Optimized Dictionaries: A Semi-Automated Workflow of Concept Identification in Text-Data

Röth, Leonce; Kaftan, Lea; Saldivia Gonzatti, Daniel

Preprint / Preprint
Arbeitspapier / working paper

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

Empfohlene Zitierung / Suggested Citation:
Röth, L., Kaftan, L., & Saldivia Gonzatti, D. (2024). *Optimized Dictionaries: A Semi-Automated Workflow of Concept Identification in Text-Data.*. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-91666-3

Optimized Dictionaries - A Semi-Automated Workflow of Concept Identification in Text-Data

Leonce Röth*, Lea Kaftan+ & Daniel Saldivia Gonzatti‡[1]

* LMU University of Munich, + GESIS — Leibniz-Institute for the Social Sciences,
‡ WZB Berlin Social Science Center

Abstract: Identifying social science concepts and measuring their prevalence and framing in text data has been a key task of scientists ever since. Whereas debates about text classifications typically contrast different approaches with each other, we propose a workflow that generates optimized dictionaries that are based on the complementary use of expert dictionaries, machine learning, and topic modeling. We demonstrate our case by identifying the concept of "territorial politics" in leading newspapers vis-à-vis parliamentary speeches in Spain (1976-2018) and the UK (1900-2018). We show that our optimized dictionaries outperform singular text-identification techniques with F1-scores around 0.9 for unseen data, even if the unseen data comes from a different political domain (media vs. parliaments). Optimized dictionaries have increasing returns and should be developed as a common good for researchers overcoming costly particularism.

Words: 8,030

Acknowledgments: We thank Keno Röller-Siedenburg, Saskia Gottschalk, Nermin Abbassi, and Moritz Raykowski for their valuable help with collecting parts of the data used in this article. We are very thankful for the helpful comments by André Kaiser, Sven-Oliver Proksch, and participants of the 2019 CES Madrid conference, the 2019 Zurich Text-As-Data workshop, the 2020 ECPR Online General Conference, and the CCCP Research Seminar at the University of Cologne.

[1] Corresponding author: daniel.saldivia-gonzatti@wzb.eu. Funding: this research was partly supported by the DFG grant KA 1741/10-2.

1 Introduction

Concepts such as inequality, decentralization, or neoliberalism are fundamental for the social sciences and the understanding of our societies. Content analysis has been routinely applied in social science to identify and map the prevalence and composition of those concepts, predominantly using methods of hand coding by experts. However, technological advances have pushed the analysis of political communication, and text, in particular, in new directions and allow the computer-assisted identification and classification of relevant concepts in the social sciences.

We follow the recent development in comparing the performance of computer-assisted methods in the identification of text (Widmann & Wich 2022; Kroon et al. 2022; Grimmer et al. 2022) but with a systematic focus on the fundamental and technical sources of error. We discuss the general advantages and disadvantages of the most frequently used methods in this domain (dictionaries, machine-learning, and topic models). Based on this overview we follow those who have recently indicated the potential to exploit the comparative advantages of using different approaches complementarily (Nelson et al 2021; Watanabe & Zhou 2022; Radford 2021). As a result, we portray a universal workflow that generates ***optimized dictionaries***, an approach that arguably combines the strengths of the singular methods of dictionaries, supervised machine learning, and topic models and is in principle extendable to exploiting the benefits of methods such as word embeddings and transformer models.

Based on earlier studies highlighting the importance of active learning and careful keyword or seed word selection (King, Lam, and Roberts 2017; Miller, Linder, and Mebane 2020; see also Druck, Mann, and McCallum 2008; Mahl et al. 2022; Rinke et al. 2022), we argue that optimized dictionaries have at least four clear advantages over existing singular approaches. (1) The classifications of optimized dictionaries have higher validity than every singular method alone. (2) Optimized dictionaries are transferable to new and unseen text and thereby allow for the straightforward comparison of results across sources. For example, optimized dictionaries can be developed on a newspaper corpus but ultimately applied to compare the prevalence of concepts in newspapers, parliamentary speeches, and manifestos (as we will show with an example). (3) Optimized dictionaries can be applied in contexts where only partial access to full-text data is given. For example, when researchers are confronted with keyword-search engines – a common situation in the social sciences. (4) Optimized dictionaries are characterized by increasing returns. Their developments require many resources, but application costs are low. Once optimized dictionaries are developed they become a common good. In short, optimized dictionaries identify concepts in text data with high precision and recall, they are transferable to new

sources of text data, they can detect concepts in data where only keyword access is granted, and they could boost the comparison of concept prevalence in the social sciences by preventing costly particularism. The combination of these four advantages enriches research in scenarios where access to textual resources is constrained. This is particularly pertinent in instances where portals like Nexis or specialized search engines tailored for restricted sources come into play, with their use increasingly gaining prominence in contemporary research practices.

To demonstrate our case, we assess the attention to *territorial politics*[2] across parliaments and the media in Spain (1976–2019) and the UK (1900–2019). This is an abstract social science concept whose prevalence arguably varies substantially across countries, regions, and over time. It is a concept organized around the power/authority distribution across levels of government and incorporates issues that range from technical issues such as fiscal competencies to highly salient and politicized ones such as political violence in disputes over secessionism. The variegated nature and our long-time series provide a hard test for optimized dictionaries because blind spots of experts are almost unavoidable and false positives and negatives are very likely when words have different meanings over time. For both reasons. We agree with the conventional view that in such a case naïve off-the-shelf dictionaries or a couple of seed words would produce considerable bias (Barberá et al. 2020; King, Lam, and Roberts 2017; Mahl et al. 2022; Rinke et al. 2022). As a consequence, some have abandoned dictionaries and encouraged the use of alternative techniques such as machine learning (Barberá et al. 2020). Our workflow envisages supervised machine learning as a complement to detect the blind spots of experts and allows the minimization of false positives and false negatives by revealing the discriminatory character of words with the provision of an informed keyword selection mechanism.[3]

Our workflow, which remains accessible for new practitioners of quantitative text analysis, results in the valid identification of our key concept within newspapers and parliamentary speeches. In the most extreme scenario, our workflow improves F1-scores (the balance between specificity and precision) from low levels such as 0.1 using a first naïve dictionary up to 0.9 after the application of all necessary steps. Thereby, our results outperform state-of-the-art text classification approaches (see for example Druck, Mann, and McCallum 2008; Miller, Linder, and Mebane 2020; Radford 2021), as well as singular approaches like dictionaries without supervised learning improvements or machine learning without

[2] We define territorial politics in line with the respective section of the ECPR: „Territorial politics is about the effects of the territorial structure of the state on issues such as citizens' attitudes towards multilevel government, voting behaviour and accountability, public policy, policy divergence and the distribution of resources between levels and across units" (ECPR 2022).

[3] See for a similar argument and application based on entropy-based word diagnostics Watanabe & Zhou (2022).

expert-driven fine-tuning (F1-score from 0.34 to 0.60). Similar levels of performance (F1-scores of 0.91 on average) are measured using unseen text as well as sources with only limited access via search engines. These are encouraging results for the transferability of optimized dictionaries to new sources.

In the following, we first review the benefits and pitfalls of hand coding, dictionaries, machine learning, and topic models applied in isolation. Second, we introduce optimized dictionaries, combining the benefits of dictionaries, hand coding, and supervised learning. Third, we use topic models to assess the variety of sub-issues in pre-identified text passages. Fourth, as an application case, we apply our approach to territorial politics in Spain and the UK.

2 Potential sources of bias in concept identification

The identification of concepts in text corpora demands the minimization of three fundamental sources of error as well as the minimization of errors that arise from technical difficulties. The fundamental sources of error are false negatives, false positives, and endogeneity bias. False positives arise when, for example, keywords identify the concept of interest but also text passages without any seeming relation to the concept. False negatives arise when text passages related to the concept are not identified. Endogeneity bias is a special case of false negatives. It is introduced if researchers are more informed or aware of some elements of the concepts in comparison to others. For example, historical bias can be introduced by developing dictionaries with keywords related to more recent events, ignoring past concepts, wordings, or debates, and resulting in a systematic production of false negatives.

Additionally, technical issues can introduce bias. A text might suffer from spelling or Optical Character Recognition (OCR) errors. Efficient measurements should take this into account. Furthermore, restricted access to text data provided by platforms such as Nexis or ProQuest, e.g. through keywords has implications for the utility and efficiency of different identification approaches.

We now discuss the four most frequently used methods to identify concepts in texts focusing on the main sources of error. Table 1 provides an overview of the benefits and pitfalls of each of the methods. Succeeding, we portray a workflow that generates optimized dictionaries, an approach that synthesized the strengths of the different approaches and avoids their pitfalls. In the following section, we discuss potential sources of bias in the most prominent approaches of concept identification.

3 State of the art

Researchers have invested substantial resources to identify and map the prevalence of abstract concepts in the social sciences. The vast majority of measures are based on expert surveys or derived from hand-coded texts. These procedures are, in our view, still the benchmark for (internal) validity, because humans are still the best coders of text (Hutter 2014). Hand coding can handle low-quality text corpora and minimizes false negatives and false positives (Atteveldt, Velden, and Boukes 2021). Humans can read texts with OCR or typing errors. They (often) detect irony, negations, and metaphors. It is useful for both exploratory and confirmatory analyses. In short, hand coding is the gold standard for concept identification in a text (Benoit, Laver, and Mikhaylov 2009; Grimmer and Stewart 2011; Nelson et al. 2018).

Nonetheless, hand coding is subject to at least three problems. The first and most obvious problem is the resource intensity that limits the scope of hand-coding endeavors. For example, the Comparative Agenda Project (CAP) manually codes newspaper content (Barberá et al. 2020; Baumgartner, Breunig, and Grossman 2019; Jacobi, Atteveldt, and Welbers 2016), but resource constraints mean only front pages are coded.

Second, expert surveys and hand-coded text from different projects are difficult to compare. Even the most transparent hand-coding is not easy to replicate, adapt or apply to new corpora because it cannot be changed ex-post. For example, expert surveys use different concepts of salience from scholars carrying out media or manifesto content analysis (see Chaqués-Bonafont, Palau, and Baumgartner 2015; Helbling and Tresch 2011; John et al. 2013). Third, hand-coding does not work well with keyword-restricted access to corpora, since researchers need to select texts with keywords before they can start coding. When issues of interest are rare and the corpus is large, hand coding can become extremely costly.

Researchers have developed alternative methods to overcome these shortcomings, most importantly keyword searches with dictionaries (Barberá et al. 2020; Hayes and Weinstein 1990; King, Lam, and Roberts 2017; Radford 2021), machine learning (Zhou and Goldberg 2009), and topic models (Blei and Lafferty 2006; Blei, Ng, and Jordan 2003; Roberts et al. 2013; Roberts et al. 2014), word embeddings and increasingly transformer-based models (Watanabe 2021; Kroon et al. 2022; Widmann & Which 2022). Some have compared the performance of these approaches (Nelson et al. 2018; Kroon et al. 2022; Radford 2021) and others have started to combine them for the study of sentiment and coverage of topics in text corpora (Watanabe 2021; Watanabe & Zhou 2022).

Our approach moves beyond previous studies that combine different methods by specifically addressing technical restrictions and the transferability of dictionaries designed to be common goods for the study of specific phenomena. We discuss the most commonly used approaches along the fundamental and technical sources of error because it enables a more systematic view of potential complementarities. Besides the sources of error, we evaluate the methods in their ability to be transferred to new sources of text. Transferability is a key attribute of a technique because it allows not only for easy replications but also to extend procedures to new sources with results that are ultimately comparable. Without transferability, a research community is stuck in costly particularism where the findings of one project face difficulties to be compared to the findings of another. So how do the most frequently applied methods of concept identification in text fare along the dimension of error reduction (validity) and transferability?

Dictionaries are a set of words describing a specific concept and are typically used for confirmatory analyses (Albugh, Sevenans, and Soroka 2013). Dictionaries are easy to replicate and easy to apply to new sources of text but also suffer from all three sources of error. If dictionaries are not carefully validated, they easily include ambiguous words that increase false positives, and leave out important words leading to false negatives (Barberá et al. 2020; Bozarth and Budak 2022; Dobbrick et al. 2021). The same can be said for seed words used for automated text identification (see Mahl et al. 2022; Rinke et al. 2022). Since humans are good at recognizing relevant keywords but bad at recalling all relevant keywords (King, Lam, and Roberts 2017), the keywords they think of are most likely endogenous to personal expertise, restricted knowledge of historic debates, and recent readings. On top, dictionaries perform badly if texts have typing or OCR errors. Although optimizing dictionaries is of utmost importance, we are unaware of a systematic discussion on how to exactly improve dictionaries.[4] Instead, some researchers discourage the usage of dictionaries in favor of machine learning (Barberá et al. 2020; Watanabe & Zhou 2022) or validate the performance of single keywords rather than dictionaries as a whole (King, Lam, and Roberts 2017). But single keywords will only in rare cases detect meaningful concepts, and, as we discuss next, machine learning bears its usual weaknesses, too.

Machine learning (ML) only requires hand-coding of a subset of texts and is especially useful for confirmatory analyses. Since it is based on hand-coding, it should perform comparably well in terms of false negatives, false positives, and endogeneity bias, as long as the hand-coded sample is large

[4] There are, however, recent contributions about the appropriateness of search terms (see Mahl et al. 2022 or Watanabe & Zhou 2022).

enough. In addition, ML can handle errors and ambiguous words comparatively well, since it calculates the probability that a text mentions a specific issue based on the joint distribution of words in the text. Problematic words simply do not have much predictive power.

However, if the issue of interest is rare and the corpus is large, hand coding becomes costly (Aggarwal 2018; Cieslak and Chawla 2008). In response, Druck, Mann, and McCallum (2008) propose to code features instead of texts, meaning that researchers should code whether single words predict the occurrence of the issue of interest rather than whether the issue of interest is mentioned in a specific text. Alternatively, using a preliminary keyword search to increase the balance in the training set between texts that mention the issue and those that do not strongly improves the performance of ML (Miller, Linder, and Mebane 2020; Markus et al. 2023). Yet, researchers must have full access to the corpus if they want to apply ML, and ML is not easily applied to new texts from unseen corpora.

