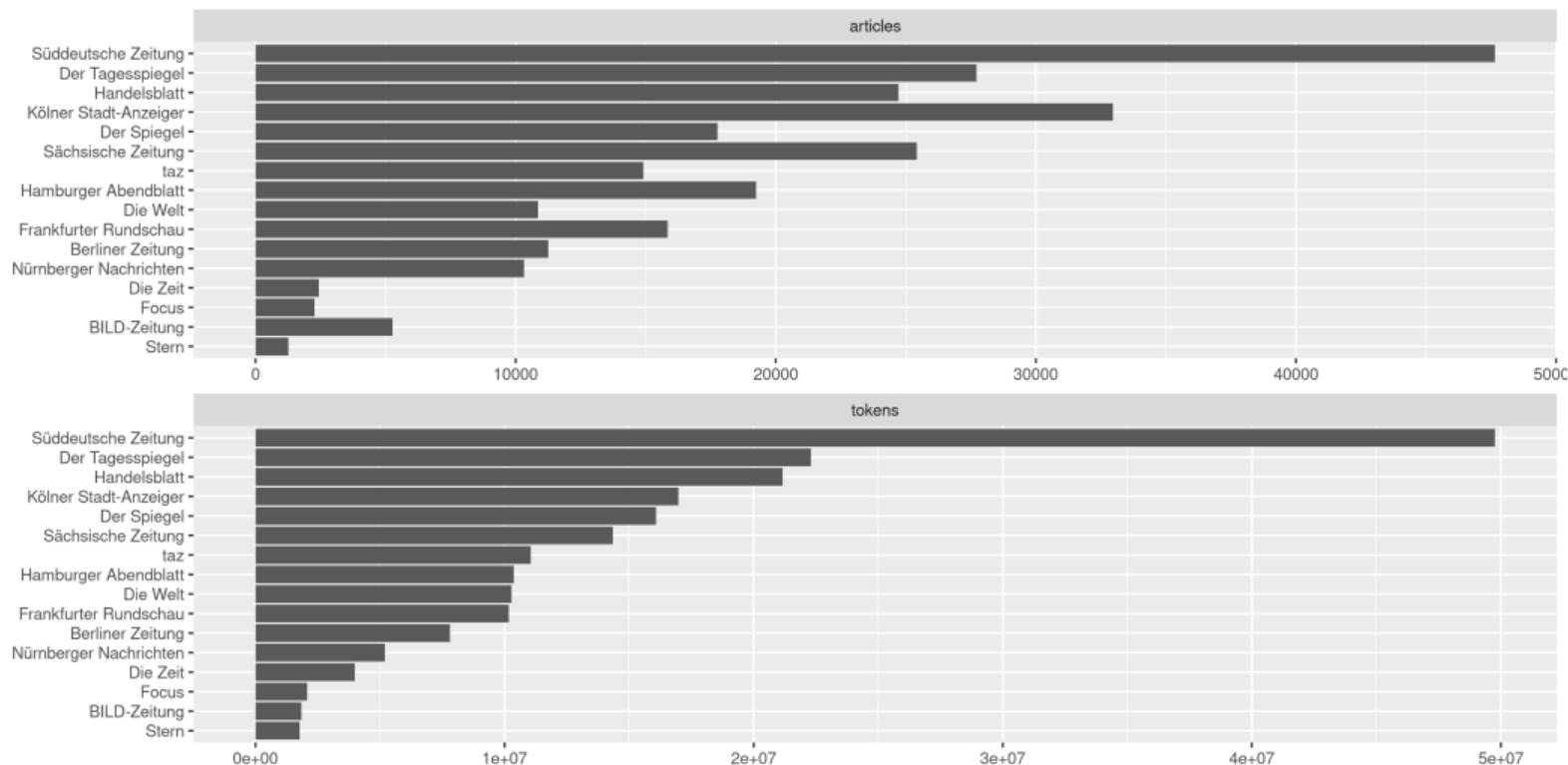
# Textmenge pro Zeitungsart (deutsches Korpus)



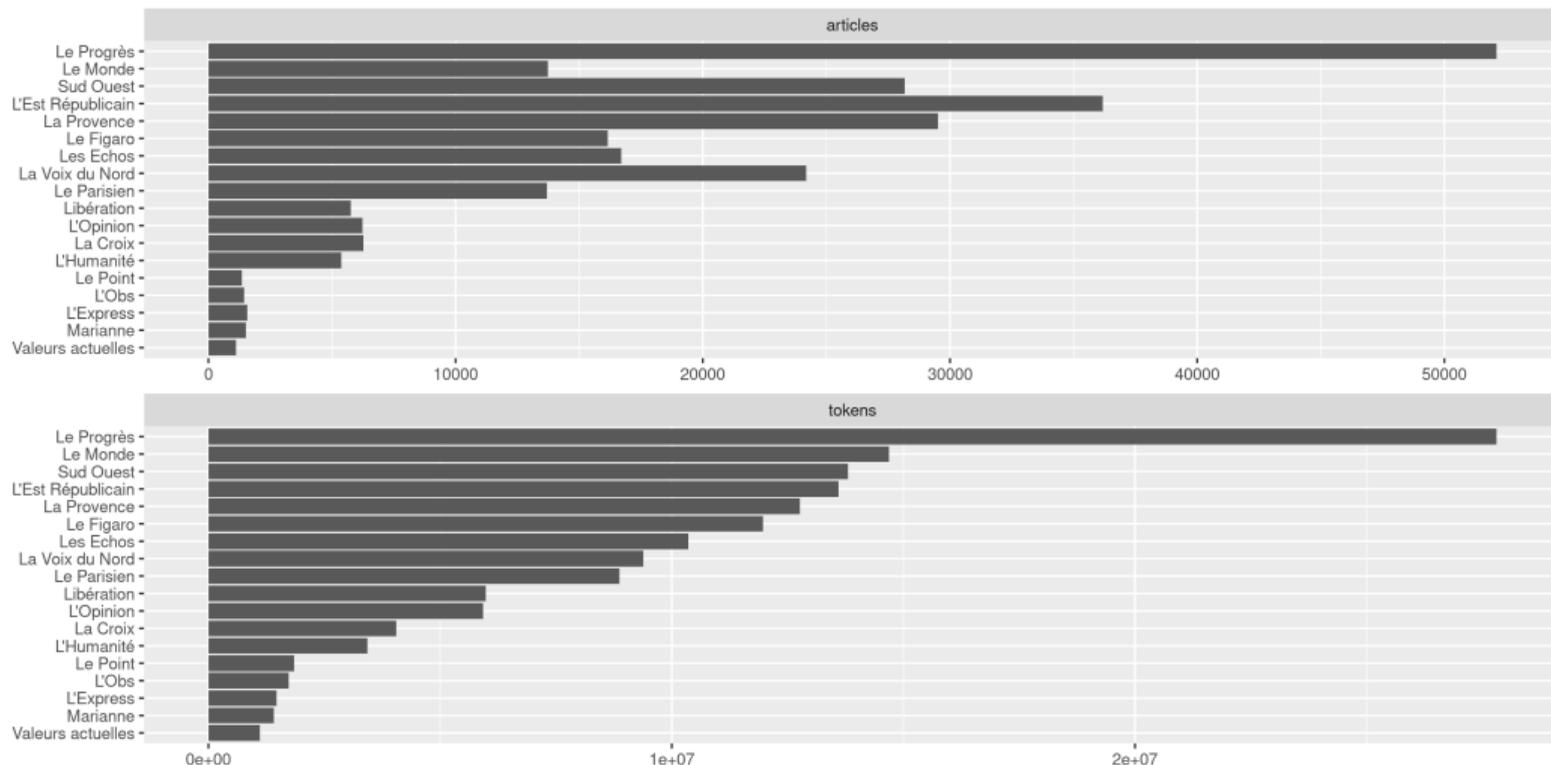
# Textmenge pro Zeitungsart (französisches Korpus)



