Texts and Images of Austerity in Britain. Deduplication Process

General Procedure

After removing duplicates that are due to download errors, we are left 23,687 published in 2010 or later for which we parse relevant meta data from the text, namely the date, source (Telegraph or Guardian), its journal code, the section, page, length (in words), and its edition. In order to improve the quality of our corpus and its metadata, we (1) process the meta variable and (2) remove duplicates and further unwanted texts from the corpus, resulting in a final corpus of 18,353 texts.

The following three meta variables are added: (1) a binary meta variable for the distinction between medium (online vs. print). We base this distinction on four easy rules: Firstly, articles published on guardian.com according to the text are "online"; secondly, all Telegraph articles are "print" (as a manual inspection shows); thirdly, all Guardian articles published after 2014 are online content. Last but not least, articles that do not have any page number in text are classified as "online" (this is consistent with the first three rules). (2) We also add a dummy variable showing whether a given article contains an image whose caption contains the token "austerity" (323 articles), and (3) introduce coarse-grained sections by means of collapsing the sections found in the text.

The removal of unwanted articles is slightly more complex. In a first step, we remove any article that is easily identifiable as unwanted online content: articles that contain "society daily" in the headline (an online column that we are not interested in) and articles that contain the tokens "cribsheet" or "block-time" in their body (obvious markers for online blog content gained from qualitative inspection of the articles); this procedure removes a total of 910 articles.

We then proceed to perform a rule-based deduplication based on text similarity. In a first step, we therefore perform pairwise comparisons of all texts: For a given text A, we calculate the proportion of tokens of A that appear in sentences which also appear in any other text B of the corpus. (If A is totally contained in B, this value is 1, even if B is much longer than A). Afterwards, we create a stratified sample of text pairs with different similarity values and interpret these qualitatively with regard to their (near-)duplicate status in order to decide upon the cut-off value for the similarity measure and to develop a more sophisticated decision rule for the detection of duplicates with the help of the meta-data. We leave out duplicate-pairs in which one article appears on page 1 and the other on a following page in the same newspaper (since these "duplicated" articles represent teasers on the front-page which we want to keep for qualitative reasons).

All other article-pairs with (unilateral) similarity-score of 0.2 or higher are (1) checked for their medium: if one of the articles is from a print medium and the other from an online medium, we remove the online articles (this removes 474 articles). (2) We check for editions: if one of the articles appears in an edition with a lower count (say, edition

1 compared to edition 3), the article of the later edition is being kept. Likewise, we remove any article in duplicate-pairs for which there exists a version of the 'national' edition compared to local (Scottish, Irish) edition. This edition-wise deduplication removes a further 1,705 articles. Lastly, (3) we build transitive clusters of duplicates (i.e. we include any article in a given cluster that shows a similarity score of at least 0.2 to some other article in the cluster) and chose articles based on two simple rules: (a) We keep articles with images and (b) we remove any but the longest article (based on the number of tokens) of each cluster. This last rule removes 2,244 articles.

The whole corpus preparation thus reduced the number of articles in the corpus from 24,743 that have been downloaded from LexisNexis to 18,353 (274 of which are linked to one of 271 unique images). For other ways to deal with duplicates in LexisNexis data see Grundmann et al. (2017: 100) and Baker (2014: 160f.)

Codebook for the qualitative coding of possible (near-duplicates)

This codebook explains the qualitative steps that have been taken with respect to the detection of (near)duplicates that are described above. The coding was conducted by Sophie-Marie Himmler, Karolina Kohl and Tim Griebel during February 2018. Please contact Tim for more information about the de-duplication process.

Definition of (near-)duplicates:

"Two documents are regarded as duplicates if they comprise identical document content. Documents that bear small dissimilarities and are not identified as being "exact duplicates" of each other but are identical to a remarkable extent are known as near duplicates" (http://www.ijcscn.com/Documents/Volumes/vol2issue2/ijcscn2012020201.pdf)
Similarities in content concern the *forms of texts on the linguistic level*, not on the argumentative or ideological level. The basic question is: Does a pair of texts contain (basically) the same article twice? We are *not* concerned here with the much broader question if a pair of texts deals with the same topic or contains the same arguments or ideologies *although they do not* show similarities in their linguistic forms.

Coding Rules with Regard to the Detection of Duplicates

1. Open the (near-)duplicate database and the excel file "Duplicate detection_manual_annotation".
2. Look at each pair of texts and decide if they contain a (near-)duplicate.
    a. If the pair does *not* contain a (near-)duplicate, both texts should be kept within the corpus. So, put an "x" into both columns "keep_A" "keep_B" in the excel file.
    b. If the pair does contain a (near-)duplicate, only one text should be kept within the corpus. Keep the longer article and put an "x" into the column of this text.
    c. Special rules for the Guardian for the steps 2a and 2b:
        i. If a pair of (near)duplicates contains both printed and online content, keep always the text for the printed version

    ii. Take care about the year that an article has been published. Do not keep articles that contain online content for the Guardian in the years 2010-2014. The marker here is “Guardian.com” in the meta data
3. Indicate any observation that might be interesting for the evaluation of the quality of the (near-)duplicate detection in the column “remark”.