Topic models. In contrast to the previously discussed methods, topic models are designed for exploratory research. Thus, researchers can use topic models if they are interested in the identification of unknown categories in texts. They conceptualize texts as mixtures of topics, and topics as clusters of words that often appear together (Blei and Lafferty 2006; Blei, Ng, and Jordan 2003; Roberts et al. 2013; Roberts et al. 2014). Since topic models are unsupervised exploratory methods, the logic of false positives and false negatives does not apply. In addition, endogeneity bias is non-existent, as long as the corpus is a large enough random sample of texts from the universe of texts. Highly ambiguous words and errors are not as problematic as for dictionaries, since words can be part of different word clusters or will lose their predictive power. Moreover, when choosing a large enough number of topics, topic models can also detect infrequent issues.

Due to their exploratory nature, topic models are difficult to validate (Ying, Montgomery, and Stewart 2021). Furthermore, the predictions made by topic models are not easily transferred to new texts. If researchers want to apply topic models in a meaningful way, they need at least a large random sample from the universe of texts. Therefore, we only use topic models to explore the constituent elements of a concept when text passages of a corpus are already identified. For example, if we extract all references in newspapers to the welfare state, topic models might be able to group them into meaningful categories such as pensions and unemployment.

Table 1 summarizes the benefits and pitfalls of existing approaches. None of them can on their own provide valid, comparable, and transferable measures of concept prevalence in text data. In the following section, we argue that we can exploit the different strengths of each of these methods to

arrive at optimized dictionaries that combine the strength of different methods by avoiding their pitfalls.

Table *1.* Properties of measurement approaches for concept identification in text corpora

Method:	**Hand-coding**	**Dictionaries**	**Machine learning**	**Topic models**	**Optimized dictionaries**
			Fundamental sources of bias		
False negatives	Low	High	Medium	-	Low
False positives	Low	High	Medium	-	Low
Endogeneity bias	Low	High	Medium	Low	Low
			Transferability		
Transfer-ability	Low	High	Low	Low	High
			Technical features		
Required access	Full corpus needed	Keyword access sufficient	Large random sample	Large random sample	Keyword access and a small random sample
Research interest	Explorative & Confirmatory	Confirmatory	Confirmatory	Explorative	Confirmatory
Resource intensity	High	Low	Medium	Low	High but increasing returns
OCR or typing errors	Not problematic	Very problematic	Problematic	Problematic	Problematic

4 The optimization of dictionaries

The bad properties of dictionaries can be substantially diminished if the complementary strength of other methods is exploited (see also Atteveldt et al. 2021, Nelson et al. 2018; Rice & Zorn 2021; Watanabe 2021; Watanabe & Zhou 2022 for similar arguments). We propose to compensate for the weaknesses of dictionaries with the strength of hand-coding and machine learning, with the ultimate aim of creating optimized dictionaries for the valid and efficient identification of concepts in texts

from different sources. Such dictionaries must be optimized in a way that they minimize as well as balance false positives and false negatives, avoid endogeneity, and allow transferability. Although similarities prevail with automated keyword selections, active learning, and optimized ML algorithms (Druck, Mann, and McCallum 2008; Miller, Linder, and Mebane 2020; King, Lam, and Roberts 2017; Di Natale, and Garcia 2023), optimized dictionaries are validated in their entirety, transferable to other corpora, and can be also used as straightforward Boolean search queries online.

We develop and test optimized dictionaries in five steps: 1) Create the initial dictionary and, 2) with the help of hand-coding, minimize false positives and negatives and correct endogeneity bias when identifiable. 3) Further minimize endogeneity bias more intensively as well as false negatives and false positives with the help of supervised machine learning. 4) Test and maximize transferability and finally, 5) explore the constituent elements of a concept with the help of topic models.

Along steps 2 through 4, we use F1-scores to assess the performance of dictionary optimization. The F1-score indicates the joint performance of sensitivity (true positive rate) and precision (true negative rate) (Derczynski 2016), balancing false positives and false negatives. We perceive F1-scores as optimal for the quantification of validity in concept identification:

$$F1 = 2 * \left(\frac{Precision * Sensitivity}{Precision + Sensitivity}\right) \text{with}$$

$$Precision = \frac{TP}{TP+FP} \text{ and } Sensitivity = \frac{TP}{TP+FN}$$

A value of zero indicates only false negatives (FN) and false positives (FP), whereas a value of 1 indicates only true positives (TP) and true negatives (TN). The comparison of F1-scores across the different iterations of dictionary optimization shows the progress of the procedure and allows comparisons to applications of singular methods.[5] Once a satisfactory F1-score is achieved with the usage of only one corpus (steps 2 and 3), we can turn to an assessment of the key strength of dictionaries, their transferability to other text corpora (step 4). The final step 5 allows to assess the degree to which the concept is discussed similarly across time and corpora.

Step 1: Initial dictionary. First and foremost, researchers need to carefully define and describe their concept of interest. Based on existing similar dictionaries, expert knowledge, and/or secondary

[5] Precision is typically high for off-the-shelf dictionaries. However, optimizing dictionaries means to increase sensitivity without reducing precision.

sources, researchers can compile a first set of keywords that covers, for example, important events, groups or specific terms related to the concept of interest. Although such a dictionary is more than what Grimmer and Stewart (2013, p. 274) describe as rudimentary, any bias it bears is still unknown.

Step 2: Hand-coding to reduce false positives and negatives. In step 2, researchers apply their preliminary dictionary version 1 to a corpus for the first time. Here, and in step 3, researchers need access to at least a large random sample of a corpus, if not the whole corpus of one source. However, be reminded that this first corpus primarily serves the purpose of optimization, and that optimized dictionaries can be applied to other corpora without full access (see step 4). We recommend to sample texts from the first corpus in a way that the sample is not too large but still encompasses texts, for example, from different time periods, different authors or different sources. Such a guided random sampling procedure increases the likelihood that the sample includes texts that speak about the concept in different ways. If the concept is rare, it is also advisable to randomly select one set of texts that contains one of the dictionary terms from dictionary version 1 to increase the likelihood to spot FP (Miller, Linder, and Mebane 2020; see Markus et al. 2023).

Using this sample, researchers can assess the performance of their first dictionary. Researchers should single out texts that mention the concept but are not detected by the dictionary. Based on these texts, researchers can include keywords to the dictionary that they have previously not thought about but deem to be important for identifying the concept in texts (King, Lam, and Roberts 2017). Researchers should also delete keywords from the dictionary that often flag texts that nevertheless do not mention the concept. This step mirrors a fully automatized procedure proposed by King, Lam and Roberts (2017), but differs in two important aspects. First, it is more accessible to researchers new to statistical text analysis. Second, we encourage researchers to assess F1-scores of different dictionary versions they create to assess whether the dictionary in its entirety measures the prevalence of the concept better than a previous dictionary version. In short, the second step is a manual optimization procedure that systematically looks at false positives and false negatives. Thereby, the procedure can already substantially reduce endogeneity bias because in particular false negatives can systematically cluster around issues that have not been envisaged in step 1.

Step 3: Machine-learning. While step 2 might be suitable to reduce FP and FN and already might detect endogeneity bias, a combination of dictionaries and machine-learning (ML) is well suited to put the reduction of FP, FN and endogeneity bias on a systematic footing. Researchers can use their hand-

coded sample of texts from step 2 as a training set for an appropriate ML algorithm.[6] Applied to the unseen texts from the first corpus, this algorithm will detect texts with a high probability to mention the concept. Similarly, the dictionary version from step 2 will also flag texts that most likely mention the concept in the unseen part of the first corpus. We argue that researchers can in particular learn about their endogeneity bias from the mismatch between ML and dictionary predictions, because the ML algorithm is good in detecting similar texts to the ones the researcher is already aware of while the dictionary still focuses on the sub-issues mentioned in the few texts the researcher has manually coded. Again, researchers should single out texts that are categorized differently by both methods, select missing keywords and delete under-performing keywords in order to substantially increase the F1-score of their dictionary. Additionally, we encourage researchers to look at highly predictive words from the ML model. Keywords should be kept and selected that increase TP without increasing FP (increased precision) and without reducing sensitivity by increasing FN. In short, every selected word should increase the F1-score in comparison to its counterfactual omission from the dictionary. Overall, the third step reduces potential biases in researchers' keyword selection due to their focus on only some elements of a concept and its manifestations and provides a numerical basis for keyword selection and deselection.[7] This is the final dictionary version to study transferability and potential sub-issues relevant for discussions around the concept.

Step 4: Transferability. The dictionary is now optimized and can be applied to all types of corpora, but researchers should validate whether this is appropriate. The easiest way to do this is to hand-code another random sample of a new corpus and calculate the F1-scores of the optimized dictionary applied to the new corpus. This allows researchers to assess the quality of the dictionary predictions in previously unseen corpora, enables inferences about the comparability of the measurement, and dramatically decreases issues of overfitting. If needed, researchers can again exclude words that lead to FP and include words that lead to TP, and test the final version of their dictionary again on another small set of new hand-coded articles from all sources. In principle, researchers might run into trade-offs between the internal validity of the identification in the training corpus and the external validity of the identification on a new corpus when the semantics of a concept differs across sources.

[6] We tested several and used the best performing one for each language. We discuss the potential application of word embeddings and transformer models in the final section.

[7] We see a lot of potential here to automatize key-word selection by identifying the relative contribution to F1-scores in contrast to a counterfactual omission of the word.

Step 5: Semantic differences across corpora. Finally, we can use topic models to assess the variation in how different sources frame the concept, the main sub-issues they mention and whether there are substantial differences across time and corpora. We recommend to apply topic models to all texts flagged by the optimized dictionary to reduce computational costs. This also allows to include texts from corpora with keyword restricted access. If researchers are interested in the context in which a concept is mentioned, we recommend to use the whole text. If researchers are rather interested in studying the sub-issues of a concept, we recommend to only use parts of texts, e.g. sentences, that mention any of the keywords included in the optimized dictionary. Either procedure will provide researchers with insights into the elements of the concept. If the topic models find topics related to the concept of interest in meaningful ways, this provides further evidence for the validity of the dictionary. Finally, topic models might also find issues that researchers have not been aware of but that meaningfully connect to the concept, thus revealing instances of previous endogeneity bias.

To demonstrate an example for our workflow, we assess the prevalence of territorial politics in parliaments and the media in Spain (1976–2019) and the UK (1900–2019). As we will see, in our example, semantics are different across sources, but optimized dictionaries build upon the instructions of step 1 through 4 still perform well across different types of corpora.

5 Application case: territorial politics in Spain and the UK

As we seek to identify references to territorial politics in Spain and the UK over a long period, we see it as a hard test for an optimized dictionary. The concept's prevalence varies substantially within and across both countries and elements range from low salient and technical issues such as fiscal competencies to highly conflictual and salient elements such as political violence associated with demands for secessionism. In both countries, we have full access to one newspaper (El País, 1976-2019 and The Times, 1900-2013) and keywords-based access to another newspaper via Nexis (El Mundo, 2002-2019 and The Guardian, 1985-2017). We used the newspapers with full access to optimize the dictionary. Since articles from The Times were highly affected by OCR errors, using these newspapers as optimization corpus makes our case an even harder test for optimized dictionaries. To test the transferability, we collected all article titles from newspapers and randomly selected a set of articles to download from Nexis that mention one of the keywords in their title, and another set of articles that do not mention any of the keywords. We also tested the transferability of our optimized

dictionary to Spanish and British parliamentary speeches from ParlSpeech (Rauh and Schwalbach 2020) for the periods 1996–2018 and 1989–2019, respectively. Finally, we applied topic models to assess the issues mentioned in the debates about territorial politics in the newspapers and parliaments.

Our application case demonstrates that optimized dictionaries can identify concepts such as territorial politics with very high precision and sensitivity. Even with keyword search engines and new text corpora, our optimized dictionaries identify relevant text passages with F1-scores of around 0.9. On those pre-identified text passages, topic models perform well to disaggregate territorial politics into meaningful sub-issues such as fiscal politics, secessionism, or political authority across levels. Finally, by constructing prevalence measures for those issues across newspapers and parliaments, we show noteworthy differences in political communication across the media and parliaments.

5.1 Optimized and transferable dictionaries for territorial politics

We first created initial dictionaries for territorial politics in Spain and the UK, each, based on historical and political science research, party manifestos, and homepages of non-governmental organizations (step 1). We structured the country-specific history of territorial politics into different debates for the period of interest and identified keywords for each of the periods and issues to avoid historical bias. We then drew the first round of random newspaper articles for hand-coding. We compared the hand-coded with the dictionary-based identification (991 articles from El País and 570 from the Times) and adjusted the dictionary by deleting keywords that caused many FP and adding new ones that increase TP (step 2).[8]

Two things became clear with the first round of hand-coding. First, while sensitivity from the dictionary and hand-coding was highly satisfactory (0.64 in average across both countries), the initial dictionary produced a very high number of FP. Second, in some periods territorial politics is rarely discussed, and thus only very few articles in our hand-coded sample referred to it, especially in the Times. Given our experiences with the first round of step 2, we further increased our hand-coded samples for both newspapers. Overall, this increased the hand-coded set for El País to 2,535 articles. We used this hand-coded sample to further adapt the dictionary, following again step 2. In the case of The Times, we drew two additional random samples, enlarging our hand-coded sample to 1,368 articles, and simultaneously adjusting the balance in the hand-coded sample to include more articles

[8] We also used the first random sample to check the inter-coder reliability of two out of three different coders (result for Spain: 78.3 per cent).

concerning territorial issues. Again, we adjusted the dictionary until we saw that further including or excluding keywords would not substantially increase the performance of the optimized dictionary in its entirety.

In the case of Spain, the second step increased our awareness of how newspapers discuss territorial politics. Instances and repercussions of violent expressions associated with Basque separatism have been the most prevalent element in the Spanish newspapers in particular in the 1980s, but our initial dictionary was not ready to capture the semantics of it. Thus, the random selection of text for hand-coding based on an initial dictionary can reduce endogeneity bias already. Since we worked on the UK after Spain, we learned that lesson and captured similar elements, as for example in the Northern Irish case already in the initial dictionary for the UK.

Table 2. Performance of dictionaries throughout steps 1 to 3 in the corpus for optimization

	Dictionary after step 1	Dictionary after step 2	Dictionary after step 3	Machine learning based on dictionary after step 2	N of hand-coded documents
El País	0.52	0.77	0.78	0.80	2,534
The Times	0.42	0.52	0.55	0.58	1,596

Note: Dictionaries were applied to all hand-coded articles. The performance scores of the ML algorithms are based on Leave-One-Out Cross-Validation (LOOCV) without model tuning and with defaults. In the case of El País, we used Random Forest as ML algorithm. In the case of the Times, we used SVM. SVM failed to categorize six articles. RF and SVM were the best-performing algorithms for the respective newspaper and in both cases outperformed NB estimators. We show F1-scores using all ML algorithms in section SM4 of the Supplementary Material.

