Kelih, E. et al. (eds.), Issues in Quantitative Linguistics, Vol. 2, 75-87

Frequency structure of New Year's presidential speeches in Czech. The authorship analysis

Radek Čech

1. Introduction

New Year's presidential speeches represent a very specific genre. They mix both political and festive aspects and, contrary to common political speeches, they usually do not have persuasive character. The New Year's speeches can be viewed as a very homogeneous genre because of their a) aim, b) form, and c) tradition. As for the aim (a), the goal of the speech is usually to summarize main events of the past year, mention perspectives of a near future, and express best wishes to the inhabitants of the state. The speeches have a very steady form (b), they are prepared in advance and read by the president. The tradition (c) of this kind of speeches is very long in Czech Republic (former Czechoslovakia) – it has started since 1949 and continued up to now (except of 1993 when no president was in office). Obviously, the strong homogeneity of the genre facilitates the authorship analysis because a host of boundary conditions is eliminated.

One can expect two contradictory "powers" which should have the greatest impact on the frequency structure of presidential speeches. On the one hand, the official and ceremonial character of this event should lead to the high uniformity of texts and, consequently, to the high similarity of frequency structures. On the other hand, presidents are usually persons with a strong individuality; as politicians, they have to be able to express their uniqueness and specificity, so, one can expect great differences among them.

For the measurement of frequency characteristics two methods were used: 1) the lambda measurement proposed by Popescu et al. (2010, 2011) and 2) the vocabulary richness index *R1* (Popescu et al. 2009); both methods are presented in the next section.

2. Methodology

The lambda-indicator expresses one aspect of frequency structure of text. In short, it takes into account both the frequency of words and the relationships among individual frequencies. It can be viewed as an indicator of frequency *technique* used by an author. One of the biggest advantage of this measurement is

the independence of lambda-indicator on the text length (cf. Popescu et al. 2011, pp. 10-12). It is defined as

$$\Lambda = \frac{L(\log_{10} N)}{N} \tag{1}$$

where L is the arc length between the ranked frequencies defined as

$$L = \sum_{r=1}^{V-1} [(f_r - f_{r+1})^2 + 1]^{1/2} \tag{2}$$

where N is the text size (in tokens), f_r is the frequency at rank r and V is the highest rank. The variance of Λ is a complex formula and it is presented in detail in Popescu et al. (2010, 2011).

The vocabulary richness index R_1 is defined as

$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right) \tag{3}$$

where h is the h-point (Popescu, Altmann 2006) and $F(h)$ is the cumulative relative frequency up to the h-point. H-point is defined as

$$h = \begin{cases} r, & \text{if there is an } r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & \text{if there is no } r = f(r) \end{cases} \tag{4}$$

i.e. that point at which $r = f(r)$, or, if there is no such point, it is computed by means of the second part of formula (4).

The variance of R_1 are computed as follows

$$Var(R_1) = \frac{F(h)\left[1 - F(h)\right]}{N}. \tag{5}$$

Let us illustrate the procedure of comparison of authors in the case of three presidents, namely Gottwald, Novotný, and Klaus, by using lambda-indicator. First, from Table 7 in Appendix both the mean lambdas and variances of means are computed, see Table 1.

Table 1.
Mean lambdas and variances of lambdas of New Year's presidential speeches. Presidents are ordered according to the magnitude of mean lambda.

President	**year**	**n**	**mean(Λ)**	**s^2(Λ)**	**s^2 (Λ)/n**
Klaus	2004-2011	8	1.9292	0.003834	0.000479
Husák	1975-1989	15	1.9211	0.008357	0.000557
Havel	1990-2003	13	1.8882	0.004031	0.000310
Zápotocký	1954-1957	4	1.8818	0.001945	0.000486
Svoboda	1968-1974	6	1.8769	0.005683	0.000026
Gottwald	1979-1953	5	1.8714	0.005268	0.001054
Novotný	1958-1968	11	1.7564	0.004349	0.000395

As can be seen in Table 1, Klaus has the highest mean lambda, while Gottwald and Novotný obtain the lowest lambda values. The first task is to observe, whether the differences of mean lambdas are significant. Because the texts of the same genre in the same language are analyzed, we use the asymptotic *u*-test

$$u = \frac{\Lambda_1 - \Lambda_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad . \tag{6}$$