# Textmenge pro Zeitung (deutsches Korpus)



# Textmenge pro Zeitung (französisches Korpus)



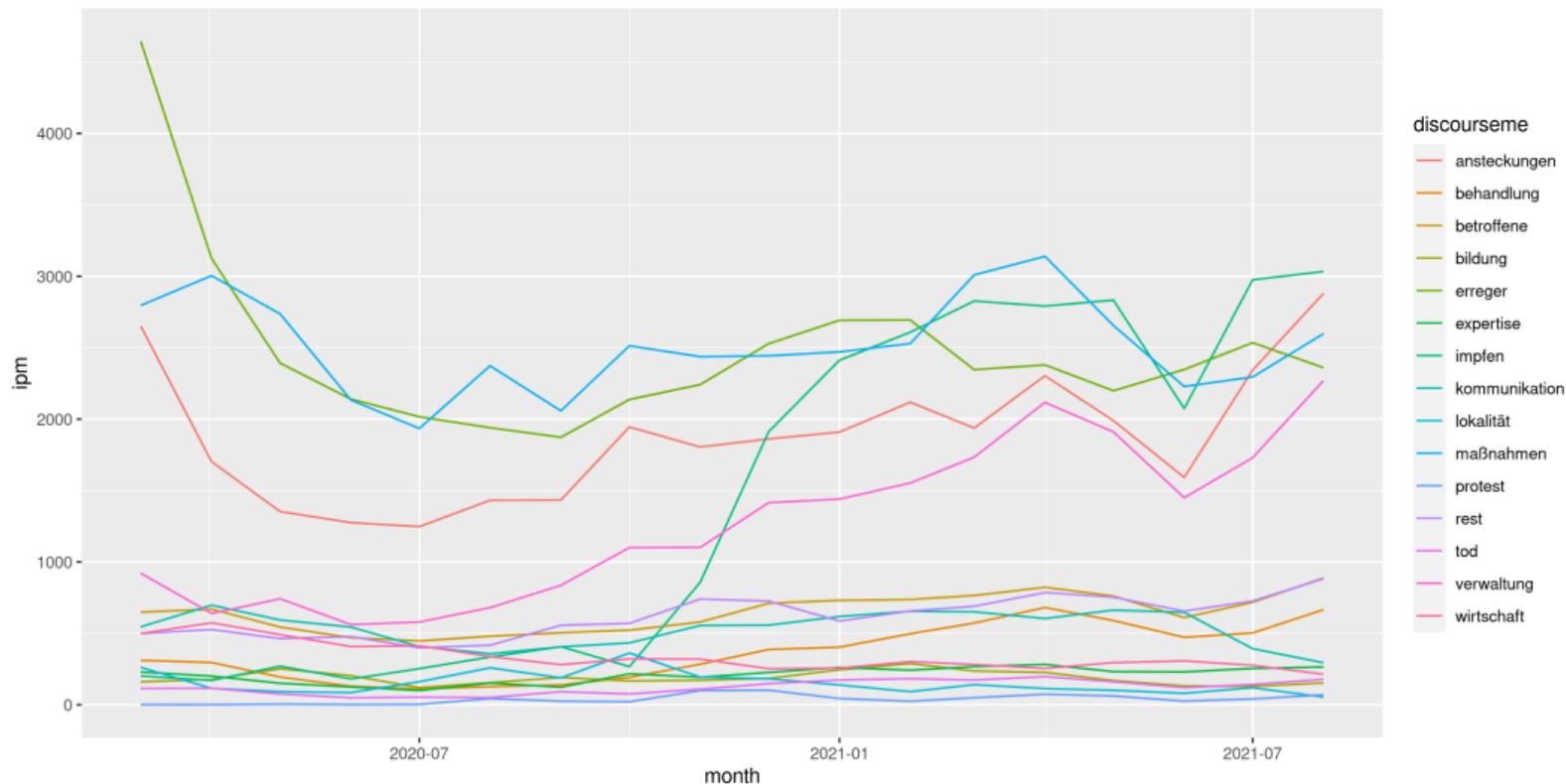
# Keyword-Analyse

- Vergleich mit Referenzkorpus
  - ▶ deutsch: Süddeutsche Zeitung 2011-2014
  - ▶ französisch: Sammlung von Zeitungen -2016
- Assoziationsmaß: *conservative log ratio*
  - ▶ untere Schranke des *relative risk*
  - ▶ Kombination von Effekt & Signifikanz
- Kategorisierung von POS-disambiguierten Lemmata
  - ▶ je (ca.) 150
  - ▶ Ausschluss von Personen & Ortsnamen
  - ▶ überwiegend mittels *distant reading*

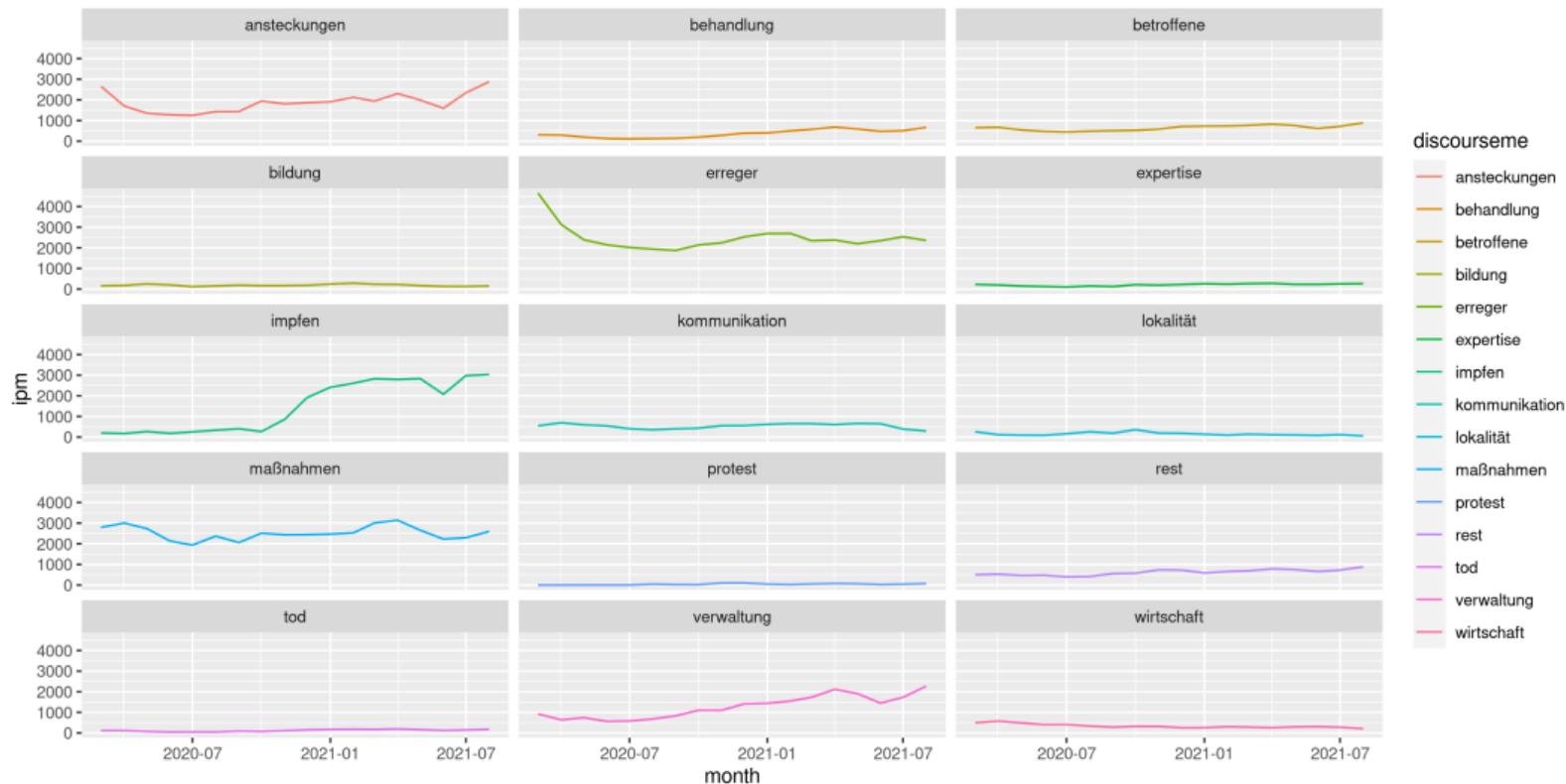
## Häufigkeiten in den Korpora

Diskursem	<i>F</i> (de)	<i>F</i> (fr)	<i>V</i> (de)	<i>V</i> (fr)	<i>f</i> (de)	<i>f</i> (fr)
erreger	533811	520092	39	52	2601.82	3462.13
maßnahmen	523655	353013	157	128	2552.32	2349.92
ansteckungen	395322	171760	91	72	1926.82	1143.36
impfen	296616	163102	75	57	1445.72	1085.73
verwaltung	258704	18310	51	9	1260.94	121.89
betroffene	133407	42951	55	18	650.23	285.91
misc	126313	158356	39	23	615.66	1054.14
kommunikation	111686	115598	32	18	544.36	769.51
behandlung	75149	62650	33	44	366.28	417.05
wirtschaft	69932	5312	31	14	340.85	35.36
expertise	43291	9716	10	25	211.00	64.68
bildung	38567		11		187.98	
lokalität	31357	29092	26	5	152.84	193.66
tod	25797	520	8	4	125.74	3.46
protest	7656	3786	6	2	37.32	25.20
politik		6446		13		42.91
desinformation		1025		3		6.82

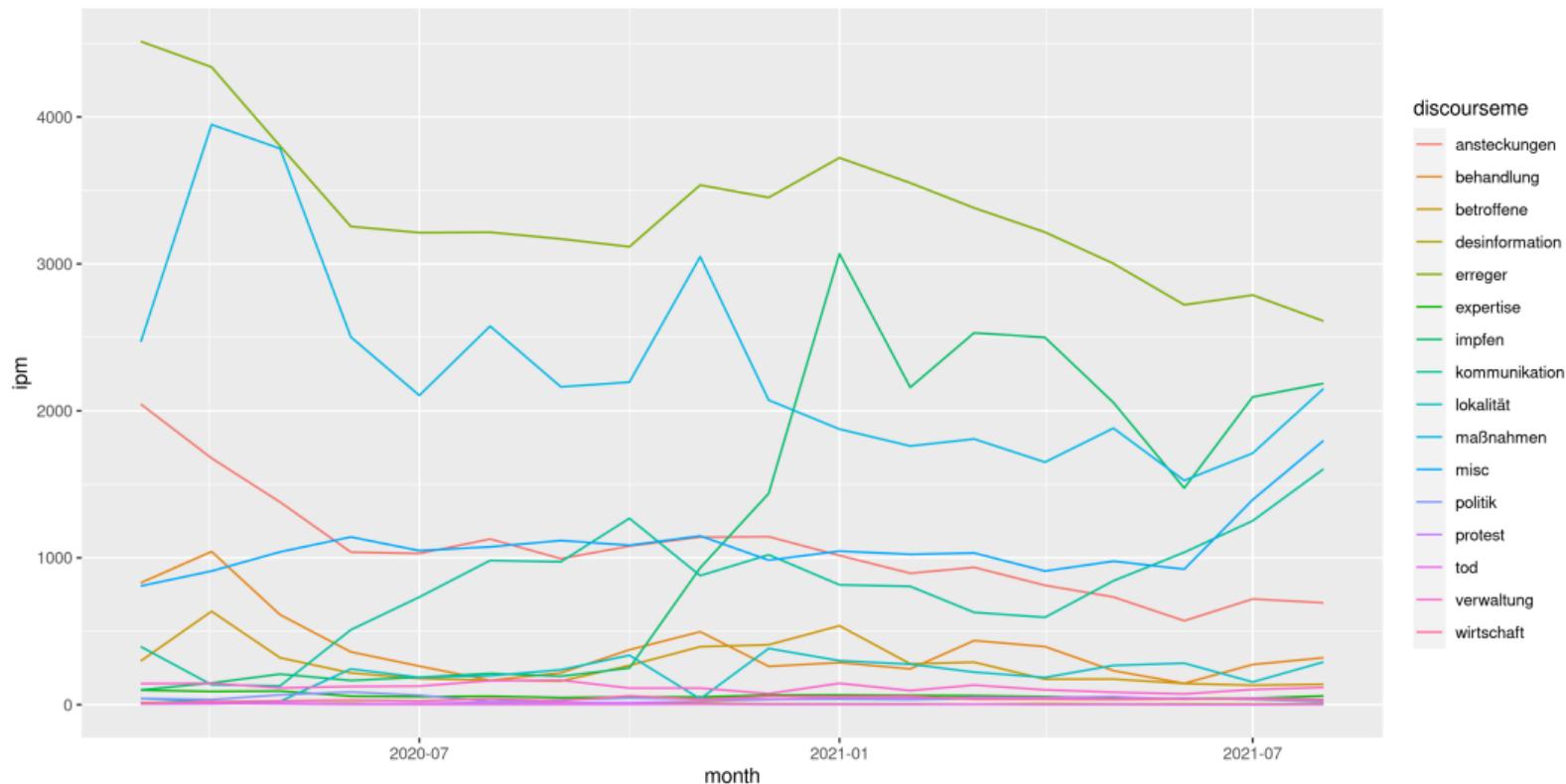
## zeitliche Ausbreitung (deutsches Korpus)