Following step 3, we applied different ML classifiers[9] using the hand-coded articles as the training set to find other articles on territorial politics in the previously unseen articles from El País and The Times. The correlation between the optimized dictionary and the ML predictions was considerably higher for El País than for The Times but significant in both cases (compare section SM3 of the Supplementary Material). We drew a final random sample of two articles per year where classifications by the optimized dictionary and the ML classifiers diverged. Our analysis of this sample made us aware of another bias in our perception of territorial politics. Whereas we had incorporated key-words to capture violent expressions of territorial conflict domestically, we have not anticipated the similarity in semantics between "domestic" territorial issues and territorial questions of the UK's oversea territories.

[9] Support Vector Machine (SVM), Random Forest (RF), and Naïve Bayes (NB).

This has not been an issue in Spain, since our Spanish newspaper data start in the late 1970s but the UKs long term data availability meant to include part of its colonial history.

Table 2 presents the F1-scores for different dictionary versions and the three different ML approaches using all hand-coded articles as a training set and only focusing on our first corpora. Dictionary version 1 mirrors the first dictionary version we created (see step 1). Dictionary version 2 refers to the dictionary created after several rounds of step 2. The optimized dictionary is the one we finally settled for after step 3. The performance of ML is assessed using Leave-One-Out Cross-Validation (LOOCV) within the hand-coded sample used in steps 1 through 3. The F1-scores for our dictionaries increased from 0.520 to 0.783 in the case of El País and from 0.417 to 0.549 in the case of The Times. Overall, the performance of each final dictionary version is comparable to the performance of the most suitable ML algorithm for each language within the first corpus we used for optimization (after step 2). The considerably lower F1-scores for The Times are due to its long-time coverage and the OCR errors. However, both ML and dictionary approaches suffer from these conditions.

Increases in the performance as shown in Table 2 might be partially due to overfitting, leading to dictionaries fitted to one specific corpus only. In step 4, we therefore examine the transferability of our dictionaries to other sources of text. The variance in performance as depicted in Table 3 delivers a better baseline to compare the performance of the individual steps, individual methods and their complementarities. As a benchmark for performance we first randomly selected a small subset of the full list of titles for each newspaper for hand-coding, assuring that the sample was more balanced than a purely random sample by selecting a set of articles whose titles mentioned any of the keywords from the optimized dictionaries, and another set that did not mention any of the keywords. We similarly selected a hand-coding sample from parliamentary speeches provided by ParlSpeech (Rauh and Schwalbach 2020). In the Spanish case, we see how poorly an initial expert dictionary can perform, even though the developers have been experts in the field of territorial politics in Spain. Step 2 helped to refine the dictionary and erased FP and FN considerably. Additionally step 2 helped to close blind spots such as violent expressions of secessionism (endogeneity bias). In short, step 2 provided a substantial optimization of the dictionary in Spain. Supervised learning added some marginal improvements on top in step 3. Machine learning alone shows a considerably lower performance (F = 0.55 versus F1 = 0.90).

Table 3. Transferability: Performance of dictionaries in unseen corpora

	Dictionary version 1	Dictionary version 2	Optimized dictionary	Machine learning	N of hand-coded documents
El Mundo	0.10	0.87	0.90	0.55	100
Congreso de los Diputados	0.42	0.86	0.89	0.40	100
Guardian	0.82	0.89	0.93	0.60	100
House of Commons	0.76	0.87	0.88	0.47	100

Note: Dictionaries were applied to all hand-coded texts. The performance scores of the ML algorithms refer to models using hand-coded articles from El País and the Times as training sets without model tuning and with defaults. We use the algorithms with best performance from previous predictions. We show F1 scores using all ML algorithms in section SM4 of the Supplementary Material.

Differences in the performance between dictionary versions 1 and 2 for the Spanish case indicate that a systematic revision of dictionaries by experts using stratified randomized sampling has an important effect, whereas supervised ML algorithms add only marginally to this already high performance. Since we learned from the Spanish case before studying the UK, performance between different dictionaries does not increase to the same extent than for the Spanish dictionary versions. Moreover, the British optimized dictionary performs as well as the Spanish optimized dictionary, increasing confidence that dictionaries optimized with corpora including spelling or OCR errors can still be transferred to new corpora.

The optimized dictionaries perform exceptionally well in comparison to other text classification methods in the social science. For example, the labeling algorithm by Miller, Linder, and Mebane (2020, p.544) reaches an F1-score of 0.75 based on a comparable sample balance of 0.05 (see also Druck, Mann, and McCallum 2008; King, Roberts, and Lam 2019). Moreover, they do not only outperform more rudimentary dictionaries, but also the most suitable ML algorithm in identifying texts with territorial politics from new corpora not used during the optimization.[10] This encourages us to claim that the increase in performance seen during the proposed optimization strategy is not driven by overfitting. Therefore, we are confident to put substantial weight on our optimization strategy. In the

[10] See section SM5 of the Supplementary Material for a detailed description of the transferability of different dictionary versions and ML algorithms.

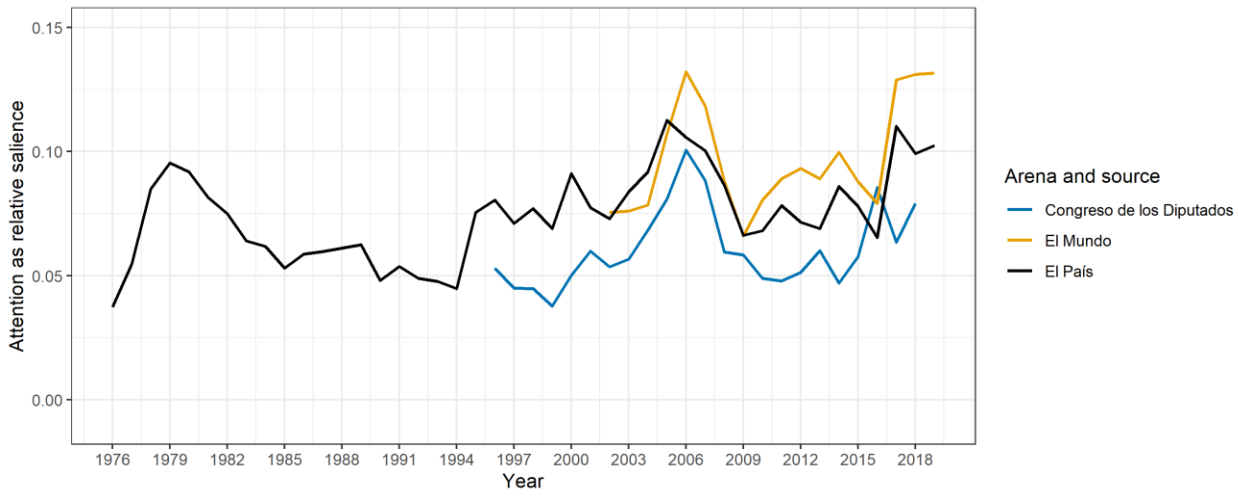
following section, we compare the prevalence and composition of territorial politics in Spanish and British parliaments and media.

5.2 Territorial politics in Spain

Figure 1 shows that territorial politics has long been a dominant issue in Spanish democratic history. From the most recent democratization process in 1976 until 2019, 7.4 percent of articles in El País and 9.7 percent of articles in El Mundo refer to territorial politics. The Congreso de los Diputados devotes less attention to territorial politics: from 1996 to 2019, 6.0 percent of parliamentary interventions and speeches alluded to it. Although the level of attention differs slightly across arenas, Figure 1 shows that their developments co-vary over time (correlation of 0.66, statistically significant at $p < 0.01$).

The strong media attention around 1980 reflects the rise of regional democracies when the constitutional architecture of Spain was being set up and fundamental questions of regional autonomies were debated. Shortly after, several decentralization laws were passed to transfer political autonomy or competencies to the Comunidades Autónomas. Such a high degree of attention was only reached again with the reform of the Catalan statute in 2006, and in 2017 with the Catalan referendum and declaration of independence by former Catalan president Carles Puigdemont. In short, parliamentary and newspaper attention to territorial politics mirrors key developments in struggles over authority in Spain.

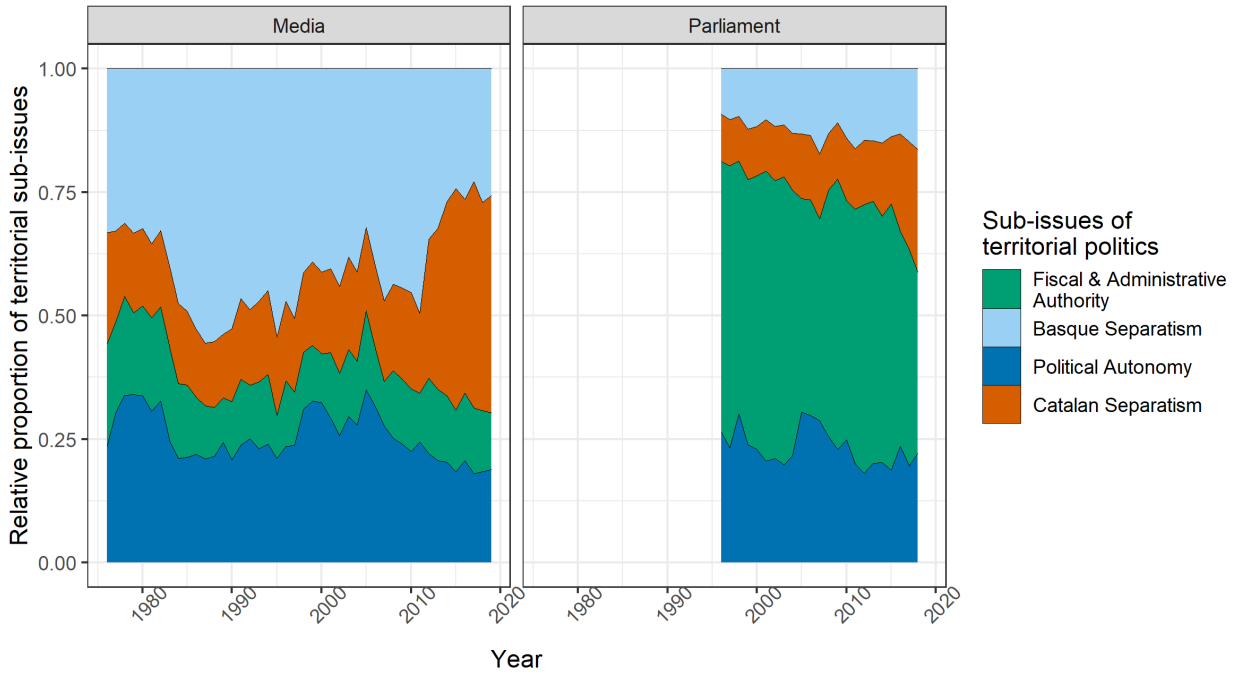
Figure 1. Attention to territorial politics across political arenas in Spain



Note: Parliamentary speeches and media articles; sources: Rauh and Schwalbach 2020; El País (1976–2019); El Mundo (2002–2019). Average saliences: El País 7.4 %; El Mundo 9.7 %; Congreso de los Diputados 6.1 %. Overall Pearson's correlation between media and parliament arena: 0.66 (statistically significant at $p < 0.01$).

The high F1-scores discussed in the previous section indicate that prevalence levels of territorial politics in the Spanish media and Parliament are comparable. However, equal values might still reflect very different discourses within the subject of territorial politics. Text passages well identified by optimized dictionaries are an ideal set-up for topic models to explore such differences.[11] Based on topic models, we infer that four main issues capture the vast majority of territorial politics references (issues are defined based on the words highlighted by the topic models).

Figure 2. Attention to territorial sub-issues in Spain



Note: Proportions reflect the yearly aggregation of sub-issues on the basis of sentences. The sum of substantially relevant topics equals 1.

Figure 2 depicts the relative salience of those issues over time and arena. While we initially believed to find references to fiscal, administrative and political autonomy in all arenas, Figure 2 shows that newspapers and parliaments differ to great extent in the way in which they discuss territorial politics. Basque separatism is a dominant theme in the media until Catalan separatism takes over. Representatives in parliament have rather discussed fiscal and administrative issues as well as shifts of political authority. In short, although the prevalence of territorial politics is similar across newspapers and Parliament, newspapers and Parliament discuss territorial politics differently: While the media

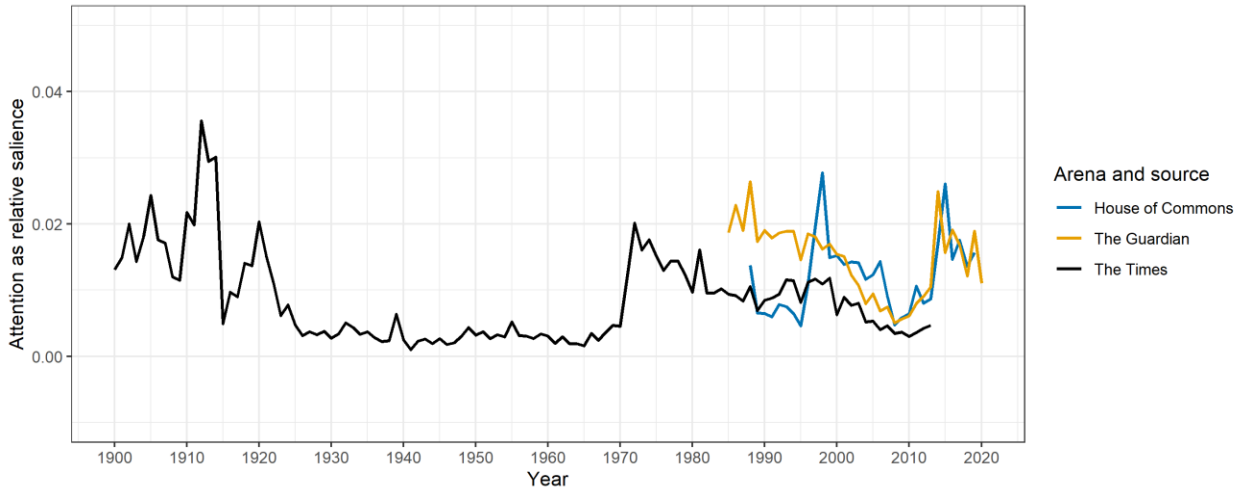
[11] We use Structural Topic Models (Roberts et al. 2013). See Sections SM6 through SM8 in the Supplementary Material for details on topic model selection, sub-issue aggregation, pre-processing and the content of the single topics. Note that sub-issue analysis with topic models is conducted only with the article-sentences identified by the optimized dictionary.

focus on issues related to violence and regional independence claims, Parliament focuses much more on technical issues.

