**Online Appendix for “How the coalition governments legislate? Drafting and negotiating bills in the executive phase of law-making”**

The final dataset for replication of analysis includes only relevant variables employed in the statistical models for the period of the two regular governments 2010–2017. Nevertheless, on request, we can provide the original dataset. The dataset consist of all the government bills introduced between years 2009–2017, including the ones initiated by the two technical governments of Jan Fisher and Jiří Rusnok (both excluded), also it consist of larger number of descriptive and text variables (e.g. name of the bill, proposer of the bill, link to the online databases, etc.). Here we provide more information on data collection and choices of model specifications as well as supporting information to provide a better picture of drafting and negotiating executive bills in the Czech Republic, and to enhance the transparency and replicability of the study. Additionally, we produced tables in R (graphical figures) and STATA 16.

**Table A**: Main regression model results from the Figure 3

|  |  |
| --- | --- |
| **Variable** | **Cabinet phase** |
| Distance from coalition compromise (H1) | -0.024 |
|  | (0.022) |
| Saliency of a bill to other coalition parties (H2) | -0.110\*\*\* |
|  | (0.023) |
| Government's Legislative Plan (H3a) | -0.026 |
|  | (0.020) |
| Ministerial phase change (H3b) | 0.176\*\*\* |
|  | (0.018) |
| Involvement of LCG (H4) | 0.295\*\*\* |
|  | (0.023) |
| Amendment | -0.149\*\*\* |
|  | (0.020) |
| Days to next parliamentary elections (N days logged) | -0.134\*\*\* |
|  | (0.019) |
| Days in the cabinet (N logged) | 0.251\*\*\* |
|  | (0.024) |
| Bill´s length (N words logged) | -0.345\*\*\* |
|  | (0.021) |
| (Constant) | 3.335\*\*\* |
|  | (0.129) |
|  |  |
|  |  |
|  |  |
| Residual deviance | 4803.000 |
| N | 427 |
| \*\*\*p < 0.01, \*\*p < 0.05, \*p < 0.1 Standardized coefficients by 2SD. Control dummy variables for government, ministerial portfolio, policy dimensions. See R script.  |
|  |

Exploratory analysis showed that the average of the changes to bills in the cabinet phase between the Nečas and the Sobotka governments was not statistically significant (t-test). Likewise, the average of the changes to bills does not differ across initiating ministries of the two coalition governments, which is mirrored in the zero value of the intra-class correlation coefficient obtained from the null model when specified with a random intercept (or simple ANOVA). This suggests that there is no need to model data as hierarchically clustered by ministries. Still, we use control dummy variables for ministerial portfolios to avoid confoundedness, as some predictors may be correlated with specific portfolios. The policy field of bills are also included in the model as control variables because some may be inherently associated with higher changes and also with selected predictors.

**Table B**: Descriptive statistics

*Raw data as collected (selected variables)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable |  Obs |  Mean |  Std.Dev. |  Min |  Max |
| Government's Legislative Plan (H3a) | 427 | .513 | .5 | 0 | 1 |
| Ministerial phase change (H3b) | 427 | 35.55 | 19.075 | 0 | 97.636 |
| Amendment (control) | 427 | .775 | .418 | 0 | 1 |
| Involvement of LGC (N meetings) (H4) | 427 | .52 | .706 | 0 | 3 |
| Days to election (N days) (control) | 427 | 938.518 | 305.543 | 325 | 1448 |
| Days in cabinet (N) (control) | 427 | 132.609 | 85.958 | 9 | 551 |
| Draft Length (N words in thousand) (control) | 427 | 7.581 | 13.383 | .056 | 167.708 |

*Transformed and standardized by two standard deviations as entering in the main regression model*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable |  Obs |  Mean |  Std.Dev. |  Min |  Max |
| Distance from coalition compromise (H1) | 427 | 0 | .5 | -.655 | 1.295 |
| Saliency of a bill to other coalition parties (H2) | 427 | 0 | .5 | -1.266 | 1.365 |
| Government's Legislative Plan (H3a) | 427 | 0 | .5 | -.512 | .487 |
| Ministerial phase change (H3b) | 427 | 0 | .5 | -.932 | 1.627 |
| Involvement of LGC (N meetings) (H4) | 427 | 0 | .5 | -.368 | 1.755 |
| Amendment (control) | 427 | 0 | .5 | -.927 | .269 |
| Days to election (N days logged) (control) | 427 | 0 | .5 | -1.32 | .659 |
| Days in cabinet (N logged) (control) | 427 | 0 | .5 | -1.861 | 1.217 |
| Draft Length (N words logged) (control) | 427 | 0 | .5 | -1.505 | 1.449 |