Specifically, for the comparison of Gottwald and Novotný we obtain

$$u = \frac{1.8714 - 1.7564}{\sqrt{0.001054 + 0.000395}} = 3.02$$

which expresses a significant difference. So, we can state that the frequency structure, expressed by mean lambda, of Gottwlad's speeches is significantly different from those of Novotný. Analogously, if we compare Klaus and Novotný, we obtain $u = 5.85$ which is significant too, however, for Gottwald and Klaus we obtain a non-significant $u = 1.48$. The results reveal that Novotný's speeches have significantly different frequency structure in comparison with both Gottwald and Klaus, while the frequency structures of Gottwald's and Klaus' speeches express similarities.

3. The lambda structure of presidential speeches

Following the procedure presented in the previous section, we obtain the results presented in Table 2.

Table 2
Comparison of mean lambdas in New Year's presidential speeches of Czech or Czechoslovak Presidents (two-sided *u*-test). Bold values express significant differences (significance level $u \leq 1.96$).

President	Gottwald	Zápotocký	Novotný	Svoboda	Husák	Havel	Klaus
$\overline{\Lambda}$	1.8714	1.8818	1.7564	1.8769	1.9211	1.8882	1.929
$s^2(\overline{\Lambda})$	0.001054	0.000486	0.000395	0.000026	0.000557	0.000310	0.000479
Gottwald	x						
Zápotocký	0.27	x					
Novotný	**3.02**	**4.22**	x				
Svoboda	0.17	0.22	**5.87**	x			
Husák	1.24	1.22	**5.34**	1.83	x		
Havel	0.45	0.23	**4.96**	0.62	1.12	x	
Klaus	1.48	1.53	**5.85**	**2.33**	0.25	1.46	x

For the sake of better lucidity, the relationships among the presidents can be expressed graphically. Figure 1 represents a small network based on Table 2 in which two presidents are connected, if there is non-significant difference between their mean lambdas (i.e., $u \leq |1.96|$). Presidents with the same number of similarities are put at the same level – for Havel, Gottwald, Husák, and Zápotocký, each obtains five similarities, Klaus and Svoboda four, and Novotný has zero.

At a first sight, the extraordinary position of Novotný is evident – there are no similarities between Novotný's mean lambda value and any other president. Further, Klaus and Svoboda can be seen as counterparts because their frequency structures differ significantly, while they both are connected to the same other presidents. Havel, Husák, Zápotocký, and Gottwald represents the most uniform cluster of this genre with regard to lambda.

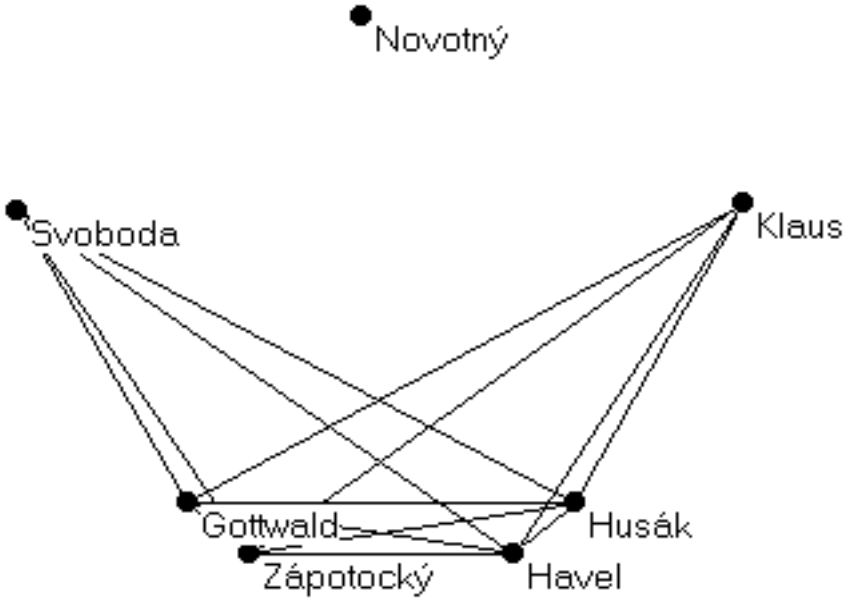

Figure 1 The network in which two presidents are connected, if there is a non-significant difference between their mean lambdas (i.e., $u \leq 1.96$)

For more detailed comparison it is possible to measure a weighted u_w differences among presidents

$$(7) \quad u_{wi} = \frac{\sum u_i}{\sqrt{k}}$$

where k is a number of comparisons. The results based on the formula (7) are presented in Table 3.

