# A keyword categorisation study on COVID-19 conspiracy discourse

Keyword analysis is central to corpus-assisted discourse studies (CADS); which is plausible given that it provides a relatively low-barrier way to comparing two corpora. For instance, comparing a specialised corpus can be compared to a larger reference corpus can serve to determine starting points for a more fine-grained analysis. The rationale behind this idea is that prominent words, when grouped into more abstract categories, will likely also indicate central discourse strands (cf. Baker 2004). However, there is no best practice as to how these categories are formed, and this question has so far received little attention. More generally, only few CADS-related projects explore different strategies directly on the same data (e.g. Baker and Levon 2015, Marchi and Taylor 2009)

This study compares two different approaches to keyword categorisation CADS. We explore how two different keyword-based approaches studying the same data result in similar, complementary, or even conflicting findings. We investigate keywords for a German Telegram channel that is known for spreading conspiracy beliefs on the COVID-19 pandemic. In order to compare our categorisation methods at different granularities, keywords are computed in comparison to two different reference corpora: first, we compare a channel to a large, topic-agnostic reference corpus and second, to the rest of our Telegram corpus, containing other conspiracy-related channels. In both cases we use conservative log ratio as an association measure (Evert et al. 2018).

The top keywords in both lists are categorised by separate researchers using one of the annotation strategies that are to be compared.

The first approach is comparable to the classic procedure in CADS. It relies on close familiarity with the overall discourse on conspiracy narratives and their relation to the pandemic. Common conspiracy theories and more general narratives that play into these (e.g. COVID-19 is harmless or vaccination causes sterility) have been identified beforehand for a different case-study (primarily intended for text-level annotation). Since the association of certain words with a given narrative or theme often depends on the context in which they are used, we expect only a relatively low number of keywords to be unambiguously related to these categories, however. The actual labels used to categorise keywords are therefore formed ad-hoc by researchers by studying the keywords and their concordances under the lens of this background knowledge

In the second approach, the annotation scheme stays very close to the linguistic surface, focusing on general semantic, lexical and morphological properties such as word formation patterns, jargon, and proper nouns. Instead of focusing on discourse-specific distinctions as in the first strategy, we annotate general-purpose semantic categories (Rayson et al. 2004). In this approach, there are no false positives because every keyword will have some annotatable property.

We hypothesise that the first strategy achieves deeper insights into specific discourse phenomena due to more fine-grained distinctions in cases where unambiguous categorisation is possible. The second strategy, on the other hand, is expected to be more

robust and transferable to other discourses, while still covering most major discourse strategies identified in the first strategy.

## 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.

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

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

Rayson, P., Archer, D., Piao, S., & McEnery, A. M. (2004). The UCREL semantic analysis system. *Proceedings of LREC*.