FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

PHILOSOPHISCHE FAKULTÄT
UND FACHBEREICH THEOLOGIE

A keyword categorisation study on COVID-19 conspiracy discourse

Nathan Dykes, Andreas Blombach, Linda Havenstein, Philipp Heinrich, Besim Kabashi, Stephanie Evert, Fabian Schäfer

FAU Erlangen-Nürnberg

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
PHILOSOPHISCHE FAKULTÄT UND FACHBEREICH THEOLOGIE

Tracking the Infodemic

- Computational Corpus Linguistics / Japanese Studies, FAU
- Corpus Linguistic methods combined with domain knowledge on far-right discourses and misinformation

Aims:

- Trace the distribution of conspiracy narratives on social media
- Analyse typical linguistic patterns and discursive strategies, especially w.r.t. the overlap with right-wing extremist and populist discourses
- Automatable methods to enable application to the spread of conspiracy theories and misinformation in the future

https://github.com/fau-klue/infodemic
https://www.linguistik.phil.fau.de/projects/tracking-the-infodemic/

Methodological background

- Keywords as entry points to CADS (cf. Baker 2004)
- Discussion on procedures mostly focused on appropriateness of association measures (e.g. Lijffijt et al. 2016)
- Categorisation usually not reproducible / transferrable
- Focus on content > linguistic patterns

Research questions

How can we use keywords to systematically explore discourse by important actors in the German COVID-19 conspiracy discourse?

- What do different annotation strategies uncover – do we get comparable results?
- Role of formal vs. content-based exploration (linguistic markers vs. domain expertise)
- Similarities and differences in topics and style of two central actors

Data and methods

- 130+ public channels (known conspiracy theorists and "alternative media"), 20 public chat groups
- > 4.5m posts, 150m tokens
- Focus today:
  - Eva Herman (former news anchor)
  - Boris Reitschuster (journalist)
- Association measure: LogRatio (conservative estimate – LRC; cf. Evert et al. 2018, Evert 2022) – intends to capture mid-frequency keywords as a compromise between significance and effect size

Data and methods

- Annotation of top lexical 200 keywords compared to a general reference corpus (*DWDS Kernkorpus*)
  (POS: NN; NE; ADJA; ADJD; HST; FM; VVFIN; VVPP; VINF; VVIMP; VVIZU)

- Three independent annotators – researcher triangulation (cf. Marchi & Taylor 2009; Baker & Levon 2015)
  - Two researchers with extensive domain knowledge; main contributors to the overall project; developed an inventory of prominent conspiracy narratives: Traditional CADS-style grouping process; developed their own groups based on their domain knowledge[1]
  - One researcher familiar with the discourse, but less involved: annotation of linguistic properties instead of content-based keyword groups

1 https://github.com/fau-klue/infodemic

Results – content-based categorisation

Annotator 1:

top categories EH: advertisement, prepping, COVID-19, interaction, vaccine, esoterics, counter-measures, other, German politician/authority, elites, social media, US-politics/scandals, channel operator

top categories BR: German politician/authority, media criticism, counter-measures, COVID-19, anti-left/Green, channel operator, vaccine, interaction, anti-measures, pseudo-pandemic, other, government communication

Other shared categories: media criticism, pseudo-pandemic, social media, US-politics/scandals, far right, Russia, title, company, conspiracy, scientist, scientist („enemy"), scientist („ally")

Results – content-based categorisation

Annotator 1:

Unique to BR: AfD, clickbait, pseudo-democracy, hypocrisy, measures more harmful than COVID, anti-left / Green, anti-measures, media, lack of free speech, victimisation, politician (ally), government communication, Belarus

Unique to EH: alternative remedies, alternative medicine, threats, elites, esoterics, EU politicians/ authorities, Gates, children, harming of children and women, prepping, conspiracy theorist, telegram, polling results, advertisement

Results – content-based categorisation

Annotator 1:

There is a strong focus on COVID-19 and counter-measures, particularly the vaccine. Political figures and other authorities play a central role in both channes. They also commonly refer to social media channels by conspiracy spreaders and social media platforms. Both channels mention 'good' and 'bad' scientists and make references to specific conspiracy narratives as well as far-right discourses.

Annotator 2:

Both channels focus on COVID itself and associated measures, especially vaccines. They routinely reference central politicians and other actors that belong to the (German) public sphere who are considered authorities, as well as counter-public spheres, who organise on social media channels, and point out specific individuals who are known conspiracy spreaders. This suggests a "strong us vs. them" framing.

Results – content-based categorisation

Annotator 1:

EH has a stronger focus on social media and direct interaction with her audience. She also mentions US-inspired themes more frequently – including references to Gates, and QAnon-related narratives such as women and/or children being targeted. Unique to her channel are mentions of alternative medicine and alternative treatments to COVID-19, esoterics, central actors on a European level, prepping and advertisements.

Annotator 2:

EH has a stronger emphasis on her community, as well as narratives referencing US discourse such as prepping and religious cues. She promotes esoterics and alternative medicine to help against COVID; suggesting that she does not necessarily deny the disease itself.