5.3 Territorial politics in the United Kingdom

Attention to territorial politics in Britain is substantially lower compared to Spain, although prevalence levels vary significantly over the last century (see Figure 3). On average, 1.2 percent of parliamentary speeches, 0.8 percent of articles in the Guardian, and 1.5 percent of articles in the Times mention territorial politics. Again, the measurements for the newspapers and the measurements for the House of Commons co-vary (correlation of 0.45, statistically significant at $p < 0.05$).

Figure 3. Attention to territorial politics across political arenas in the UK



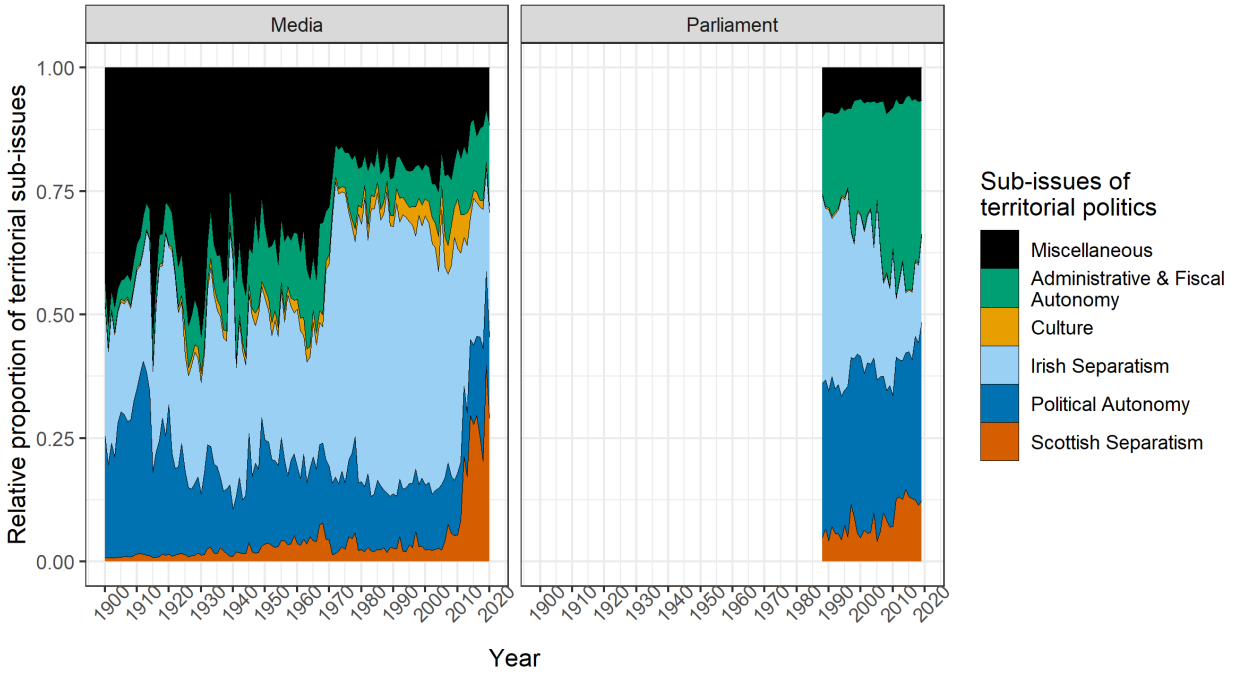
Note: Parliamentary speeches and media articles; sources: Rauh and Schwalbach 2020; the Times (1900–2013); Guardian (1985–2020). Average saliences: The Guardian 0.8 %; the Times 1.5 %; House of Commons 1.2 %. Overall Pearson's correlation between media and parliament arena: 0.45 (statistically significant at $p < 0.05$).

During the 1910s, attention to territorial politics was at its historical height, especially due to the Scottish Home Rule Bill of 1913 and the Irish War of Independence in 1916. However, the Government of Ireland Act of 1920 and the Anglo-Irish Treaty of 1921 capped a turbulent decade of state-building and were followed by nearly half a century of non-salient territorial politics. The 1970s saw a revival in territorial politics with reinvigorated struggles over authority in Scotland, Wales, and particularly in Northern Ireland (the "Troubles").[12] The Times and the Guardian began to put more emphasis on territorial politics, with around 1.3 percent of articles addressing the issue.

[12] Parts of this steep increase in attention to territorial politics is driven by a change in the newspapers design. In the 1970s, the Times started to put more emphasis on reports and articles instead of notices.

Between 1995 and 1998, attention to territorial politics again increases, due to the Good Friday Agreement ending the Northern Irish conflict and the devolution process, which culminated in the establishment of Welsh and Scottish Parliaments. Afterward, attention to territorial politics in both newspapers decreased until around 2014 when the Scottish independence movement gained more traction.

Figure 4. Attention to territorial sub-issues in the UK



Note: Proportions reflect the yearly aggregation of sub-issues on the basis of sentences. The sum of substantially relevant topics equals 1.

Like in the Spanish case, references to territorial politics can be dis-aggregated into sub-issues via topic modeling. Figure 4 shows that the media put much more emphasis on Irish and Scottish separatism whereas the parliament prioritizes fiscal and administrative debates. The category "miscellaneous" captures in large part territorial issues on oversea territories.[13] Without surprise, we see a semantic overlap between territorial issues discussed on overseas territories and those discussed within the UK. Since our definition explicitly ruled out oversea issues and the dictionaries are not developed accordingly, we cannot validly call it "territorial issues overseas" but call it "miscellaneous". Nevertheless, due to the semantic overlap, neither ML nor dictionaries were good in differentiating between territorial politics covering overseas or British "homeland" territories.

[13] See Sections SM6 through SM8 in the Supplementary Material for details on topic model selection, sub-issue aggregation, pre-processing and the content of the single topics.

Concluding, our demonstration cases of Spain and the UK showed that prevalence levels of territorial politics resonate with the history of territorial politics in both countries, providing further face validity to the high F1-scores. The findings further illustrate that the communication of territorial politics in the media and parliaments is systematically distinct. The media highlights conflict over technical issues whereas parliaments do the opposite. Thus, optimized dictionaries measure prevalence of social science concepts such as territorial politics exceptionally well, even in corpora not used for optimization and systematically different from the corpus used for optimization. Nevertheless, the topic modeling also shows that *how* these concepts are framed can still differ substantially across different corpora.

6 Discussion

Identifying social science concepts in text data has been a key task of scientists ever since. Identifying concepts in texts allows for the study of their prevalence over time and space. Identified text fragments serve as a starting point to assess the concepts' composition, framing, and positional character and allow to relate actors such as individuals and parties to concepts. The history of this task has been predominantly on the shoulders of experts as skilled humans that have made annotations manually (for example Baumgartner, Breunig, and Grossman 2019).

However, the variety of techniques has broadened with technological advances. As a consequence, many have started to compare the performance, strengths, and weaknesses of old and new approaches to text identification (Nelson et al 2021; Radford 2021; Watanabe & Zhou 2022). We follow the recent development in comparing the performance of computer-assisted methods in the identification of text (Grimmer et al. 2022; Kroon et al. 2022; Widmann & Wich 2022) but with a focus on measuring the prevalence of concepts in texts across time and corpora. We portrayed a workflow that generates ***optimized dictionaries***, an approach that arguably combines the strengths of the singular methods of dictionaries, supervised machine learning, and topic models

Our demonstration case of references to *territorial politics* across parliaments and the media in Spain (1976–2019) and the UK (1900–2019) indicates that our proposed four-step workflow to develop optimized dictionaries yields valid and comparable results. It outperforms the valid identification of text fragments as achieved by every singular method alone. F1-scores around 0.9 even for unseen text corpora not only show the validity of identified text fragments but also that optimized dictionaries can be transferred to new sources of text even with restricted access – a common endeavor in social science research.

Whereas the development of optimized dictionaries is resource intense, high validity and good properties in terms of transferability allow for increasing returns in the future because scientists can use optimized dictionaries to identify relevant text passages with very low resource input once they are developed. In short, increasing returns of optimized dictionaries make optimized dictionaries a valuable common good that can outperform costly particularism in the identification of concepts in text data and substantially promote accumulative and comparative research agendas.

The workflow rests on the complementary use of techniques such as dictionaries, machine learning, and topic models. Although we test different ML algorithms, these might already be outperformed by models such as transformer-based neural networks (Goldberg 2017; He and Choi 2020; Thai et al. 2018). Transformer models allow for token-level identification and contextualization that are often multi-word expressions within sentences. Accordingly, the optimization of dictionaries might be achieved more efficiently by transformer models instead of SVM or Random Forest in the future. Furthermore, well-identified text passages as achieved with optimized dictionaries can be used to train transformer-based neural network models to boost their precision and recall.

The most important message of our study is that different methods should not be seen as competitors but as complements for the sake of the valid identification of text. Optimized dictionaries exploit that fact and deliver easy-to-use dictionaries that can validly identify abstract concepts in text data, and be applied to unseen text even in a situation where access to text is restricted through search engines. Thus, we think of optimized dictionaries for relevant concepts in the social sciences as a common good to overcome costly particularism.

References

Aggarwal, C. 2018. Machine Learning for Text. New York: Springer.

Albugh, Q., J. Sevenans, and S. Soroka. 2013. Lexicoder Topic Dictionaries, June 2013 versions. Montreal: McGill University.

Atteveldt, W. van, M. van der Velden, and M. Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." Communication Methods and Measures 15 (2): 121–140.

Barberá, P., A. Boydstun, S. Linn, R. McMahon, and J. Nagler. 2020. "Automated Text Classification of News Articles: A Practical Guide." Political Analysis (online first), 1–24.

Baumgartner, F., C. Breunig, and E. Grossman. 2019. Comparative Policy Agendas: Theory, Tools, Data. Oxford: Oxford University Press.

Benoit, K., M. Laver, and S. Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." American Journal of Political Science 53(2): 495-513.

Bozarth, L., and Budak, C. 2022. Keyword expansion techniques for mining social movement data on social media. EPJ Data Science, 11(1), 1-24.

Blei, D., and J. Lafferty. 2006. "Correlated Topic Models." Advances in Neural Information Processing Systems 18:147–154.

Blei, D., A. Ng, and M. Jordan. 2003. "Latent Dirichlet Allocation." Journal of Machine Learning Research 3:993–1022.

Chaqués-Bonafont, L., A. Palau, and F. Baumgartner. 2015. Agenda Dynamics in Spain. London: Palgrave Macmillan.

Cieslak, D., and N. Chawla. 2008. "Learning Decision Trees for Unbalanced Data." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer.

Derczynski, L. 2016. "Complementarity, F-score, and NLP Evaluation." Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16) 16:261–266.

Di Natale, A., and Garcia, D. 2023. LEXpander: Applying colexification networks to automated lexicon expansion. Behavior Research Methods, 1-16.

Dobbrick, T., Jakob, J., Chan, C. H., & Wessler, H. 2022. Enhancing theory-informed dictionary approaches with "glass-box" machine learning: The case of integrative complexity in social media comments. Communication Methods and Measures, 16(4), 303-320.

Druck, G., G. Mann, and A. McCallum. 2008. "Learning from Labeled Features Using Generalized Expectation Criteria." In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 595–602. SIGIR '08. Singapore, Singapore: Association for Computing Machinery. https://ecpr.eu/Events/Event/SectionDetails/572

ECPR (2022). Comparative Territorial Politics. https://ecpr.eu/Events/Event/SectionDetails/448.

Goldberg, Y. (2017). "Neural network methods for natural language processing." Synthesis lectures on human language technologies, 10(1): 1-309.

Grimmer, J., and B. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." Political Analysis 21 (3): 267–297.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). Text as data: A new framework for machine learning and the social sciences. Princeton University Press.

Hayes, P. J., and S. P. Weinstein. 1990. "Construe-TIS: A System for Content-Based Indexing of a Database of News Stories. In: *IAAI*. 1990: 49-64.

He, H., & Choi, J. 2020, May. "Establishing strong baselines for the new decade: Sequence tagging, syntactic and semantic parsing with BERT." In The Thirty-Third International Flairs Conference.

Helbling, M., and A. Tresch. 2011. "Measuring Party Positions and Issue Salience from Media Coverage: Discussing and Cross-validating New Indicators." Electoral Studies 30(1): 174–183.

Hutter, S. 2014. "Protest Event Analysis and Its Offspring." In Methodological Practices in Social Movement Research, edited by D. della Porta. Oxford: Oxford University Press.

Jacobi, C., W. van Atteveldt, and K. Welbers. 2016. "Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling." Digital Journalism 4(1): 89–106.

