Cross-National and Sub-National Diffusion of Issue Definition: The Case of Smoking Bans in Switzerland*

Fabrizio Gilardi† Charles R. Shipan‡ Bruno Wueest§ Lucien Baumgartner¶

Work in progress
August 27, 2018

Abstract

We study how the definition of a policy issue diffuses cross-nationally and sub-nationally relying on an an original, comprehensive corpus of newspaper coverage of smoking bans in Switzerland. Previous research has shown that the way policy issues are defined varies depending on how widespread that policy is within a given unit's diffusion network. In this paper, we further this research question by studying how issue definition in the French and German-speaking regions refer to the experience of neighboring countries (Germany, Austria, France, Italy). Second, at the sub-national level, we examine how the first adoption of smoking bans in Switzerland, in the Italian-speaking canton of Ticino, affected the definition of the issue in French and German-speaking cantons. The analysis relies on structural topic models and Named Entity Recognition techniques, which are applied to about 30,000 German and French newspaper paragraphs covering smoking bans.

*We thank Katrin Affolter, Nina Bader, Nina Buddeke, Sarah Däscher, Andrea Häuptli, Sabrina Lüthi, Adriano Meyer, Christian Müller, and Thomas Willi for excellent research assistance, Klaus Rotherhäusler for technical assistance, and Fridolin Linder for the evaluation of topic models. The financial support of the Swiss National Science Foundation (grant nr. 100017_150071/1) is gratefully acknowledged.

†Department of Political Science, University of Zurich (http://www.fabriziogilardi.org/).
‡Department of Political Science, University of Michigan (cshipan@umich.edu).
§Department of Political Science, University of Zurich (http://www.bruno-wueest.ch/).
¶Department of Political Science, University of Zurich.

1 Introduction

Policy diffusion is the process whereby policies in one unit such a state or country are shaped by the policies of other units. There is a vast literature studying this phenomenon, spanning several subfields such as international relations, American politics, federalism, and public policy (Graham et al., 2013, 2014). Research in this area is very diverse in terms of the specific policies and cases that are investigated, but what most studies share is a focus on the decisions that are made or the policies that are adopted or implemented. However, Gilardi et al. (2018) recently argued that policy diffusion research could benefit from shifting the focus from adoption or implementation to a prior stage of the policy cycle, namely, issue definition. One of the basic ideas of agenda setting is that the policies are adopted and, later, implemented depend crucially on how those policies, and the problems they aim to address, are defined and understood. The argument put forward by Gilardi et al. (2018) is that diffusion processes can operate already at the issue definition stage. Looking at smoking bans in the US, Gilardi et al. (2018) found that the way smoking bans are framed in a given state is related to the policies previously enacted by other states within that state's diffusion network. Specifically, normative frames (such as health or freedom) are less responsive to diffusion than more practical frames related to smoking bans' concrete aspects such as specific regulations and establishments. Moreover, frames tend to become more complex as the policy becomes more widespread within the diffusion network.

In this paper, we build on these ideas by looking more specifically at how issue definition is connected with specific units that could be potential sources of diffusion. Concretely, we focus on Switzerland and analyze how smoking bans were framed in the 26 cantons. We leverage two aspects of our case. First, three main languages are spoken in Switzerland (German, French, and Italian), which correspond to the languages of Switzerland's neighboring countries (Germany, Austria, France, and Italy). Therefore, we look at how mentions of these countries co-vary with different ways of framing smoking bans. Second, the spread of smoking bans in Switzerland was initiated by a clear leader, the Italian-speaking canton of Ticino. Therefore, we consider how mentions of this canton co-vary with the frames used to describe smoking bans.

Similar to Gilardi et al. (2018), our analysis relies on structural topic models (Roberts et al., 2016), which we apply to an original corpus of about 30,000 texts in German and French and complement with named entity recognition tools to identify automatically the units mentioned in the texts.

The next section briefly discusses the theoretical basis for our study. We then explain the methodology, including case selection, the corpus, the structural topic model we estimate, and the covariates we include. Finally, we present our preliminary results. The short conclusion outlines the next steps for this project.

2 Policy Diffusion and Issue Definition

The vast majority of studies in the policy diffusion literature focuses on the adoption of policies or their implementation. This makes sense for several reasons. Policy adoption and implementation are obviously important and, from a practical perspective, can be observed and measured with relative ease. However, there is more than adoption and implementation to policy making. A basic insights of policy analysis, very explicit in the policy cycle, is not only that many issues do not reach the decision stage (Kingdon, 2003), but also that agenda setting is closely related to issue definition (Schattschneider, 1960). Therefore, a strict focus on decisions misses those aspects of policy making that are less visible but essential to understand both non-decisions (Bachrach and Baratz, 1962) and the specific form of the policies that do reach the adoption and implementation stages.

The study of issue definition has been at the center of many influential studies (Baumgartner and Jones, 1993; Baumgartner et al., 2008), but very few have included a diffusion angle. Boushey (2016) makes the connection but from a perspective opposite to ours—in his work, issue definition is considered an explanatory variable for policy diffusion, whereas for us it constitutes the dependent variable. Here, we build on our previous work Gilardi et al. (2018), which established a connection between policy adoption within a state's diffusion network and issue definition in that state. We extend that analysis using a new dataset to account for the units that could serve as sources for the diffusion process, an aspect that was not considered in Gilardi et al. (2018).

As we explain below, our empirical approach is strongly inductive. Therefore, formulating clear hypothesis *a priori* is not straightforward. However, based on Gilardi et al. (2018) the expectation is that issue definition will be more strongly connected with diffusion for topics that are practical rather than normative. Moreover, we expect the connections to reflect the linguistic commonalities among Swiss cantons and neighboring countries, as well as Italy's role as a leader in this policy area.

3 Methodology

3.1 Case Selection

Our analysis of policy frames as a part of the diffusion process concentrates, as noted earlier, on the adoption of antismoking policies in Swiss cantons. Swiss cantons historically have had considerable autonomy in public health areas, and smoking restrictions are no exception. Although smoking-related issues have been often discussed by politicians at the national level, most policymaking has taken place within the cantons. Thus, the issue of anti-smoking laws at the state level provides an excellent forum for examining the process of diffusion.

Our choice of policy area is also motivated by several other considerations. First, it is well established that smoking bans have exhibited a diffusion process. This allows us to concentrate on the nature of the process—in particular, the ways in which this issue is defined as a function of references to geographical units—rather than the mere existence of diffusion. Second, smoking bans have been adopted in a convenient time frame in Switzerland—roughly a ten-year period—that is long enough to detect variations and to supply sufficient information, but short enough to be practically manageable.