**Table C**: Alternative models – unstandardized coefficients

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Poisson** | **Negative****Binomial** | **OLS** | **OLS (logged DV)** | **Binomial** | **Logit** | **Zero-inflated beta** |
| Distance from coalition compromise (H1) | 0.005 | -0.004 | -0.409 | -0.062 | -0.007 | 0.367 | -0.026 |
|  | (0.021) | (0.078) | (2.523) | (0.093) | (0.126) | (0.327) | (0.109) |
| Saliency of a bill to other coalition parties (H2) | -0.185\*\*\* | -0.160 | -4.686 | -0.123 | -0.254\* | -0.576 | -0.155 |
|  | (0.029) | (0.107) | (3.416) | (0.126) | (0.148) | (0.459) | (0.147) |
| Government's Legislative Plan (H3a) | -0.033\* | -0.040 | -1.055 | -0.056 | -0.053 | -0.123 | -0.049 |
|  | (0.020) | (0.070) | (2.275) | (0.084) | (0.112) | (0.324) | (0.099) |
| Ministerial phase change (H3b) | 0.184\*\*\* | 0.190\*\*\* | 5.704\*\*\* | 0.211\*\*\* | 0.285\*\* | 0.429 | 0.245\*\* |
|  | (0.018) | (0.065) | (2.201) | (0.081) | (0.132) | (0.289) | (0.096) |
| Involvement of LCG (H4) | 0.262\*\*\* | 0.260\*\*\* | 8.021\*\*\* | 0.189\* | 0.393\*\*\* | 1.012\*\*\* | 0.303\*\*\* |
|  | (0.022) | (0.083) | (2.642) | (0.097) | (0.122) | (0.374) | (0.114) |
| Amendment | -0.169\*\*\* | -0.161\*\* | -5.085\*\* | -0.102 | -0.257\*\* | -0.355 | -0.198\*\* |
|  | (0.019) | (0.072) | (2.300) | (0.085) | (0.115) | (0.314) | (0.098) |
| Days to next parliamentary elections (N days logged) | -0.116\*\*\* | -0.120\* | -2.891 | -0.107 | -0.159 | -0.236 | -0.138 |
|  | (0.019) | (0.069) | (2.193) | (0.081) | (0.116) | (0.302) | (0.095) |
| Days in the cabinet (N logged) | 0.267\*\*\* | 0.250\*\*\* | 7.621\*\*\* | 0.329\*\*\* | 0.393\*\*\* | 0.405 | 0.367\*\*\* |
|  | (0.023) | (0.084) | (2.599) | (0.096) | (0.134) | (0.367) | (0.113) |
| Bill´s length (N words logged) | -0.359\*\*\* | -0.368\*\*\* | -10.747\*\*\* | -0.353\*\*\* | -0.539\*\*\* | -1.166\*\*\* | -0.435\*\*\* |
|  | (0.021) | (0.076) | (2.390) | (0.088) | (0.123) | (0.354) | (0.104) |
| (Constant) | 4.159\*\*\* | 4.103\*\*\* | 51.658\*\*\* | 3.818\*\*\* | 0.247 | 0.425 | 0.082 |
|  | (0.079) | (0.297) | (9.658) | (0.356) | (0.426) | (1.253) | (0.417) |
| Log likelihood | -3680.424 | -1795.300 | -1899.122 | -489.621 | -185.150 | -157.142 | 129.097 |
| Pseudo-R2/R2 if applicable | 0.161 | 0.026 | 0.188 | 0.171 |  | 0.191 |   |
| N | 427.000 | 427.000 | 427.000 | 427.000 | 427.000 | 395.000 | 427.000 |

\*\*\*p < 0.01, \*\*p < 0.05, \*p < 0.1

We also present alternative model specifications. The first Poisson model is the final model presented in the paper. Other models are listed as follows: Negative binomial regression, standard OLS, OLS with logged dependent variable, binomial regression for proportion, logit regression when 0<40% change, and 1≥ 40% change, and zero inflated beta regression models (DV as proportion). In the other model specification, the political variable has no statistical significant effect, however, the direction of the coefficient remains the same. Table C was produced in STATA do file (provided) as not all the models can be produced in R.

