Online Supplement

This supplement provides the technical detail of the two analyses of the chapter “Topic Modeling, Parliamentary Records, and Knowledge Accumulation in Ideational Policy Studies”. Topic modeling in both examples was conducted using the R package STM. In addition, R packages tidyverse, tidytext, and stopwords were used in preprocessing and result presentation. The R codes are available from the first author per request.

Example 1

In our first example, we applied correlated topic model (CTM) to ParliamentSampo data in .csv form to explore the different meanings of the idea of wellbeing in Finnish parliamentary speeches in 1999-2023.[1] We used the STM package in R to conduct CTM. When no covariates are included to the model, structural topic model (STM) reduces to CTM.[2] The first data preprocessing steps were the combining of the separate .csv data files into one data frame, and removal of speeches held other than MPs as well as the interruptions and speaker’s comments. As we based the synthetical bounding of sentence corpus on the pattern of “(\\. |\\! |\\? |:)((?=([[:upper:]]))|(?=([[:digit:]])))”, we found Finnish and Swedish abbreviations ending with full stop problematic. Therefore, we located the cases where abbreviations ending with full stop were not at the end of the sentence, causing a faulty tokenization, and removed the full stops in those cases.

We then built three corpora with varying document bounds using a search string “hyvinvoi” OR “Hyvinvoi”. The first corpus consisted of whole speeches, the second of paragraphs (bounded by the line change in the speeches) and the third of sentences (bounded as explained above). After the corpora were built, we converted all the text into lowercase, and removed numbers, punctuation, Finnish and Swedish terms in the stopwords ISO list, and words that appeared only in one document.

We used search function to find the best number of topics. First, we ran the function with 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 topics on each corpus using Spectral initialization. Based on the diagnostics we then ran another round of searchK with 10-50 topics on speech corpus (Fig. 1), and 5-40 topics on paragraph (Fig. 2) and sentence corpora (Fig. 3). We chose the best number of topics for the three corpora by balancing the maximum semantic coherence and held-out likelihood: 12 for

[1] Hyvönen et al. 2024
[2] Roberts, Stewart, and Tingley 2019

speech corpus, 12 for paragraph corpus, and 12 for sentence corpus. The identical number of topics are coincidental.

