Qual Quant (2014) 48:899–910
DOI 10.1007/s11135-012-9811-3

Language and ideology: quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011)

Radek Čech

Published online: 7 December 2012

Abstract The relationship between ideology and language is analyzed by using quantitative linguistic methods to measure the thematic concentration of texts. The assumption is that totalitarianism and democracy represent radically different types of ideology and that this difference will be reflected in different levels of thematic concentration in texts of the same genre. The analysis focuses on the New Year speeches given by Czechoslovak and Czech presidents; these texts were chosen because they represent a relatively clearly delineated genre with a long tradition and because they are one of the most important outlets for the public expression of political opinions by the head of state. The results of statistical tests show that there exist significant differences between the thematic concentrations in the speeches of presidents from the totalitarian period and the period of democracy. The analysis also revealed that the largest differences in thematic concentrations were between the speeches made by the presidents representing the most ideologically polarized positions: the leader of the 1948 communist coup Klement Gottwald and the public face of the 1989 democratic Velvet Revolution Václav Havel.

Keywords Ideology · Language · Thematic concentration · Statistical testing · Presidential speeches

1 Introduction

The relationship between language and ideology has been at the forefront of researchers' interest for many years (e.g., Orwell 1946; Klemperer 1947; Fairclough 1989; Stubbs 1994; Van Dijk 1995, 2006; Li 2010; Al Ali 2011; Charteris-Black 2011; Salama 2011). During this time, many approaches have been developed (cf. Wodak and Meyer 2001; Van Leeuwen 2006) to analyze the relationship between these two important phenomena. Despite the great variety of these approaches, the vast majority of them has one common denominator: they

R. Čech (✉)
Palacký University Olomouc, Olomouc 77180, Czech Republic
e-mail: cechradek@gmail.com

involve description and interpretation. This means that analyzes of this kind are burdened with subjectivity as a result of personal attitudes, political preferences, life experiences and so on. Even when some statistical characteristics are taken into account (e.g., word frequency, word keyness, collocation), the impact of subjectivity is, due to the descriptive character of analyzes, "only" reduced (cf. Baker 2006). But is it actually possible to go beyond the boundaries of descriptive/interpretative methodologies and examine the language-ideology relationship experimentally, i.e. by a hypothesis testing? In the following lines we show that it is indeed possible—by applying methods of quantitative linguistics (Köhler and Altmann 2011). To our knowledge, only Stubbs (1994) has analyzed the language-ideology relationship experimentally, in the sense of a deductive hypothesis testing, so far.

The analysis is based on the following assumptions: (1) most texts concern a theme or a set of themes; (2) in each text it is possible to quantify the extent to which the text is focused on its central theme or themes (this extent or degree is termed 'thematic concentration' (Popescu et al. 2009; Popescu and Altmann 2011); (3) totalitarianism and democracy represent radically different types of ideologies (Tannenbaum 2011) and this difference is expected to be reflected in different levels of thematic concentration in different texts of the same genre type.

The analysis of the relationship between thematic concentration and ideology is performed on New Year speeches given by Czechoslovak and Czech presidents from the period 1949–2011. The aim is to determine if, and if so how, the levels of thematic concentration reflect the influence of (a) ideology, (b) authorship. We hypothesize (a) that the levels of thematic concentration in the texts of totalitarian presidents will be (significantly) higher than the levels of the democratic presidents due to the influence of totalitarian ideology (see Part 4), (b) that the levels of thematic concentration result from the influence of a range of factors, of which one of the most important is authorship, and that therefore there will be important differences between the speeches of the individual presidents (see Part 5).

2 Method for measuring thematic concentration in texts

Perhaps only with the exception of Dadaistic texts or the speech of people with mental handicaps (e.g. Wernicke's aphasia), the vast majority of instances of language use concern a theme or set of themes (in the broadest sense of the word). One can commonly speak of the main theme(s) or topic(s) of a conversation, interview, book, poem, and so on, or of secondary themes or topics. We know intuitively that e.g. texts of a non-literary character are thematically more closely delineated than literary texts; we know that a particular person "had something to say", while another "didn't really say anything". One of the ways of moving beyond the boundaries of intuitive and (always to a certain extent) vague evaluations of the thematic characteristics of texts is to apply the method of measuring a text's 'thematic concentration' (TC). This method was introduced by Popescu (2007) and elaborated further by Popescu et al. (2009) and Popescu and Altmann (2011). It enables so-called 'thematic words' to be detected within a text, their thematic weight to be measured, and ultimately the TC of the entire text to be determined. The differences between the TCs of individual texts can subsequently be tested using statistical tests.