John, P., A. Bertelli, W. Jennings, and S. Bevan. 2013. Policy Agendas in British Politics. Basingstoke: Palgrave Macmillan.
King, G., M. Roberts, and P. Lam. 2019. Systems and Methods for Keyword Determination and Document Classification from Unstructured Text; Patent / Trademark Office: United States of America US 10,275,516 B2 (U.S Patent / Trademark Office).
King, G., P. Lam, and M. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." American Journal of Political Science 61(4): 971–988.
Kroon, A. C., van der Meer, T., & Vliegenthart, R. 2022. "Beyond Counting Words: Assessing Performance of Dictionaries, Supervised Machine Learning, and Embeddings in Topic and Frame Classification." Computational Communication Research, 4(2): 528-570.
Mahl, D., von Nordheim, G., & Guenther, L. 2022. "Noise pollution: A multi-step approach to assessing the consequences of (not) validating search terms on automated content analyses." Digital Journalism, 1-23.
Markus, D. K., Mor-Lan, G., Sheafer, T., & Shenhav, S. R. (2023). Leveraging Researcher Domain Expertise to Annotate Concepts Within Imbalanced Data. Communication Methods and Measures, 1-22.
Miller, B., F. Linder, and W. R. Mebane. 2020. "Active Learning Approaches for Labeling Text: Review and Assessment of the Performance of Active Learning Approaches." Political Analysis 28(4): 532–551. https: //doi.org/10.1017/pan.2020.4.
Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. 2021. "The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods." Sociological Methods & Research, 50(1): 202-237.
Radford, B. J. 2021. "Automated Dictionary Generation for Political Event Coding." Political Science Research and Methods 9(1): 157–171.
Rauh, C., and J. Schwalbach. 2020. "The ParlSpeech V2 Data Set: Full-text Corpora of 6.3 Million Parliamentary Speeches in the Key Legislative Chambers of Nine Representative Democracies." Harvard Dataset, V2 1.
Rice, D. R., and C. Zorn. 2021. "Corpus-based Dictionaries for Sentiment Analysis of Specialized Vocabularies." Political Science Research and Methods 9(1): 20–35.
Rinke, E. M., Dobbrick, T., Löb, C., Zirn, C., & Wessler, H. 2022. "Expert-informed topic models for document set discovery." Communication Methods and Measures, 16(1): 39-58.
Roberts, M., B. M. Stewart, D. Tingley, and E. M. Airoldi. 2013. "The Structural Topic Model and Applied Social Science." Conference paper: Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.
Roberts, M., B. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand. 2014. "Structural Topic Models for Open-ended Survey Responses." American Journal of Political Science 58(4): 1064–1082.
Thai, D., Ramesh, S. H., Murty, S., Vilnis, L., & McCallum, A. 2018. "Embedded-state latent conditional random fields for sequence labeling." arXiv preprint arXiv:1809.10835.
Watanabe, K. 2021. "Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages." Communication Methods and Measures, 15(2): 81-102
Watanabe, K., & Zhou, Y. 2022. "Theory-driven analysis of large corpora: Semisupervised topic classification of the UN speeches." Social Science Computer Review, 40(2): 346-366.
Widmann, T., & Wich, M. 2022. "Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text." Political Analysis, 1-16. doi:10.1017/pan.2022.15
Ying, L., J. M. Montgomery, and B. M. Stewart. 2021. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." Political Analysis, 1–20. https://doi.org/10.1017/pan.2021.33.
Zhou, X., and A. B. Goldberg. 2009. "Introduction to Semi-Supervised Learning." Synthesis Lectures on Artificial Intelligence and Machine Learning 3(1): 1–130.

Optimized Dictionaries - A Semi-Automated Workflow of Concept Identification in Text-Data

Supplementary Material

Leonce Röth*, Lea Kaftan+ & Daniel Saldivia Gonzatti‡[14]

* LMU University of Munich, + GESIS — Leibniz-Institute for the Social Sciences,
‡ WZB Berlin Social Science Center

SM1 Data description

Spain	El País	El Mundo	Congreso de los Deputados
Source	Website archive scrape & LexisNexis	LexisNexis	ParlSpeech V2 (Rauh and Schwalbach 2020)
Coverage period	1976-2019	2002-2019	1996-2018
Documents: articles or speeches	1,972,504	735,792	262,276
Articles mentioning territorial politics	145,778 (7.4%)	70,477 (9.6%)	15,695 (5.98%)
Sentences mentioning territorial politics (total)	377,618	195,213	37,434
United Kingdom	The Times	The Guardian	House of Commons
Source	Times Digital Archive	LexisNexis	ParlSpeech V2 (Rauh and Schwalbach 2020)
Coverage period	1900-2013	1985-2020	1988-2019
Documents: articles or speeches	7,293,823	2,571,009	1,956,223
Articles mentioning territorial politics	56,982 (0.8%)	34,421 (1.34%)	23,629 (1.21%)
Sentences mentioning territorial politics (total)	102,911	79,877	73,716

[14] Corresponding author: daniel.saldivia-gonzatti@wzb.eu. Funding: this research was partly supported by the DFG grant KA 1741/10-2.

SM2 Optimized dictionaries

All dictionaries were applied ignoring lower or upper cases.

SM2.1. Spain

Dictionary 1: First expert knowledge-based territorial politics dictionary for Spain – version 1

preautonomía, preautonom*, preautonómico, preautómica, vía rápida, pactos autonómicos, autonomía regional, regionalism* espa*, LOAPA, loapa, loap*, vi* rapid*, vía rapida", independentismo, regional, plurinacional, plurilingüistic*, uniformación, nación única, pluricult*, pluricultural, derechos históricos, derech* historic*, espa* federal, federalizar espa*, plurinac*, descentra*, regionalist*, reforma regional, competencias regionales, poder regional, negociaciones regionales, negociaciones con las comunidades autónomas, reforma de las comunidades, competencias de las comunidades, reforma de competencias fiscales, reforma de competencias regionales, reforma de competencias locales.

Dictionary 2: First optimized territorial politics dictionary for Spain – version 2

Note: categorization in bold only for orientation

Recentralization keywords: loap*, nación única, lengua oficial del estado, imposición lingüistica, castellano como lengua vehicular, desafíos rupturistas, unidad de españa, gal, lengu* vehicul*.
Decentralization keywords: preautonom*, pactos autonómicos, autonomía regional, regionalism* espa*, independentismo, plurinacional, plurilingüistic*, pluricult*, pluricultural, derechos históricos, espa* federal, federalizar espa*, plurinac*, descentra*, regionalist*, reforma regional, competencias regionales, poder regional, negociaciones con las comunidades autónomas, reforma de las comunidades, competencias de las comunidades, protagonismo de las comunidades autónomas, estado de las autonomías, estado autonómico, equiparación competencial, distribución de competencias, traspaso de competencias, transferencia de competencias, organización territorial, traspasos a las comunidades autónomas, historia autonómica, modelo autonómico, solidaridad interterritorial, identidades de nuestras nacionalidades y regiones, estructura territorial, pactos locales autonómicos, marco estatuario, pacto de ajuria enea, españa de las autonomías, modelo de financiación autonomico, acuerdos autonómicos de 1992, concierto económico, pluralidad de españa, sistema autonómico, descentralización política, descentralización fiscal, gestión descentralizada, administraciones territoriales, conferencia general de cooperación autonómica, autonomía de las nacionalidades, autonomía de las regiones, lengua cooficial, lengua común, pluralidad lingüistica, pluralidad cultural, derecho a elegir el idioma, bilingüismo equilibrado, españa autonómica, cohesión territorial, marco competencial, marco autonómico, soberanía regional, desafíos territoriales, diálogo autonómico, derechos forales, estatuto de sau, declaración de barcelona, nación sin estado, espíritu de ermua, catalanismo, catalanista, estatutos de autonomía, estatuto de autonomía, financiación autonómica, autogobierno, independentistas, independentista, eta, secesionist*, reconocer a las autonomías, referéndum de autonomía, demandas autonómicas, demandas regionales, rupturista*, exigencias autonóm*, reconocimiento autonómico, solidaridad autonómica, diversidad ling*, fomento autonóm*, presupuest* de las autonom*, etarra, terra lliure, reivindicación territorial, reivindicación autonómica, soberanía autonómica, nacionalidades, devolución, financiar a las autonomías, autodeterminacionist*, normalización ling*,

estatuari*, abertzal*, lizarr*, andalucist*, territorialidad, antiespañol*, autonomismo, autonomist*, aragonesista, antiautonomista, proetarra, federalización, vasquist*, catalanidad, cosoberanía, transferencia a las autonomías, soberanista, soberanism*, descentralizador, alta inspección.

Optimized dictionary: Last optimized territorial politics dictionary for Spain – version 3

Note: categorization in bold only for orientation

Decentralization keywords: secesionist*, recono cer a las autonomías, referéndum de autonomía, demanda* autonómica*, demanda* regional*, exigencias autonóm*, reconocimiento autonómico, solidaridad autonómica, diversidad ling*, fomento* autonóm*, presupuest* de las autonom*, terra lliure, reivindicaciones territoriales, reivindicación territorial, reivindicación* autonómica*, soberanía autonómica, financiar a las autonomías, autodeterminacionist*, autogobierno, estatutari*, abertzal*, nlizarr*, territorialidad, antiespañol*, autonomismo, autonomist*, aragonesista*, proetarra*, federalización, vasquist*, catalanidad, cosoberanía, transferencia* a las autonomías, soberanista*, soberanism*, preautonóm*, preautonom*, pacto* autonómico*, autonomía regional, regionalism* espa*, independentismo*, plurilingüistic*, pluricult*, pluricultural*, derecho histórico, derechos históricos, espa* federal, federalizar espa*, plurinac*, descentra*, nregionalist*, reforma* regional*, competencia* regional*, poder* regional*, negociaciones con las comunidades autónomas, reforma* de las comunidades, competencia* de la* comunidad*, protagonismo de las comunidades autónomas, estado de las autonomías, estado autonómico, equiparación competencial, distribución de competencias, traspaso de competencias, transferencia de competencias, organización territorial, traspasos a las comunidades autónomas, historia autonómica, nmodelo* autonómico*, solidaridad interterritorial, identidades de nuestras nacionalidades y regiones, estructura territorial, pacto* local* autonómic*, marco* estatutario*, pacto de ajuria, españa de las autonomías, modelo de financiación autonómico, acuerdos autonómicos de 1992, concierto económico, pluralidad de españa, sistema* autonómico*, gestión descentralizada, administraciones territoriales, conferencia general de cooperación autonómica, autonomía de las nacionalidades, nautonomía de las regiones, lengua cooficial, pluralidad lingüistica, pluralidad cultural, derecho a elegir el idioma, bilingüismo equilibrado, españa autonómica, cohesión territorial, marco competencial, cohesión autonómica, marco* autonómico*, soberanía* regional*, desafío* territorial*, diálogo* autonómico*, derecho* foral*, estatuto de sau, declaración de barcelona, nación sin estado, espíritu de ermua, catalanismo*, catalanista*, estatutos de autonomía, estatuto de autonomía, financiación autonómica, independentistas, independentista, plurilingüismo, estatut* autonómic*, estatutos autonómicos, reforma* estatutaria*, reforma* de los estatutos, reforma* del estatuto, reformar el estatuto, estatut d'autonomia, plan ibarretxe, nacionalismo catalán, nacionalismo vasco, competencias territoriales, procesos autonómicos, acceso a la autonomía, estatuto de cataluña, estatuto catalán, estatuto del país vasco, estatuto vasco, nestatuto valenciano, estatuto de valencia, estatuto de galicia, estatuto gallego, estatuto de andalucía, estatuto andaluz, estatuto de madrid, estatuto madrileño, estatuto de murcia, estatuto murciano, reintegración y amejoramiento del régimen foral de navarra, estatuto de extremadura, estatuto extremeño, estatuto de la rioja, estatuto riojano, estatuto asturiano, estatuto de asturias, estatuto de aragón, estatuto aragonés, estatuto canario, estatuto de islas canarias, nestatuto de las islas canarias, estatuto de canarias, estauto de cantabria, estatuto cántabro, estauto de castill*, estatuto de castilla y león, estatuto balear, estatuto de las islas baleares, estatuto de islas baleares, estatuto de baleares, estatuto de ceuta, estatuto de melilla, competencia* autonómica*, amejoramiento del fuero, conferencia de presidentes de las comunidades autónomas, autonomía catalana, autonomía vasca, autonomía andaluza, autonomía aragonesa, nautonomía riojana, autonomía valenciana, autonomía gallega, autonomía murciana, autonomía balear, autonomía madrileña, autonomía extremeña, autonomía asturiana,

autonomía canaria, autonomía cántabra, autonomía castellana, procés, independencia de catalunya, independencia de cataluña, independencia catalana, antiespañolis*, estatuto de catalunya, estatuto de euskadi.

(Re-)Centralization keywords: nación única, lengua oficial del estado, imposición lingüistica, castellano como lengua vehicular, alta inspección, lengua común, unidad de españa, lengu* vehicul*, nor malización ling*, antiautonomista, loapa, desafíos rupturistas, anticatalanist*, antiregionalist*, antiindependentis*, españolis*, andalucist*.

Territorial terrorism keywords: eta, gal, etarra.

SM2.2. United Kingdom

Dictionary 1: First expert knowledge-based territorial politics dictionary for the United Kingdom – version 1

confederalism, devolution, federalism, secession, self-government, separation, unionism, unionist, anti-treaty, anti-Treatyite, IRA, Orangeman, Orangemen, pro-treaty, Treatyite, Unionists, Anglo-Irish Treaty, Bloody Friday, devolved powers, Easter Rising, federal system, home rule, independent Ireland, Ireland's independence, Irish Bill, Irish independence, Irish nationalism, Irish problem, Irish question, Irish troubles, Irish Troubles, Kilbrandon Report, powers delegated, republican Ireland, Scotland Act, Scotland Bill, Scottish Assembly, Scottish independence, tax-varying powers, Ulster Covenant, Welsh disestablishment, Welsh independence, Crowther Commission, Kilbrandon Commission, Orange order, Scottish Parliament, separate aspirations, Smith Commission, Speaker's Conference, United Ireland, Independence referendum, Ireland act, British union, Welsh devolution, Cymru Fydd, Welsh Board, Welsh Acts, Welsh affairs, Welsh Office, Lord Crowther, Lord Kilbrandon, Welsh Assembly, Richard Commission, Welsh Government, Silk Commission, Adminsitrative devolution, Devolution Referendum, Scottish Parliament, Reserved Powers Model, Scottish Board, independence of Scotland, independence of Wales, Irish Citizen Army, Irish Republican Army, Irish Republican Brotherhood, Scottish Covenant Association, Stone of Destiny, Council of Wales, Welsh Assembly Government, Wales Act, disestablishment in Wales, Independence of Ireland, Irish Parliamentary Party, Joint Ministerial Committees, Regional Assemblies Bill, Welsh Assembly, Yorkshire parliament, Northern assembly campaing, Royal Commission on the Constitution, Campaign for a Northern Assembly, Regional Assemblies Preparation Bill, Regional economic planning board, Regional economic planning council, Your Region, Your Choice, Secretary of State for Wales, Secretary of State for Scotland, Council for Wales and Monmouthshire, A Voice for Wales, Government of Wales Act 1998, Government of Wales Act 2006, Powers for a Purpose in 2015, Tax Collection and Management Act.