Third, smoking ban policies have have well-defined characteristics and are comparable across units.

3.2 Corpus

The time period we examine begins in 2003, which is two years before the first canton-wide smoking ban was adopted in Ticino. Moreover, 2003 was also the year in which a nationwide smoking ban was approved in Italy (implemented two years later)—one of the main countries of interest in this study. To analyze public discussions and identify policy frames within a canton, we processed articles published in 23 Swiss newspapers covering 22 cantons (see Appendix A). Our goal was to include one of the largest newspapers in terms of circulation for each canton in the German and French-speaking regions, which we achieved in all except two cases. For the canton of Jura, we could not get access to any of the relevant newspapers. In addition, we could only retrieve one newspaper for the two half cantons of Appenzell (*Appenzeller Zeitung*). The corpus covers the full time period for most newspapers. We use print media rather than television or radio programs partly for technical reasons but especially because they generally report more extensively on political matters than do on-air media (Druckman, 2005, 469). We can therefore expect that newspapers convey rich information on local debates on smoking bans. One important question that arises is whether the media coverage we examine reflects how policies are framed, or whether it influences the frames. On this question we are agnostic. Regardless of whether this coverage reflects or influences frames, media coverage can be used as an accurate source for identifying the ways in which smoking bans are framed and, more generally, "as an indicator of the nature of public discussion" (Baumgartner et al., 2008, 20).

We retrieved newspaper texts using a simple, broad keyword search from different database providers. Then we split the texts into paragraphs of a similar length and removed duplicate paragraphs, which produced a corpus containing 340,824 paragraphs (German corpus) and 99,227 paragraphs (French corpus). We provide more details on these procedures in Appendix B. A manual evaluation of a random sample of paragraphs revealed a very low share of paragraphs actually covering smoking bans, most likely due to the looseness of our keyword search, which was aimed at minimizing the number of articles on smoking bans escaping our search. To remove irrelevant paragraphs, we conducted a supervised text classification. First, we used the crowd-sourcing platform Crowdflower (now Figure Eight[1]) to annotate a sample of about 10,000 paragraphs as relevant or irrelevant. We followed the procedures explained in Benoit et al. (2016) and found that the crowd annotation produced results comparable with three expert codings. Appendix D discusses the coding instructions given to the crowd-workers and the design of the crowd-coding.

Second, using the information obtained through crowd annotation, we then classified all paragraphs in our corpus as relevant or irrelevant using a machine-learning classifier built with the Python module `scikit-learn`. Prior to the classification, we pre-processed all documents with standard procedures.[2] Next we evaluated seven algorithms[3] on 100 bootstrapped training samples and optimized the output

[1] https://www.figure-eight.com/.

[2] Text segmentation, tokenizing, removal of punctuation, stemming, part-of-speech tagging, and conversion of all words to lowercase.

[3] Ada boost, Bernoulli naïve Bayes, Gaussian naïve Bayes, K-nearest neighbors vote, random forest, support vector machines, and logistic regression.

in terms of the ratio between true positives and false positives (i.e., the receiver operating characteristic). An ensemble of two random forest and one support vector machine classifiers proved to be most effective on the French corpus, while an ensemble of two support vector machines and one random forest classifier outperformed all other ensembles on the German corpus (see Appendix E for the evaluation). These ensembles produced a final German corpus of 26,077 paragraphs and a final French corpus of 4,093.

3.3 Structural Topic Model

We identify policy frames inductively with a structural topic model (STM) (Roberts et al., 2014b, 2016). The STM builds on well-established generative topic models, namely the Correlated Topic Model (CTM) (Blei and Lafferty, 2007), which is itself an extension of the well-known Latent Dirichlet Allocation (LDA) model (Blei et al., 2003). These models are mixed-membership, that is, they assume that each unit of text (i.e., in our case, each paragraph) consists of a mixture of topics (Grimmer and Stewart, 2013, 283–285). As a consequence of the logistic-normal distribution underlying these models, topic prevalences always add up to 1 for each document. Therefore, if a topic has a higher-than-average prevalence in a document, it lowers the prevalence of the other topics. This assumption is very realistic: if a given topic takes up more space in a text, it necessarily reduces the attention given to other topics. Moreover, the assumption is consistent with the strategy used by Baumgartner et al. (2008) to manually code all existing component parts of frames for every document.

The STM's major innovation is that the prior distribution of topics can vary as a function of covariates (Roberts et al., 2014b, 2016). The inclusion of covariates in the topic model makes it possible to test hypotheses in a regression-like framework, that is, to uncover covariation between topic prevalence and variables of interest. Concretely, in our study, the STM's ability to include covariates means that we can examine directly our main expectation, namely, that topic prevalence within a canton—our measure of issue definition—is linked to references to neighboring countries and other cantons. Moreover, the STM allows us to control for other factors that may be related with topic prevalence. We discuss the covariates that we include in our analysis in Section 3.4.

We estimate our topic models using the `stm` package in R (Roberts et al., 2014a). We initialize the models with the spectral algorithm, which is robust to changes in several CTM parameters and starting values (Roberts et al., 2016). For both the German and the French corpora, we evaluated 47 models using `word2vec` (O'Callaghan et al., 2015), varying the number of topics from 3 to 50, and found that the 12-topic model performed best for the German corpus and the 9-topic model performed best for the French corpus. We describe this step of the analysis in more detail in Appendix C.

3.4 Covariates

The most important covariates in our analysis measure references to neighboring countries and the canton of Ticino. To construct these variables, we first conducted a Named Entity Recognition (NER) to extract all geographical terms such as the names of cities, countries, regions, continents, neighborhoods and administrative divisions in the paragraphs. We use the `polyglot` NER for this task, which identifies geographical terms with word embedding models generated by neural networks that are trained on

Wikipedia sites (Al-Rfou et al., 2015). We inspected samples of the extracted references and found that this NER in general worked well. In a second step, we manually recoded all extracted references to the level of cantons and countries when the geographical units did not refer to Switzerland. Hence, we are able to include the information on which canton and foreign country is mentioned in a paragraph into the STMs.

