Methods of analysis of a thematic concentration of the text

Radek Čech, University of Ostrava, Ostrava, Czech Republic
Ioan-Iovitz Popescu, The Romanian Academy, Bucharest, Romania
Gabriel Altmann, Ruhr-Universität Bochum, Bochum, Germany

Keywords: thematic concentration, language unit, word-form, lemma, hreb

1. Introduction

Thematic concentration (TC) is a kind of laying stress on some textual entities. Obviously, it can be considered from an infinite number of points of view. However, only some of them are useful. Popescu et al. (2009) and Popescu, Altmann (2011) proposed methods of quantitative analysis of TC based on the frequency of meaningful units which form the central thematic entities of the text. It should be emphasized that these meaningful units are not prescribed or codified and that the choice of these entities has a great impact on a result of the analysis of TC. In this study, three approaches to analysis of TC are presented, each of them taking into account different language units: *word-forms*, *lemmas*, *hrebs*.

First, the approach based on *word-forms* represents obviously the easiest way of the analysis of TC. However, in highly synthetic languages one may obtain weaker concentration because here possibly each occurrence of a thematic word appears in other grammatical form. In strongly analytic languages where word-forms are at the same time lemmas (i.e., the canonical word forms), one obtains a different result. The comparison of these results may be used typologically to express the degree of synthetism of a language.

Second, *lemmatization* is a more adequately focused approach eliminating the effect of synthetic morphology and enabling us to make textological comparisons even between two different languages. Most probably the same text translated into any two languages should display the same value of TC after lemmatization. However, the problem is that the particular theme is not associated with a single lemma in the text.

Third, not only the given word but also the references to it belong to the same theme, e.g. pronouns are always parts-of-speech referring to nouns. Thus it is possible to join words, their synonyms and their references to a greater set (or list), called usually *hreb* (cf. Ziegler, Altmann 2002).

The article is organized as follows: in *Section 2* the method of measurement of TC is presented; the three approaches to analysis of TC are exemplified in *Section 3*; in *Section 4* the measurement of so called diffuseness of TC is proposed; and the article is closed by conclusion and further research proposals.

2. Method for measuring TC in texts

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* (Popescu et al.

2009).[1] If one ranks the observed units 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 unit under consideration (i.e., word-form, lemma, hreb) is equal to its frequency, i.e.

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

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

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

where i and j are the unit ranks and $f(i)$ and $f(j)$ are their frequencies, given that $i<j$. Specifically, for the hypothetical rank-frequency distribution

rank	frequency
1	7
2	5
3	3
4	1
5	1
6	1

formula (1) yields a clear *h*-point = 3. Further, in the distribution

rank	frequency
1	7
2	5
3	2
4	1
5	1
6	1

there is no rank which equals its frequency, so, formula (2) has to be used

$$h=\frac{5\cdot 3-2\cdot 2}{3-2+5-2}=2.75 \quad .$$

If there are units with the same frequency in the rank frequency distribution, the mean rank is computed. Specifically, for the distribution

1 The introduction of the h-point in linguistics (Popescu 2007) was inspired by the h-index used in scientometrics (Hirsch 2005).

rank	frequency
1	7
2	5
3	2
4	2
5	1
6	1

we obtain

rank	frequency
1	7
2	5
3.5	2
3.5	2
5.5	1
5.5	1

and, further, the h-point is computed as follows

$$h=\frac{5\cdot 3.5-2\cdot 2}{3.5-2+5-2}=3 \quad .$$

It has been shown by Popescu (2007) and Popescu et al. (2009) that the *h-point* can be interpreted as a fuzzy boundary between synsemantic and autosemantic words. Consequently, autosemantic words as well as the other autosemantic units (i.e., lemmas, hrebs) whose rank is lower than the h-point (and which thus occur in the synsemantic 'area') are units which, due to their frequency characteristics, can be considered as the units expressing the main theme of the text.

The calculation of the TC of a unit is based on both the frequency of the unit and the distance between the h-point and the rank of the unit; further, it is normalized by dividing each thematic unit by the sum of all the weights of all units above the h-point and the highest frequency of the unit in the text $f(1)$, i.e.

$$(3) \quad TC_{unit}=2\frac{(h-r')f(r')}{h(h-1)f(1)} \quad ,$$

where h is h-point, r' is the rank of autosemantic unit above the h-point, $f(r')$ is the frequency of r'.

The thematic concentration of the entire text is then given by the sum of values of thematic concentrations of the individual thematic units, i.e.

$$(4) \quad TC_{text}=2\sum_{r=1}^{T}\frac{(h-r')f(r')}{h(h-1)f(1)} \quad ,$$

where *T* is the number of thematic words above the h-point.