Dictionary 2: First optimized territorial politics dictionary for the United Kingdom – version 2

H Block, H-Block, IRA , UAC , UDA , UFF , UVF , a Parliament in Northern Ireland, Air Passenger Duty, All Wales Convention, All-Wales Convention, Anglo-Irish Agreement, Anglo-Irish Treaty, Anti-treaty, Anti-treatyite, assemblies for Scotland, assemblies for Wales, assembly for England, assembly for Scotland, assembly for Wales, Belfast Agreement, Blanket protest, Bloody Friday, Bloody Sunday, Border poll , Campaign for a Northern Assembly, Catholic areas, Commission on the Powers and Electoral Arrangements of the National Assembly for Wales, Confederalism, Constitutional convention, Crowther Commission, Cymru Fydd,

Devolution, Devolved powers, Easter Rising, Fair Employment Act, Fair Employment Agency of Northern Ireland, Fair Employment Northern Ireland Act, Free Derry, Fresh Start Agreement, Good Friday Agreement, Hillsborough Agreement, Hillsborough Castle Agreement, Home rule, I. R. A., I.R.A., Independence of Ireland, Independence of Scotland, Independence of Wales, Independence referendum, Independent Ireland, Independent Scotland, Independent Wales, International Body on Arms Decommissioning, Ireland Act, Ireland's independence, Irish bill, Irish Citizen Army, Irish Free State Act, Irish Free State Consequential Provisions Act, Irish independence, Irish National Liberation Army, Irish nationalism, Irish nationalist, Irish Parliamentary Party, Irish problem, Irish question, Irish Republican Army, Irish Republican Brotherhood, Irish troubles, Irish unity, Irish Volunteers, Irish War of Independence, Joint Ministerial Committees, Kilbrandon Report, Long Kesh, Maze Prison, Mitchell principles, Mitchell report, Northern Assembly Campaign, Northern Ireland Act, Northern Ireland Constitution, Northern Ireland's Fair Employment Agency, Orange order, Orangeman, Orangemen, Parliament for Northern Ireland, peace in Northern Ireland, peace in Ulster, Powers for a Purpose, Pro-treaty, Republican Ireland, Reserved Powers Model, Richard Commission, Rome rule, Royal Commission on the Constitution, Scotland Act, Scotland Bill, Scotland's independence, Scottish Covenant Association, Scottish independence, Scottish local government, Scottish nationalism, Scottish referendum, Separate aspirations, Silk Commission, Smith Commission, Special category status, St Andrews Agreement, Status of Northern Ireland, Status of Scotland, Status of Wales, Stone of Destiny, Stormont House Agreement, Sunningdale Agreement, Suspensory Act, Tax Collection and Management Act, Tax Collection and Management Wales Act, Tax-varying powers, Treatyite, U. A. C., U. D.A., U. F. F., U. V. F., U.A.C., U.D.A., U.F.F., U.V.F., Ulster Army Council, Ulster Covenant, Ulster crisis, Ulster Defence Association, Ulster Freedom Fighters, Ulster unionism, Ulster Volunteer Force, Ulster Workers' Council strike, Ulster's Solemn League and Covenant, Unified Ireland, UWF strike, Voice for Wales, Wales act, Wales bill, Wales referendum, Wales' independence, Welsh devolution, Welsh Government, Welsh independence, Welsh nationalism, Welsh nationalist, Welsh referendum, Welsh taxes, West Lothian question, Yorkshire parliament, Your region, your choice.

Optimized dictionary: Last optimized territorial politics dictionary for the United Kingdom – version 3

Note: categorization in bold only for orientation

(Northern) England and Cornwall keywords: Campaign for a Northern Assembly, Cornish assembly, assembly for England, Northern Assembly Campaign, Yorkshire parliament, Your region, your choice.
Northern Ireland (and Ireland) keywords: Irish born loyalists, Irish Free State Act, Irish Free State Agreement, Irish Free State Consequential Provisions Act, Irish independence, Irish War of Independence, Loyalists in Southern Ireland, Parliament for Southern Ireland, H Block, H-Block, IRA , UAC , UDA , UFF , UVF , a Parliament in Northern Ireland, Anglo-Irish Agreement, Anglo-Irish Treaty, Belfast Agreement, blanket protest, Bloody Friday, Bloody Sunday, border poll , Catholic areas, constitutional convention, direct rule, Easter Rising, Fair Employment Act, Fair Employment Agency of Northern Ireland, Fair Employment Northern Ireland Act, Free Derry, Fresh Start Agreement, Good Friday Agreement, Hillsborough Agreement, Hillsborough Castle Agreement, I. R. A., I.R.A., Independence of Ireland, Independent Ireland, International Body on Arms Decommissioning, Ireland Act, Ireland Bill, Ireland's independence, Irish bill, Irish Citizen Army, Irish National Liberation Army, Irish Parliamentary Party, Irish Republican Army, Irish Republican Brotherhood, Irish troubles, Irish Volunteers, Kilbrandon Report, Long Kesh, Maze Prison, Mitchell principles, Mitchell report, Northern Ireland Act, Northern Ireland Constitution, Northern Ireland's Fair Employment Agency, Orange Order, orangemen, Parliament for Northern Ireland, peace in Northern Ireland, peace in Ulster,

pro-treaty, republican Ireland, Rome rule, Special category status, St Andrews Agreement, Status of Northern Ireland, Stormont House Agreement, Sunningdale Agreement, Suspensory Act, Treatyite, U. A. C., U. D. A., U. F. F., U. V. F., U.A.C., U.D.A., U.F.F., U.V.F., Ulster Army Council, Ulster Covenant, Ulster crisis, Ulster Defence Association, Ulster Freedom Fighters, Ulster Volunteer Force, Ulster Workers' Council strike, Ulster's Solemn League and Covenant, Unified Ireland, UWF strike, anti-treaty, anti-treatyite, Irish loyalist, Irish nationalism, Irish nationalist, Irish problem, Irish question, Irish unity, Loyalists in Ireland, Parliament for Ireland, peace in Ireland, Ulster loyalist, Ulster unionism.

Scotland keywords: Air Passenger Duty, assembly for Scotland, Crowther Commission, dispute resolution process, Independence of Scotland, Independent Scotland, powers to Scotland, Scotland Act, Scotland Bill, Scotland's independence, Scottish Covenant Association, Scottish independence, Scottish nationalism, Scottish referendum, Status of Scotland, Stone of Destiny, tax-varying powers, Wales Bill, West Lothian question.

Scotland, Wales and England keywords: assemblies for England, assemblies for Scotland, assemblies for Wales.

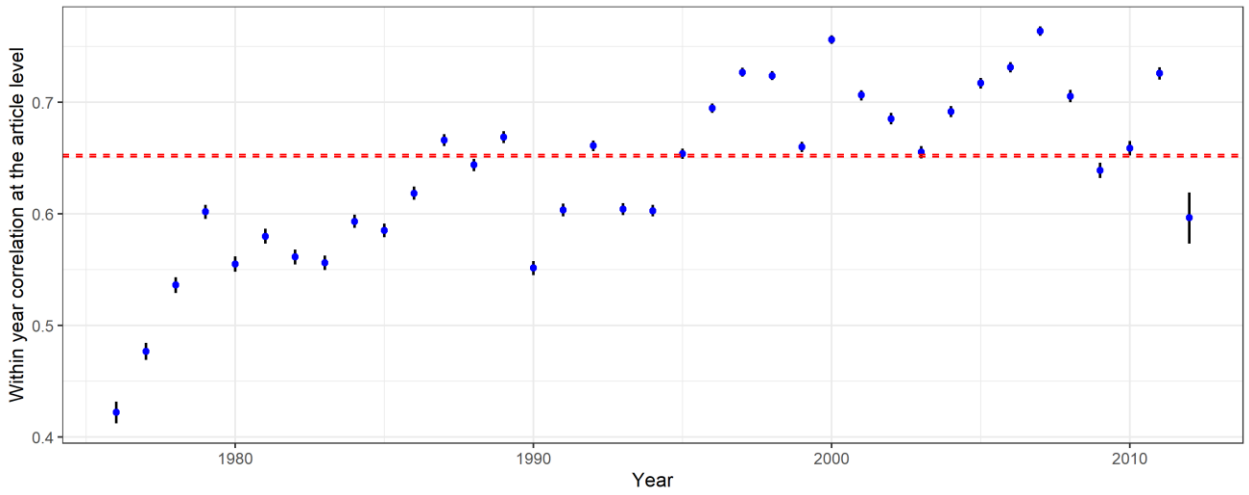
United Kingdom keywords: act of the Union, Barnett floor, Barnett formula, devolve powers, devolved government, devolved institutions, devolved powers, federal Britain, federal constitution, federal UK, federal United Kingdom, fiscal powers, home rule, Independence referendum, Parliaments for Ireland, Parliaments for Scotland, Parliaments for Wales, Reserved Powers Model, Royal Commission on the Constitution, Separate aspirations, Smith Commission.

Wales keywords: All Wales Convention, All-Wales Convention, Commission on the Powers and Electoral Arrangements of the National Assembly for Wales, Cymru Fydd, devolution in Wales, Holtham Commission, Independence of Wales, Independent Wales, Powers for a Purpose, powers to Wales, Richard Commission, Silk Commission, Status of Wales, Tax Collection and Management Act, Tax Collection and Management Wales Act, Voice for Wales, Wales Act, Wales' devolution, Wales referendum, Wales' independence, Welsh devolution, Welsh independence, Welsh nationalism, Welsh nationalist, Welsh referendum, Welsh taxes.

SM3 Territorial issue attention using machine learning and optimized dictionaries

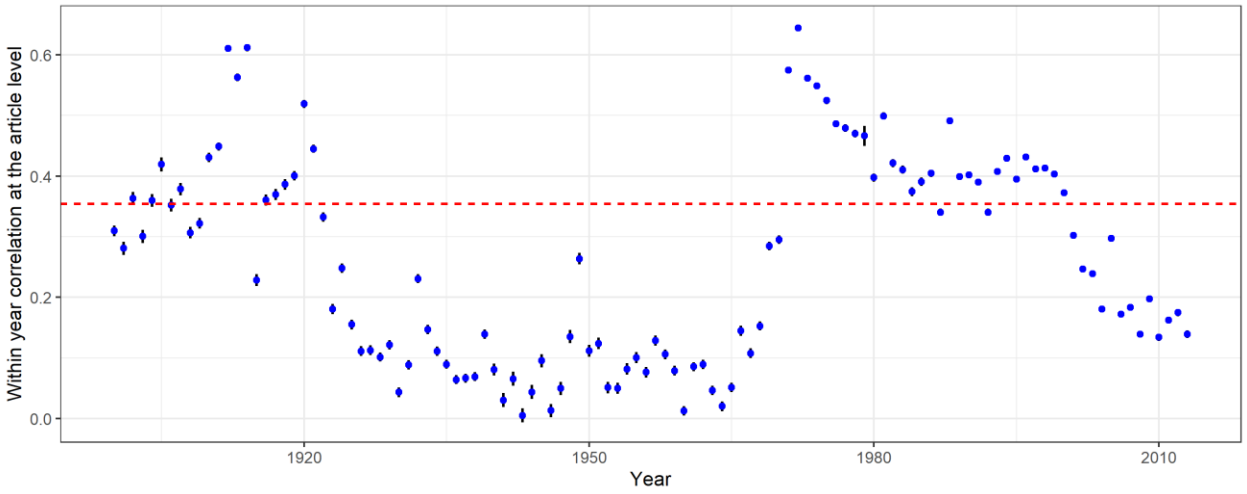
We test how much the measurement of territorial issues based on our optimized dictionaries correlates with a measurement based on the best-performing machine-learning algorithm at the level of articles. Thus, we compare how much machine-learning algorithms and optimized dictionaries coincide in their decision whether a single article contains territorial issues. Three aspects are worth mentioning: 1) When territorial politics is more salient, both methods converge more strongly; 2) Both methods correlate less in earlier periods both in the UK (The Times) and in Spain (El País). This might be related to historical biases of both measures; 3) Correlations are overall lower in the UK (The Times). We explain this with the high prevalence of OCR errors in The Times. Salience measures based on daily data correlate substantially more than measures based on the article level.

Figure SM.3.5. Correlation of territorial issue attention within articles in Spain, El País 1976-2012



Note: Pearson's correlations across years, unit of analysis: article (El País); N = 1,555,553, training set N = 2,535. Average correlation: 0.65 (0.95-CI: 0.65; 0.65); Horizontal red lines show 0.95-confidence intervals for overall correlation of both methods across time; Vertical black lines show 0.95-confidence intervals for correlations within each year. Machine learning algorithm used: Random Forest. Average salience across whole period: 6.1% (ML), 7.1% (optimized dictionaries).

Figure SM.3.6. Correlation of territorial issue attention within articles in in the UK case, The Times 1900-2013



Note: Pearson's correlations across years, unit of analysis: article (The Times); N = 7,292,227, training set N = 1,368. Average correlation: 0.35 (0.95-CI: 0.35; 0.35); Horizontal red lines show 0.95-confidence intervals for overall correlation of both methods across time; Vertical black lines show 0.95-confidence intervals for correlations within each year. Machine learning algorithm used: Support Vector Machine. Average salience across whole period: 1.1% (ML), 0.8% (optimized dictionaries).

SM4 Dictionary performance compared to the performance of machine learning algorithms

Note: All performance tests are conducted with the whole period under investigation. However, if we only test performance on exclusively overlapping periods (for Spain 2002-2018 and for UK 1985-2013), both dictionary and machine learning prediction performances increase slightly.

Table SM4.5. Territorial politics prediction confusion matrix and performance parameters of different dictionary versions and different ML algorithms in the Spanish case, El País (N = 2,535)

Prediction method	TN	FN	FP	TP	Sensitivity	Precision	Accuracy	F1
Dictionary 1	1991	236	116	191	0.45	0.62	0.86	0.52
Dictionary 2	1964	70	143	357	0.84	0.71	0.92	0.77
Optimized dictionary	1962	59	145	368	0.86	0.72	0.92	0.78
SVM	2018	125	89	302	0.71	0.77	0.92	0.74
Random forest	2052	105	55	322	0.75	0.85	0.94	0.80
Naïve Bayes	1832	54	275	373	0.87	0.58	0.87	0.69

Table SM4.6: Territorial politics prediction confusion matrix and performance parameters of different dictionary versions and different ML algorithms in the UK case, The Times (N = 1,368)

Prediction method	TN	FN	FP	TP	Sensitivity	Precision	Accuracy	F1
Dictionary 1	883	39	486	188	0.83	0.28	0.67	0.42
Dictionary 2	1203	88	166	139	0.61	0.46	0.84	0.52
Optimized dictionary	1242	93	127	134	0.59	0.51	0.86	0.55
SVM	1302	106	61	121	0.53	0.66	0.89	0.59
Random forest	1356	164	13	63	0.28	0.83	0.89	0.42
Naïve Bayes	557	18	812	209	0.92	0.20	0.48	0.33

SM5 Transferability of optimized dictionary to new sources and political arenas

SM5.1. Spain

Table SM5.7. Transferability of an optimized dictionary of territorial politics to a new media source, Spain (El Mundo)

Prediction method	TN	FN	FP	TP	Sensitivity	Precision	Accuracy	F1
Dictionary 1	45	52	0	3	0.06	1.00	0.48	0.10
Dictionary 2	43	11	2	44	0.80	0.96	0.87	0.87
Optimized dictionary	43	8	2	47	0.86	0.96	0.90	0.90
Random forest	28	28	17	27	0.49	0.61	0.55	0.55

Note: Prediction confusion matrix and performance parameters of different dictionary versions and different ML algorithms for transferring a territorial politics optimized dictionary to a new newspaper outlet (El Mundo, N = 100) based on El País (N = 2,535).