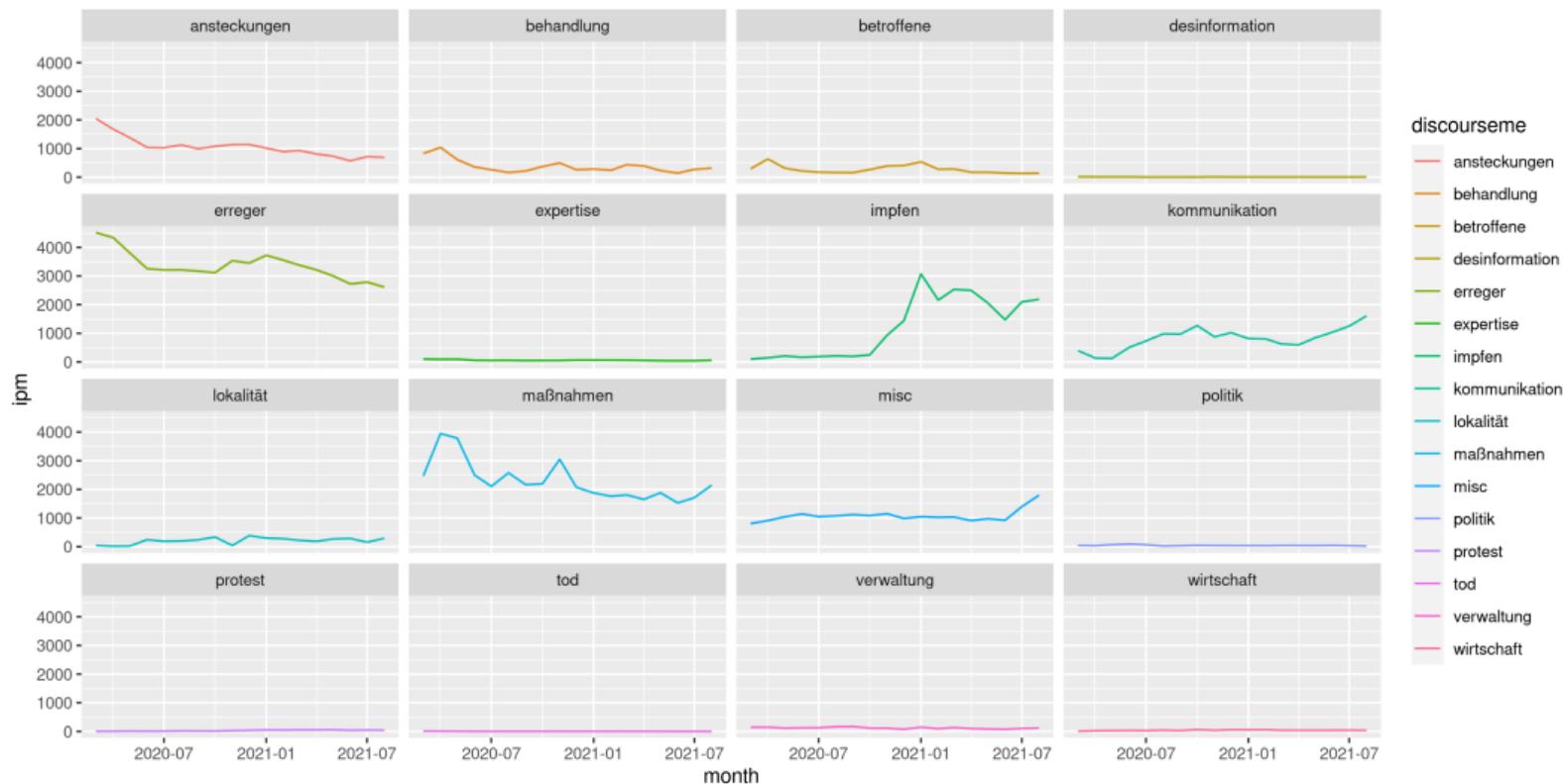
## zeitliche Ausbreitung (deutsches Korpus)



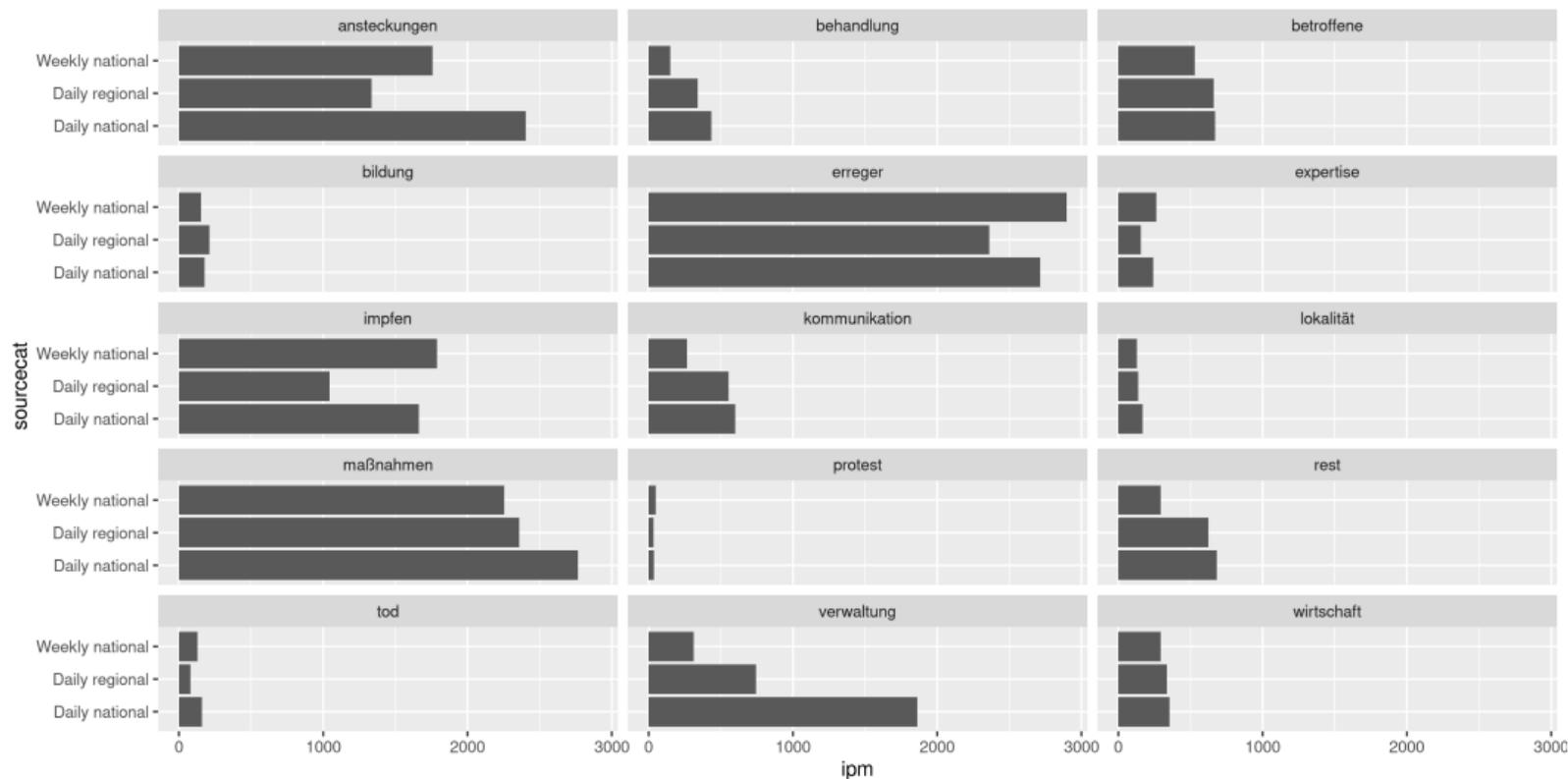
## zeitliche Ausbreitung (französisches Korpus)