3. Word-forms, lemmas, hrebs

The poem *Iba neha* written by the Slovak poet Eva Bachletová (see the Appendix) will be used for an illustration of all the above mentioned approaches to analysis of TC. Let us first consider the word-forms in the poem. The frequencies are presented in Table 1. Forms having the same frequency have been ordered alphabetically due to respective programming and ranks have been simply converted to mean ranks. Using formula (2) we can compute h-point for frequency distribution of word-forms

$$h_{Ibaneha} = \frac{5 \cdot 6 - 3 \cdot 2.5}{6 - 2.5 + 5 - 3} = 4.5455 \quad .$$

Because there is no autosemantic word above of h-point (cf. Table 1), the TC of the poem based on word-forms equals zero.

Table 1
Ranks and frequencies of word-forms in the poem *Iba neha*

r	mean (r)	word form	f_i	r	mean (r)	word form	f_i	r	mean (r)	word form	f_i
1	1	a	12	32	59.5	dobre	1	63	59.5	príde	1
2	2.5	sa	5	33	59.5	dotýkaš	1	64	59.5	prídeš	1
3	2.5	v	5	34	59.5	dúfame	1	65	59.5	skúmať	1
4	6	či	3	35	59.5	ide	1	66	59.5	slovami	1
5	6	ktoré	3	36	59.5	je	1	67	59.5	slovom	1
6	6	neviem	3	37	59.5	keď	1	68	59.5	smejem	1
7	6	som	3	38	59.5	ľahko	1	69	59.5	spätá	1
8	6	že	3	39	59.5	láska	1	70	59.5	spojená	1
9	17	bližšie	2	40	59.5	láske	1	71	59.5	stávam	1
10	17	cítim	2	41	59.5	lásku	1	72	59.5	ťa	1
11	17	čo	2	42	59.5	lebo	1	73	59.5	tebou	1
12	17	dúfam	2	43	59.5	ľúbim	1	74	59.5	tichu	1
13	17	ešte	2	44	59.5	mi	1	75	59.5	tíšiš	1
14	17	hlasom	2	45	59.5	mne	1	76	59.5	to	1
15	17	ja	2	46	59.5	nádej	1	77	59.5	tom	1
16	17	ma	2	47	59.5	nás	1	78	59.5	toto	1
17	17	na	2	48	59.5	naša	1	79	59.5	tvojou	1

18	17	neuveriteľne	2	49	59.5	nehou	1	80	59.5	unesie	1
19	17	niečom	2	50	59.5	nemôžeme	1	81	59.5	uväznená	1
20	17	o	2	51	59.5	neskutočnom	1	82	59.5	veľa	1
21	17	s	2	52	59.5	než	1	83	59.5	viac	1
22	17	tak	2	53	59.5	obaja	1	84	59.5	vieme	1
23	17	tu	2	54	59.5	objatie	1	85	59.5	vo	1
24	17	tvojím	2	55	59.5	otvorí	1	86	59.5	všetko	1
25	17	už	2	56	59.5	označiť	1	87	59.5	všetkom	1
26	59.5	ako	1	57	59.5	perami	1	88	59.5	závislá	1
27	59.5	bojím	1	58	59.5	plačem	1	89	59.5	závratnom	1
28	59.5	budem	1	59	59.5	počítam	1	90	59.5	zneistení	1
29	59.5	budeme	1	60	59.5	povedať	1	91	59.5	zovretá	1
30	59.5	čakanie	1	61	59.5	prebúdzaš	1	92	59.5	zvláštne	1
31	59.5	dávno	1	62	59.5	prekvapená	1	93	59.5	ženu	1

Thus we perform the second step and lemmatize the poem. We automatically obtain a smaller number of lemmas because Slovak is a highly synthetic language. They can be found in Table 2