**Comparison of ministerial phase and cabinet phase (Figure 3 in the manuscript)**



In the ministerial phase, the dependent variable is the change between the initial version and the version after the IMC that is forwarded to the cabinet. Standardized coefficients.

**Table D**: Governments in the Czech Republic

|  |  |  |  |
| --- | --- | --- | --- |
| **Government (Prime minister)** | **Type** | **Period** | **Party composition** |
| Jan Fisher\* | technical | 08.05.2009 – 13.07.2010 | - |
| Petr Nečas (ODS) | majority coalition | 13.07.2010 – 10.07.2013 | ODS, TOP09, VV |
| Jiří Rusnok\* | technical | 10.07.2013 – 29.01.2014 | - |
| Bohuslav Sobotka (ČSSD) | majority coalition | 29.01.2014 – 13.12.2017 | ČSSD, ANO 2011, KDU-ČSL |

*\*Excluded technical governments*

We have excluded Jan Fisher and Jiří Rusnok governments due to the data limitation on political variables. The governments were installed after government crisis by the presidents Václav Klaus and Miloš Zeman, respectively. The Nečas’s government was a center-right government consisting of typical right-wing conservative party ODS of the prime minister and coalition partner conservative right TOP09 and new populist movement VV (Public Affairs). The Sobotka’s government was formed of traditional Social Democrats (ČSSD), new populist movement ANO 2011 of Andrej Babiš (current PM) and traditional Christian Democrats (KDU-ČSL, one of the oldest parties in the Czech Republic). For more details on the Czech party systems, see Haugton & Deegan-Krause (2015) and Havlík & Voda (2018).

**Table E**: Portfolio allocations

Table E presents the portfolio allocation. During both governments, there were changes of ministers; however, the allocation to the parties remained the same.

|  |  |  |  |
| --- | --- | --- | --- |
| **Portfolio** | **Petr Nečas (ODS)** | **Bohuslav Sobotka (ČSSD)** |  |
| Prime minister | ODS | ČSSD |  |
| Industry (MPO)\* | ODS | ČSSD |  |
| Justice (MSP) | ODS | ANO2011 |  |
| Interior (MV) | VV | ČSSD |  |
| Agriculture MZE  | ODS | KDU-ČSL |  |
| Transportation (MD) | VV | ANO2011 |  |
| Finance (MF) | TOP 09 | ANO2011 |  |
| Culture (MK) | TOP 09 | KDU-ČSL |  |
| Defence (MO) | ODS | ANO2011 |  |
| Labour and social affairs (MPSV) | TOP 09 | ČSSD |  |
| Regional Development (MMR) | VV | ANO2011 |  |
| Education (MŠMT)  | VV | ČSSD |  |
| Foreign Affairs (MZV) | TOP 09 | ČSSD |  |
| Health (MZD)  | TOP 09 | ČSSD |  |
| Environment (MŽP) | ODS | ANO2011 |  |

*\*the Czech abbreviations for portfolios in brackets.*

**Table F:** Seats allocations and portfolios

|  |
| --- |
| **Petr Nečas (ODS)** |
| **Party** | **Seats in parliament** | **(%)\*** | **Portfolios** | **(%)\*\*** |
| ODS | 53 | 26.5 | 6 | 40,0 |
| TOP09 | 41 | 20,5 | 5 | 33.3 |
| VV | 24 | 12 | 4 | 26.7 |
|  |  |  |  |  |
| **Bohuslav Sobotka (ČSSD)** |
| **Party** | **Seats in parliament** | **(%)** | **Portfolios** | **(%)** |
| ČSSD | 50 | 26.5 | 7 | 46.7 |
| ANO 2011 | 47 | 20.5 | 6 | 40.0 |
| KDU-ČSL | 14 | 12 | 2 | 13.3 |

*\*Divided by the total seats in the parliament (200)*

*\*\*Divided by the total number of portfolios (15) including 14 ministries and 1 PM.*

**Table G**: Correlation matrix (Spearman)

 

*Note: Using “corrplot” package in R.*

We use Spearman rank coefficient because most of the variable are over-dispersed or do not follow normal distribution. VIF factor is low in the models (under 2), but some variables are highly correlated with control dummies (ministries, policy areas). In the R script, you find other regression diagnostic that is specific to Poisson models.