In addition to sentiment and the share of prior policy adoptions within a state's diffusion network, the analysis includes several other covariates, which we use to control for important factors that might affect the way smoking bans are framed: (1) the share of neighboring cantons that already have implemented a government buildings or restaurant smoking ban; (2) newspaper IDs, to identify the states in which newspapers are based; (3) the number of months before and after the enactment of smoking bans (with a B-spline of order 10), since the framing of smoking bans is likely to change before and after their introduction; (4) the shares of the most important parties (the Swiss People's Party, the Social Democrats, the Liberals and the Christian Democratic People's Party) in the cantonal legislatives and executive, because these parties tend to have different views on smoking restrictions; and (5) whether a newspaper is published in a tobacco-producing canton, since this variable might be related to the popularity of smoking bans.

4 Results

The discussion of our results proceeds in three steps. First, in Section 4.1 we present the topics identified by our STMs (Figures 1 and 2). Second, in Section 4.2 we show how policy frames are related to cross-national and sub-national diffusion—that is, how topic prevalence is linked to the coverage of smoking bans in the four neighboring countries (Figures 3 and 4) and the canton of Ticino (Figure 5).

4.1 Topics

For the reasons explained in Section 3.3, we present here the results of models assuming 22 topics for the German corpus and 10 Topic for the French corpus. Figures 1 and 2 show the top-30 words associated with each topic, along with labels that we assigned to each topic based on both the top-30 words themselves and also a reading of the most relevant paragraphs for each topic. The interpretation of the topics is quite straightforward and their connection with smoking bans clear—indeed, the frames that public health experts had previously identified all emerge from the data, lending strong support to the validity and value of our approach. Overall, our model identifies relevant and meaningful topics, and does so to a surprising extent, given that they were produced purely inductively, without human input apart from the selection of the number of topics.

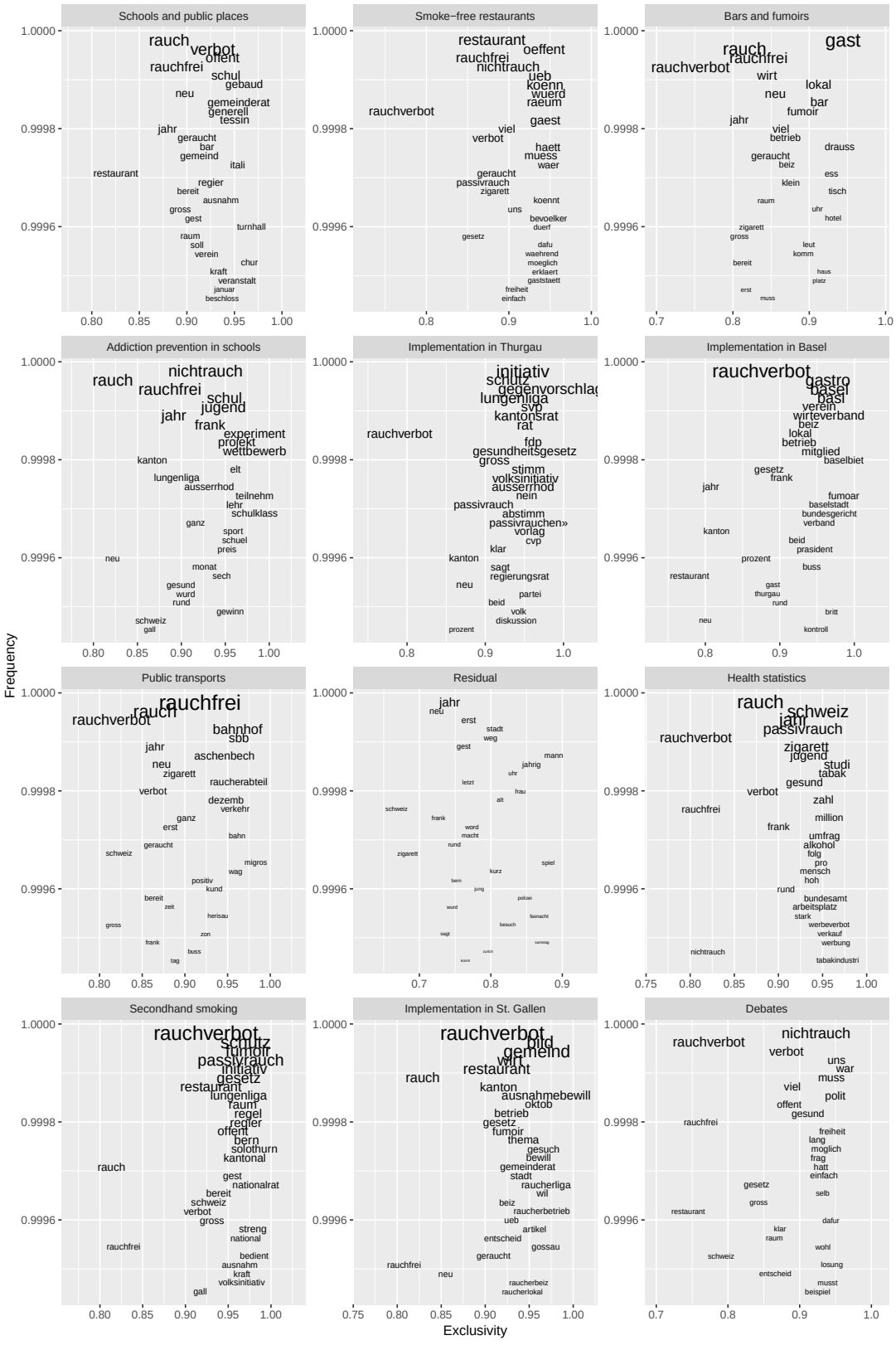

Figure 1: *German corpus: top-30 words for the 12-topic model. Exclusivity refers to the frequency with which words occur for one topic, compared to the occurrence for all other topics, estimated following Bischof and Airoldi (2012).*