Table 2
Lemmas of the poem *Iba neha* and their frequencies

r	**mean (r)**	**lemma**	**f_i**	**r**	**mean (r)**	**lemma**	**f_i**	**r**	**mean (r)**	**lemma**	**f_i**
1	1	a	12	25	21	tak	2	49	51.5	otvoríť sa	1
2	2.5	byť	6	26	21	to	2	50	51.5	označiť	1
3	2.5	v	6	27	21	tu	2	51	51.5	pera	1
4	4.5	ja	5	28	21	už	2	52	51.5	plakať	1
5	4.5	ty	5	29	21	veľa	2	53	51.5	počítať	1
6	6	vedieť	4	30	21	všetko	2	54	51.5	povedať	1
7	9	či	3	31	51.5	ako	1	55	51.5	prebúdzať	1
8	9	dúfať	3	32	51.5	báť sa	1	56	51.5	prekvapený	1
9	9	ktorý	3	33	51.5	čakanie	1	57	51.5	skúmať	1
10	9	láska	3	34	51.5	dávno	1	58	51.5	smiať sa	1
11	9	že	3	35	51.5	dobre	1	59	51.5	spätý	1
12	21	bližšie	2	26	51.5	dotýkať sa	1	60	51.5	spojený	1
13	21	cítiť	2	37	51.5	isť	1	61	51.5	stať sa	1
14	21	čo	2	38	51.5	keď	1	62	51.5	ticho	1
15	21	ešte	2	39	51.5	ľahko	1	63	51.5	tíšiť	1
16	21	hlas	2	40	51.5	lebo	1	64	51.5	toto	1

17	21	my	2	41	51.5	ľúbiť	1	65	51.5	unesie	1
18	21	na	2	42	51.5	môcť	1	66	51.5	uväznená	1
19	21	neuveriteľne	2	43	51.5	nádej	1	67	51.5	závislý	1
20	21	niečo	2	44	51.5	neha	1	68	51.5	závratný	1
21	21	o	2	45	51.5	neskutočný	1	69	51.5	žena	1
22	21	prísť	2	46	51.5	než	1	70	51.5	zneistený	1
23	21	s	2	47	51.5	obaja	1	71	51.5	zovretý	1
24	21	slovo	2	48	51.5	objatie	1	72	51.5	zvláštne	1

Here the *h*-point is $r = 4.8$, and up to 4.8 we have again no autosemantics. Nevertheless, there are two lemmas forming the core of the poem, namely pronouns "ja" (*I*) and "ty" (*you*). They represent the author and her beloved. If we accept these two lemmas as thematic units, we can compute TC of the lemmatized poem

$$TC_{Iba\ neha\ (lemmatized)} = TC_{ja} + TC_{ty} = 2\frac{(4.8-4.5)\cdot 5}{4.8(4.8-1)\cdot 12} + 2\frac{(4.8-4.5)\cdot 5}{4.8(4.8-1)\cdot 12} = 0.02741228$$

which is not a very high value.

The fact that the whole poem concentrates on the relation of two persons and in spite of this it has a very small thematic concentration, is a sign of insufficient depth of the analysis. If we translated the poem into English, it would have a higher concentration because the two pronouns (*I*, *you*) must be expressed explicitly in each case, while in Slovak they are parts of the verbs. This means that in languages like Slovak simple lemmatization does not have to bring sufficient results; one must take into account also individual morphemes. For example, the first person (*I*) is contained in the following words: *ja, neviem, som, citim, dúfam, bojím sa, budem, ľúbim, mi, mne, plačem, počítam, smejem sa, stávam sa,* and semantically, it is also part of some plural forms *budeme, dúfame, nás, naša, nemôžeme, obaja, vieme*. Hence a hreb-analysis seems to be the most adequate for this purpose.

The hreb analysis can be performed at different levels according to what units we consider: morphs, lemmas, word-forms, phrases, clauses, sentences or verses. Since the analyzed text is very short, we begin with morphs and omit those of declination and those referring only grammatically but not semantically. Thus the morpheme of third person will be omitted in this poem becauses it refers only to general object, while those of first and second person refer specifically to the speaker and the hearer, the core of the poem. Further, prepositions can be left out because they belong to the noun (just as articles in some languages); conjunctions have merely grammatical meaning and can be omitted. Thereby we obtain a still smaller inventory of units, here 53. Some of the units can be elements of several hrebs, e.g. *we* means *you* and *I*, hence it can be part of the hreb {I} and {you}. Details on establishing hrebs can be found in Ziegler, Altmann (2002). In Table 3 we present both the hrebs and the position of their elements in the poem – for the sake of easier finding. In some words, the referring morphemes are marked with bold letters; suppletivism has not been marked.