**Supporting Material H:** Coding justification and data sources in detail

*Dataset*

The Czech case provide an opportunity to study executive phase of drafting bill due to the publicly partly-open online system of tracking government bill – VeKLEP, available at: <https://apps.odok.cz/veklep>. Data from the parliamentary phase can be obtain from the website of the Czech Chamber of Deputies, available at: <https://www.psp.cz/sqw/sntisk.sqw>. The raw dataset was coded by instructed students and controlled by the authors. We have extracted all the bills from 2009 until the end of the Sobotka government 2017. We started data collection and between the years 2018–2019 (N=828), so we were not able to code most recent two Babiš’s minority governments (13.12. 2017 till 27.6. 2018; 27.6. till now). The VeKLEP provides information on the name of the bill, policy areas it regulates, date of initiation, dates of interministerial consultation period, dates of the Legislative Council of the Government (LGC) meetings, date of final government approval, reference to EU regulations and directives if applies, type of a bill (amendment vs. self-standing bill), presence in the Government´s Legislative Plan. Besides, most importantly, the drafts can be download in three stages. The initial, interim (after IMC) and final. Besides, the document that list all the comments from relevant actors to particular articles and section of a draft bill is provided as well. From this document, we have extracted number of comments, their initiators, and their status (accepted, declined, explained and disputed).

We have excluded bills that were drafted by central agencies and not by the ministries. As total, there were 650 initiated ministerial drafts between 2010–2017 for Nečas and Sobotka governments. Unfortunately, there were errors for two cases as the documents were not available in some of the executive phase. Overall, 154 bills were terminated in the IMC, so we are unable to compute any change as we have just initial version. Additionally, we have excluded also 4 budgetary bills (technical drafts). This gives 492 drafts bill for the first ministerial phase. In the second cabinet phase, 65 drafts were terminated. We decided to present models only for the bills that completely passed the executive phase (N=427). The question of bills terminations is not the subject of the article. However, the model for complete 492 cases in the first ministerial phase is included below.

*Dependent Variable*

The dependent variable is the change of the number of words between the pair of text. 0 means no change, 100 means the draft was completely changed. We have programmed a function in R software (see at the end of the appendix) that is part of this appendix. The script was developed with cooperation with the Department of Linguistics (hidden for review). It can be easily used for comparing any two text that are converted into plain txt file with the same encoding of characters (ideally utf-8). A Word document can be easily saved as txt file. Pdf files, if provided by VeKLEP, were transformed to txt files by conversion online. There are several options how to compute the change of a text based on the methods analysing texts in a vectorised manner. First, the choice of the number of grams (uni-grams, bi-grams, etc.) and distance between them (e.g. Jaccards distance). We have opted for the bi-gram algorithm as proposed and tested by Gava et al. (2020). Much simpler operationalization could have been used. For example, the absolute difference in the number of words standardized by the total number of words of the pair text. However, this method does not count the difference between particular words. Interestingly, such simple method correlates quite substantially with our index (0.78).

*Independent variables*

 Table below show variable type. Nevertheless, more complex variables deserve more detailed explanation.

*Table CH1: List of variables*

|  |  |
| --- | --- |
| Variable | Type |
| Drafter: Main ruling party (H1) | dummy |
| Disagreement comments - coalition partner (N) (H2) | count |
| Saliency of a bill (H3) | index |
| Distance from coalition compromise (H4) | index |
| Distance from legislative median (H5) | index |
| Ministerial phase change (control) | index |
| Complexity (N of policy areas) (control)  | count |
| Amendment (control) | dummy |
| Number of LGC meetings (control) | count |
| Implementation of EU legislation (control) | dummy |
| Days to election (N days) (control) | count |
| Days in cabinet (N) (control) | count |
| Draft Length (N words in thousand) (control) | count |
| Government's Legislative Plan (control) | dummy |