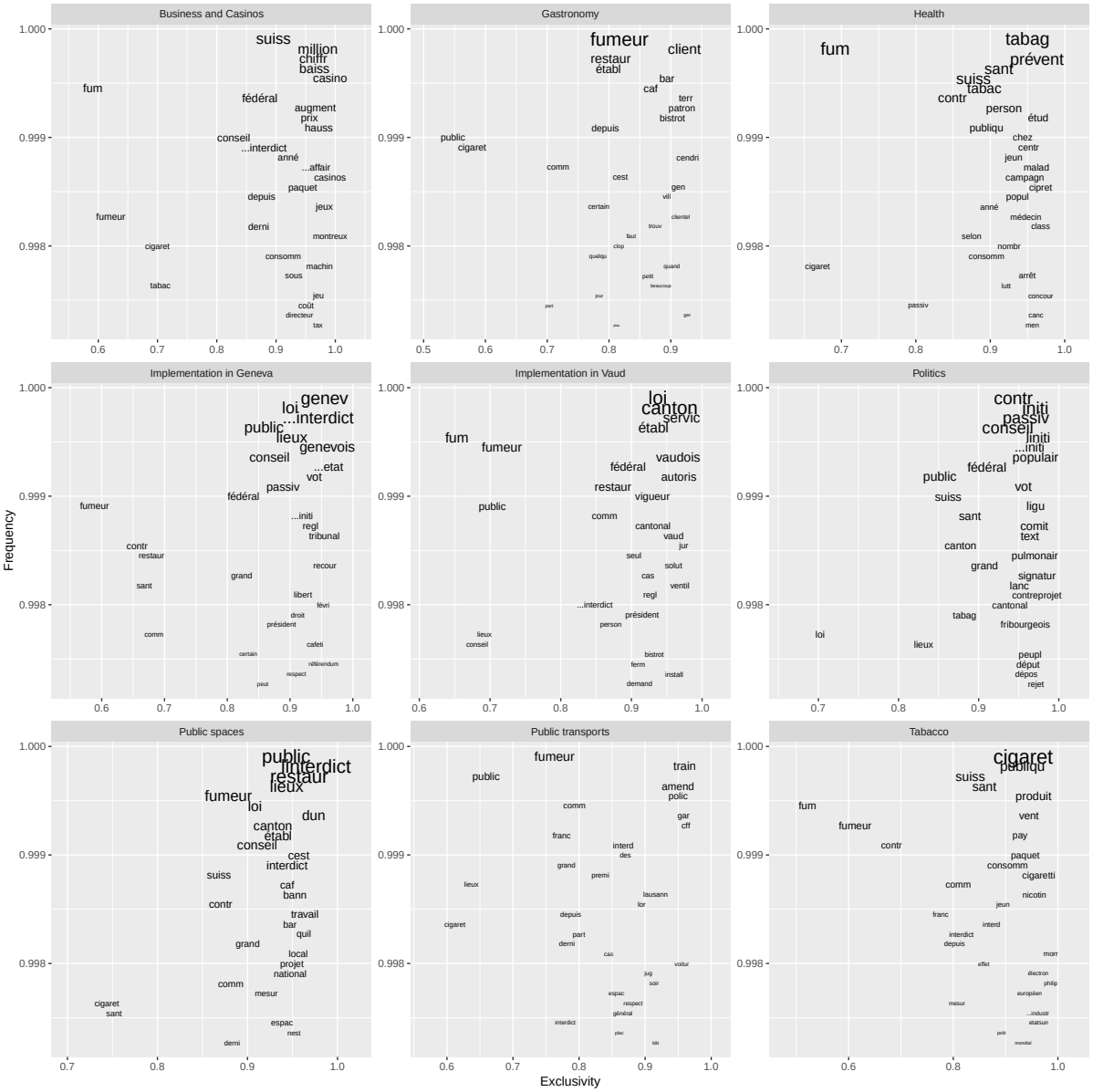

Figure 2: *French corpus: top-30 words for the 9-topic model. Exclusivity refers to the frequency with which words occur for one topic, compared to the occurrence for all other topics, estimated following Bischof and Airoldi (2012).*

In the German corpus (Figure 1), topics are related to public health (*Secondhand smoking*, *Health statistics*, and *Addiction prevention in schools*), locations affected by smoking bans (*Schools and public places*, *Smoke-free restaurants*, *Bars and fumoirs*, *Public transport*), implementation in specific cantons (*Implementation in Thurgau*, *Implementation in Basel*, *Implementation in St. Gallen*), and debates surrounding smoking bans (*Debates*). We also have a residual topic that was difficult to categorize.

In the French corpus (see Figure 2), topics only partially overlap with those found in the German corpus. Health, restaurants, public transport, and public places all show up, as well as implementation in two French-speaking cantons (Geneva and Vaud), but we also see a stronger emphasis on politics, as well as the negative effects of smoking bans on casinos, which we did not identify in the German corpus.

The biggest difference with the analysis of U.S. states in Gilardi et al. (2018) is the absence, in the Swiss corpus, of references to individual freedom, although the topic may still emerge as we refine the analyses.

4.2 Cross-National and Cross-Cantonal References and Issue Definition

In addition to determining the content of each topic, our approach also allows us to investigate whether these topics are discussed more or less prominently if the newspapers cover smoking bans in other geographical units—that is, it allows us to identify the correlation between the topic prevalences and the references to neighboring countries and the canton of Ticino found in each paragraph. Figure 3 (German corpus) and 4 (French corpus) show the extent to which each topic is related to one of Switzerland's neighboring countries, measured by the percentage of paragraphs on that topic with a geographical reference minus the percentage with no geographical reference. Although we are careful with interpreting the results at this early stage of the analysis, we think that figure provides several insights into the topics we have identified.

In the German corpus, references to smoking bans in Italy correlate strongly with *School and public places* and *Health statistics*, as well as *Debates*. This suggests that the Italian example was discussed with reference to the consequences of smoking and smoking bans (*Health statistics*) and when debating the merits of the policy (*Debates*). Interestingly, *Smoke-free restaurants* correlates equally with all four countries. *Second-hand smoking* correlates strongly with Austria, a country that to date has not implemented nation-wide smoking bans. It seems that Italy stands out compared to the other three countries. Germany correlates strongly only with a topic related to implementation in one specific canton.

In the French corpus, surprisingly France does not stand out, although it correlates with the topic *Gastronomy* roughly to the same extent as Italy. Also surprisingly, Germany strongly correlates with *Politics*, while Austria is strongly related to *Public spaces*. Overall, the correlations in the French corpus are less intuitive than those in the German corpus.