Again, the extraordinary position of Novotný is even more obvious. A comparison of weighted differences u_w and mean lambdas reveals that Novotný's position is given by the simplest frequency structure of his speeches, as is illustrated in Figure 2. The values of u_w of the other presidents are located within a relatively small interval <2.70, 5.26> which indicates high homogeneity of this genre with regard the frequency structure expressed by lambda-indicator.

Table 3
The weighted differences u_w of presidents

President	Λ	u_w
Gottwald	1.8714	2.70
Zápotocký	1.8818	3.13
Havel	1.8882	3.61
Husák	1.9211	4.49
Svoboda	1.8769	4.52
Klaus	1.929	5.26
Novotný	1.7564	11.95

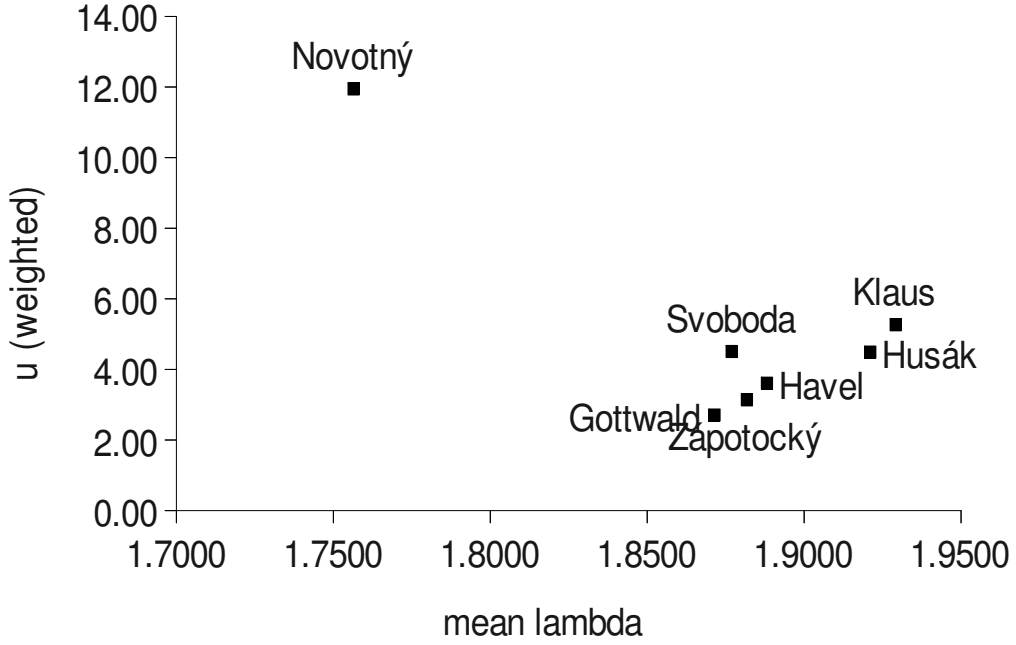

Figure 2. The weighted differences u_w of presidents

4. The vocabulary richness R_1

The computation of vocabulary richness R_1 reveals different ranking of presidents, and very small differences among them – all mean values of R_1 are in the interval <0.8546, 0.8770>, see Table 4.

Table 4
Mean vocabulary richness R_1
and variances of R_1 of New Year's presidential speeches.
Presidents are ordered according to the magnitude of mean R_1.

President	**year**	**n**	**mean(R_1)**	**$s^2(R_1)$**	**$s^2(R_1)$/n**
Svoboda	1969-1974	6	0.8770	0.000556	0.000094
Klaus	2004-2011	8	0.8727	0.000125	0.000016
Husák	1975-1989	15	0.8724	0.000207	0.000014
Havel	1990-2003	13	0.8607	0.000311	0.000024
Zápotocký	1954-1957	4	0.8555	0.000052	0.000013
Novotný	1958-1968	11	0.8552	0.000079	0.000007
Gottwald	1949-1953	5	0.8546	0.000548	0.000110

Performing *u*-tests among all presidents we obtain the results presented in Table 5 and graphically expressed differences in Figure 3.