The measurement of TC is based on the analysis of the frequency characteristics of a text; specifically, it is based on the properties of the so-called h-point; the introduction of the h-point in linguistics (Popescu 2007) was inspired by the h-index used in scientometrics (Hirsch 2005). If we rank the words of a given text according to frequency in descending order, the value of the h-point is determined by the point at which the rank of the word is equal to its frequency, i.e.

Table 1 Twenty most frequent words in the speech given by K. Gottwald, 1950

Rank	Frequency	Word	Rank	Frequency	Word
1	93	*a*[*and*]	11	26	*ten* [*the, this*]
2	84	*být* [*be*]	12	22	*který* [*which*]
3	59	*můj* [*my*]	13	22	*i* [*and*]
4	55	*v* [*in*]	14	21	***lid*** [***people***]
5	53	*o* [*about*]	15	19	*s*[*with*]
6	53	***rok*** [***year***]	16	19	*hodně* [*much]*
7	37	***procento*** [***percent***]	17	15	*než* [*than*]
8	34	*se* [*onself*]	18	14	*všechen* [*all*]
9	29	*na* [*on*]	19	13	*tento* [*this*]
10	28	*že* [*that*]	20	12	*práce* [*work*]

Word in *bold font* represents thematic words of the text

$$r = f(r), \tag{1}$$

where r is the rank of the word and $f(r)$ the frequency of the word at the given rank. If no such value occurs in the frequency distribution, the h-point is calculated as follows:

$$h = \frac{f(i)\,j - f(j)\,i}{j - i + f(i) - f(j)}, \tag{2}$$

where i and j are the word ranks and $f(i)$ and $f(j)$ are their frequencies, given that $i < j, i < f(i)$, and $j > f(j)$ To illustrate, let us take the calculation of the h-point in the frequency distribution of words (actually, lemmas. i.e. canonical word forms, called lemmas, are determined; for example lemma *do* represents word forms *do*, *does*, *did*, *done*, and *doing*) in the speech given in 1950 by K. Gottwald – see Table 1.

Given that in Table 1 there is no word for which the rank equals the frequency, $r \neq f(r)$, we use Eq. (2) and obtain

$$h_{Gottwald1950} = \frac{19 \cdot 17 - 15 \cdot 16}{17 - 16 + 19 - 15} = 16.6.$$

Popescu (2007) and Popescu et al. (2009) show that the h-point can be interpreted as a fuzzy boundary between synsemantic and autosemantic words. Autosemantic words whose rank is lower than the h-point (and which thus occur in the synsemantic 'region') are words which, due to their frequency characteristics, can be considered as the words expressing the main theme of the text. In view of the fact that the bearers of a text's theme are usually nouns, we shall consider thematic words to be nouns, plus also their predicates of the first rank, adjectives and verbs. If we take into consideration the words in Table 1, we see that there are three thematic words according to this definition whose rank is lower than the value of the h-point: the words are *rok* [*year*], *procento* [*percent*] and *lid* [*people*].

To compare the thematic weights (TW) of individual words and TCs of entire texts, it is necessary to quantify these weights. The calculation of the TW as presented by Popescu et al. (2009) is based on both the frequency and rank of the word. At first sight it is clear that the lower the rank number of a word, the higher its TW. Based on this, the TC of a text is represented by thematic words above the h-point (the rank of these words is indicated by the symbol r'), while the TW can be characterized as the distance between the h-point and the rank of a word above the h-point multiplied by its frequency $f(r')$, i.e.

$$\left(h-r'\right)\cdot f\left(r'\right). \tag{3}$$

Having thus calculated the TW, it is normalized by dividing each thematic word by the sum of all the weights of all words above the h-point and the highest frequency of the word in the text $f(1)$. The sum of all weights is calculated as follows:

$$\sum_{r=1}^{h}(h-r)=h\,(h)-\sum_{r=1}^{h}r=h^2-\frac{h\,(h+1)}{2}=\frac{h\,(h-1)}{2}. \tag{4}$$

If we divide the TW of the word, i.e. $(h-r')\cdot f(r')$, by this sum, we can calculate the index of a word's TW as

$$TW_{word}=2\frac{\left(h-r'\right)f(r')}{h\,(h-1)\,f(1)}. \tag{5}$$

For the word *rok* [*year*] in Table 1 the TW is thus calculated as follows:

$$TW_{rok[year]}=2\frac{(16.6-5.5)\,53}{16.6\,(16.6-1)\,93}=0.048855.$$

In view of the fact that Table 1 contains two words with the frequency $f=53$, i.e. the fifth and sixth ranked words, the value of r' is given by the mean rank of two words with the same frequency, i.e. $r'_{(rok[year])}=5.5$.

