Rauh, Christian

Clear messages to the European public? The language of European Commission press releases 1985–2020

Journal of European Integration

Clear messages to the European public?
The language of European Commission press releases 1985-2020

APPENDIX

Christian Rauh

WZB Berlin Social Science Center

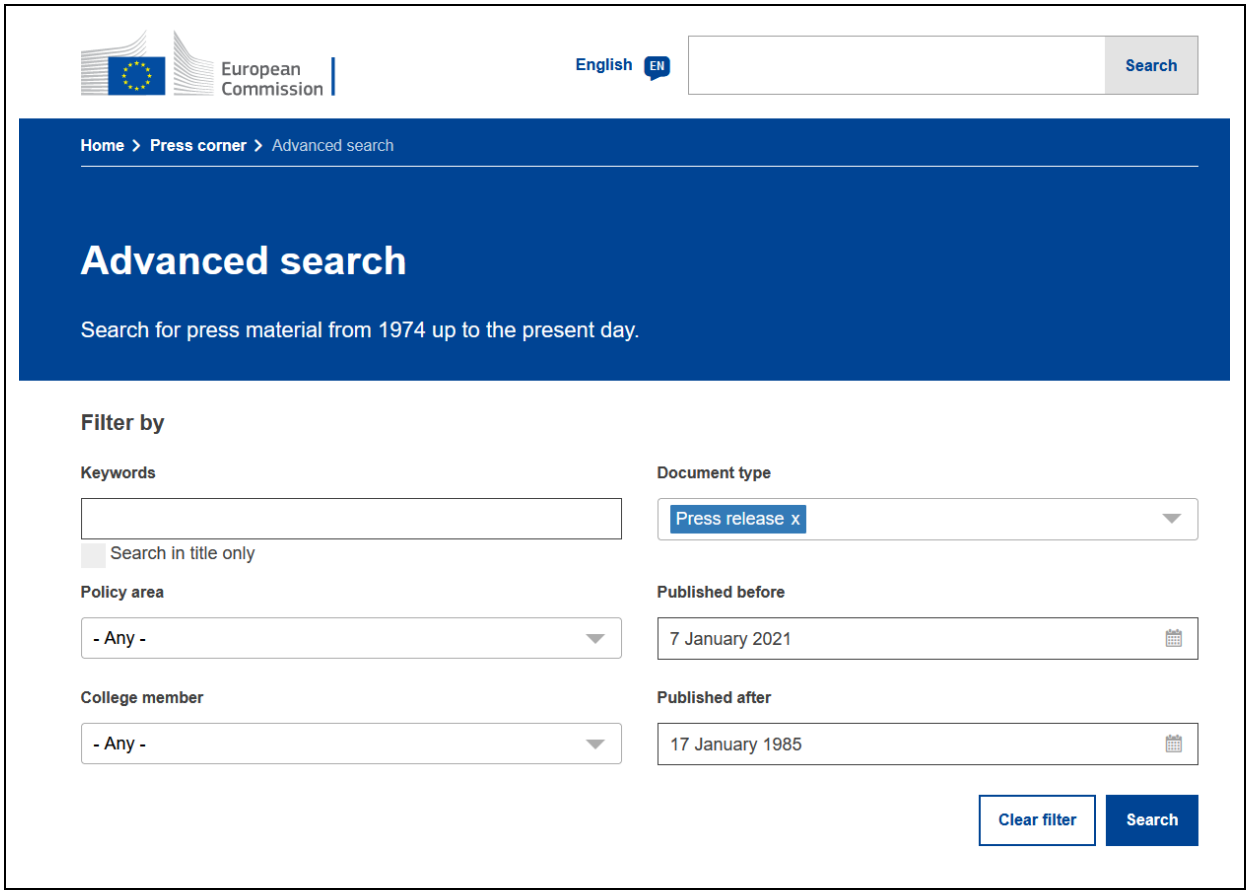
christian.rauh@wzb.eu

www.christian-rauh.eu

0000-0001-9357-9506

Appendix A: Constructing the EC press release corpus

As a first step of data collection, a manual search for all press releases was initiated in the online archive of the Commission's press services (accessed January 8, 2021, https://ec.europa.eu/commission/presscorner/advancedsearch/en). The figure below indicates the chosen search parameters.