Table 5
Comparison of vocabulary richness R_1 in New Year's presidential speeches
of Czech or Czechoslovak Presidents (two-sided *u*-test).
Bold values express significant differences (significance level $u = |1.96|$)

President	**Gottwald**	**Zápotocký**	**Novotný**	**Svoboda**	**Husák**	**Havel**	**Klaus**
$\overline{R_1}$	0.8546	0.8555	0.8552	0.877	0.8724	0.8607	0.8727
$s^2(\overline{R_1})$	0.00011	0.000013	0.000007	0.000094	0.000014	0.000024	0.000016
Gottwald	x						
Zápotocký	0.08	x					
Novotný	0.06	0.07	x				
Svoboda	1.57	**2.08**	**2.17**	x			
Husák	1.60	**3.25**	**3.75**	0.44	x		
Havel	0.53	0.85	0.99	1.50	1.90	x	
Klaus	1.61	**3.19**	**3.65**	0.41	0.05	1.90	x

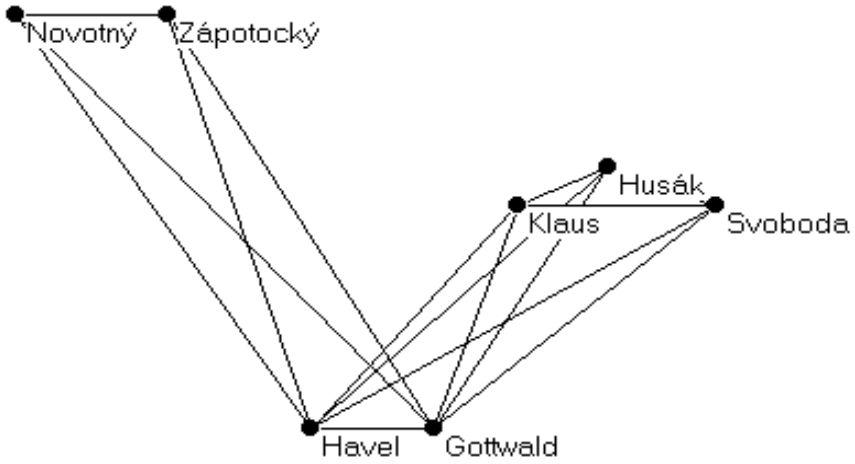

Figure 3. The network in which two presidents are connected,
if there is non-significant difference between their mean R_1 (i.e., $u \leq |1.96)$

Analogously to Figure 1, presidents with an equal number of links are put at the same level and, further, presidents who connect the same other presidents are clustered. As is seen in Figure 3, Novotný is again the president with the lowest similarities (with Zápotocký). Contrary to lambda-measurement, Klaus and Svoboda have non-significant difference of R_1, so, their speeches differ because of frequency technique they used. Finally, Havel and Gottwald are connected to all presidents which means that they are the most conformal authors with regard to vocabulary richness (the author's conformity is analysed in more detail in Section 5)

The computation of weighted u_w differences of R_1 reveals very small differences among presidents, cf. Table 6 and Figure 4.

Table 6.
The weighted differences u_w of presidents

President	R_1	u_w
Gottwald	0.8546	2.22
Havel	0.8607	3.13
Svoboda	0.8770	3.34
Zápotocký	0.8555	3.89
Novotný	0.8552	4.36
Klaus	0.8727	4.41
Husák	0.8724	4.49