The TC of the entire text is then given by the sum of values of TWs of the individual thematic words, i.e.

$$TC=\sum_{r'=1}^{T}2\frac{\left(h-r'\right)f(r')}{h\,(h-1)\,f(1)}, \tag{6}$$

where T is the number of thematic words above the h-point. In the case of K. Gottwald's speech given in 1950, the TC is as follows:

$$\begin{aligned}TC_{Gottwald1950}&=TW_{rok[year]}+TW_{procento[percent]}+TW_{lid[people]}\\&=\left(2\frac{(16.6-5.5)\,53}{16.6\,(16.6-1)\,93}\right)+\left(2\frac{(16.6-7)\,37}{16.6\,(16.6-1)\,93}\right)\\&\quad+\left(2\frac{(16.6-14)\,21}{16.6\,(16.6-1)\,93}\right)\\&=0.048855+0.029498+0.004534=0.082887.\end{aligned}$$

The TC values vary by several orders of magnitude, and so for clarity of expression the TC unit *tcu* is used to express TC multiplied by 1000, i.e.

$$tcu=1000\,(TC)\,. \tag{7}$$

Popescu and Altmann (2011) derive an equation for calculating the variance of the TC of a text, which is necessary for the statistical testing of TC differences:

$$VAR\,(TC)=\left(\frac{2}{h\,(h-1)\,f(1)}\right)^2\cdot\left(\sum_{r'=1}^{T}f\left(r'\right)\right)\cdot m_{2r'}, \tag{8}$$

where $m_{2,r'}$ is the variance (the second central moment) of thematic words above the h-point, i.e.

$$m_{2r'}=\frac{\sum_{r'=1}^{T}\left(r'-m_{1r'}\right)^2 f\left(r'\right)}{\sum_{r'=1}^{T}f\left(r'\right)}, \tag{9}$$

where $m_{1,r'}$ is the first central moment, i.e.

$$m_{1r'} = \frac{\sum^{r'} \cdot f\left(r'\right)}{\sum f\left(r'\right)}. \tag{10}$$

For Table 1 the calculation of variance is as follows: the value of the first central moment is

$$m_{1r'} = \frac{(5.5\,(53) + 7\,(37) + 14(21)}{111} = 7.6081,$$

on which basis we obtain the variance (second central moment)

$$m_{2r'} = \frac{(5.5 - 7.6081)^2 53 + (7 - 7.6081)^2 37 + (14 - 7.6081)^2 21}{111} = 9.9748.$$

The value of the variance of TC in the 1950 speech by K. Gottwald (Table 1) is

$$VAR\,(TC_{Gottwald1950}) = \left(\frac{2}{16.6\,(16.6 - 1)\,93}\right)^2 \cdot 111 \cdot 9.9748 = 0.000007636.$$

Now an asymptotic u-test can be used to test the differences between the individual texts:

$$u = \frac{TC_1 - TC_2}{\sqrt{VAR(TC_1) + VAR(TC_2)}}. \tag{11}$$

For illustration we can compare the speeches by K. Gottwald from 1949 (TC = 0.063116, Var(TC) = 0.000001729) and 1950 (TC = 0.082887, Var(TC) = 0.000007636). On the basis of Eq. (11) we obtain

$$u = \frac{0.082887 - 0.063116}{\sqrt{0.000007636 + 0.000001729}} = 6.46068,$$

which means that there is a significant difference between the thematic concentration in both speeches because the calculated value of u is higher than 1.96 (significance level $\alpha = 0.05$).

If we want to compare mean values of TC, we take Eq. (11) and replace VAR(TC) with the value of the ratio of variance of means TC s^2 and the number of measurements n, i.e.

$$u = \frac{\overline{TC}_1 - \overline{TC}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{12}$$

Specifically, comparing speeches by K. Gottwald and A. Zápotocký, (see Table 4) we obtain

$$u = \frac{0.062688 - 0.016297}{\sqrt{\frac{0.0001408}{5} + \frac{0,0000598}{4}}} = 7.07,$$

which means that there is a significant difference in TC between both presidents.