On January 8, 2021, this search produces 46,507 results. My first scraper then automatically traverses through the results pages, to collect any link to an IP ('information presse') document in the English language (marked by the /en/ element in the URL to an individual IP document). For 46,437 documents (99,85%) such a link is available.

A second scraper then calls each of these URLs individually, downloads the full HTML page, to then extract headline, lead, and body text of each individual press release. In 1,459 cases (3.14% of all search results) this procedure failed for technical reasons: most are IP documents in the very

early investigation period 1985/6 that are provided only as non-machine-readable images, a few links in the Commission's archive are dead (HTTP response 404), and in a few instances the wrong language version was stored within '/en/' URL. In sum, the resulting corpus covers 96.86% of all press releases that the Commission archive offers for the period from January 17, 2985, to January 7, 2021.

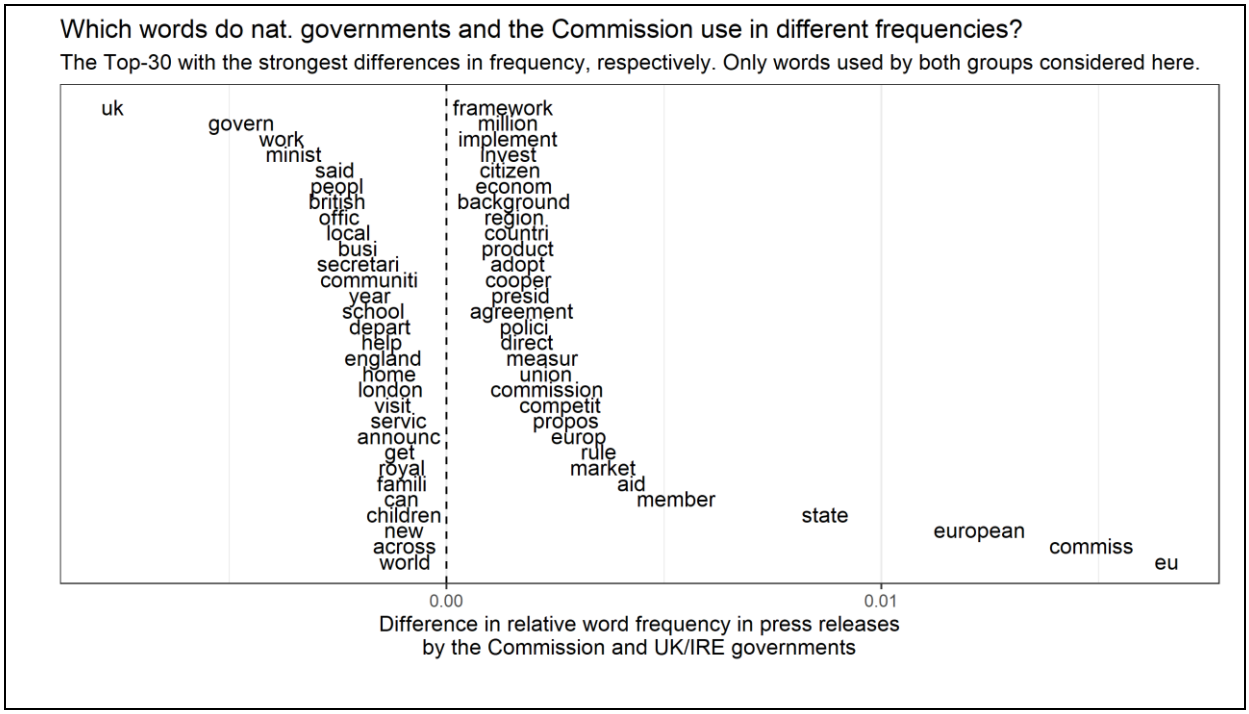
For scholars interested in expanding the corpus either in terms of time or in terms of other EU languages, I provide the scraping scripts in the replication materials for this article at https://doi.org/10.7910/DVN/UGGXUF

Appendix B: Details on the topic matching approach

The text matching approach described in the main text is restricted to the period for which my data provide both Commission and national press releases (January 2010 to January 2021) and comprises 103,443 documents in total.