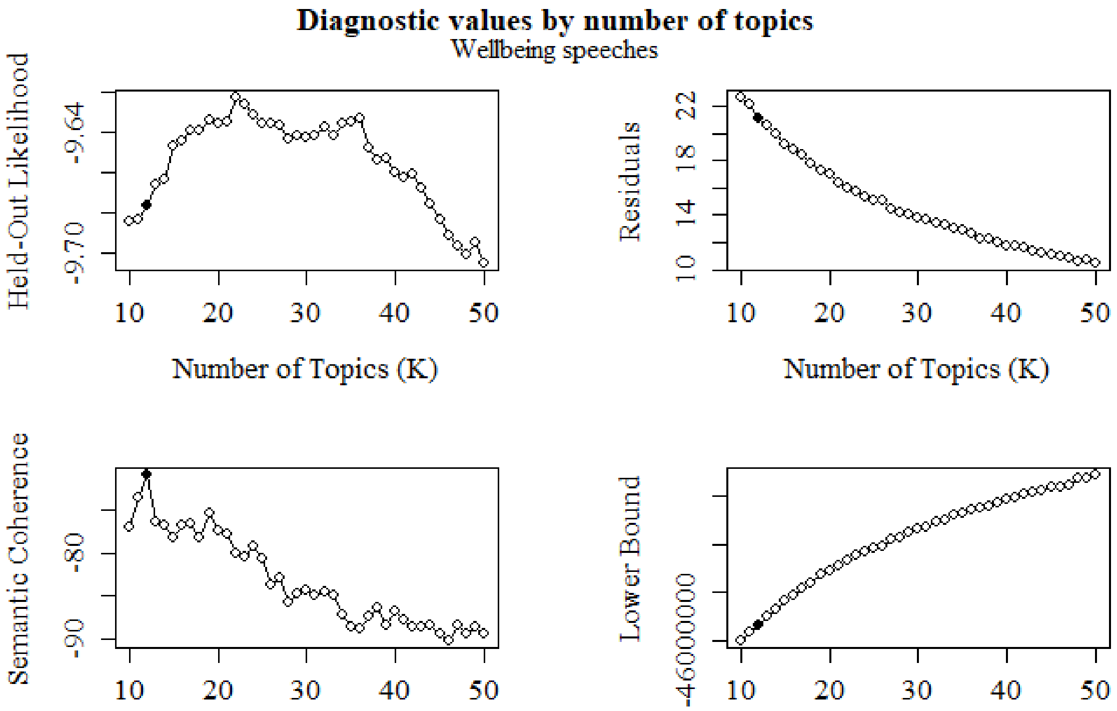

Figure 1 SearchK results for wellbeing speeches. Model with 12 topics is marked in black

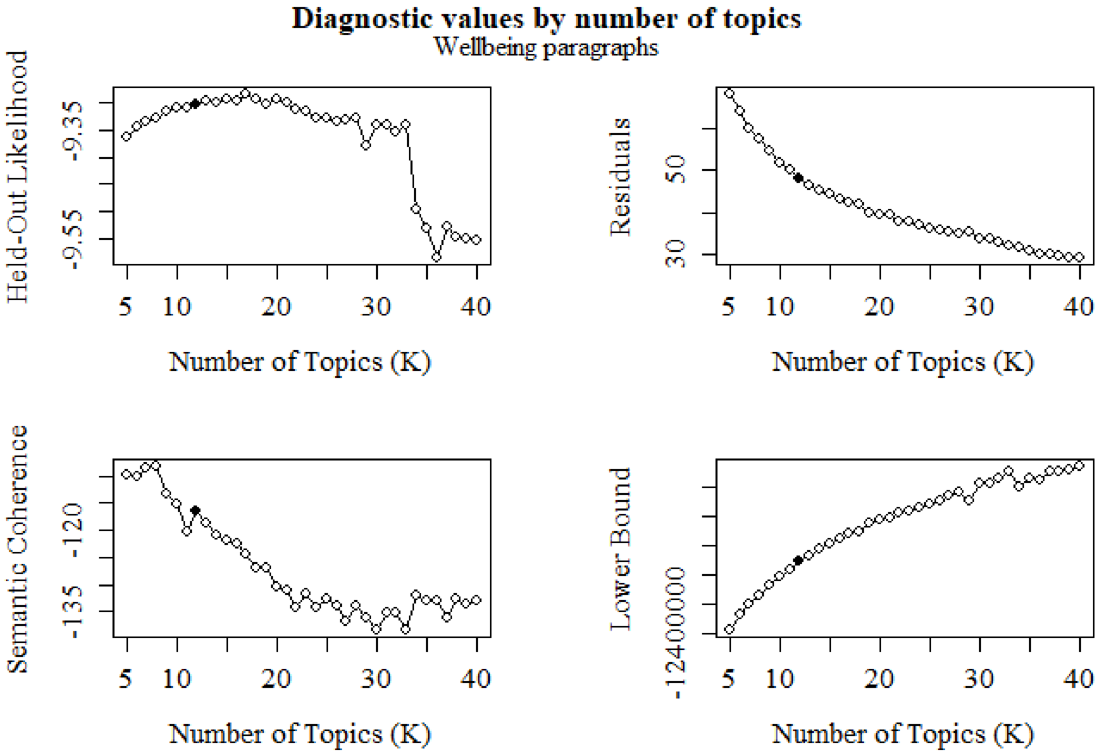

Figure 2 SearchK results for wellbeing paragraphs. Model with 12 topics is marked in black.