3 The nature of the linguistic material

The analysis of thematic concentration was based on transcripts of New Year speeches by Czechoslovak and Czech presidents from the period 1949–2011, which were lemmatized and morphologically tagged so that each word/lemma was assigned information regarding its word class. New Year speeches represent suitable material for analysis particularly because they are a specific, relatively clearly delineated genre: they are pre-prepared spoken speeches targeted at the nation's citizens; they have certain ceremonial characteristics; they are always essentially political speeches. Given that these speeches belong among the most important means of public expression open to the head of state, we expect the influence of ideology and authorship to be reflected in them. Of course it is not always clear whether the president has written the speech himself (except in the cases of V. Havel and V. Klaus, about whom it is known that they did or do indeed write their own speeches). On the other hand, each president is politically responsible for the speech and thus exercises a major influence on the text even if he may not be its actual author: for our purposes, authorship can be viewed as the bearing of responsibility for the text.

4 Thematic concentration and ideology

In the Part 1 we expressed the hypothesis that the TC of speeches given by totalitarian presidents will be higher than the TC of democratic presidents' speeches. This hypothesis is based on the following reasoning: totalitarian (or authoritarian) regimes have a tendency to view the world from the perspective of authoritarian ideologies which are not based on the critical discourse of 'competing' opinions. This type of regime is typified by ideological propaganda in which the complexity of the world and its problems is simplified via ideological slogans and clichés such as 'the battle for peace', 'the battle against the enemies of the people and the revolution', 'war', 'fulfilling the plan', 'racial purity' and so on. These slogans represent themes which the regime used as part of its propaganda efforts to 'explain' its political attitudes and actions. In view of the fact that totalitarian and authoritarian regimes are typified by the suppression of all free discussion and the monopolistic control of access to information (cf. Tannenbaum 2011), the range of themes that may appear in public (or possibly even private) discourse is under the control of the state. It is understandable that totalitarian and authoritarian regimes desire to limit this range of themes and to subject the themes to the maximum possible degree of central control (a masterly analysis of such an approach taken by totalitarian regime is presented in Orwell's novel 1984, particularly his description of 'newspeak'). Given that the New Year speech – a speech aimed at the broadest possible public – undoubtedly played an important propaganda role in communist Czechoslovakia, we expect that the propagandistic attitudes of the authoritarian regime will be reflected in higher TC values (of course not without exceptions, as this is expected to be a tendency only).

Tracing the historical development of TC in individual speeches, we obtain the following data as presented in Table 2 and Fig. 1.

Figure 1 shows that the highest TC values are found in speeches from the early 1950s and the 1970s; these correspond with the harshest period of totalitarianism (early 1950s) and a time known as the period of 'normalization' (1970s), under G. Husák, when the communist ideology felt the need to reinforce itself via increased repression (e.g. against members of *Charter 77* or members of the *Committee for the Defence of Those Unjustly Prosecuted*) and ideological pressure (e.g. the so-called *'Anti-Charter 77' movement*). At the other end of the spectrum are the speeches by V. Havel, which (with a few exceptions) show zero TC values. This is a reflection of the thematic diversity of the texts: Havel's speeches do not contain any central theme, but instead centre around many smaller-scale themes. The low TC values can be viewed as a reflection of the president's attempt to reflect the complexity and diversity of the real world in which we live.

Table 2 Thematic concentration (TC) of presidential speeches (1949–2011)