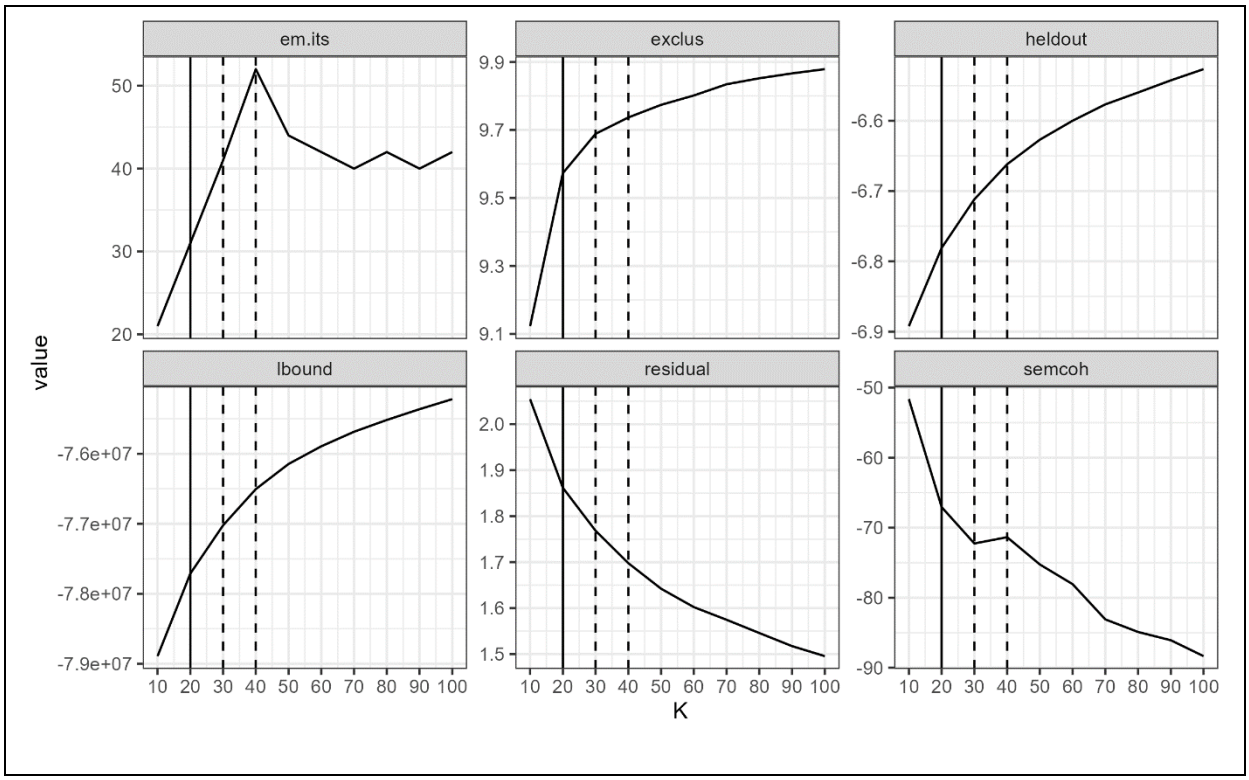
The primary interest of constructing this joint corpus lies in estimating whether the clarity indicators differ between the Commission and national press even when one controls for the topics covered by individual press releases. Given the large data size, I opted for an automated approach of topic classification resorting to the well-known structural topic model (Roberts et al., 2014), which optimizes the distribution of recurring and partially overlapping word clusters across documents towards a pre-specified number of topics.

In my application, one initial concern for estimating this model is the possibility that the Commission and national executives use partially different words even when speaking about the same topics. I thus first inspected differences in relative word frequencies across Commission and national documents. The figure below highlights that these differences stay well within the range of +/- 5 percentage point range. However, a few words are used differently. Notably the European Commission speaks consistently more about itself, about the EU and about member states.



To avoid that such differences load into separate topics for Commission and national documents, I estimate the STM with a content covariate so that slight changes in word frequency within topics are allowed to differ by sender.

While the STM algorithm learns word cluster from data itself, the researcher must specify the number of k topics beforehand. The figure below provides the standard parameters that are usually used to for a data-driven choice in this regard across different numbers of k topics to optimized for.