The final step of the analysis focuses on sub-national diffusion, which is why references to the canton of Ticino—the first mover in terms of discussing, deciding on and implementing a canton-wide smoking ban in Switzerland—are under scrutiny. Figure 5 accordingly shows the correlations of the topics with references to the canton of Ticino in both corpora. In the German corpus, the prevalence

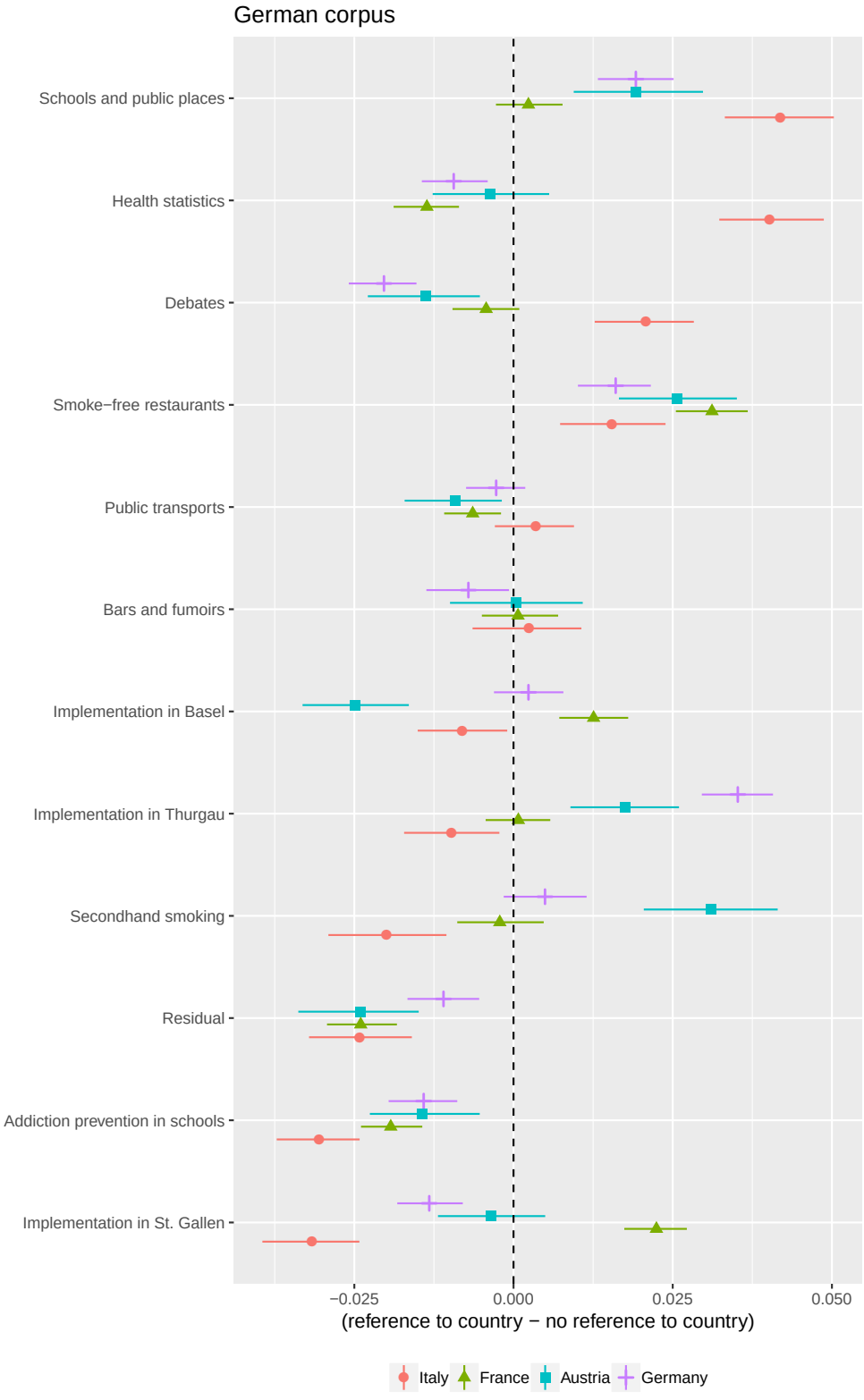

Figure 3: *Difference in topic prevalence and references to neighboring countries in the German corpus. A reference to a specific country means that the Named Entity Recognition identified a geographical term related to this country in a paragraph.*