President	Year	f(1)	h-point	Pre-h thematic words (rank/frequency)	TC	tcu	var(TC)
Gottwald, K.	1949	65	13	rok [year] (5/25); plán [plan] (7/20)	0.063116	63.12	1.73E−006
Gottwald, K.	1950	93	16.6	rok [year] (5.5/53); procento[percent] (7/37); lid [people] (14/21)	0.082887	82.89	7.64E−006
Gottwald, K.	1951	87	14.66667	rok [year] (5/51); nový[new] (13.5/16)	0.058682	58.68	1.16E−005
Gottwald, K.	1952	98	14.66667	rok [year] (4/40); nový[new] (11.5/18); americký [American] (11.5/18)	0.055048	55.05	1.10E−005
Gottwald, K.	1953	80	14	rok [year] (6.5/22); mír[peace] (9/21); sovětský [Soviet] (9/21); svaz [union](13/16)	0.053709	53.71	0.00423946
Zápotocký, A.	1954	133	17.33333	rok [year](11/25); výroba [production] (16/19)	0.009756	9.76	7.61E−007
Zápotocký, A.	1955	99	14	mír [peace] (7/20);rok [year] (9/17); lid [people] (12.5/15);	0.027473	27.47	1.63E−005
Zápotocký, A.	1956	141	19	rok [year] (10/31);republika [republic] (17/21)	0.013313	13.31	1.06E−006
Zápotocký, A.	1957	101	16	rok [year] (12/19);síla [power] (12/19); lid [people] (14.5/17)	0.014645	14.65	5.00E−007
Novotný, A.	1958	76	14.5	socialistický (10/20);rok [year] (11/19); země [country] (13.5/15)	0.023056	23.06	1.95E−006
Novotný, A.	1959	97	18	rok [year] (16/18),společnost [society] (16/18); socialistický [socialistic](16/18)	0.007277	7.28	0.00
Novotný, A.	1960	129	18	rok [year] (13/22);národní [national] (17/20)	0.006992	6.99	4.30E−007
Novotný, A.	1961	66	14	socialistický [socialistic] (7/21); rok [year] (8.5/20)	0.042791	42.79	6.39E−007
Novotný, A.	1962	112	18	rok [year] (8/35); výroba[production] (15.5/21)	0.023489	23.49	2.51E−006
Novotný, A.	1963	88	16	sjezd [convention] (11/19)	0.008996	9.00	0
Novotný, A.	1964	140	20	rok (9/39); národní[national] (15/26)	0.021015	21.02	7.94E−007
Novotný, A.	1965	115	16.5	rok [year] (10/32);socialistický [socialistic] (14/20)	0.017544	17.54	9.11E−007
Novotný, A.	1966	178	20	rok [year] (12/30);člověk [human] (19.5/20)	0.007392	7.39	5.90E−007

Table 2 continued

President	Year	f(1)	h-point	Pre-h thematic words (rank/frequency)	TC	tcu	var(TC)
Novotný, A.	1967	132	18.5	rok [year] (17.5/19);socialistický [socialistic] (17.5/19)	0.001778	1.78	0
Novotný, A.	1968	86	17.5	republika [republic] (11.5/23); národní [national] (11.5/23); rok [year](16.5/18)	0.023679	23.68	2.10E−006
Svoboda, L.	1969	110	16.33333	rok [year] (9/33)	0.017569	17.57	0
Svoboda, L.	1970	95	17.5	rok [year] (10.5/22); země [country] (15/18)	0.014509	14.51	1.07E−006
Svoboda, L.	1971	59	14	rok [year] (13/14); socialistický [socialistic] (13/14)	0.005215	5.22	0
Svoboda, L.	1972	23	7	rok [year] (4/12)	0.074534	74.53	0
Svoboda, L.	1973	27	7.5	rok [year] (5/12)	0.045584	45.58	0
Svoboda, L.	1974	22	7	rok [year] (6/8)	0.017316	17.32	0
Husák, G.	1975	72	14	rok [year] (5/38); lid [people] (11/16); práce [work] (13.5/14); dobrý [good] (13.5/14)	0.061661	61.66	0.00491536
Husák, G.	1976	82	13.5	rok [year] (7/23); lid[people] (12.5/14); socilalistický [socialistic] (12.5/14)	0.025655	25.66	7.98E−006
Husák, G.	1977	61	13.5	rok [year] (6/24);socialistický [socialistic] (8.5/16); práce [work](10.5/15)	0.059259	59.26	7.31E−006
Husák, G.	1978	87	14.75	rok [year] (5/34); život[life] (9/21); lid [people] (10.5/19)	0.060415	60.42	5.53E−006
Husák, G.	1979	71	11	rok [year] (5/22);socialistický [socialistic] (9/13)	0.040461	40.46	8.57E−006
Husák, G.	1980	81	12.5	rok [year] (5/25);další [next] (9.5/17); země [country] (12/13)	0.042083	42.08	1.39E−005
Husák, G.	1981	76	14.25	rok [year] (8/20);socialistický [socialistic] (10/18); život (12.5/16);práce [work] (14/15)	0.032509	32.51	0.00675179
Husák, G.	1982	57	11	socialistický [socialistic](5/20); rok [year] (7/18)	0.061244	61.24	3.86E−006
Husák, G.	1983	57	11	rok [year] (6/18);socialistický [socialistic] (9.5/13)	0.034928	34.93	9.41E−006
Husák, G.	1984	65	11	rok [year] (5/13);socialistický [socialistic] (9/12)	0.028531	28.53	7.81E−006
Husák, G.	1985	69	11.5	rok [year] (7/26);další [next] (10/12)	0.032406	32.41	4.26E−006