Table SM5.8. Transferability of an optimized dictionary of territorial politics to a different arena, Spain (Congreso de los Diputados)

Prediction method	TN	FN	FP	TP	Sensitivity	Precision	Accuracy	F1
Dictionary 1	51	36	0	13	0.27	1.00	0.64	0.42
Dictionary 2	47	9	4	40	0.82	0.91	0.87	0.86
Optimized dictionary	45	5	6	44	0.90	0.88	0.89	0.89
Random forest	20	29	31	20	0.41	0.39	0.40	0.40

Note: Prediction confusion matrix and performance parameters of different dictionary version and different ML algorithms for transferring a territorial politics optimized dictionary to a new political arena (Spanish parliament Congreso de los Diputados, N = 100) based on El País (N = 2,535).

SM5.1. United Kingdom

Table SM5.9. Transferability of an optimized dictionary of territorial politics to a different media source, UK (the Guardian)

Prediction method	TN	FN	FP	TP	Sensitivity	Precision	Accuracy	F1
Dictionary 1	44	8	9	39	0.83	0.81	0.83	0.82
Dictionary 2	48	5	5	42	0.89	0.89	0.90	0.89
Optimized dictionary	48	2	5	45	0.96	0.90	0.93	0.93
SVM	53	27	0	20	0.43	1.00	0.73	0.60

Note: Prediction confusion matrix and performance parameters of different dictionary versions and SVM for transferring a territorial politics optimized dictionary to a new newspaper outlet (the Guardian, N = 100) based on The Times (N = 1,368).

Table SM5.10. Transferability of an optimized dictionary of territorial politics to a different arena, UK (House of Commons)

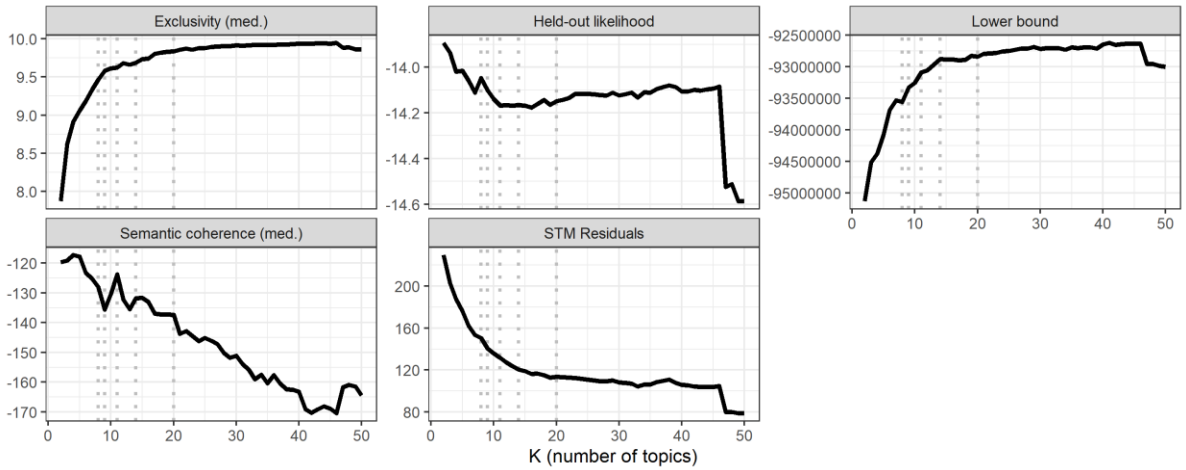
Prediction method	TN	FN	FP	TP	Sensitivity	Precision	Accuracy	F1
Dictionary 1	49	12	8	31	0.72	0.79	0.80	0.76
Dictionary 2	49	4	8	39	0.91	0.83	0.88	0.87
Optimized dictionary	48	2	9	41	0.95	0.82	0.89	0.88
SVM	54	29	3	14	0.33	0.82	0.68	0.47

Note: Prediction confusion matrix and performance parameters of different dictionary versions and SVM for transferring a territorial politics optimized dictionary to a new political arena (British parliament House of Commons, N = 100) based on The Times (N = 1,368).

SM6 STM selection

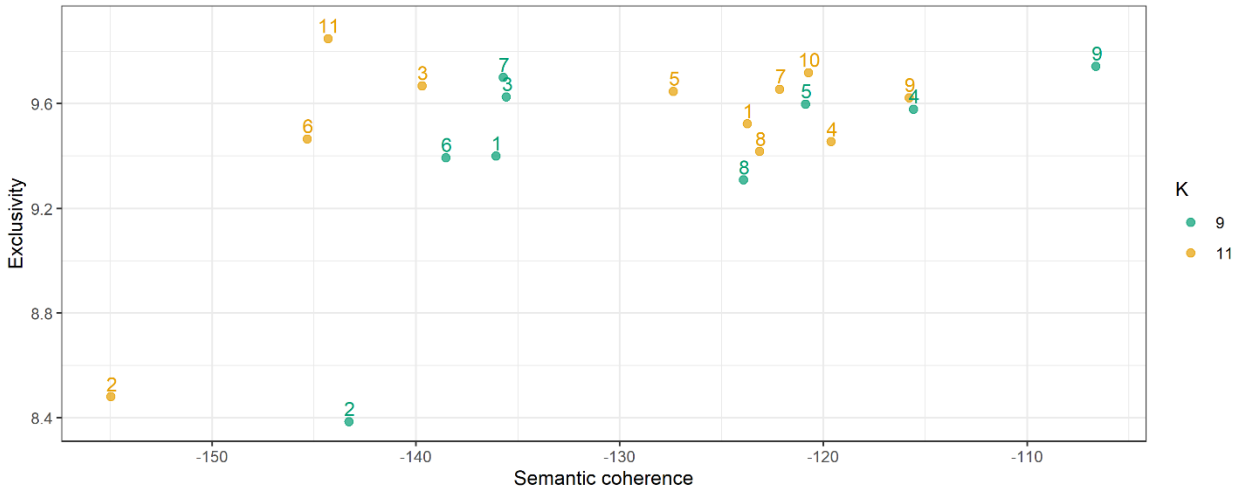
SM6.1. Spain

Figure SM6.7. STM optimization parameters for territorial sentences in newspapers and parliament, Spain



Note: K range from 2 to 50; specification: spectral initialization without covariates. Grey vertical lines mark visual local optima.

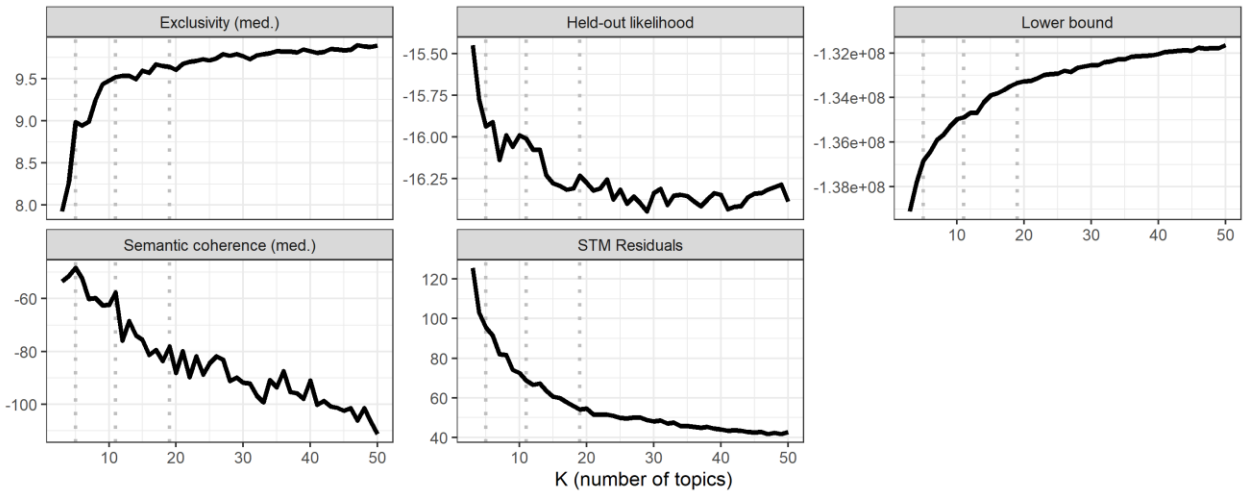
Figure SM6.8. Comparison of exclusivity and semantic coherence of STMs with K = 9 and 11, Spain



Note: Model comparison of topics with Ks 9 and 11; Model selected according to (1) local optima across the range of topics, and (2) discussions in the research group. We based our decision on the distribution of words within topics, exemplarity texts for each topic and topic correlations. We selected the model with K = 11.

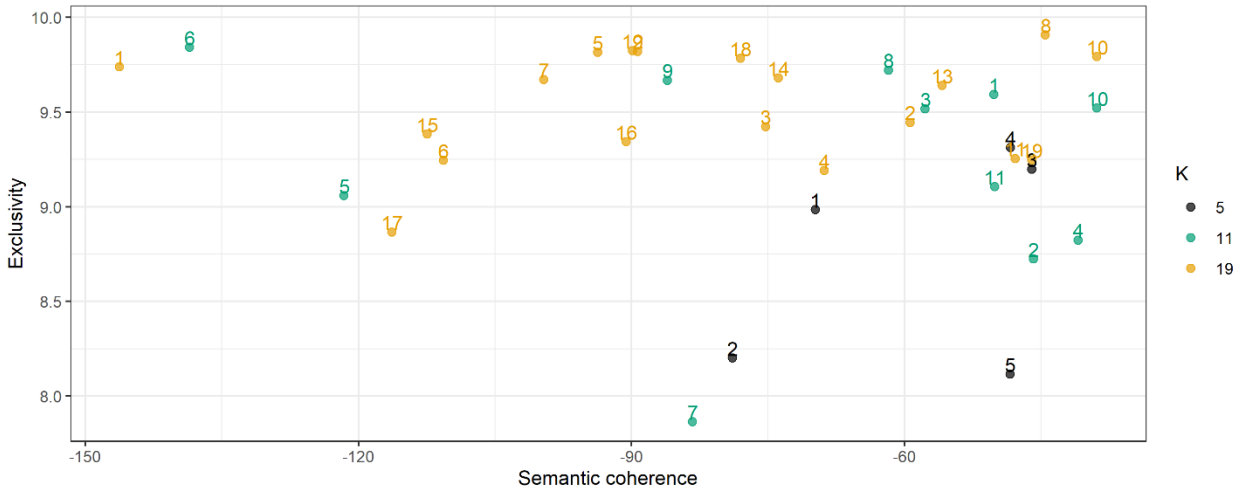
SM6.2. United Kingdom

Figure SM6.9. STM optimization parameters for territorial sentences in newspapers and parliament, UK



Note: K range from 2 to 50; specification: spectral initialization without covariates. Grey vertical lines mark visual local optima. In the case of The Times, we used paragraphs instead of sentences due to issues with sentence recognition because of OCR errors.

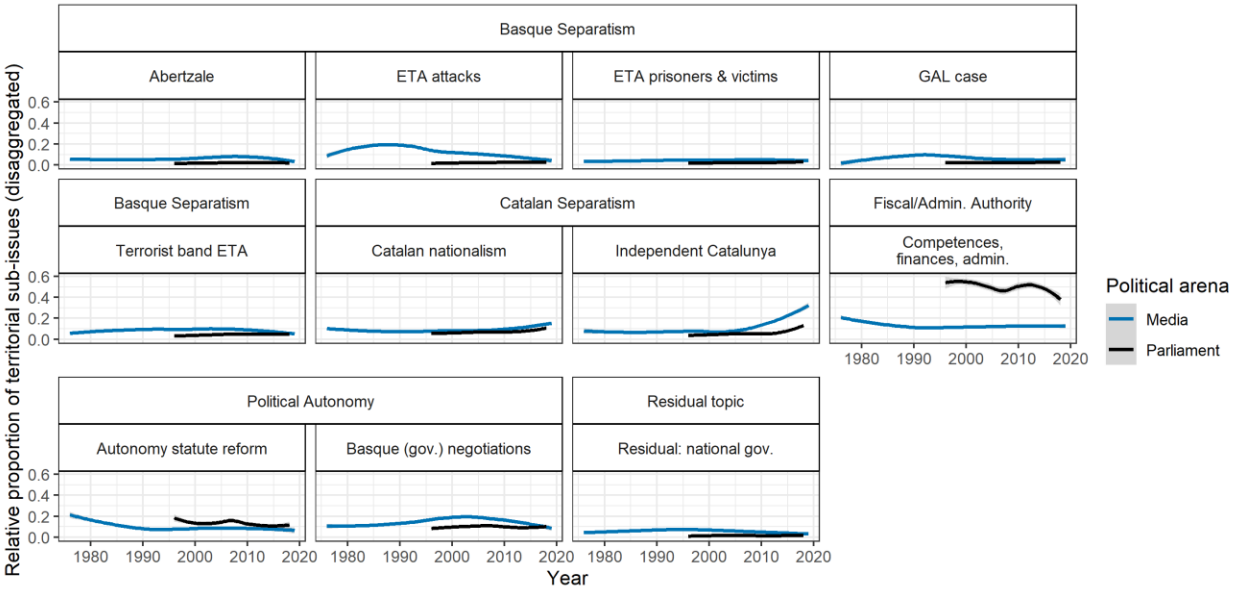
Figure SM6.10. Comparison of exclusivity and semantic coherence of STMs with K = 5, 11, and 19, UK



Note: Model comparison of topics with Ks 5, 11 and 19; Model selected according (1) local optima across the range of topics, and (2) discussions in the research group. We based our decision on the distribution of words within topics, exemplarity texts for each topic and topic correlations. We selected the model with K = 19.