Results – content-based categorisation

Annotator 1:

BR has more focus on critiquing established media outlets and political figures. The depiction of COVID-19 as a pseudo-pandemic is also among his top keyword categories. He makes favourable references to the far-right party AfD and protest movements such as Querdenken, and pronouncedly frames 'the leftists/ Greens' as an enemy. The state is criticised to be manipulative, authoritarian and imposing unreasonable measures, not allowing for counter-voices to be heard, and critics of the measures are depicted as victims of the system.

Annotator 2:

BR has more emphasis on the political measures that are taken against COVID and the way that the pandemic is covered in public media and their style of reporting. He uses emotive terms to depict things he does not agree with ('incredible', 'hate') and language typically used in media coverage ('search for clues', 'exclusive'); suggesting that he might be framing himself as a 'proper' source of media.

Procedure: formal categorisation

Pre-determined formal annotation categories, cf. Dykes et al. (2020, 2021)

- Innovation: creative word formations, neologisms etc.
- Multiple constituents: more than one lexical constituent (compounds, complex hashtags etc.)
- Markers of computer-mediated communication (CMC): hashtags, links, usernames etc.
- Specialised lexis/ jargon: words that are not part of everyday discourse
- Clippings and abbreviations (*WHO*)
- Names: person, place, institution, others
- Manually assigned semantic tags, following the UCREL tagset (Rayson et al. 2004)

Results – formal categorisation

Compounds:

Key in both (17):
only notably common constituent: *Corona-* (6), but most other first constituents clearly related to the pandemic (‚*vaccine*', ‚*infection*', *pcr-* …).
Second constituents mostly relate to political measures (‚*law*', ‚*chancellor*', ‚*crisis*',‚*politics*', …)

Results – formal categorisation

Compounds

BR only(48):

- overall focus on politics, measures, media and protests in constituents occuring more than once (‚corona/COVID'„chancellor', ‚protest', ‚leader' ‚Green')
- pointers to pseudo-pandemic narrative (‚ICU beds','COVID numbers' …) and „unconstitutional measures" (‚*constituional rights*', ‚*constitutional judge*'), right-wing (‚*migration background*'), biased media (‚*fee-funded*', ‚taxpayer', ‚*freedom of opinion*')

Results – formal categorisation

Compounds – EH only(61):

- COVID and measures (*corona/COVID, vaccine, virus*)
- Interaction (*evening prayers, voice message, greetings, channel*)
- Prepping and food-making (*filter, bread, food, chain, oven, stock*)

Specialised language

- Overlap: *COVID, lockdown, PCR test …*
- BR only: *framing, to frame, narrative, quarantine, case numbers …*
- EH only: *hollow-fibre filter(?), chlorine dioxide solution, ferment, fermentation glas, base powder, blood aura(?), herd immunity*

Results – formal categorisation

CMC features:

- Only hashtags; BR (18) seems to use them a lot more than EH (2) – contradicts annotators1 & 2‘s interpretation of more interaction if we see hashtags as creating intertextuality?

Not necessarily:

- BR‘s hashtags reference state media (#ard #zdf #tagesschau), names of politicians (#merkel #laschet) and COVID + measures (#astrazeneca #corona)
- EH‘s two key hashtags #abendgebet (evening prayer) and #stabildurchdenwandel (persisting through the change) are specific to her community: communal evening prayers and voice messages with her perspective on daily events

Conclusions

- Similar categorisations from annotators 1 and 2 despite no interaction on this particular task and no prior discussion of these actors
- … NB! (obvious) differences in label names, but also granularity (e.g. *politicians* vs. *US politicians* and *German politicians*) and level of interpretation (*social media platform* vs. *counter-public*) – might conceivably lead to different focuses in further interpretation – **we only focused on the big picture emerging from starting-points**!
- Similar results indicated by topic-agnostic annotation procedure – lacks some depth content-wise, but can help to show *how* communicative outcomes are achieved
- Both EH and BR unsurprisingly reference exactly the themes we would expect, („so what", Baker & Levon 2015: 232), but also differing perspectives
- Closer linguistic focus can help to shed light on ‚so how'

References

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. Journal of English Linguistics, 32(4), 346–359.

Baker, P., & Levon, E. (2015). Picking the right cherries? A comparison of corpus-based and qualitative analyses of news articles about masculinity. Discourse & Communication, 9(2), 221–236.

Dykes, N., Peters, J. , & Evert, S. (2020). Categorising keywords in discourse. A case study of texts on bacterial resistance. Corpora and Discourse International Conference, Brighton.

Dykes, N., Heinrich, P., Evert, S. (2021). Keywords gone viral. Exploring keyness techniques for corpus-based discourse analysis in German tweets on COVID-19. Corpus Linguistics, Limerick.

Evert, S. (2022). Measuring keyness. Digital Humanities, Tokyo. https://osf.io/cy6mw/files/osfstorage

References

Evert, S., Dykes, N., & Peters, J. (2018). A quantitative evaluation of keyword measures for corpus-based discourse analysis. Corpora and Discourse International Conference, Lancaster.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. Digital Scholarship in the Humanities, 31(2), 374–397. https://doi.org/10.1093/llc/fqu064

Marchi, A., & Taylor, C. (2009). If on a winter's night two researchers…: a challenge to assumptions of soundness of interpretation. Critical Approaches to Discourse Analysis Across Disciplines: CADAAD, *3*(1), 1-20.

Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system. Proceedings of the beyond named entity recognition semantic labelling for NLP tasks workshop, 7-12.