Table 2 continued

President	Year	f(1)	h-point	Pre-h thematic words (rank/frequency)	TC	tcu	var(TC)
Husák, G.	1986	71	12.66667	rok [year] (8/17);socialistický [socialistic] (10/15); nový [new] (12/14)	0.024526	24.53	4.48E−006
Husák, G.	1987	71	14	nový [new] (9/17); rok[year] (10/16); socialistický [socialistic] (13.5/14)	0.024145	24.15	3.99E−006
Husák, G.	1988	40	9		0.000000	0.00	0
Husák, G.	1989	50	11	rok [year] (9/11)	0.008000	8.00	0
Havel, V.	1990	114	18.5		0.000000	0.00	0
Havel, V.	1991	128	18.5	rok [year] (8/35); nový [new](16/22)	0.020391	20.39	2.01E−006
Havel, V.	1992	157	20.5	rok [year] (17/25); nový[new] (19/21)	0.003792	3.79	4.64E−008
Havel, V.	1994	122	21.5	občanský [civic] (13/27); stát [state] (19/23)	0.010675	10.68	6.19E−007
Havel, V.	1995	152	21.25		0.000000	0.00	0
Havel, V.	1996	155	18.66667		0.000000	0.00	0
Havel, V.	1997	37	8		0.000000	0.00	0
Havel, V.	1998	49	14		0.000000	0.00	0
Havel, V.	1999	71	15	zeď [wall] (12.5/19)	0.006372	6.37	0
Havel, V.	2000	79	16.5	svět [world] (15/17)	0.002524	2.52	0
Havel, V.	2001	70	15		0.000000	0.00	0
Havel, V.	2002	87	16.33333		0.000000	0.00	0
Havel, V.	2003	93	16		0.000000	0.00	0
Klaus, V.	2004		12	rok [year] (7/21)	0.035354	35.35	0
Klaus, V.	2005		11.5	rok [year] (9/15)	0.014116	14.12	0
Klaus, V.	2006		10		0.000000	0.00	0
Klaus, V.	2007		10.5		0.000000	0.00	0
Klaus, V.	2008		12.42857	rok [year] (9/17)	0.022181	22.18	0
Klaus, V.	2009		11.33333	rok [year] (7/19)	0.029916	29.92	0
Klaus, V.	2010		10	rok [year] (6/19)	0.035185	35.19	0
Klaus, V.	2011		10	rok [year] (7/17)	0.025185	25.19	0

Despite the relatively large differences between TC values in individual speeches, we can test the apparent different between totalitarian and democratic speeches by comparing their mean values; see Table 3.

Based on the data in Table 3 we obtain

$$u = \frac{0.0310534 - 0.009795}{\sqrt{0.000011605 + 0.000007821}} = 4.82, \tag{13}$$

which represents a significant difference. We can therefore state that the hypothesis on the influence of ideology on the mean TC value in presidents' speeches was not falsified.

Table 3 Mean values of thematic concentration (TC) in totalitarian and democratic speeches

	Mean (TC)	Mean (tcu)	s^2(TC)	s ^ 2(TC)/n
Totalitarian (1949–1989)	0.03105337	31.05	0.0004758007	1.16E−005
Democratic (1990–2011)	0.00979481	9.79	0.0001642398	7.82E−006

