Identifying keywords and phrases in British COVID-19 newspaper discourse

Julia Schilling & Robert Fuchs
University of Hamburg
julia.schilling@uni-hamburg.de, robert.fuchs@uni-hamburg.de

The COVID-19 pandemic has upended life around the globe, leading to intense public debate and a flurry of lexical innovation across many languages. (Socio-)Linguists quickly started to document and analyze COVID-19 discourse (Baines et al. 2021; Saraff et al. 2021), but there is as yet no systematic analysis of the lexical items and discourse patterns that characterize British COVID-19 discourse. We address this research gap through a systematic comparative analysis of public discourse during the COVID-19 pandemic. Through a big data approach, we identify not just distinct keywords and phrases linked to the pandemic but also track their development over time and across regions.

As news can offer an insight into and simultaneously influence the public's perception of the COVID-19 pandemic, our analysis focuses on discourse in regional and national English newspapers. The starting point of the analysis is a contrastive keyword analysis of the discourse of every month of 2019 with its equivalent in 2020 and 2021, comparing pandemic with pre-pandemic discourse, while filtering out seasonal effects (e.g. discussion of *snow* in January). Our data comprises 10% of all articles from 51 national and regional English newspapers published between January 2019 and October 2021, producing a corpus of approximately 386,118 articles and 229,347,771 tokens. Rather than collecting newspaper articles based on a pre-existing list of keywords, we use a data-driven approach to identify COVID-19 related n-grams ($1 \leq n \leq 4$) for each month of the pandemic based on log likelihood and log ratio. We then assign these keywords to semantic fields such as COVID-19 NAMES (e.g. *Covid-19*, *SARS-CoV-2*), PUBLIC HEALTH INSTRUCTIONS (e.g. *self-isolation*, *quarantine*), and VACCINATION and examine their development over time using statistical measures such as median, IQR, standard deviation, and skewness of the distribution.

This analysis yielded over 300 1-grams, 350 2-grams, 200 3-grams, and 100 4-grams related to the COVID-19 pandemic. Results indicate that the lexis of COVID-19 discourse in British newspapers significantly varies not only over time, but also within semantic fields of discourse and across regions.

Baines, Annalise; Ittefaq, Muhammad & Mauryne Abwao (2021) "#Scamdemic, #Plandemic, or #Scaredemic: What Parler Social Media Platform Tells Us About COVID-19 Vaccine". *Vaccines* 9 (421), pp. 1-16.

Saraff, Sweta; Singh, Tushar & Ramakrishna Biswal (2021) "Coronavirus Disease 2019: Exploring Media Portrayals of Public Sentiment on Funerals Using Linguistic Dimensions". *Frontiers in Psychology* 12:626638.