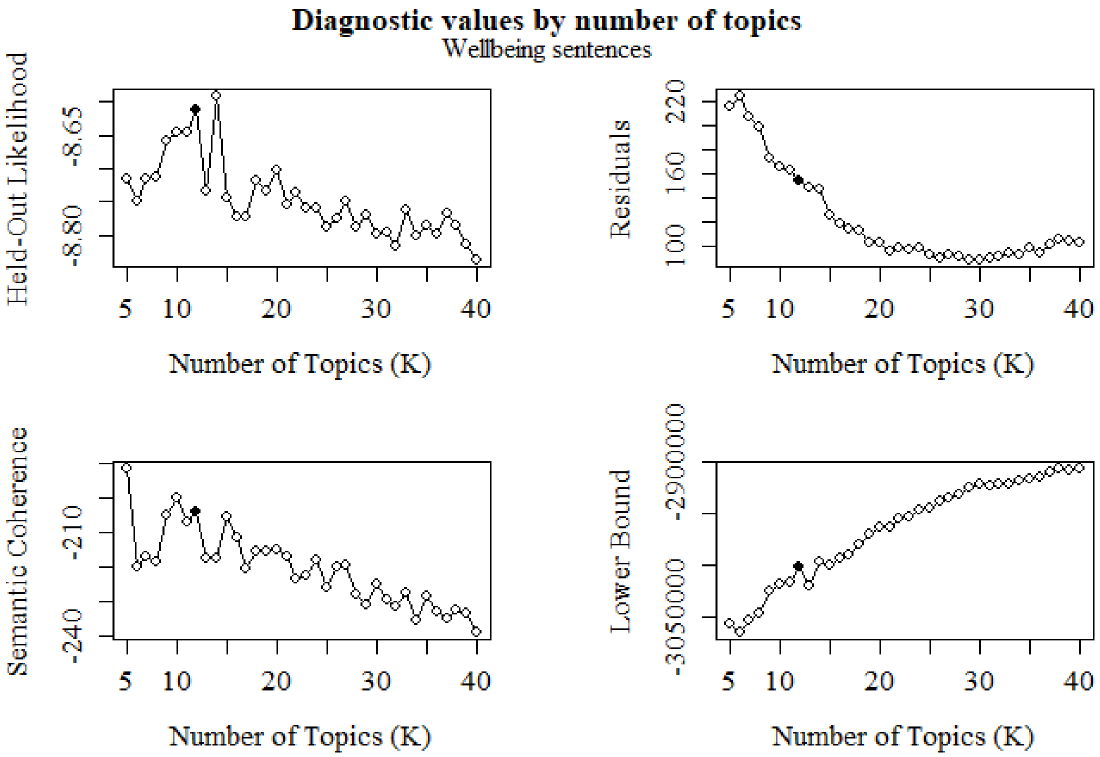

Figure 3 SearchK results for wellbeing sentences. Model with 12 topics is marked in black.

Example 2

In our second example, we explored the how the party affiliation effected the prevalence of the idea of inequality in Finnish parliamentary speeches in 2015-2022 using Parliament of Finland open datasets.[3] From the speech data, we built a corpus of synthetically bounded documents, i.e. sentences mentioning inequality, using a search string “eriarvoi” OR “Eriarvoi”. We then constructed a dummy variable: “left” includes members of parliament (MPs) affiliated to the Left Alliance and Social Democrats, while “other” consists of MPs affiliated to any other parties. Preprocessing begun with the removal of interruptions and speaker’s comments. After that, the sentences were converted to lowercase, and numbers, punctuation, and Finnish and Swedish terms in the stopwords ISO list were removed.

To determine the best number of topics, we ran searchK function with 3-20 topics using Latent Dirichlet Allocation (LDA) initialization and a random number of 911032 as the held-out seed. We used the LDA initialization instead of the recommended spectral initialization, as the performance of the spectral initialization has shown to suffer from small data sizes.[4] The best model with six topics maximized held-out likelihood while scoring also well in semantic coherence (Fig. 4).

When the spectral initialization is not used, it is to run selectModel function to overcome the shortcomings of the initialization method.[5] We ran the function with 20 models with six topics and chose the runout number one that maximized the exclusivity (Fig. 5).

To estimate the relationship between our dummy variable and topics, we ran estimateEffect function with a random number of 840223 as the seed. The results are shown in Table 1.