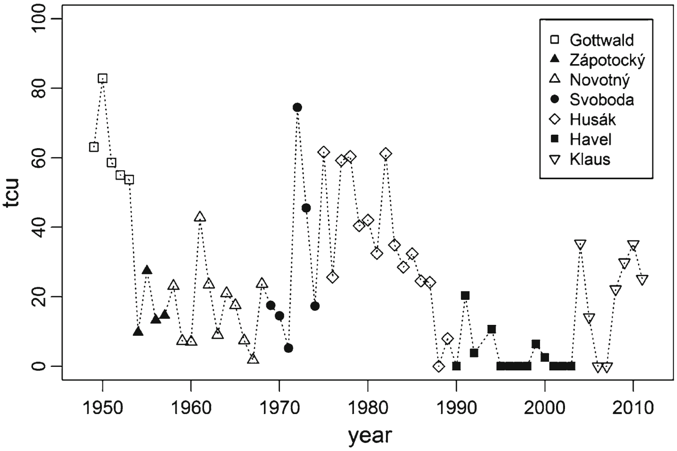

Fig. 1 Graphic depiction of the development of TC (expressed by the *tcu*) in New Year speeches

5 Comparison of individual presidents

Although the results presented above indicate at least the possibility of a relationship between ideology and the TC of presidents' speeches, Fig. 1 instantly shows that the relationship between ideology and TC is not direct, and that TC is also affected by authorship: compare the TC in speeches by K. Gottwald and A. Zápotocký, both dating from the harshest years of the 1950s, or note that the mean TC of the (democratic) V. Klaus is higher than that of the (totalitarian) A. Zápotocký and A. Novotný.

In order to obtain a more detailed picture of the relationship between TC and authorship, we tested the differences between the mean TC values of individual presidents—cf. Table 4 and Fig. 2.

Figure 2 shows the exceptional status of K. Gottwald and V. Havel—the TC values for both presidents are significantly different both in relation to each other and in relation to the TC values of all the remaining presidents. The graph also shows that the postulated difference in TC between totalitarian and democratic presidents is a consequence of the difference existing between the speeches of K. Gottwald and V. Havel. The speeches of the other presidents can be viewed (thanks to non-significant TC differences, expressed in the graph by lines joining them) as a 'core' zone of this specific genre – above all the speeches by L. Svoboda and V. Klaus, which show non-significant differences with regard to all presidents except the two extreme cases of K. Gottwald and V. Havel. The idea of a 'generic core' is also supported by the fact that in the case of L. Svoboda and V. Klaus the only thematic word (except in two of L. Svoboda's speeches) is *rok* [*year*], which represents the genre regardless of ideology or authorial preferences. The TC of other three presidents is strongly influenced by words with

Table 4 Results of test comparing mean TC values for individual presidents

President	Gottwald, K.	Zápotocký, A.	Novotný, A.	Svoboda, L.	Husák, G.	Havel, V.	Klaus, V.
Mean (TC)	0.0626884	0.01629675	0.01672809	0.02912117	0.03572153	0.003365692	0.02024212
$s^{\wedge}2(TC)/n$	2.82E−005	1.49E−005	1.26E−005	0.0001130395	2.40E−005	2.85E−006	2.55E−005
u(weighted)	6.21	2.51	2.58	1.46	3.17	4.82	2.18
Gottwald, K.	x						
Zápotocký, A.	**7.07**	x					
Novotný, A.	**7.20**	0.08	x				
Svoboda, L.	**2.82**	1.13	1.11	x			
Husák, G.	**3.74**	**3.11**	**3.14**	0.56	x		
Havel, V.	**10.65**	**3.07**	**3.40**	**2.39**	**6.25**	x	
Klaus, V.	**5.79**	0.62	0.57	0.75	**2.20**	**3.17**	x

Values in *bold font* express a significant difference

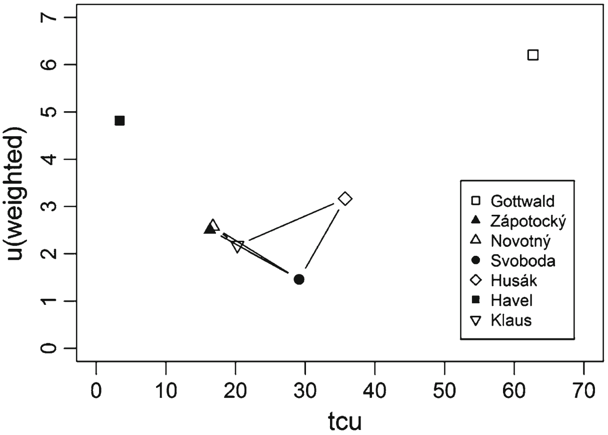

Fig. 2 Weighted differences ($u_{weighted} = \Sigma u/n$) versus the TC (expressed by *tcu*) of presidents' speeches; the line joins the presidents between which there is no significant difference of TC

ideological connotations, such as *socialistický* [*socialist*], *práce* [*work*], *lid* [*people*] etc. (cf. Table 2).