The number of comments was collected for each ministry. Then we summed the “important comments” (excluding the “recommended comments” category) initiated by the ministries that were held by a coalition partner, or by the same party that drafted the commented bill. We did not differentiate between accepted, explained, rejected and disputed comments. The reason is, surprisingly, that usually bills that attracted most of the comments had also high number of both accepted and rejected comments. The categories are highly correlated. Regarding the disputed comments, which are qualitatively interesting, there were just couple of bills that featured disputed comments. We have decided to transform the number of comments using natural logarithm. This makes a scale from 0 (no comments) to a maximum 5.89 of logged number of comments. Alternatively, we could use the share ratio of the number of comments from the same party/coalition partner on the total number of comments. Such measure overestimates ratio value when the number of comments is small. For example, if a draft bill gets one comment from a coalition partner, and one from the ministry holding the same party, the ratio would be 50%. In extreme case, a one comment may result in 100% ration of comments from the coalition partner, thus overestimate the coalition control. Additionally, we have standardized the number of comments by the length of the bill (number of words on average per a comment) as lengthy bill naturally attracts more comments. This transformation gives the similar results and do not change the fact that bills that are commented more from the coalition partner are changed more in all phases.

 The policy position and saliency of the bill are the variables that are linked in terms of coding and operationalization. Partially, a ministry issues bills that belongs to specific policy area or dimension. So, there is an overlap in ministerial portfolios and policy dimension. Technically, there are two possible datasets from which we can construct the variables. Each has different methodology. While the Comparative Manifesto Project (CMP) relies on human coding of party manifestos, the Chappell Hill survey (CHES) relies on evaluation of national experts (see Benoit & Laver 2006). We decided to follow current literature that employs CHES to code policy position and saliency toward bill that regulates it (Goetz & Zubek 2017; Martin & Vanberg 2014: 985; Zubek & Klüver 2015). CHES dataset is similar to data from an extensive study of party policy positions by Benoit & Laver (2006) which we could not use because it is the outdated. We have used CHES 2010 and CHES 2014 datasets (available at: <https://www.chesdata.eu/our-surveys>). However, in 2014 there are no saliency measures of some of the dimensions, so we have used the saliency from the year 2010. The first step was to assign each bill to a specific CHES category. We have used *civlib\_laworder, deregulation, environment, redistribution, regions, urban\_rural, sociallifestyle, spendvtax*. Each bill was assigned to one of the policy dimensions. This was done by human coders (students of Law) who highlighted difficult and problematic cases that were double-checked by the authors. However, not all the bills could be assigned. Technical norms and other bills were labeled as a new category “*others*”. In order not to lose cases, we imputed the mean position and mean value of saliency. While this influences the variable *Saliency of a bill* (H3), the CHES party positions were not used independently but were part of an index variable *Distance from coalition compromise* (H4) as computed by Martin & Vanberg (2014: 985). The other category has virtually value of 0 as there is no measurable disagreement between the parties. The distance was computed as follows. First, we computed the coalition seat weighted saliency weighted mean position. From this value, we have subtracted the seat weighted and saliency weighted position of a party holding the drafting ministry. The distance is in an absolute value of a difference between party position and coalition mean position. As a result, we did not need to transform the scales of the position (which would be impossible because there are more than one left-right dimensions), the index simply shows the disagreement on the particular policy issue between coalition partners. Thus, the variable serves as an imperfect proxy variable for party positions to a drafted bill.

 *Figure H1: CHES party positions*

 *Figure H2: CHES issue saliency*

*Table H4: CHES dimensions frequency*

|  |  |  |
| --- | --- | --- |
| CHES dimensions | Frequency | Percent |
| civlib\_laworder | 114 | 26,7 |
| deregulation | 115 | 26,9 |
| environment | 36 | 8,4 |
| other | 36 | 8,4 |
| redistribution | 48 | 11,2 |
| regions | 12 | 2,8 |
| sociallifestyle | 10 | 2,3 |
| spendvtax | 36 | 8,4 |
| urban\_rural | 20 | 4,7 |
| Total N | 427 | 100 |

*Table H5: Bills by drafter (ministry)*

|  |  |  |
| --- | --- | --- |
| Ministry | Frequency | Percent |
| Finance (MF) | 95 | 22,2 |
| Justice (MSP) | 57 | 13,3 |
| Interior (MV) | 45 | 10,5 |
| Labour and social affairs (MPSV) | 41 | 9,6 |
| Industry (MPO) | 37 | 8,7 |
| Health (MZD)  | 33 | 7,7 |
| Agriculture MZE  | 29 | 6,8 |
| Environment (MŽP) | 25 | 5,9 |
| Transportation (MD) | 16 | 3,7 |
| Regional Development (MMR) | 15 | 3,5 |
| Education (MŠMT)  | 13 | 3,0 |
| Culture (MK) | 11 | 2,6 |
| Defence (MO) | 10 | 2,3 |
| Total | 427 | 100 |

**References:**

Benoit, K., & Laver, M. (2006). *Party Policy in Modern Democracies*. London: Routledge.