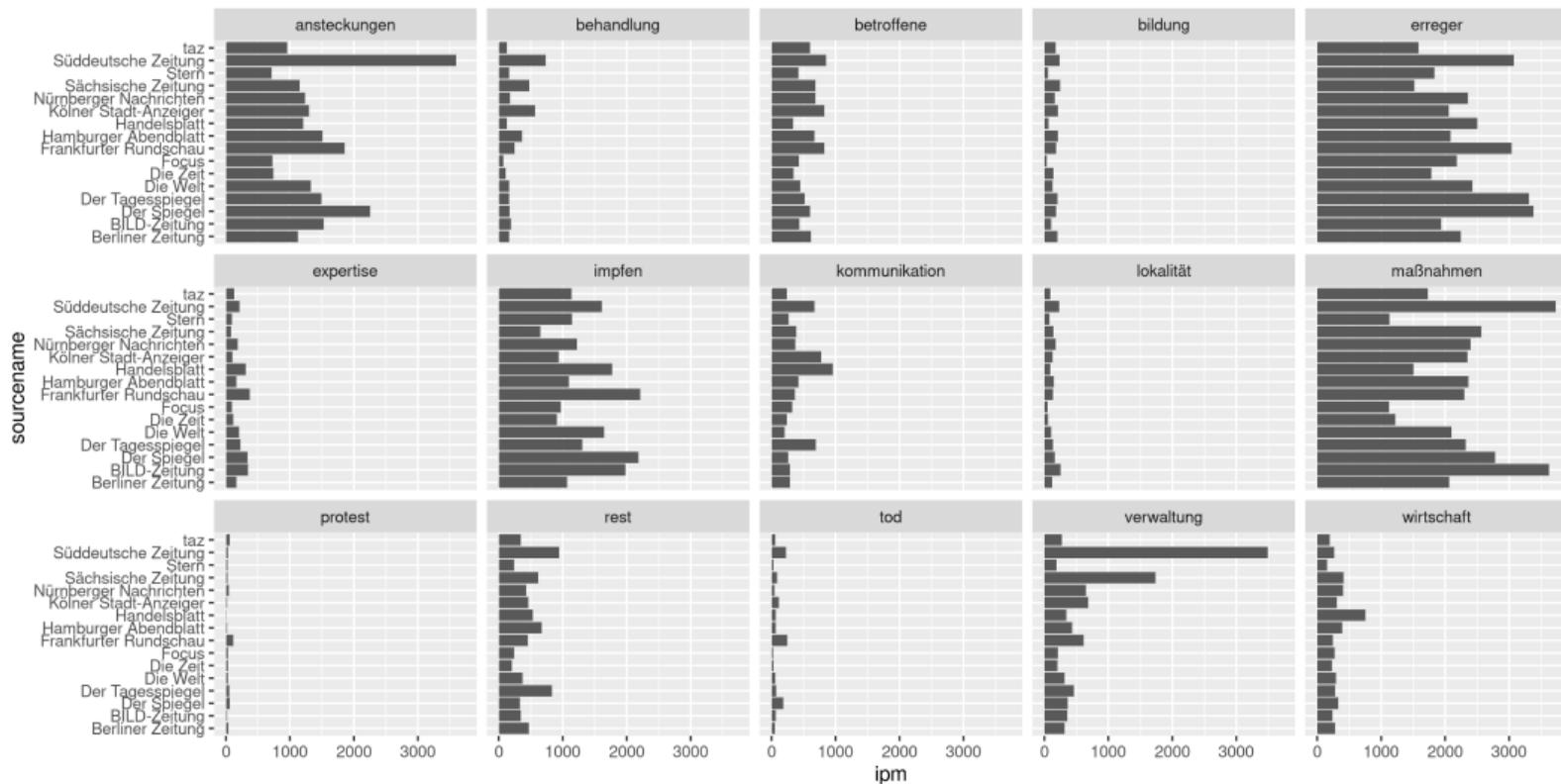
## zeitliche Ausbreitung (französisches Korpus)



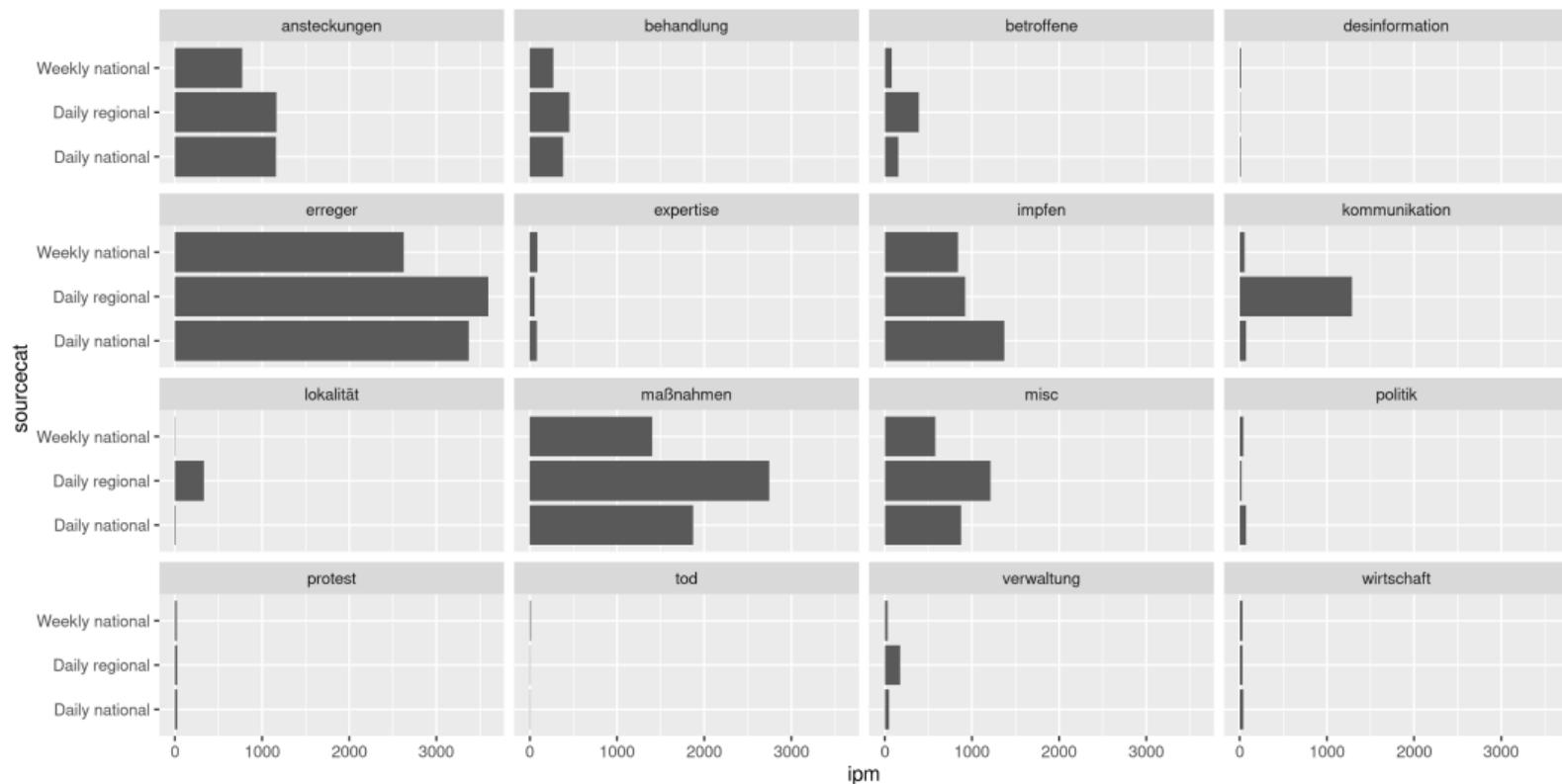
# Verteilung über die Zeitungsarten (deutsches Korpus)



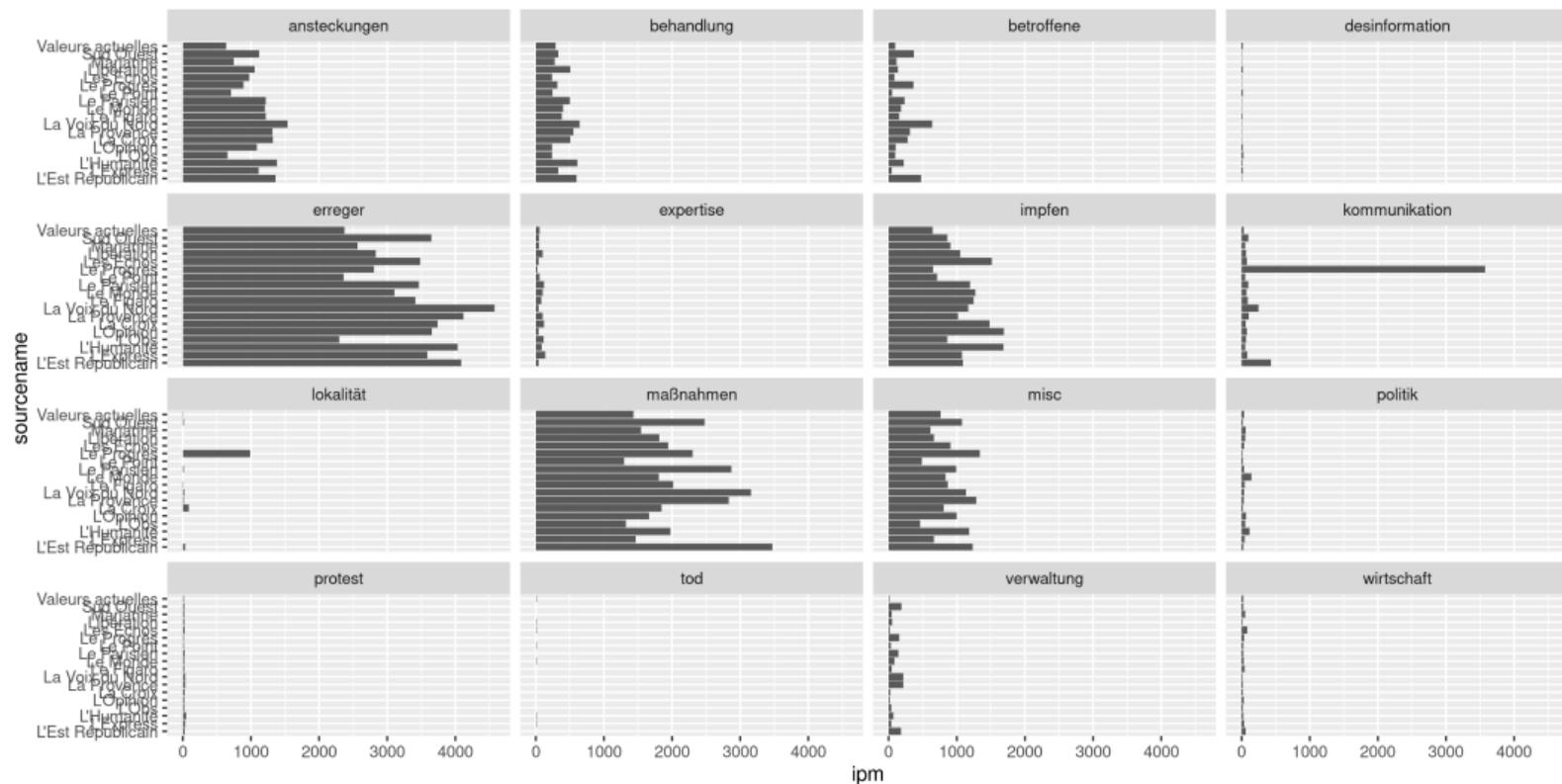
## Verteilung über die Zeitungen (deutsches Korpus)