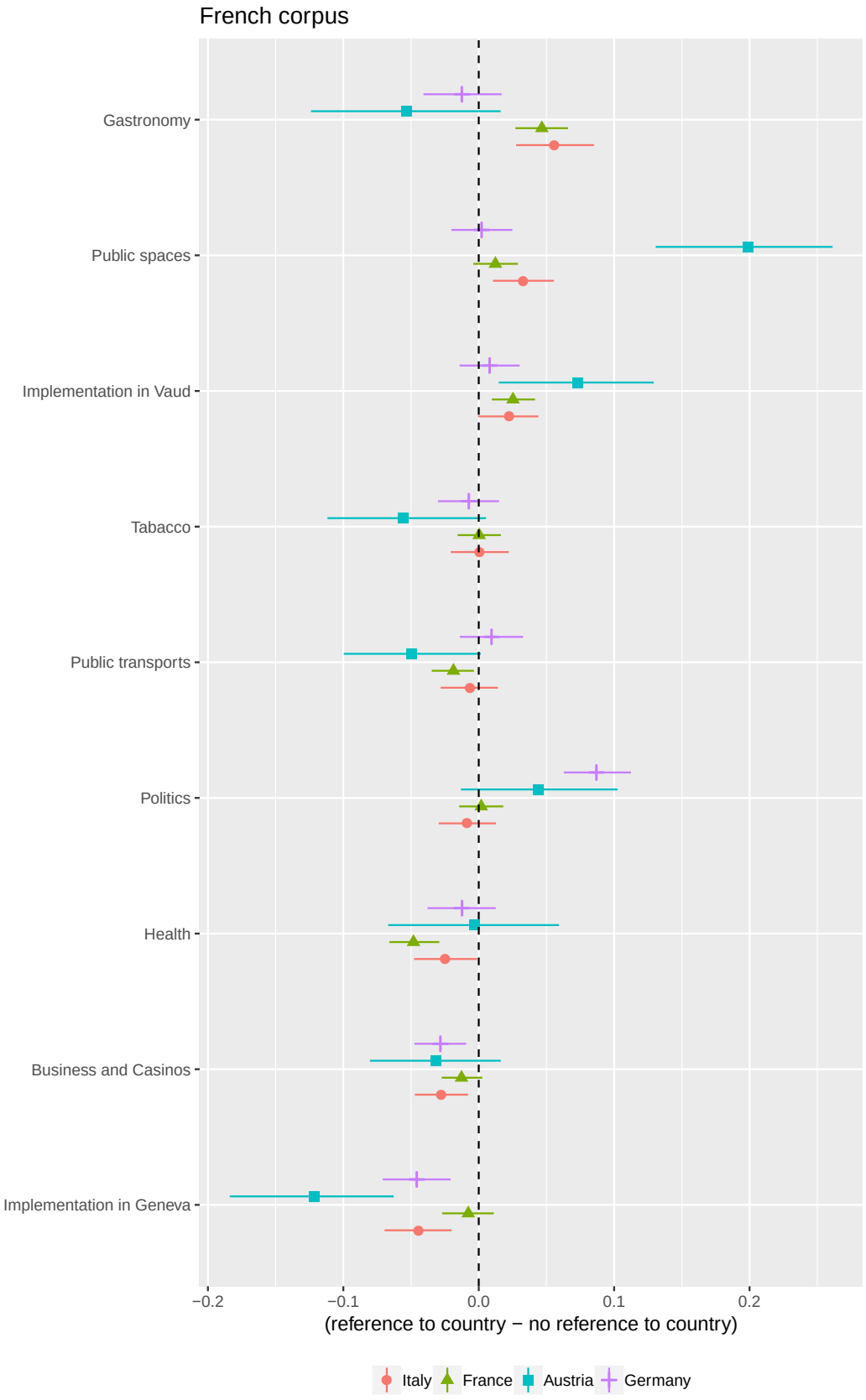

Figure 4: *Difference in topic prevalence and references to neighboring countries in the French corpus.*

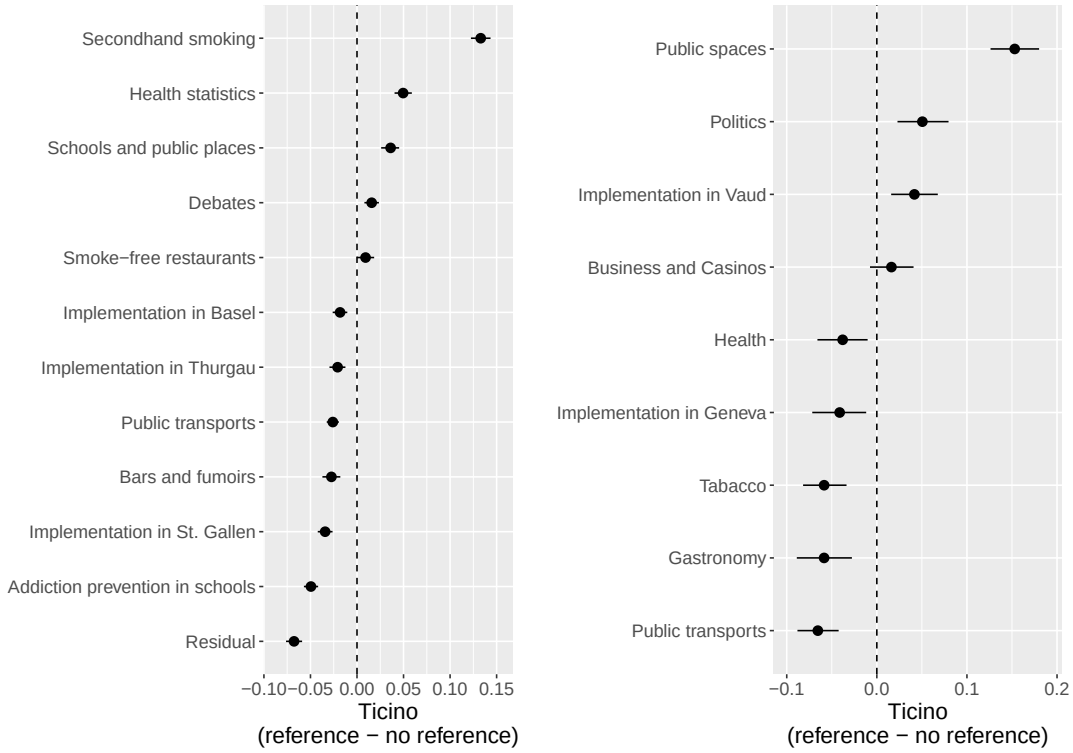

Figure 5: *Difference in topic prevalence and references to the canton of Ticino. A reference means that the Named Entity Recognition identified a geographical term related to this canton in a paragraph.*

of topics related to the health consequences of smoking bans (*Secondhand smoking*, *Health statistics*), as well as locations (*Schools and public places*) and, to a lesser extent, *Debates*, is significantly higher in articles mentioning the canton of Ticino. In the French corpus, we find the same pattern for *Public spaces* and *Politics*, but not for *Health*. Based on these results, it seems that a couple of topics are clearly related to mentions of the first canton to introduce smoking bans in Switzerland, but we have not yet uncovered a very coherent pattern.

5 Conclusion

In this paper, we have presented preliminary evidence on how issue definition in a given canton is related to mentions of neighboring countries as well as of the first canton to adopt smoking bans in Switzerland (Ticino). The analysis is preliminary and the results are bound to change. The main next steps for this project are the following:

- Include mentions to more cantons and countries.
- Consider how many cantons or countries are mentioned in a text.
- Identify the main focus of the text (smoking bans in the same canton, in another canton, or abroad).
- Consider translating all texts into English.
- Optimize the number of topics.

References

Al-Rfou, R., Kulkarni, V., Perozzi, B., and Skiena, S. (2015). Polyglot-ner: Massive multilingual named entity recognition. *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada, April 30-May 2, 2015*.

Bachrach, P. and Baratz, M. S. (1962). Two faces of power. *American Political Science Review*, 56(3):947–952.

Baumgartner, F. R., De Boef, S. L., and Boydstun, A. E. (2008). *The Decline of the Death Penalty and the Discovery of Innocence*. Cambridge University Press, New York.

Baumgartner, F. R. and Jones, B. D. (1993). *Agendas and Instability in American Politics*. University of Chicago Press, Chicago.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., and Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, forthcoming.

Bischof, J. and Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 201–208.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022.

Boushey, G. (2016). Targeted for diffusion? how the use and acceptance of stereotypes shape the diffusion of criminal justice policy innovations in the american states. *American Political Science Review*, 110(1):198–214.

Collingwood, L. and Wilkerson, J. (2012). Tradeoffs in accuracy and efficiency in supervised learning methods. *Journal of Information Technology & Politics*, 9(3):298–318.

Druckman, J. N. (2005). Media matter: How newspapers and television news cover campaigns and influence voters. *Political Communication*, 22(4):463–481.

Gilardi, F., Shipan, C. R., and Wüest, B. (2018). Policy diffusion: The issue-definition stage. University of Zurich and University of Michigan.

Graham, E. R., Shipan, C. R., and Volden, C. (2013). The Diffusion of Policy Diffusion Research in Political Science. *British Journal of Political Science*, 43(3):673–701.