[3] Eduskunta 2021; 2023
[4] Arora et al. 2013
[5] Roberts, Stewart, and Tingley 2019

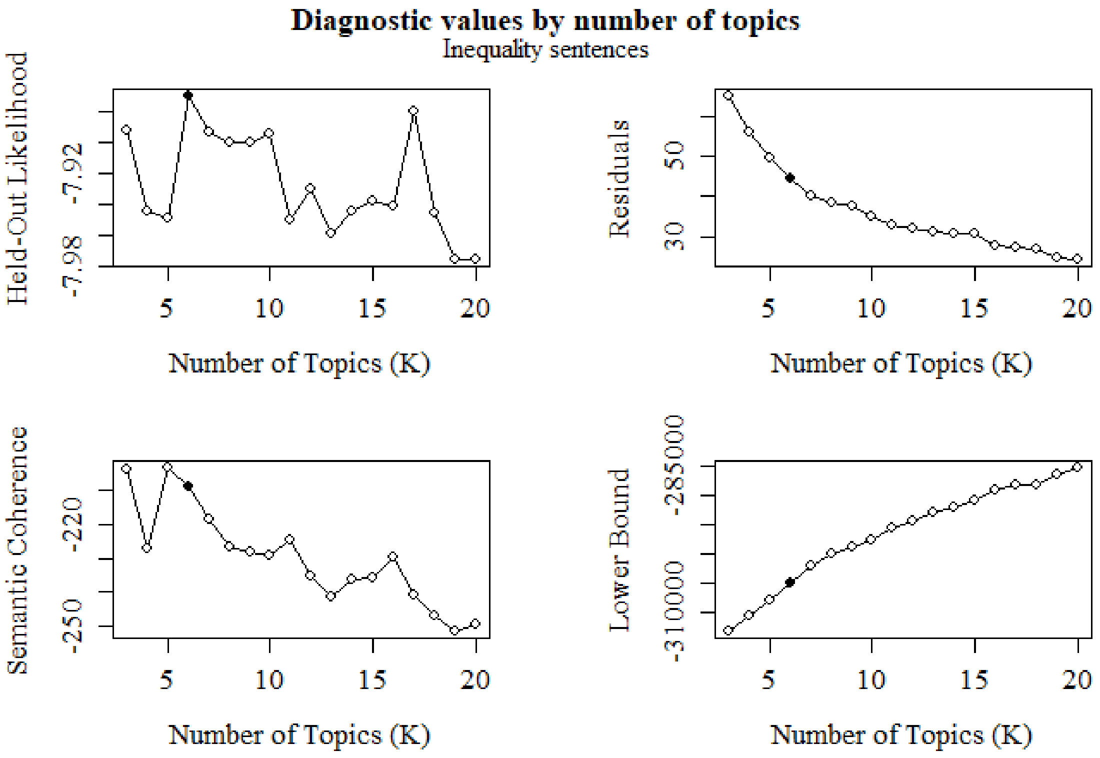

Figure 4 Diagnostic values by number of topics. SearchK results for inequality sentences. Model with 6 topics is marked in black.