# Verteilung über die Zeitungsarten (französisches Korpus)



# Verteilung über die Zeitungen (französisches Korpus)



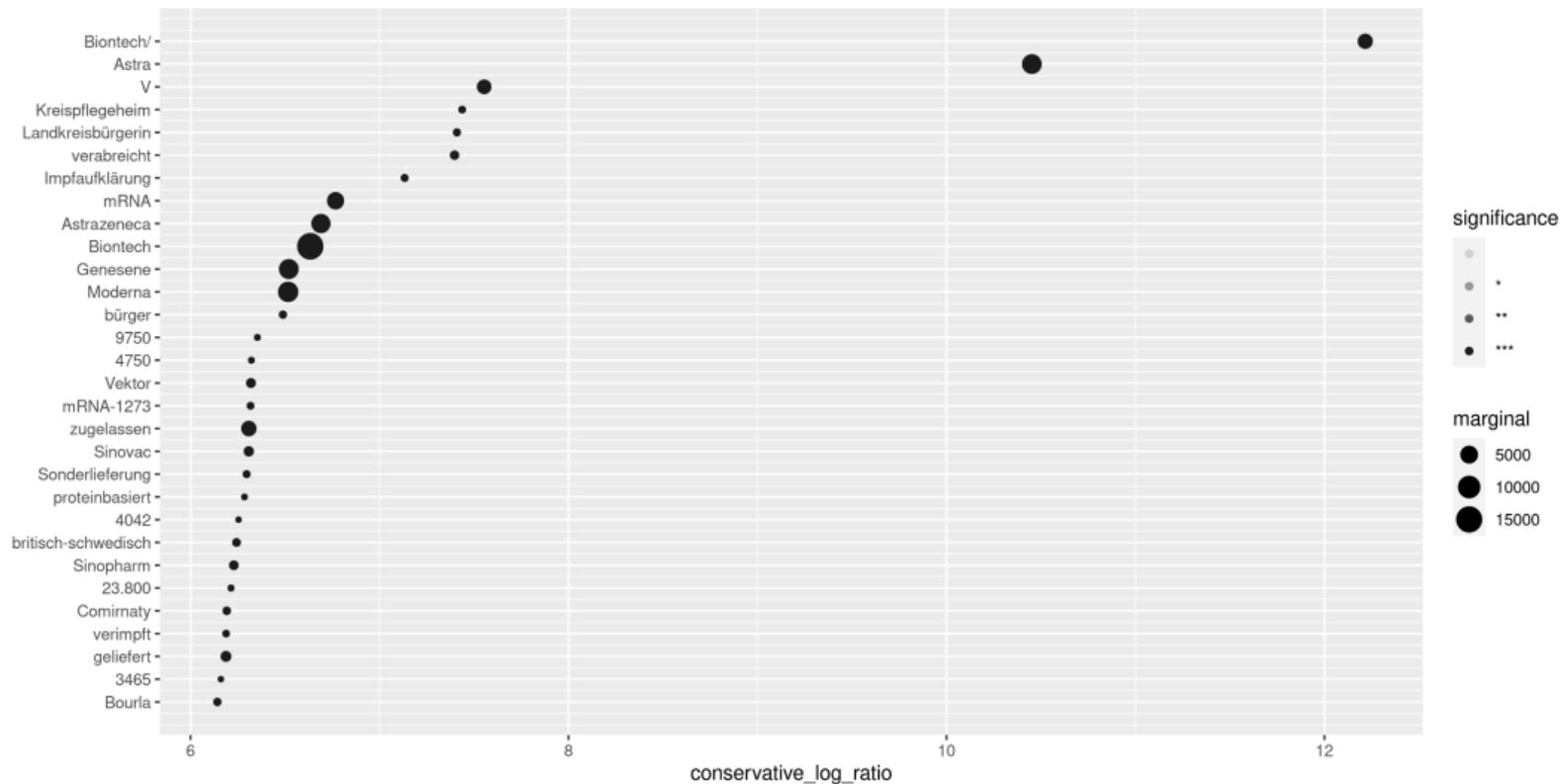
## Textuelle Assoziationen (deutsches Korpus)

	node	candidate	conservative log-ratio	log-likelihood
1	ansteckungen	tod	4.25	17727.56
2	expertise	tod	2.99	8606.10
3	ansteckungen	expertise	2.88	16900.73
4	tod	expertise	2.77	8606.10
5	maßnahmen	lokalität	2.69	11297.46
6	behandlung	tod	2.59	6901.61
7	ansteckungen	lokalität	2.45	11656.97
8	erreger	tod	2.45	4975.13
9	impfen	expertise	2.36	11106.81
10	tod	behandlung	2.30	6901.61
11	maßnahmen	tod	2.07	5286.09
12	verwaltung	tod	2.06	5755.35
13	lokalität	tod	2.00	2830.40
14	impfen	tod	1.98	4639.48
	⋮	⋮	⋮	⋮

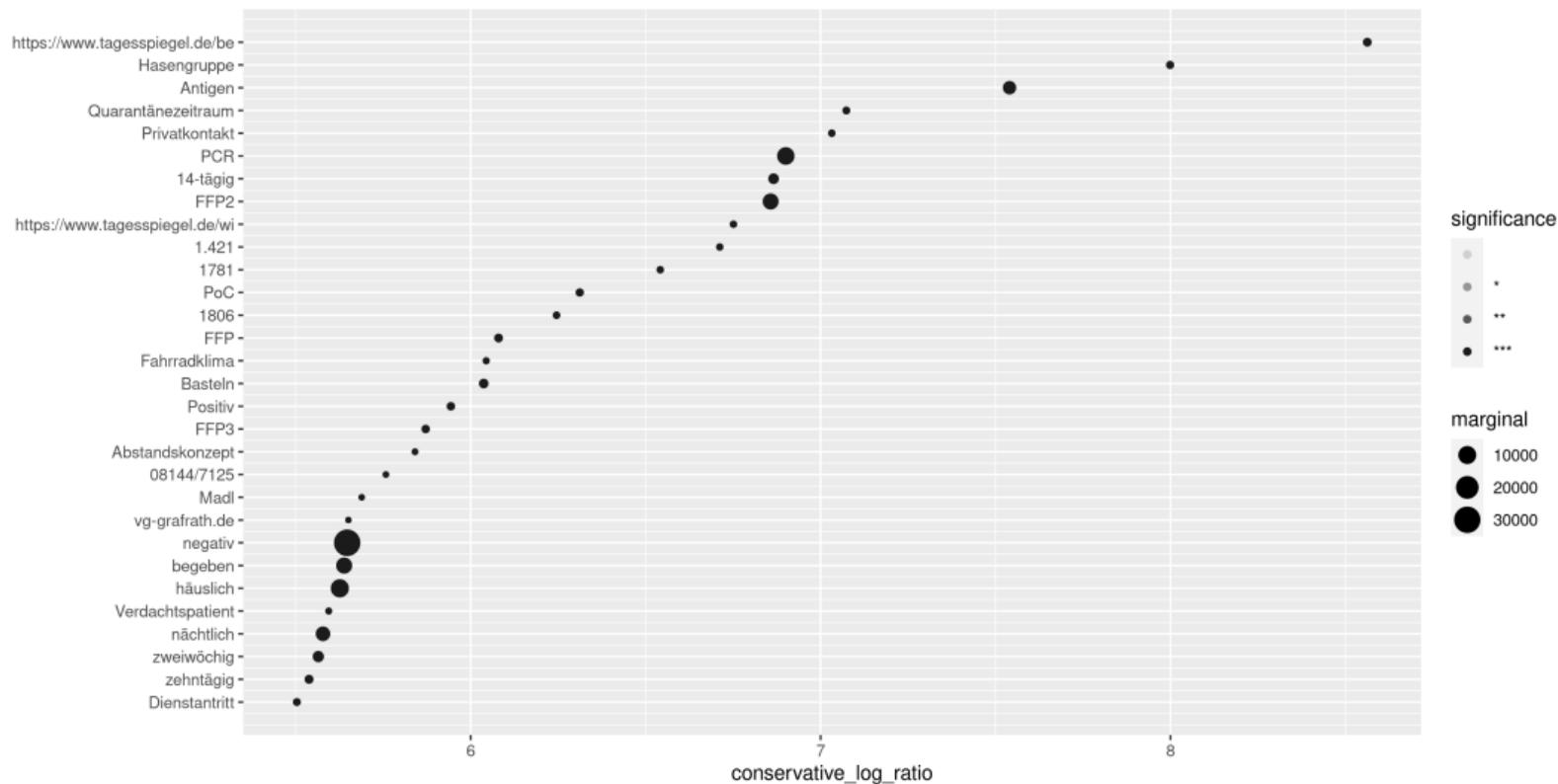
## Textuelle Assoziationen (französisches Korpus)

	node	candidate	conservative log-ratio	log-likelihood
1	kommunikation	lokalität	4.53	5874.69
2	expertise	tod	3.91	500.74
3	lokalität	kommunikation	3.59	5874.69
4	tod	expertise	3.53	500.74
5	impfen	protest	3.41	4768.43
6	kommunikation	wirtschaft	3.00	2363.84
7	ansteckungen	tod	2.85	438.66
8	ansteckungen	expertise	2.68	5401.01
9	wirtschaft	kommunikation	2.65	2363.84
10	behandlung	expertise	2.46	4165.98
11	protest	impfen	2.43	4768.43
12	behandlung	betroffene	2.27	8617.46
13	behandlung	verwaltung	2.23	5312.39
14	betroffene	verwaltung	2.17	3552.26
⋮	⋮	⋮	⋮	⋮