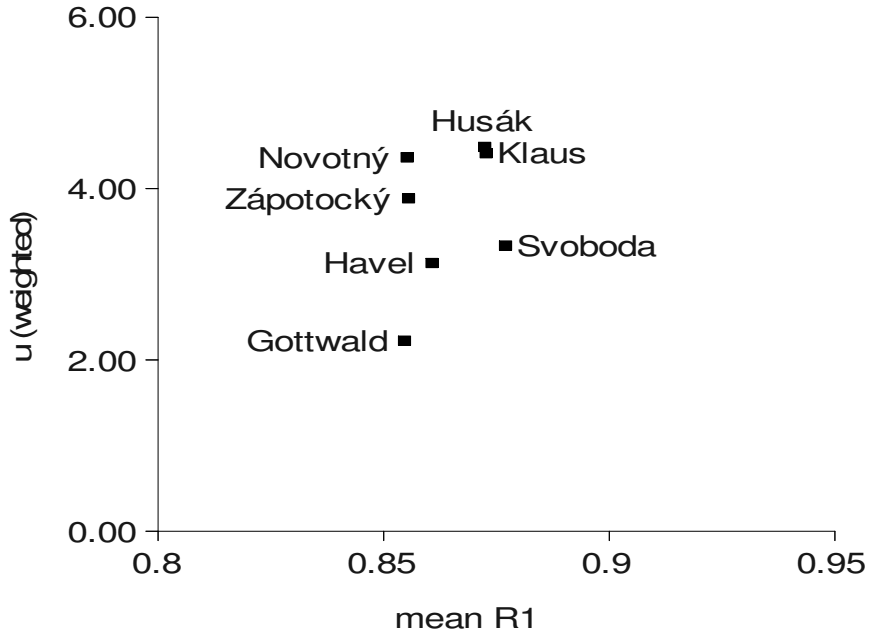

Figure 4. The weighted differences u_w of presidents

The small interval in which all weighted differences u_w lie reflects a high similarity of vocabulary richness. So, the authorship's differences of presidents are caused mainly by the different frequency techniques (i.e. expressed by lambda) which particular presidents used, as is clearly seen in Figure 5.

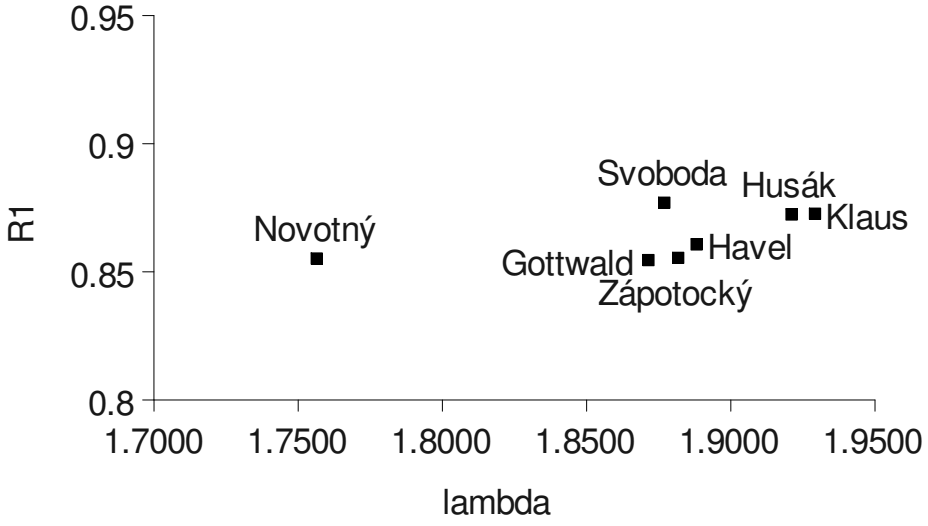

Figure 5. Comparison of lambda-values and R_1.

In Figure 5, all presidents lay in almost horizontal line which means high similarities of R_1. The authorship differences are caused by dissimilar frequency techniques which is indicated by lambda-differences.

5. The measurement of author's conformity within the genre

The properties of graphs (see Figure 1 and 3) enable to propose a *conformity indicator*. It is given by the relative degree of node representing the author, i.e. by the relative number of its links

$$(7) \quad d_{i\,(rel)} = \frac{\sum l_i}{l_{i\,(max)}}$$

where l_i is an observed number of links of the node and $l_{i(max)}$ is the maximum number of links which can the node obtain

$$(8) \quad l_{i\,(max)} = (n-1)x$$

where n is a number of nodes in the network and x is the number of particular networks used for measurement. For illustration, for Havel (based on graphs in Figure 5 and 3, i.e. $x = 2$) we obtain

$$d_{Havel\,(rel)} = \frac{11}{(7-1)2} = 0.92\,.$$

The conformity indicator of all presidents is shown in Table 6

Table 6
Conformity indicator of Czech presidents, based on lambda-measurement and vocabulary richness R_1, expressed by the relative degree of node d_{rel}. The lower d_{rel} , the more original is an author and vice versa.