Although we expected ideology to influence TC values, the extent of the TC differences between K. Gottwald and V. Havel is quite surprising. It is clearly not merely by chance that extreme TC values were found in presidents who represent radically different worldviews and who were the main protagonists of the most important post-war social changes in Czechoslovakia – the communist coup of 1948 and the 'Velvet Revolution' of 1989.

6 Conclusion

In contrast to previous research analyzing the relationship between language and ideology – which are descriptive and interpretative in nature – we applied quantitative methods which enable us to statistically test the differences between observed phenomena. The advantages of this approach are that it enables research to move beyond the boundaries of description and helps to eliminate the influence of subjective evaluations of language data. The analysis revealed that there is a relationship between the chosen property of language (thematic concentration) and ideology, however this relationship is not simple and direct. Instead it is a result of the radically different nature of the speeches given by two presidents—K. Gottwald and V. Havel—who represented entirely opposite ideological worldviews.

Acknowledgments This work was supported by the Czech Science Foundation [grant number P406/11/0268], and by the Council of Czech Government [ESF OPVK 2.2 - The Inovation of the General linguistics and Theory of communication in cooperation with the natural sciences (CZ.1.07/2.2.00/28.0076) and ESF OPVK 2.3 - Linguistic and lexicostatistic analysis in cooperation of linguistics, mathematics, biology and psychology (CZ.1.07/2.3.00/20.0161)].

References

Al Ali, G.: Hero or terrorist? A comparative analysis of Arabic and Western media depictions of the execution of Saddam. Discour. Commun. **5**, 301–335 (2011)

Baker, P.: Using Corpora in Discourse Analysis. Continuum, London (2006)

Charteris-Black, J.: Politicians and Rhetoric: The Persuasive Power of Metaphor, 2nd edn. Palgrave Macmillan, New York (2011)

Fairclough, N.: Language and Power. Longman, London (1989)

Hirsch, J.E.: An index to quantify an individual's scientific research output. Proc. Nat. Acad. Sci. USA **102/46**, 16569–16572 (2005)

Klemperer, V.: LTI. Lingua Tertii Imperii. Die Sprache des Dritten Reiches. Reclam, Leipzig (1947)

Köhler, R., Altmann, G.: Quantitative lnguistics. In: Hogan, C.P. (ed.) The Cambridge Encyclopedia of the Language Sciences, pp. 695–697. Cambridge, New York (2011)

Li, J.: Transitivity and lexical cohesion: press representations of a political disaster and its actors. J. Pragmat. **42**, 3444–3458 (2010)

Orwell, G.: Politics and the English language. Horizon **13**, 252–265 (1946)

Popescu, I.-I.: Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.) Exact Methods in the Study of Language and Text, pp. 557–567. Mounton de Gruyter, Berlin (2007)

Popescu, I.-I., Altmann G.: Thematic concentration in texts. In: Kelih, E., Levickij, V., Matskulyak, Y. (eds.) Issues in Quantitative Linguistics, vol. 2, pp. 110–116. RAM Lüdenscheid (2011)

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

Salama, A.H.Y.: Ideological collocation and the recontexualization of Wahhabi-Saudi Islam post-9/11: a synergy of corpus linguistics and critical discourse analysis. Discour. Soc. **22**, 315–342 (2011)

Stubbs, M.: Grammar, Text, and ideology: computer-assisted methods in the linguistics of representation. Appl. Linguist. **15**, 201–223 (1994)

Tannenbaum, D.: Totalitarism. In: Kurian, G.T. (ed.) The Encyclopedia of Political Science, pp. 1673–1674. CQ Press, Washington (2011)

Van Dijk, T.A.: Discourse semantics and ideology. Discour. Soc. **6**, 243–289 (1995)

Van Dijk, T.A.: Ideology and discourse analysis. J. Polit. Ideol. **11**, 115–140 (2006)

Van Leeuwen, T.: Politics and language critical discourse analysis. In: Brown, K. (ed.) Encyclopedia of Language and Linguistics, 2nd edn. pp. 290–294. Elsevier, Oxford (2006)

Wodak, R., Meyer, M. (eds.): Methods of Critical Discourse Analysis. Sage Publications, London (2001)