Gava, R., Jaquet, J.M. & Sciariny, P. (2020), Legislating or rubber‐stamping? Assessing parliament's influence on law‐making with text reuse. *European Journal of Political Research*. Doi:[10.1111/1475-6765.12395](https://doi.org/10.1111/1475-6765.12395)

Goetz, K.H. & Zubek, R. (2007). Government, Parliament and Law-making in Poland. The Journal of Legislative Studies 13(4): 517–538.

Haughton, T., & Deegan-Krause, K. (2015). Hurricane Season. *East European Politics and Societies: And Cultures*, 29(1), 61–80. <https://doi.org/10.1177/0888325414566072>

Havlík, V., & Voda, P. 2018. Cleavages, Protest or Voting for Hope? the Rise of Centrist Populist Parties in the Czech Republic. *Swiss Political Science Review*, 24, 161–186. doi:10.1111/spsr.2018.24.issue-2.

Martin, L.W. & Vanberg, G. (2014), Parties and Policymaking in Multiparty Governments: The Legislative Median, Ministerial Autonomy, and the Coalition Compromise. *American Journal of Political Science* 58(4): 979–996

Zubek, R. & Klüver, H. (2015). Legislative pledges and coalition government. *Party Politics* 21(4): 603–6014.

**Supporting Material CH: R script for comparing change of the bills text**

*########################################################*

*# Script for computing differences between two texts:*

*# How do coalition governments legislate?*

*# Drafting and negotiating bills in the executive phase of law-making*

*########################################################*

*# Date 30.7.2020*

*# General comment:*

*############################*

*# the first part of the script is the function used in computing the diff change*

*# the dependent variable ranging from 0 (no change) to 100 (complete change)*

*# the diff algorithm (CompareTexts) can be used for any two comparable versions of text*

*# in the script the first text is A.txt and second B.txt - so just rewrite these or the text*

*# documents in your file.*

*########################################################*

*# Comparison Definition*

*########################################################*

*## Jacard as in Gava 2020, pp. 10 (binary)*

*diffScoreJaccardLogic <- function(textA, textB){*

 *textA <- pmin(textA, 1)*

 *textB <- pmin(textB, 1)*

 *M11 <- sum(textA & textB)*

 *M10 <- sum(textA & !textB)*

 *M01 <- sum(!textA & textB)*

 *1 - ( M11 / (M11 + M10 + M01))*

*}*

*## Jaccard as in Gava 2020, pp. 10 (binary; faster implementation, text vectorization is not needed)*

*diffScoreJaccardLogicSet <- function(tokens\_a, tokens\_b){*

 *setSize <- function(x) length( unique(x) )*

 *1 - ( setSize( intersect(tokens\_a, tokens\_b) ) / setSize( union(tokens\_a, tokens\_b) ) )*

*}*

*# Text vectorization:*

*# a = squirrel ate moose*

*# b = squirrel ate pancake*

*# v = {squirrel, ate, moose, pancake}*

*# vec\_a = 1, 1, 1, 0*

*# vec\_b = 1, 1, 0, 1*

*getWordFreq <- function(word, freqTable){*

 *if (word %in% names(freqTable)){*

 *return( as.numeric(freqTable[word]) )*

 *}*

 *return(0)*

*}*

*# Vectorize texts*

*# a <- c("squirrel", "ate", "moose")*

*# b <- c("squirrel", "ate", "pancake")*

*# textVectors <- makeTextVecs(a, b)*

*# vec\_a <- textVectors[[1]]*

*# vec\_b <- textVectors[[2]]*

*makeTextVecs <- function(tokensA, tokensB){*

 *v <- unique( c(tokensA, tokensB) )*

 *vec\_a <- rep(0, length(v))*

 *vec\_b <- rep(0, length(v))*

 *freqs\_a <- table(tokensA)*

 *freqs\_b <- table(tokensB)*

 *for(i in 1:length(v)){*

 *type <- v[i]*

 *vec\_a[i] <- getWordFreq(type, freqs\_a)*

 *vec\_b[i] <- getWordFreq(type, freqs\_b)*

 *}*

 *list( vec\_a, vec\_b )*

*}*

*# reads content of file as plain text*

*GetFileContent <- function(fileName){*

 *readedText <- readLines(fileName, encoding="UTF-8")*

 *readedText <- do.call(paste, c(as.list(readedText), sep=" "))*

 *return(readedText)*

*}*

*# Basic plain text tokenizer by regular expression split mask*

*# Eg.: TokenizeText("Hey, this is an example") returns "hey", "this", "is", "an", "example"*