Presidents	d_{rel}
Gottwald, Havel	0.92
Husák	0.75
Klaus, Svoboda, Zápotocký	0.67
Novotný	0.25

Novotný is evidently the most original author among presidents, with regard to both lambda-structure and vocabulary richness. On the other hand, Gottwald and Havel are the most conformal ones. As for Havel, this result is a little surprise – one could expect that Havel, as the world-famous dramatist, should strive for the greatest language originality. However, if Havel's frequency technique appears to be conformal also in the other genres, it should mean that frequency conformability can be taken as a characteristic feature of his language usage. Of course, the conformability or frequency technique itself has nothing to do with a content and literary quality of his texts. Contrariwise, the same frequency technique can be used for extremely different purposes, as our results clearly manifest – it is striking that the most conformal authors (i.e., Gottwald, Havel) are persons which can be viewed as political and personal counterparts: Gottwald was a professional politician, leader of communist coup, dictator, while Havel has been a writer, long-term leader of democratic opposition in communist Czechoslovakia, democrat, humanist.

6. Conclusion

The analysis of lambda-structure and vocabulary richness of presidential speeches reveals surprisingly high number of similarities among presidents. This indicates that the tendency to uniformity prevails the need to express individuality of particular persons. Moreover, both measurements do not unveil the impact of period (the speeches do not reflect the changes in sixty years of language development) or political regime (Gottwald, Zápotocký, Novotný, Svobooda, and Husák are representatives of communist totality, while Havel and Klaus represent democracy).

Acknowledgement
I thank Jaroslav David for providing me the data which have been used for the analysis. This work has been also supported by the Czech Science Foundation, grant no. P406/11/0268.

References

Popescu, I.-I., Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics 13, 23-46.*

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style and typology. *Glottotheory 3(1), 89-96*

Popescu, I.-I., Čech, R., Altmann, G. (2011). *The lambda-structure of texts.* Lüdenscheid: RAM.

Appendix

Table 7
New Year's presidential speeches.

President	year	N	V	h	L	Λ	var(Λ)	R1	var(R1)
Gottwald	1949	1413	828	10	881.4247	1.9650	0.000395	0.8882	0.000089
Gottwald	1950	2205	1115	15	1188.516	1.8021	0.000377	0.8342	0.000077
Gottwald	1951	2211	1121	13.5	1186.033	1.7941	0.000352	0.8359	0.000074
Gottwald	1952	1817	971	11.67	1053.653	1.8901	0.000240	0.8454	0.000085
Gottwald	1953	1651	911	11.5	977.842	1.9058	0.000399	0.8692	0.000086
Zápotocký	1954	2590	1272	14	1387.437	1.8285	0.000316	0.8502	0.000059
Zápotocký	1955	1570	846	9	931.7357	1.8966	0.000200	0.8627	0.000087
Zápotocký	1956	2904	1446	14.5	1567.306	1.8690	0.000153	0.8485	0.000053
Zápotocký	1957	2486	1331	13	1415.178	1.9329	0.000219	0.8606	0.000058
Novotný	1958	1592	863	10.5	925.0773	1.8606	0.000391	0.8701	0.000086
Novotný	1959	2119	1066	14.25	1145.578	1.7982	0.000231	0.8563	0.000073
Novotný	1960	2755	1300	14.67	1409.963	1.7606	0.000234	0.8550	0.000055
Novotný	1961	1583	876	11.5	920.397	1.8603	0.000457	0.8630	0.000093
Novotný	1962	2709	1289	14	1374.971	1.7423	0.000276	0.8520	0.000056
Novotný	1963	1954	973	13	1033.525	1.7407	0.000404	0.8570	0.000078
Novotný	1964	2915	1338	17	1399.033	1.6628	0.000304	0.8472	0.000055
Novotný	1965	2290	1092	14.5	1162.03	1.7049	0.000456	0.8664	0.000064
Novotný	1966	3263	1494	17	1575.538	1.6965	0.000209	0.8390	0.000050
Novotný	1967	2572	1210	15	1281.981	1.6998	0.000212	0.8513	0.000060
Novotný	1968	2293	1157	14.5	1224.202	1.7941	0.000195	0.8505	0.000069
Svoboda	1969	2059	1081	13.67	1146.529	1.8452	0.000317	0.8462	0.000078
Svoboda	1970	2186	1097	14.5	1171.538	1.7898	0.000276	0.8505	0.000073
Svoboda	1971	1554	883	11	929.011	1.9079	0.000407	0.8755	0.000088
Svoboda	1972	454	295	6	310.0376	1.8145	0.000963	0.8965	0.000270
Svoboda	1973	508	342	6	358.4382	1.9092	0.000432	0.8937	0.000239
Svoboda	1974	429	311	6	325.1071	1.9949	0.000639	0.8998	0.000284
Husák	1975	1520	789	10.5	845.9319	1.7708	0.000248	0.8613	0.000095
Husák	1976	1486	815	10.67	884.6006	1.8883	0.000612	0.8573	0.000100
Husák	1977	1281	682	10	730.9015	1.7731	0.000514	0.8735	0.000108
Husák	1978	1598	825	13.25	897.1975	1.7986	0.000464	0.8509	0.000102
Husák	1979	1322	778	9.667	836.7562	1.9756	0.000346	0.8757	0.000101
Husák	1980	1377	803	9.5	871.703	1.9871	0.000464	0.8636	0.000102
Husák	1981	1548	863	11	924.6461	1.9053	0.000370	0.8647	0.000093
Husák	1982	1155	671	9	716.5085	1.8999	0.000440	0.8818	0.000112