This initially shows that the marginal gains in the statistical goodness-of-fit measures, such as the lower bound of the maximum likelihood (lbound), the heldout likelihood, or the estimation residuals, start to decrease above a number of 20 topics. The same holds for the more content-oriented parameters, such as the word-to-topic exclusivity (exclus) or the frequency by which the top words per topic occur frequently together (semantic coherence, semcoh). Overall, these patterns suggest that the optimal number of topics for this particular corpus lies somewhere between 20 and 40 topics. As I am not primarily interested in interpreting topic content

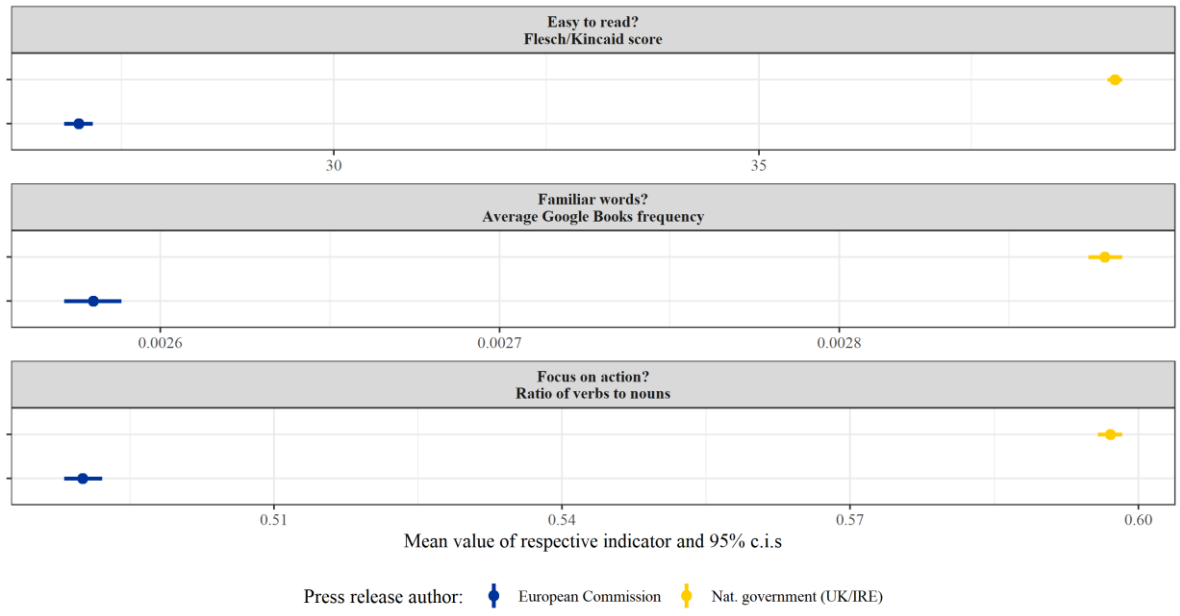
substantially here but rather want to exploit it for efficiently presenting a topic-based comparison, I finally opted for the 20-topic solution.

Having estimated this model, I then extract the distribution of topics per document (the so-called theta values). The first comparison described in greater detail in the main text then simply groups the corpus by documents that have the highest prevalence of the same topic. Put differently, I label each PR from both the Commission and the national executives with the topic that has the highest estimated theta value within each document. Afterwards I split the corpus into European and national PRs to then compare the three clarity indicators across Commission and national senders within texts that apparently focus on the same main topic.

However, each document is described by a distribution of all 20 topics and there is the possibility that the full topic distribution within and across documents may vary systematically between Commission and national PRs, which would confound my main comparison in Fig. 2 of the main text. I thus also employ Roberts et al (2020: 893) topical inverse regression matching (TIRM) approach to ensure that not only the full theta distribution within documents is by-and-large identical (coarsened, using the thresholds provided by Roberts et al) but that also the likelihood of being treated (=being a Commission PR) is identical in the analyzed population (which reduces the size of the joint corpus from ~103k to ~82k documents, indicating some variation in topic prevalence across both types of authors). Yet and still, in this balanced/matched population the differences in my three clarity indicators are statistically still highly significant and numerically similar to the results in the full sample (Figure below), thereby topic confounding can be ruled out.

All of these steps can be reproduced, using the scripts 'X-SearchK.R' and ' 5-PRs_TextMatching.R' provided in the replication package to this article at https://doi.org/10.7910/DVN/UGGXUF

Comparing language of Commission and national press releases in a sample matched on topic distributions in full texts



Press release author: European Commission Nat. government (UK/IRE)

Coarsened exact match on topic distributions from a k=20 STM with a content covariate accounting for wording differences between Commission and national press releases.
82,132 press releases (Commission: 11,321; Nat. gov.: 70,811) published between January 2010 and January 2021.