*# TokenizeText("Hey, this is an example", regexPattern="\\b\\w{3}\\b", regexIsMask=TRUE) returns "hey"*

*library(stringr)*

*library(stringi)*

*TokenizeText <- function(text, regexPattern="\\W+", regexIsMask=FALSE, convertToLowerCase=TRUE){*

 *if (regexIsMask){*

 *tokens = str\_extract\_all(text, regexPattern)*

 *}else{*

 *tokens = stri\_split\_regex(text, regexPattern)*

 *}*

 *tokens = unlist(tokens)*

 *if (convertToLowerCase){*

 *tokens = tolower(tokens)*

 *}*

 *tokens = tokens[ tokens != "" ]*

 *return(tokens)*

*}*

*# Makes n-grams from tokens.*

*# If Glue is a string, vector of n-gram strings are returned, otherwise a matrix*

*# Eg.: MakeNGrams( c("a", "b", "c", "d", "e", "f", "g"), n=3, glue=" ")*

*MakeNGrams <- function(tokens, n=2, glue="->"){*

 *m <- t( sapply( 1:(length(tokens)-n+1), function(i) tokens[i:(i+n-1)] ) )*

 *if (is.character(glue)){*

 *m <- apply(m, 1, function(r) paste(r, collapse=glue))*

 *}*

 *return( m )*

*}*

*CompareTexts <- function(fileNameA, fileNameB, ngrams){*

 *plain\_text\_a <- GetFileContent(fileNameA)*

 *plain\_text\_b <- GetFileContent(fileNameB)*

 *tokens\_a <- TokenizeText(plain\_text\_a, "\\p{L}+", regexIsMask=T) ## only letter-strings*

 *tokens\_b <- TokenizeText(plain\_text\_b, "\\p{L}+", regexIsMask=T) ## only letter-strings*

 *if (ngrams >= 2){*

 *tokens\_a <- MakeNGrams(tokens\_a, ngrams)*

 *tokens\_b <- MakeNGrams(tokens\_b, ngrams)*

 *}*

 *vectors <- makeTextVecs(tokens\_a, tokens\_b)*

 *list( "jaccard\_logic" = diffScoreJaccardLogic(vectors[[1]], vectors[[2]]) )*

*}*

*########################################################*

*## READ ALL PAIRS AND COMPARE*

*########################################################*

*#setwd() # to your source file location*

*#CompareTexts("PRED\_A.txt", "PO\_B.txt")*

*#CompareTexts(fileNameA="D:/Temp/A.txt", fileNameB="D:/Temp/B.txt", ngrams=2)*

*folders <- list.dirs(path = "E:/data/zákony", full.names = TRUE, recursive = FALSE)*

*filesStart <- paste0(folders, "/dosenatu.txt")*

*filesFinal <- paste0(folders, "/konecnytext.txt")*

*files <- cbind(filesStart, filesFinal)*

*resultsAll <- data.frame( matrix(NA, length(filesStart), 2))*

*#files <- files[1:10]*

*for(i in 1:nrow(files)){*

 *cat(i, "/", length(filesStart), " -- Working on: ", filesStart[i], " pair ...\n")*

 *fileA <- files[i, 1]*

 *fileB <- files[i, 2]*

 *if (all( file.exists(files[i, ]) )){*

 *unigramResults <- CompareTexts(fileNameA=fileA, fileNameB=fileB, ngrams=1)*

 *bigramResults <- CompareTexts(fileNameA=fileA, fileNameB=fileB, ngrams=2)*

 *resultsAll[i, ] <- c( unlist( unigramResults ) ,*

 *unlist( bigramResults ) )*

 *}else{*

 *cat("\n[!] Could not find pair: ", fileA, " and ", fileB, "\n")*

 *}*

*}*

*colnames(resultsAll) <- c( names(unigramResults), paste0("2-gram ", names(bigramResults) ))*

*rownames(resultsAll) <- folders*

*## EVAL*

*boxplot(resultsAll)*

*########################################################*

*# End of the script*