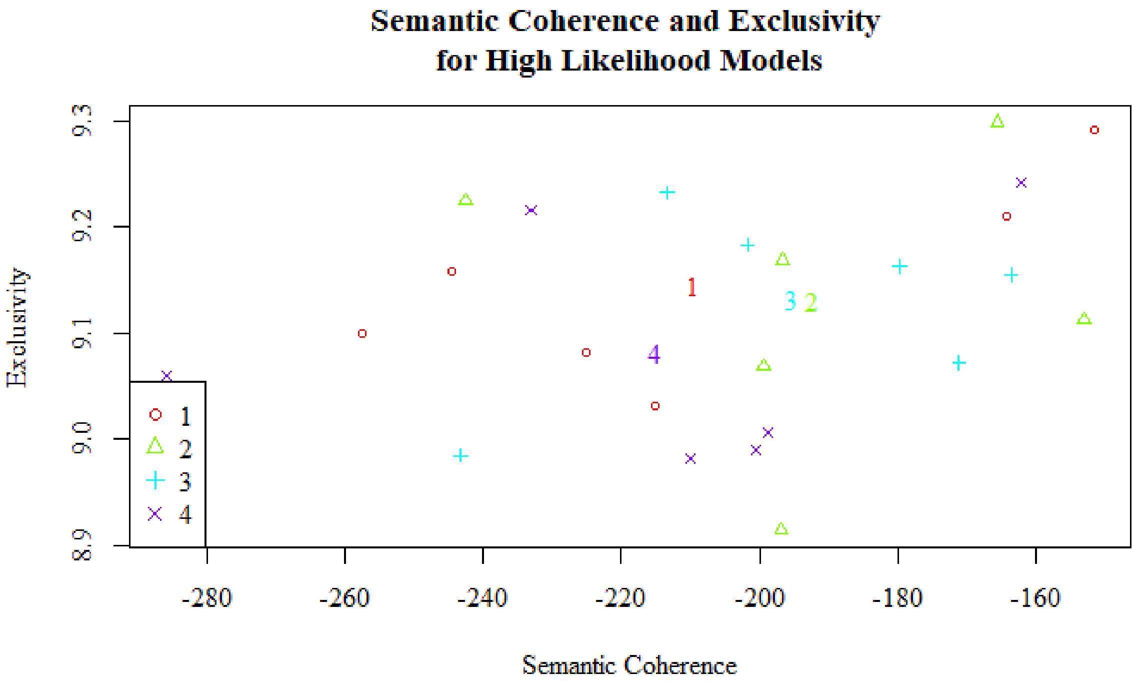

Figure 5 SelectModel results for inequality sentences.

Table 1 EstimateEffect results

Topic	Coefficients	Estimate	Standard error	t value	Pr(>|t|)
1	Other	0.194148	0.007178	27.048	< 2e-16 ***
	Left	0.030224	0.009374	3.224	0.00127 **
2	Other	0.127953	0.006667	19.192	< 2e-16 ***
	Left	0.032095	0.009316	3.445	0.000577 ***
3	Other	0.207600	0.007884	26.331	< 2e-16 ***
	Left	-0.043491	0.010381	-4.189	2.86e-05 ***
4	Other	0.149838	0.006572	22.80	<2e-16 ***
	Left	0.018012	0.008703	2.07	0.0386 *
5	Other	0.145105	0.005975	24.28	<2e-16 ***
	Left	-0.007988	0.008068	-0.99	0.322
6	Other	0.17526	0.00665	26.357	< 2e-16 ***
	Left	-0.02866	0.00912	-3.143	0.00169 **

Online Supplement References

Sources

Arora, Sanjeev, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. “A Practical Algorithm for Topic Modeling with Provable Guarantees.” In *Proceedings of the 30th International Conference on Machine Learning*, 280–88. PMLR. https://proceedings.mlr.press/v28/arora13.html.

Eduskunta. 2021. “Kansanedustajien Puheenvuorot Täysistunnoissa 2015–2018 (Xlsx).” https://downloads.ctfassets.net/ihup72fnxs9n/6agOn1YXumK15vBJ4fxSbJ/7074982002f38cf6dd04801afb536879/Kansanedustajien-puheenvuorot-taysistunnoissa-2015-2018.xlsx.

———. 2023. “Kansanedustajien Puheenvuorot Täysistunnoissa 2019–2022 (Xlsx).” https://downloads.ctfassets.net/ihup72fnxs9n/1q9IyxAnMQpxp2V2hFQvZz/da3d0658b26ed99b50ec02852a409f11/Kansanedustajien_puheenvuorot_2019-2022.xlsx.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2024. “Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland.” *Semantic Web*. https://www.semantic-web-journal.net/system/files/swj3683.pdf.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. “Stm: An R Package for Structural Topic Models.” *Journal of Statistical Software* 91 (October):1–40. https://doi.org/10.18637/jss.v091.i02.