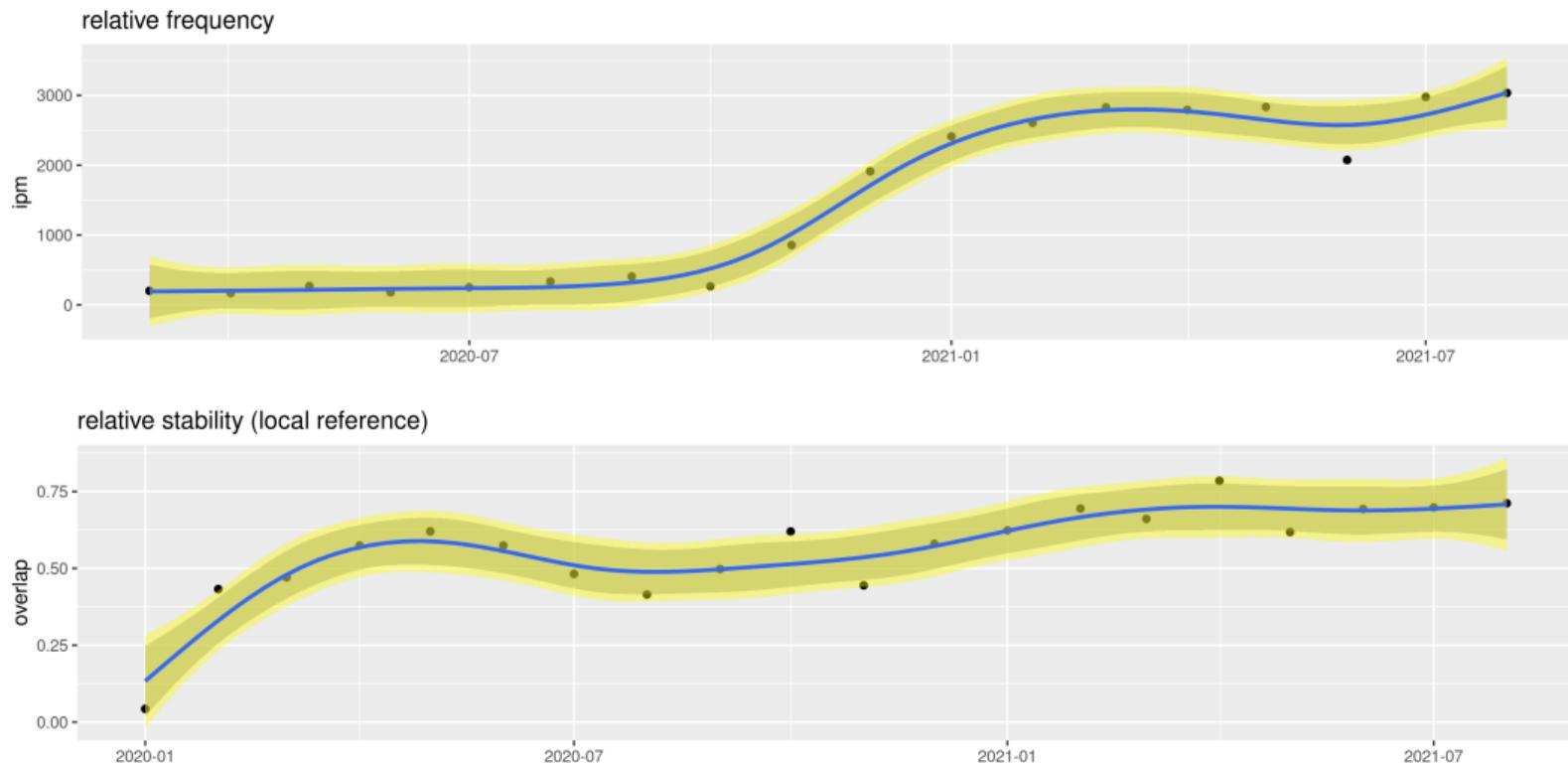
# Kollokate: Impfen (deutsches Korpus)



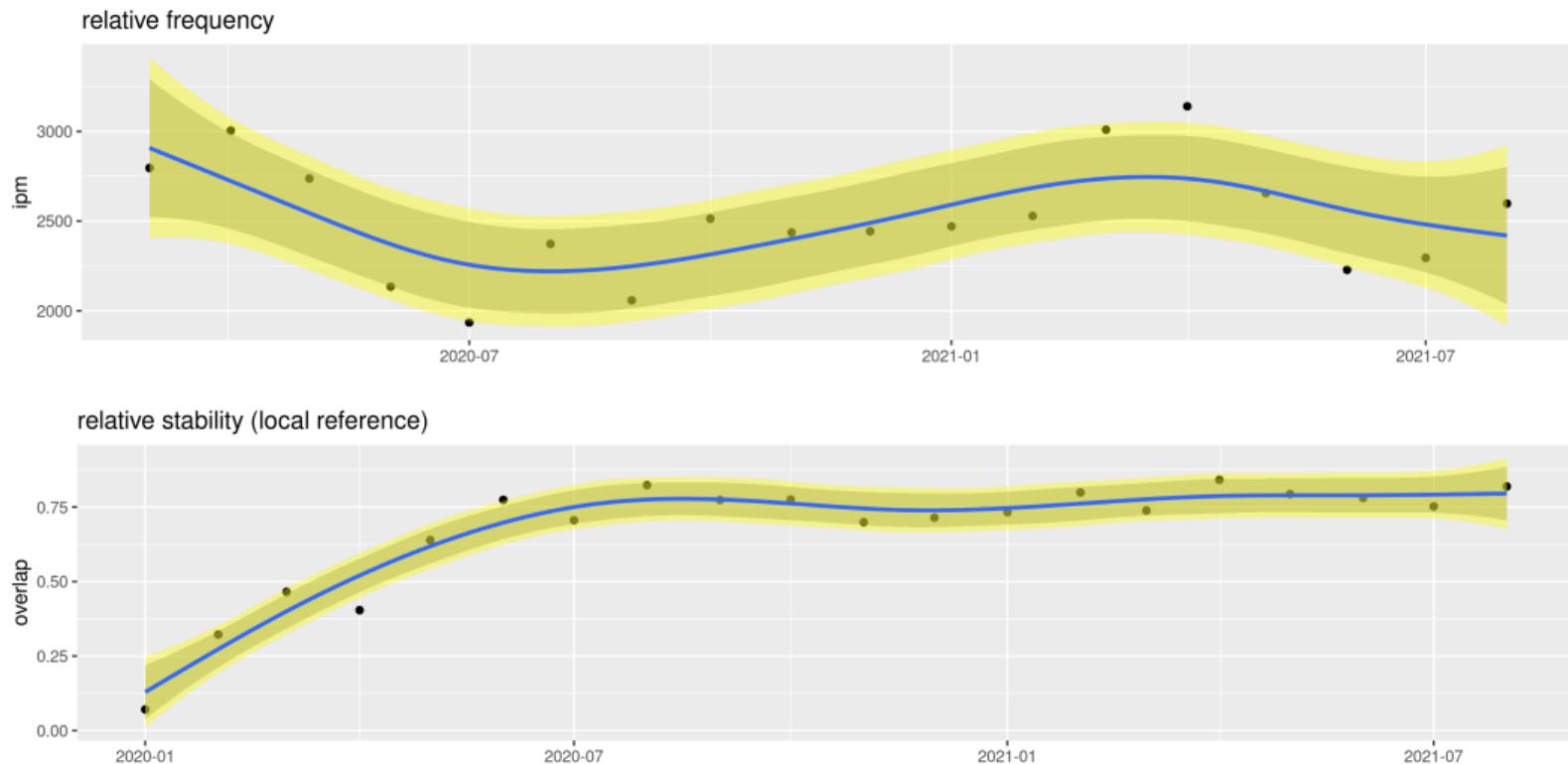
# Kollokate: Maßnahmen (deutsches Korpus)