Graham, E. R., Shipan, C. R., and Volden, C. (2014). The communication of ideas across subfields in political science. *PS: Political Science & Politics*, 47(02):468–476.

Greene, D. and Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25(1):77–94.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21:267–297.

Hurrelmann, A., Krell-Laluhová, Z., Nullmeier, F., Schneider, S., and Wiesner, A. (2009). Why the democratic nation-state is still legitimate: A study of media discourses. *European Journal of Political Research*, 48(4):483–515.

Kingdon, J. W. (2003). *Agendas, Alternatives, and Public Policies*. Longman, New York, second edition.

O'Callaghan, D., Greene, D., Carthy, J., and Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42:5645–5657.

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003.

Roberts, M. E., Stewart, B. M., and Tingley, D. (2014a). *stm: R Package for Structural Topic Models*.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S., Albertson, B., and Rand, D. (2014b). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58:1064–1082.

Schattschneider, E. E. (1960). *The Semisovereign People: A Realist's View of Democracy in America*. Holt, Rinehart and Winston, New York.

Wueest, B., Clematide, S., Bünzli, A., Laupper, D., and Frey, T. (2011). Electoral campaigns and relation mining: Extracting semantic network data from swiss newspaper articles. *Journal of Information Technology and Politics*, 8(4):444–463.

Appendix

A Newspaper corpus

Newspaper	*Canton*	*Articles*	*Paragraphs*	*Filtered*
German				
Aargauer Zeitung	Argovia	8,782	27,005	1,370
Appenzeller Zeitung	Both of Appenzell	7,757	18,121	3,311
Basellandschaftliche Zeitung[a]	Basle-Country	4,444	11,851	1,479
Basler Zeitung	Basle-City	10,990	37,994	2,410
Berner Zeitung	Bern	12,968	39,135	2,718
Der Bund	Bern	7,508	28,127	1,328
Neue Luzerner Zeitung	Lucerne	5,947	18,738	1,369
Neue Nidwaldner Zeitung	Nidwalden	625	1,619	211
Neue Obwaldner Zeitung	Obwalden	58	116	15
Neue Schwyzer Zeitung	Schwyz	785	2,003	151
Neue Urner Zeitung	Uri	736	1,897	263
Neue Zuger Zeitung	Zoug	2,081	5,248	671
Schaffhauser Nachrichten	Schaffouse	1,688	4,485	387
Solothurner Zeitung	Soleure	2,916	8,410	804
St. Galler Tagblatt	St Gall	18,758	50,656	4,282
Südostschweiz (Glarus)[a]	Glarus	1,601	4,111	347
Südostschweiz (Grisons)	Grisons	3,755	10,659	864
Tages-Anzeiger	Zurich	15,564	54,753	2,217
Thurgauer Zeitung	Thurgovia	6,295	15,896	1,880
Total		113,258	340,824	26,077
French				
24 Heures	Vaud	6,961	26,083	886
La Liberté	Fribourg	8,184	28,408	1,199
L'Express[a]	Neuchâtel	3,288	11,544	537
Tribune de Genève	Geneva	10,094	33,192	1,471
Total		28,527	99,227	4,093

[a] More than one year of coverage could not be retrieved due to access restrictions.

Table A1: *Newspaper corpus.*

B Newspaper articles retrieval

The keyword string for the different newspaper databases was an adaptation of “tobacco OR non-smoking OR anti-smoking OR smoking OR cigar! OR (lung AND cancer) OR smoker." The specific form of the keyword string depends on the language of the newspaper and options available for Boolean operators and truncation wildcards.

We then split the texts into paragraphs of a similar length. The original paragraph structure of the documents was kept, but paragraphs with fewer than 150 tokens were merged until the paragraph exceeded 150 tokens. This ensures the comparability of the texts from different newspapers and across different document formats in each newspaper.

Following many previous newspaper text analyses in political science (e.g., Hurrelmann et al., 2009; Wueest et al., 2011), we disaggregate the retrieved newspaper articles into single paragraphs. We did so for two reasons. First, newspaper articles have very different lengths. Brief news stories and lengthy background reports occur even within the same newspaper. By splitting articles into paragraphs, we construct a more balanced corpus. Second, in journalistic writings, paragraphs usually are the basic structuring elements that feature a coherent and distinct content, and not all content is relevant for our topic. Our corpus, for example, contains a lot of general reports on parliamentary sessions. The debate on smoking bans is often only one among many debates that are covered in the same news article. Therefore, for our purposes the texts covering such other debates are best discarded for the analysis, as they would just introduce noise.

Finally, we identified and removed duplicate paragraphs. Our downloads contained a considerable number of articles that are almost duplicates of other articles—about 3 to 20 percent, depending on the newspaper outlet. These almost-duplicates are generated because publishers upload different versions of the same article into the database (e.g., when small corrections are made). We found that two paragraphs with a Jaccard distance of 0.97 or higher on their word sets can be safely classified as duplicates and we kept only one of them.

C Topic model coherence and discrimination

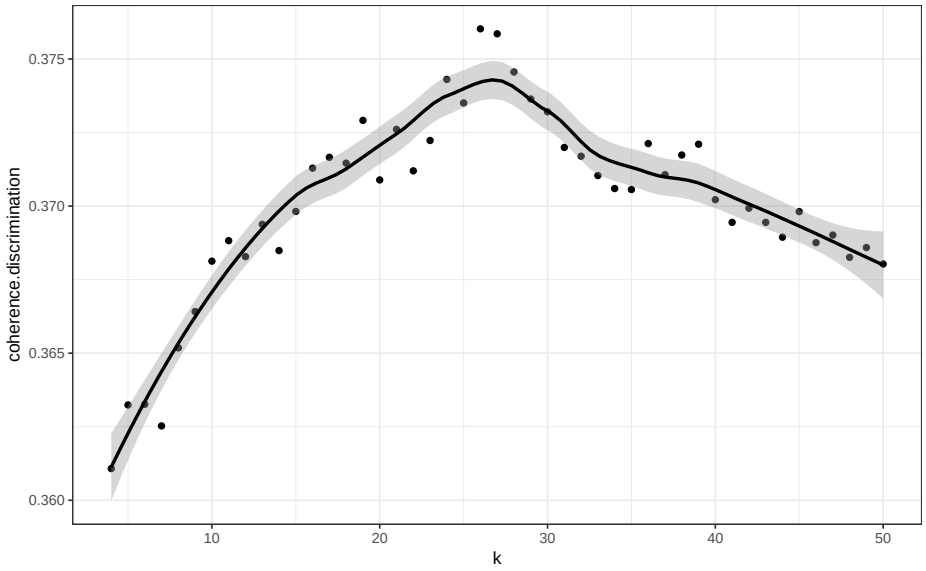

Figure C1: *Word2vec topic coherence and discrimination averages for varying numbers of topics in the German corpus.*