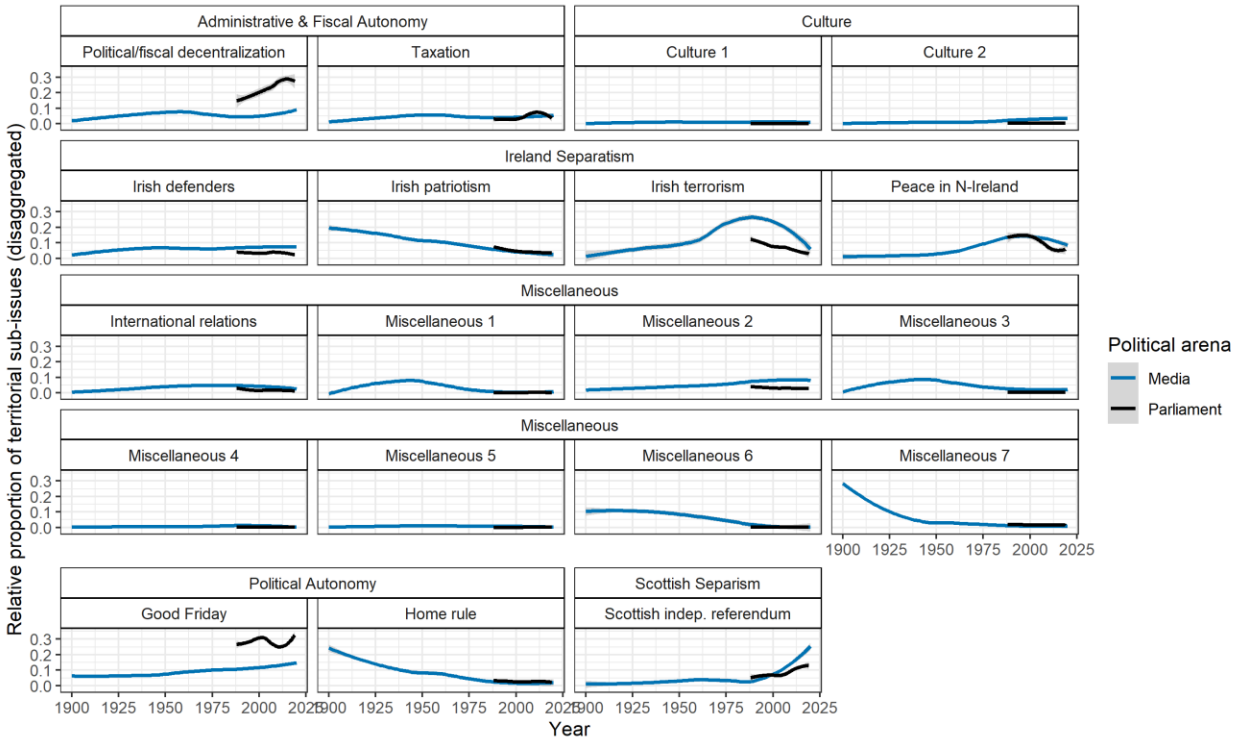
SM7 Territorial topics without aggregation across time and arenas

Figure SM7.11. Territorial topics without aggregation, Spain 1976-2019



Note: Mean prevalence of each topic in each year, based on an STM with K = 11. Aggregated issues in overarching labels.

Figure SM7.12. Territorial topics without aggregation, UK 1900-2019



Note: Mean prevalence of each topic in each year, based on an STM with K = 19. Aggregated issues in overarching labels.

SM8 Representative texts of the topics without aggregation

Note: Representative documents for each topic in the STM selected for Spain and the UK. For space reasons, we chose only the first 300 characters of each document.

Representative sentences for 11 territorial topics, Spain

[1: Terrorist attacks band ETA] "En ella dice que oyo una voz que le llamaba pronunciando dos veces su apellido, lo que le sobresalto porque, por haber recibido varias amenazas de muerte, creyó que se trataba de algun activista de ETA que iba a atentar contra su vida, pero, que al ver que quien le llamaba era un joven muy pulcro y [...]"

[2: Catalan separatism] "Entenent que la ficcio te aixopluc fins i tot quan l'autor parteix de la pròpia ida, ja que en realitat, indica Amat, ël que fan molts autors es partir d'una veritat personal, emocional i sentida i embolicar-la amb mentides, es a dir amb ficcio, Primera Persona proposa la presència¨d'escriptors [...]"

[3: GAL case] "El caso del secuestro de Larretxea fue desgajado del sumario prIncipal de los GAL y se encuentra pendiente de instruccion en el Juzgado Central numero 1 de la Audiencia Nacional hastaque el fiscal interponga la correspondiente querella, El asunto se encuentra pendiente de resolución del fiscal, qui [...]"

[4: Abertzale] "APOYO LAS LISTAS QUE EL PP NO PUDO PARAR Electoral Mendi (AEN) Alegikoalde Azkertiar Abertzalea (Atea-Alegia) Anueko Indarra (AI) Bagoaz (Zestoa Belauntzako Sustraiak (Belauntza Berriozar Baietz (BB) Branka (Hondarribia Erreil Bizirik (Errezil Herriarengatik Irun HerriaIzustarri Maeztuko Auker [...]"

[5: Residual national governmetn] "ÁBALOS MECO, Jose Luis ACEDO PENCO, Pedro AGIRRETXEA URRESTI, Joseba Andoni AGUIAR RODRÍGUEZ, Ernesto AGUIRRE RODRÍGUEZ, Ramon ALBA GOVELI, Nayua Miriam ALBA MULLOR, Maria Dolores ALBADALEJO MARTÍNEZ, Joaquin ALCONCHEL GONZAGA, Miriam ALLI MARTÍNEZ, Ínigo Jesus ALONSO ARANEGUI, Alfonso ALONSO CANTOR [...]"

[6: Independent Catalunya] ">El lider del PSC y candidato a presidir la Generalitat, Pere Navarro, acuso ayer al lider de ICV-EUiA, Joan Herrera, de «contribuir a hacer crecer la deriva independentista del presidente de la Generalitat y candidato a la reeleccion, Artur Mas (Ci En un mitin en Manresa ante 300 personas, Navarro [...]"

[7: Basque (goevernment) negotiations] "Ahora bien, pese a dejar clara la voluntad del Pacto de propiciar ese final dialogadode la violencia, Ardanza dejo claro que para llegar a ese punto ETA debe dar muestras inequivocas de querer abandonar la violencia¨ porque lo contrario ëquivaldria a provocar fracaso y frustracion."

[8: ETA attacks] "APOYOUn completo arsenal etarraEntre el material incautado a ETA figuran 180 kilogramos de nitrato amonico, 15 litros de nitrometano, un subfusil MAT con dos cargadores,

una pistola Browning con dos cargadores, un revolver Smith Wesson calibre 38 con municion, varios 'tupper' para la confeccion de b [...]"

[9: Comptences, finances and administration] "Son de especial relevancia: 1) incluir en la Constitucion, como sugirio el Consejo de Estado, mencion expresa a las comunidades autonomas; 2) regular el Senado como Camara que represente eficazmente a los territorios tanto por su composicion como por sus funciones; 3) reconocer las singularidades y [...]"

[10: Autonomy statute reform] "Los representantes de UCD y PSOE han acordado, tras una reunión conjunta, solicitar que en el orden del dia de la asamblea que el pleno del Consejo General de Castilla y Leon celebrara el sabado en Avila se incluya una peticion al Gobierno para que convoque la Asamblea de Parlamentarios y Diputacion [...]"

[11: ETA prisoners and victims] "Al igual que Bolinaga fue puesto en libertad con el pretexto de que sufria una enfermedad terminal, al igual que decenas de presos han salido a la calle por la aplicacion de la sentencia de Estrasburgo y al igual que otros muchos disfrutan de permisos y beneficios penitenciarios por la via Nanclares [...]"

Representative paragraphs for 19 territorial politics, UK

[1: Miscellaneous 1] "A Aberavon E 50,025 V 35,963 (71.9%) John Morris (Lab) 25,650 Ron McConville (LD) 4,079 Peter Harper (C) 2,835 Phil Cockwell (PC) 2,088 Peter David (Ref) 970 Captain Beany (Beanus) 341 Lab hold Maj 21,571 Swing 2.7% from LD to Lab 1992: Lab 26,877; C 5,567; LD 4,999; PC 1,919; Real Bean 707 Aberdeen [...]"

[2: Home rule] "A- meeting of the Opposition peers was held yesterday at the House of -Lords to consider the contentious Bills which the Government are sending up shortly. There were about 40 present, including Lord Lansdowne, the Duke of Devonshire, Lord Midletont, Lord Salisbury, Lord Camperdown, Lord Kenyon, Lor [...]"

[3: International relations] "Decades of discord 1951 Iran nationalises precursor of BP, the AngloIranian Oil Company, triggering a dispute with Britain 1953 The Prime Minister, Mohammed Mossadeq, deposed in a coup with British and US backing 1980 Britain closes its embassy in Tehran after the Islamic revolution 1989 The Irania [...]"

[4: Miscellaneous 2] "Remember Kia-Ora Remember Kia- Ora Remember Kia-Ora Remember ftsaOrRi be 'iaa Remem her n wMnber EKia-Ora RemeJAr . EW'cmember Kia- Ora 1 I5 l vAr Kia-Ora Remcmber Kiaa Fmember Kia-Ora Remem ber Kia-Ora Remember Kia-Ora Remember Kia-Ora Remember Kia- Ora Remember Kia-Ora Remember Kia-Ora Remember [...]"

[5: Culture 1] " mmmmMfWftmi swas«w««««'w$?mmmm!8M>mMm!m Theatres ADELPHI 0844412 4651 loveneverdies.com 'ANDREW LLOYD WEBBER AT HIS MUSICAL BEST' Times LOVE NEVER DIES Mon-Sat 730pm, Wed Sat 2.30pm ALDWYCH THEATRE 0844 847 1714 DIRTY DANCING THE CLASSIC STORY ON STAGE Mon-Thur 730, Fri 5 8.30pm, Sat 3 h 730 [...]"

[6: Miscellaneous 3] "Home Away P W D L F A W D L F A GD Pts 1 Walsall 22 9 1 1 21 3 5 5 1 14 10 22 48 2 Swindon 22 7 1 3 16 9 6 3 2 14 9 12 43 3MKDons 22 7 1 3 19 13 6 2 3 20 16 10 42 4 Lincoln City 22 6 2 3 23 14 7 0 4 20 14 15 41 5 Wycombe 22 6 4 1 15 7 5 1 5 11 12 7 38 6 Peterborough 22 5 2 4 25 21 6 3 2 [...]"

[7: Taxation] "JwÔMAN AVIATION SERVICES CO It t *Ji J+* »-A * H»-» Uff tMa tYUTOK UTKIl CO. CUOQ Excellent Career Challenge Attractive Tax Free Salary Other Benefits Oman Aviation Services Company SAOG is a growth orientated public company in the Aviation industry based in the Sultanate of Oman The company s a [...]"

[8: Good Friday] "There are very few references to the border at all in the Belfast agreement, but where there are references, they do not in any way suggest that this decision cannot take place."

[9: Scottish indep. referendum] "Here's the agenda for the day. 10am: Conference opens with announcement of the results of the deputy leadership election. 10.15am: Welcome address by Elizabeth Grant, provost of Perth and Kinross council. 10.30am: Debates on the independence referendum, the minimum wage, social justice, cycling and [...]"

[10: Miscellaneous 4] "Japan Growth 294.20 313.701 - 1.90 ... Japanindex 81.65 87.14 - 0.21 0.09 Japan Smllr Cos 37.28 39.79 -0 .60 ... international High Growth Funds Asian 57.24 61.09 + 0.63 0.15 Hong Kong Gwlh 99.13 105.80 +0.50 0.78 Spore fiMlysn Gth 78.41 83.68 -0 .13 ... Tiger Index 208.30 2223W +0.10 0.22 INVESCO F [...]"

[11: Irish patriotism] "The relief whihh a settlement would bring to right-minded people in America would be only less acute than t.hat which it woould bring to the Irish and the British. The three peoples have very strong ties of blood, culture and sympat hy, which have not been severed during the last few tragic years of [...]"

[12: Political/fiscal decentralization] "Gentleman agree with the recommendations of the final Holtham report, published today, which calls for an immediate Barnett floor to protect Wales from further convergence, the implementation of transition mechanisms towards a needs-based formula, and a place at the table for the Welsh Government in [...]"

[13: Irish terrorism] "Nine regular soldiers were injured in two attacks by the Provisional IRA near the Irish border late on Saturday and early yesterday. None was seriously hurt and only one was kept in hospital. The attacks occurred in the same area where a corporal aged 30 was killed last Thursday by a landmine laid b [...]"

[14: Irish defenders] "O'Rahilly was a born rebel: a self-described anarchist whose grandfather had died while storming Dublin's General Post Office during the 1916 Easter Rising."

[15: Culture 2] "Thandie Newton stars in The Chronicles of Riddick (Sky Movies 4, 8pm) 7.00PM 7.30 8.00 8.30 9.00 9.30 10.00 See Choice (F) available on Freeview (HD) High Definition MAIN CHANNELS SKY ONE The Simpsons Three back-to-back episodes ofthe cartoon: The Regina Monologues; Special Edna; and Goo Goo Ga/' Pa [...]"

[16: Miscellaneous 5] "B HI niyn ¨/ Wi " ¨' 2O' i 14 Locker CO A 142 ... 3.5 ... 49 3B ML Hogs 39 ... 2.7 3.5 . 67 26 MS Inll 26 ... 12.0 5.7 179 129 Mang Bronzet 141 + 1 3.5 11.0 539 412 McKecliniet 428... 4.3 15.B 120 71'=MeBQiltt 82 6.0 13.2 120 101 Metalr 106 ... 4.4 17.6 133 99 MdlSKt 131 ... II I? 589 518 Molins [...]"

[17: Miscellaneous 6] "J.D. YAXMULO.COnntrollm ,theetir of nt TADING r IT d the d 1INMY (c) N AsiCuTv 190 Ute and to the Matter ot CEItXIAItD cltm HEY 1 2 d nr tomand Apostle. Q i.ond n ckrentYEC. aOrerof the ead of adbe, date th1t (df julyl96 pruodr Scto d. sob-scutIo 11(b) or the above mentiod ct.or*equiry rD th t inh [...]"

[18: Peace in N-Ireland] "The collapse of Northern Ireland's political institutions moved closer yesterday after David Trimble, the Ulster Unionist Party leader, said that he was withdrawing his ministers from the Stormont Executive in protest at the IRA's continued refusal to decommission. He said he was pressing ahead afte[...]"

[19: Miscellaneous 7] "The Chancellor of the Exchequer had said that almost the whole of the provisions to be found in this Bill were taken from the report to which his right hon. friend the member for Wimbledon referred and that all they objected to was the machinery That seemed to him to be begging the question. On a su[...]"