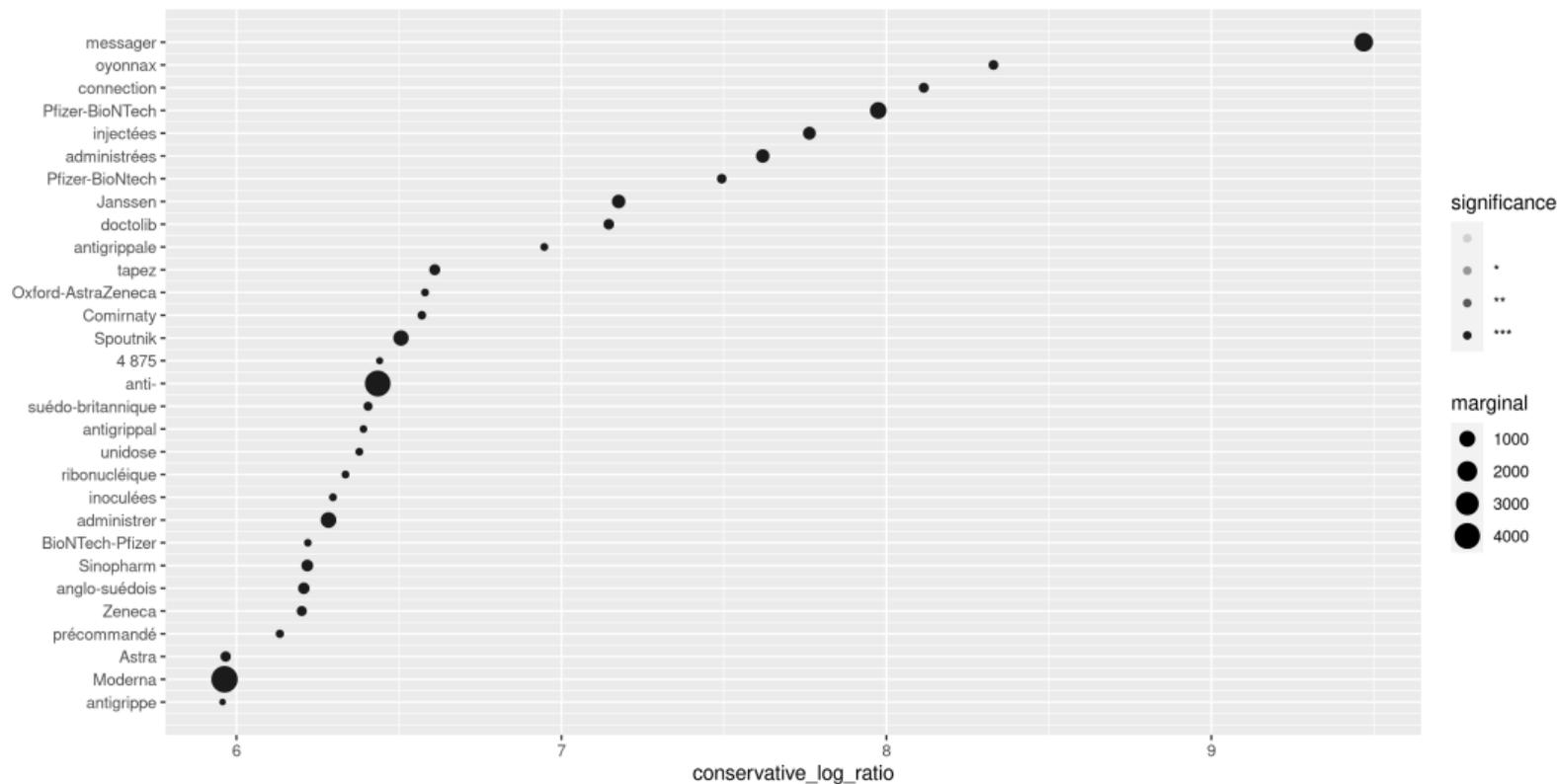
# UFA: Impfen (deutsches Korpus)



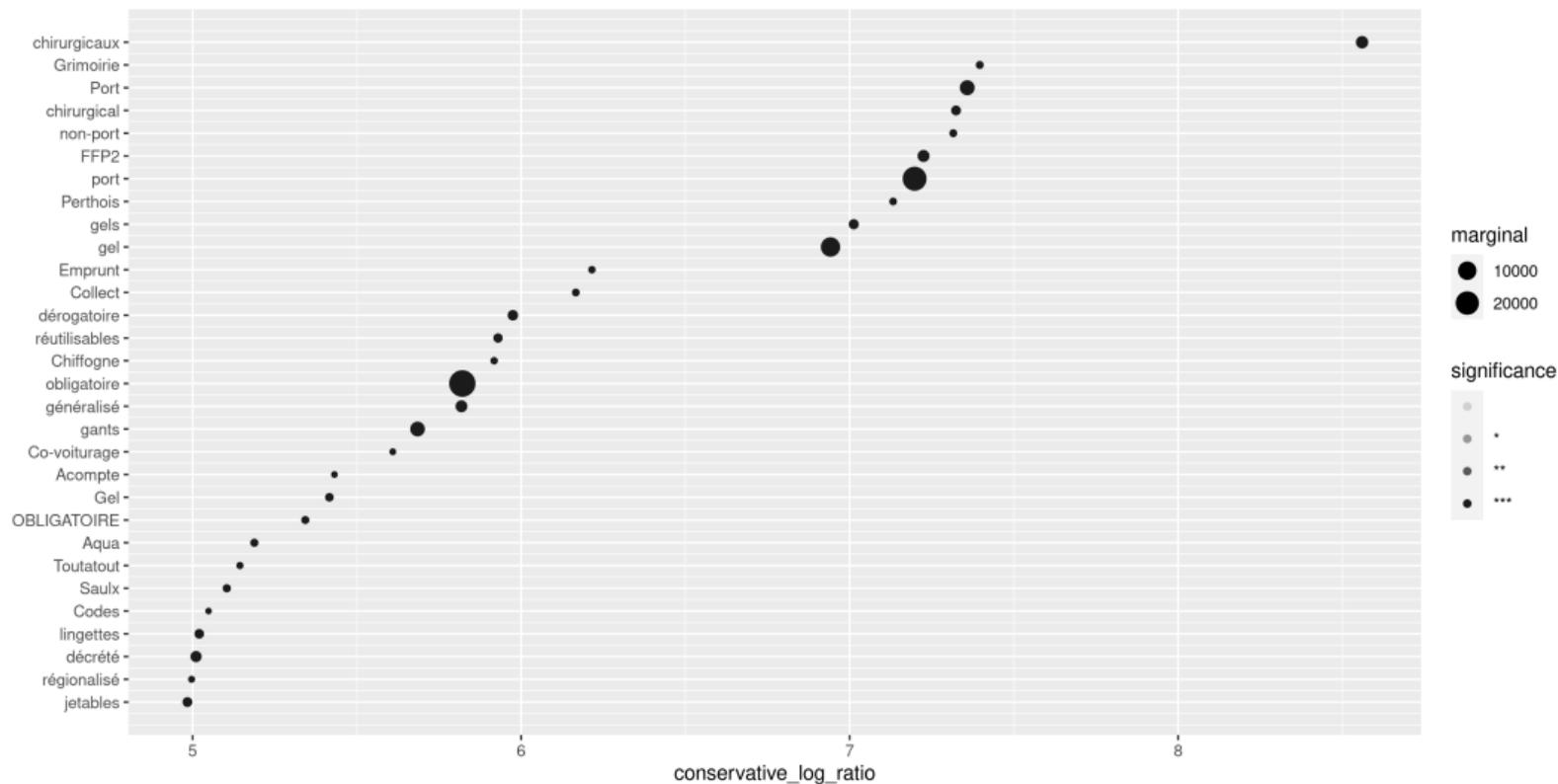
# UFA: Maßnahmen (deutsches Korpus)



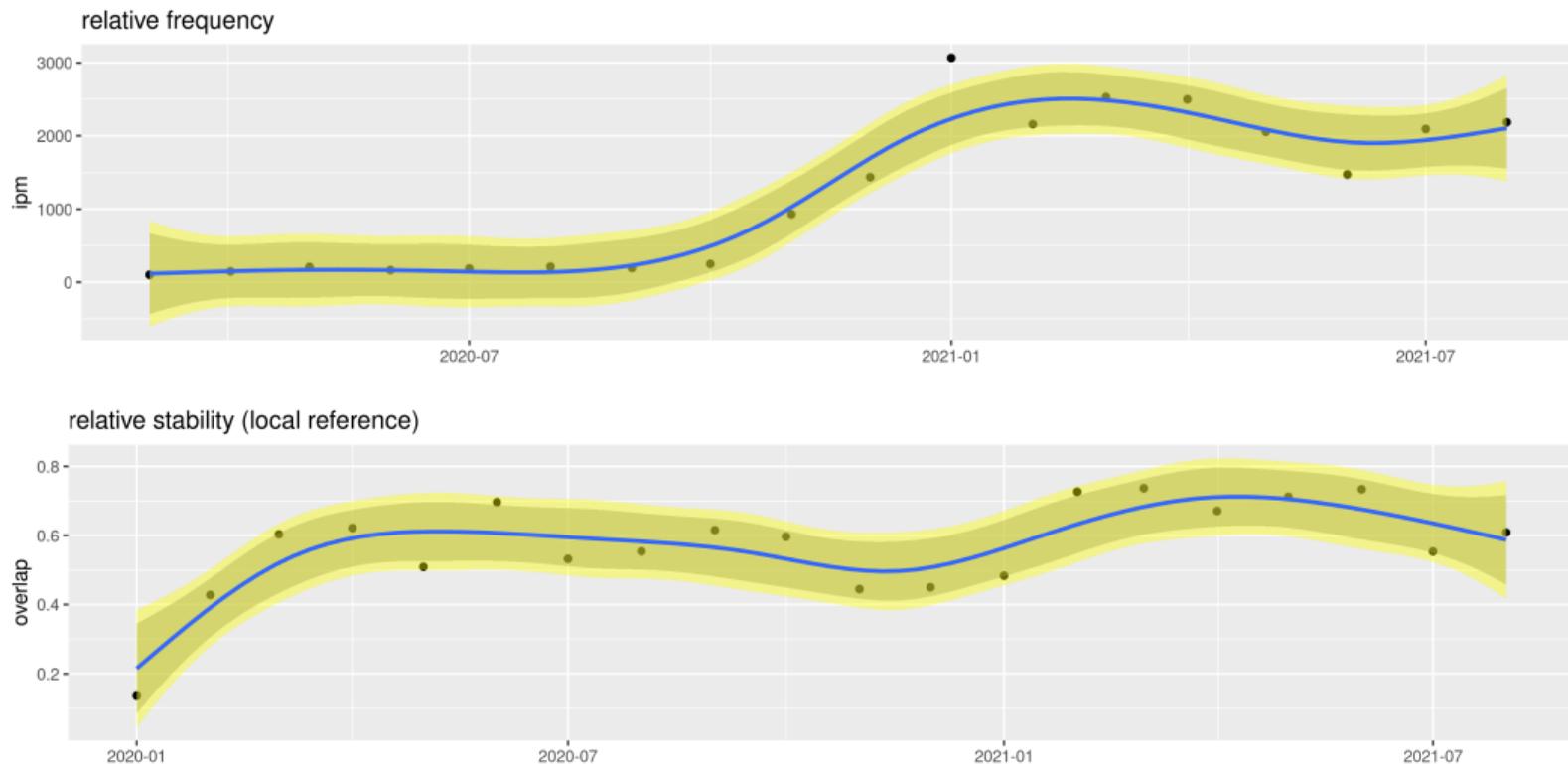
## Kollokate: Impfen (französisches Korpus)