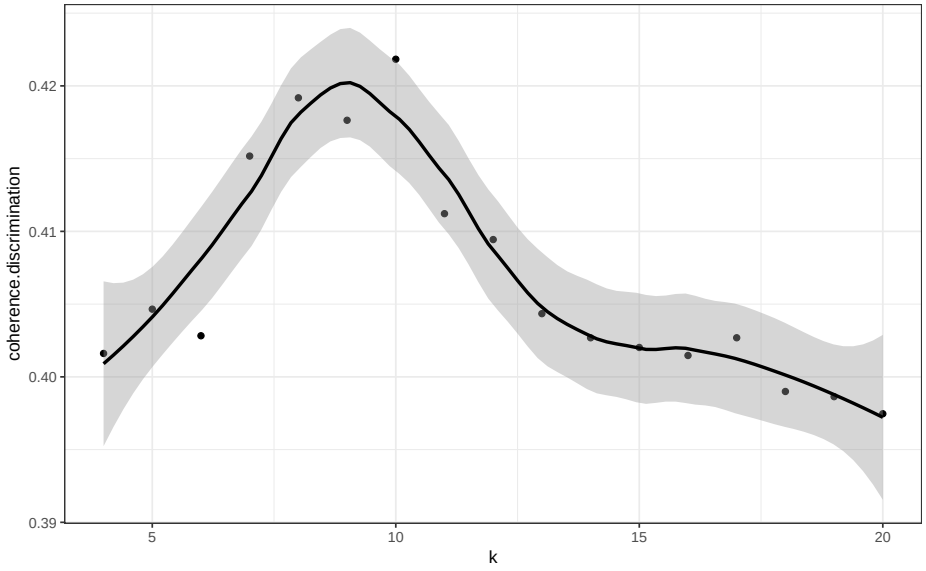

Figure C2: *Word2vec topic coherence and discrimination averages for varying numbers of topics in the french corpus.*

For this evaluation, the word2vec topic coherence and discrimination is calculated as follows. Let $T = t_1, ...t_K$ be the K topics estimated by a model and $t_i = [w_{i1}, ...w_{iP}]$ a vector of P top ranked words[4] that characterize each topic. In addition, let $w_{ij} = [d_{i1}, ...d_{iD}]$ be the D dimensional semantic space

[4]The ranking is based on the probability of observing each word in the vocabulary under a given topic. This probability,

estimated by *word2vec* for term w_j in topic i. Then, the coherence of topic t_i is the mean pairwise cosine similarity among the terms in the topic's word vector (see Greene and Cross, 2017):

$$c(t_i) = \binom{P}{2}^{-1} \sum_{m=2}^{P} \sum_{n=1}^{m-1} cos(\theta_{w_{im}, w_{in}}).$$

The discrimination between two topics t_i and t_j, in contrast, is the averaged inverse of the pairwise cosine similarity of all word pairs across the topics:

$$d(t_i, t_j) = P^{-2} \sum_{m=1}^{P} \sum_{n=1}^{P} (1 - cos(\theta_{w_{im}, w_{jn}})).$$

Our objective function for the evaluation of the topics, finally, is the average of discrimination and coherence weighted by α, which is set to 0.3 in our case:

$$f(T) = \alpha \binom{K}{2}^{-1} \sum_{i=2}^{K} \sum_{j=1}^{i-1} d(t_i, t_j) + (1 - \alpha) K^{-1} \sum_{i=1}^{K} c(t_i).$$

or β, is one of the main outputs of the STM (Roberts et al., 2016). For the most probable word lists per topic, words are simply ranked according to their topic-specific probability.

D Discussion of crowd coding

Our coding instructions indicated that relevant paragraphs are those containing information on smoking restrictions—that is, bans or limits on smoking in public places or specific workplaces. This definition includes statements about any kind of restriction of smoking ("smoking ban") in public places or businesses introduced through legislative action, executive action, or other democratic actions (e.g., direct-democratic processes). By contrast, we defined paragraphs discussing, for example, smoking bans introduced by private actors (e.g., companies, businesses), or bans of specific tobacco products (e.g., mentholated cigarettes), as irrelevant.

For establishing a development set for the classification of paragraphs into relevant or irrelevant ones in terms of coverage of smoking bans, we randomly draw around 10,000 paragraphs from the corpus and let them annotate on the crowd-coding platform *Crowdflower* (by now called *Figure eight*) as follows. First, we coded a sample of 60 paragraphs to establish the gold standard for the crowd coding. We deliberately oversampled relevant paragraphs to make sure crowd coders have enough learning material for this class. In a random sample, their share would have been negligible (around 7 percent). This gold standard was then used for an entry test as well as the continuous quality control during the annotations—every coder needed to have at least 80 percent of the gold standard questions correct. Otherwise, annotations were dropped. Second, we let five crowd coders annotate every paragraph in the full sample.

E Evaluation of the ensemble filters

The two ensembles worked well. Our evaluation of the classification on the German corpus indicates that 90 percent of the paragraphs classified as relevant, and 95 percent of those classified as irrelevant, are also identified as such in the crowd-annotated data. Moreover, the classifier is able to retrieve 82 percent of all paragraphs crowd-coded as relevant, and 97 percent of those crowd-coded as irrelevant. As for the French corpus, the N-weighted average precision is 92 percent and the N-weighted average recall is 93 percent. Finally, most classification runs we tested agreed, with an overall F1-Score[5] of 0.80 or higher—a further sign of the consistency and thus reliability of the classification (Collingwood and Wilkerson, 2012).

	Precision	Recall	N held-out set
German corpus			
Irrelevant	0.95	0.97	1,517
Relevant	0.90	0.82	457
Average	0.94	0.94	1,976
French corpus			
Irrelevant	0.94	0.98	1,713
Relevant	0.85	0.60	292
Average	0.92	0.93	2,005

Table E1: *Evaluation of the support vector classification filter. Recall is the fraction of correct classifications among the retrieved documents; precision is the fraction of correct classifications that have been retrieved over the sum of correct classifications; the held-out set is a subset of the training data that is exclusively used for evaluating the classifier.*

[5]The F1-Score is the harmonic mean of precision and recall. In addition, the overall F1-Score is inversely weighted by the number of documents in each class.