Husák	1983	1128	661	10	705.1472	1.9081	0.000366	0.8715	0.000127
Husák	1984	1032	661	8.5	715.8022	2.0903	0.000514	0.8974	0.000115
Husák	1985	1378	795	11	851.3336	1.9394	0.000720	0.8661	0.000106
Husák	1986	1288	756	10	813.2448	1.9636	0.000664	0.8672	0.000110
Husák	1987	1491	839	11.33	895.8157	1.9067	0.000321	0.8714	0.000095
Husák	1988	776	518	6.5	549.2147	2.0453	0.000536	0.9061	0.000137
Husák	1989	860	536	7	575.5355	1.9638	0.000245	0.8773	0.000149
Havel	1990	2347	1225	15	1287.602	1.8491	0.000299	0.8673	0.000063
Havel	1991	2421	1241	14.33	1323.707	1.8502	0.000343	0.8516	0.000064
Havel	1992	3278	1638	15	1774.774	1.9034	0.000138	0.8437	0.000047
Havel	1994	2746	1353	16	1428.073	1.7883	0.000339	0.8518	0.000057
Havel	1995	3240	1594	16	1724.62	1.8686	0.000187	0.8407	0.000049
Havel	1996	2750	1392	15	1471.607	1.8405	0.000255	0.8391	0.000059
Havel	1997	598	397	7	419.7517	1.9490	0.000503	0.9055	0.000196
Havel	1998	1312	720	12	754.9176	1.7940	0.000579	0.8643	0.000118
Havel	1999	1723	1012	11	1067.827	2.0057	0.000249	0.8738	0.000079
Havel	2000	2021	1111	13	1172.44	1.9177	0.000171	0.8671	0.000071
Havel	2001	1587	893	12.5	946.0883	1.9080	0.000500	0.8526	0.000100
Havel	2002	1926	1062	12	1126.907	1.9219	0.000293	0.8629	0.000075
Havel	2003	1941	1089	13	1150.816	1.9495	0.000294	0.8689	0.000074
Klaus	2004	913	527	10	560.2908	1.8168	0.000616	0.8642	0.000169
Klaus	2005	979	604	9.5	637.3513	1.9471	0.001002	0.8714	0.000147
Klaus	2006	845	526	9.5	556.1016	1.9262	0.000593	0.8676	0.000179
Klaus	2007	799	534	9	555.8242	2.0192	0.000632	0.8867	0.000172
Klaus	2008	914	559	8.75	584.8158	1.8945	0.000675	0.8635	0.000160
Klaus	2009	870	551	9	574.851	1.9423	0.000667	0.8914	0.000151
Klaus	2010	910	557	8.67	584.0399	1.8991	0.000564	0.8611	0.000162
Klaus	2011	891	568	8	600.5498	1.9900	0.000412	0.8754	0.000151