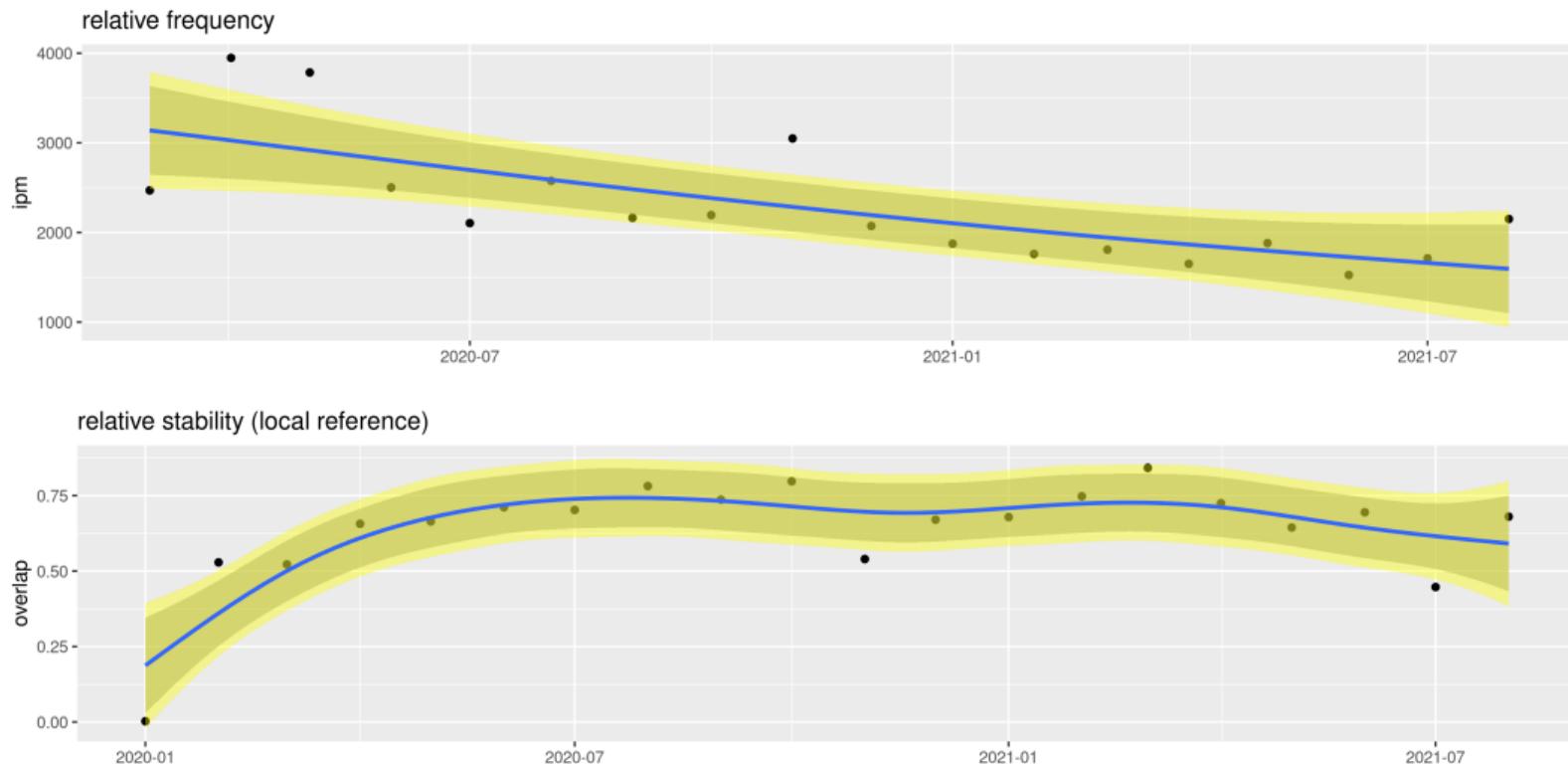
# Kollokate: Maßnahmen (französisches Korpus)



# UFA: Impfen (französisches Korpus)



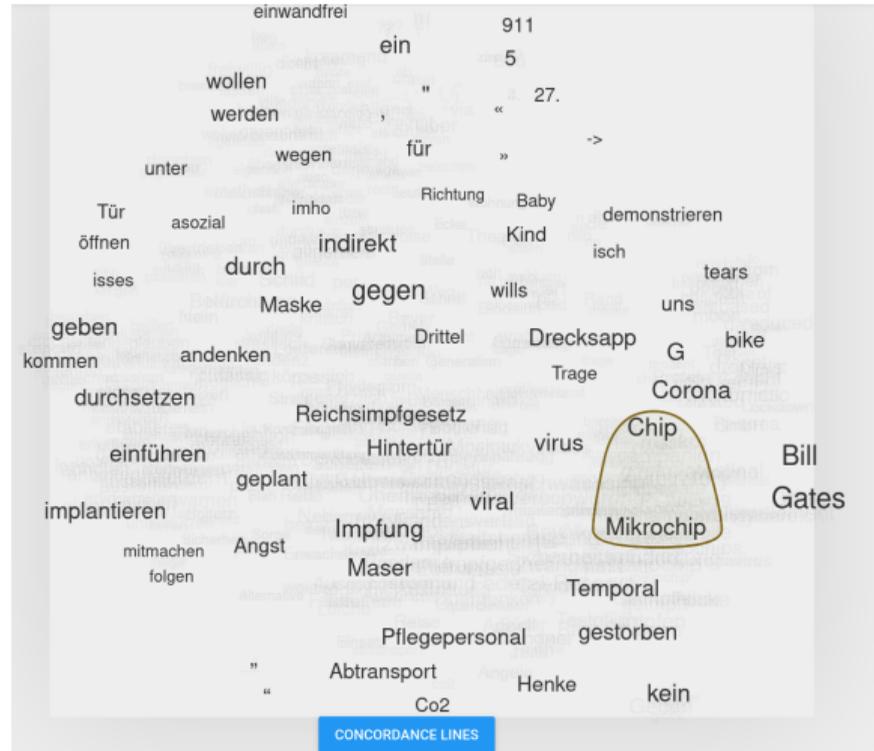
# UFA: Maßnahmen (französisches Korpus)



## Zusammenfassung & Herausforderungen

- insgesamt ähnlicher Diskurs in deutscher und französischer Berichterstattung
  - ▶ Erreger & Maßnahmen durchgehend dominant
  - ▶ Impfdiskurs wichtig ab Ende 2020
- am meisten Arbeitsaufwand:
  - ▶ manuelle Kategorisierung
  - ▶ Disambiguierung
- Ambiguität im Französischen problematischer als im Deutschen
  - ▶ insb. auf Grund von Komposita
  - ▶ bspw. *système sanitaire* vs. *Gesundheitssystem*

## MMDA toolkit



## MMDA toolkit

Mixed Methods Discourse Analysis

HOME VIGNETTE EN LOGOUT

admin 170

Keyword Analysis  
Collocation Analysis  
Discourseemes  
Constellations  
Administration

Wordcloud Options  
Collocation Parameters  
Second-Order Collocation  
Comparison  
Search Item

ID	... context	keyword	context ...
1	Toussaint gehört wie Prof. Raoult, Prof. Perronne, Dr. Bellier und Dr. <u>Henrion-Caude</u> zu den <u>Gegner</u> des	Maskenzwangs	, für den es keine wissenschaftliche Begründung gebe
2	... mal gesehen ", "Keine Freiheit ohne Aufklärung ", "Vertraut auf Gott ", "so GATES nicht ", "Impfen muss <u>freiwillig</u> bleiben ", " <u>Stoppt</u> den	Corona-Wahnsinn	", sowie <u>einige</u> ausgedruckte Überschriften , die die Validität sogenannter PCR-Tests infrage stellen , um
3	Diese deutsche	Corona-Lüge	muss endlich , endlich <u>aufhören</u> .
4	In Le Parisien haben 140 Mediziner <u>einen</u> Aufruf veröffentlicht , den	Corona-Wahnsinn	endlich zu <u>beenden</u>
5	... wie Andreas Hoff und Dirk Brandes , Reden <u>gegen</u> die Coronapandemie-Maßnahmen <u>und gegen</u> unterstellte Zwangsimpfung und	Maskenzwang	halten wollen
6	In den Reden , <u>die gehalten</u> wurden , wurde durchgängig verschwörungstheoretisch <u>gegen</u> den ,	Corona-Wahnsinn	" und <u>angebliche</u> Interessen des Microsoft-Gründers Bill Gates gewertet
7	Und hier das <u>aller , aller wichtigste</u> , was <u>Ihr</u> aus dieser <u>schwachsinnigen</u>	Pseudopandemie	mitnehmen sollten : Diese <u>Lockdowns</u> und Impfungen helfen <u>NUR</u> den Privilegierten !
8	Unter dem Motto "	Corona-Wahnsinn	<u>beenden</u> " hatte <u>der</u> NPD-Stadtrat von Aue Bad-Schlerma , Stefan Hartung , am Freitagabend zu einer Kundgebung auf den Anton-Günther...
9	Ich <u>wehre</u> <u>mich</u> auch <u>gegen</u> den	Maskenzwang	, indem ich <u>keine aufsetze</u> , <u>außer</u> ich werde <u>direkt</u> gezwungen , an Eingängen usw und wenn ich da reinmuss , aber nicht im öffentlichen Na...
10	...gt hat er muss sich fürs Christentum einsetzen oder ob jetzt irgendem gemaischer Gott gekommen ist , <u>der</u> ihn befiehlt er muss sich <u>gegen</u> die	Corona-Lüge	und 5 G -Ausbreitung einsetzen ... wo liegt <u>der</u> Unterschied ?

Rows per page: 10 1-10 of 42



Vielen Dank für Ihre Aufmerksamkeit.

**Fragen?**