Now, since in the denotative analysis we are not interested in grammatical relations, a part of synsemantics disappeared and the words were re-classified rather according to their semantic and referential contents, the thematic concentration must

become stronger. In Table 3 the *h*-point is $h = 4.6$. Using formula (2) and considering the hrebs {ja}, {ty}, {my} as thematic, we obtain

$$TC_{Iba\ neha\ (hrebs)} = TC_{\{ja\}} + TC_{\{ty\}} + TC_{\{my\}} =$$

$$= 2\frac{(4.6-1)\cdot 30}{4.6(4.6-1)\cdot 30} + 2\frac{(4.6-2)\cdot 15}{4.6(4.6-1)\cdot 30} + 2\frac{(4.6-4)\cdot 5}{4.6(4.6-1)\cdot 30} = 0.6038647$$

which is more than twenty times greater value than the TC of the same poem based on lemmas.

Table 3
Hrebs in the poem *Iba neha* by E. Bachletová

r	mean (r)	hreb	elements	f_r
1	1	ja	{počíta**m** 1, so**m** 10, stáva-**m** sa 14-15, bojí**m** sa 26-27, obaja 33, vie**me** 35, nás 39, ma 45, ja 50, cíti**m** 51, mne 54, mi 62, ma 69, ja 71, cíti**m** 73, dúfa**m** 78, dúfa**me** 81, so**m** 83, so**m** 91, nemôže**me** 100, ľúbi**m** 104, nevie**m** 107, nevie**m** 111, nevie**m** 115, bude**m** 119, bude**me** 123, plače**m** 129, smeje**m** sa 130-131, dúfa**m** 133, naša 135}	30
2	2	ty	{tvojím 3, tvojou 5, tvojím 7, obaja 33, vie**me** 35, nás 39, prebúdza**š** 55, tíši**š** 68, dúfa**me** 81, tebou 85, nemôže**me** 100, ťa 109, prí**deš** 113, bude**me** 123, naša 135}	15
3	3	byť	{som 10, je 61, som 83, som 91, budem 119, budeme 123}	6
4	4	my	{obaja 33, nás 39, dúfa**me** 81, nemôže**me** 100, naša 135}	5
5	5.5	vedieť	{vieme 35, neviem 107, neviem 111, neviem 115}	4
6	5.5	všetko	{tom 126, všetkom 127, toto 137, všetko 138}	4
7	7.5	láska	{lásku 57, láske 90, láska 136}	3
8	7.5	objatie	{to 74, objatie 75, ktoré 80, ktoré 95}	3
9	15.5	slovo	{slovom 8, slovami 46}	2
10	15.5	hlas	{hlasom 4, hlasom 47}	2
11	15.5	niečo	{niečom 18, niečom 24}	2
12	15.5	čo	{čo 25, čo 107}	2
13	15.5	tak	{tak 19, tak 24}	2
14	15.5	cítiť	{cítim 51, cítim 73}	2
15	15.5	dúfať	{dúfam 78, dúfame 81}	2
16	15.5	prísť	{príde 109, prídeš 113}	2
17	15.5	už	{už 72, už 99}	2
18	15.5	tu	{tu 117, 121}	2

19	15.5	ešte	{ešte 118, ešte 122}	2
20	15.5	bližšie	{bližšie 28, bližšie 30}	2
21	15.5	neuveriteľne	{ neuveriteľne 20, neuveriteľne 66}	2
22	15.5	veľa	{veľa 42, viac 102}	2
23	40.5	neha	{nehou 6}	1
24	40.5	prekvapený	{prekvapená 11}	1
25	40.5	ako	{ako 12}	1
26	40.5	ľahko	{ľahko 13}	1
27	40.5	stať sa	{sa stávam 14-15}	1
28	40.5	závislý	{závislá 16}	1
29	40.5	dobre	{dobre 67}	1
30	40.5	neskutočný	{neskutočnom 21}	1
31	40.5	závratný	{závratnom 22}	1
32	40.5	báť sa	{bojím sa 26-27}	1
33	40.5	označiť	{označiť 29}	1
34	40.5	skúmať	{ skúmať 31}	1
35	40.5	dávno	{dávno 34}	1
36	40.5	dotknúť sa	{dotýkaš sa 43-44}	1
37	40.5	pery	{perami 48}	1
38	40.5	prebudiť	{prebúdzaš 55}	1
39	40.5	žena	{ženu 56}	1
40	40.5	nádej	{nádej 58}	1
41	40.5	čakanie	{čakanie 59}	1
42	40.5	zvláštne	{zvláštne 64}	1
43	40.5	tíšiť	{tíšiš 68}	1
44	40.5	spojený	{spojená 86}	1
45	40.5	spätý	{spätá 87}	1
46	40.5	uväznený	{uväznená 88}	1
47	40.5	zovretý	{zovretá 92}	1
48	40.5	ticho	{ticho 94}	1
49	40.5	otvoriť sa	{sa otvorí 96-97}	1
50	40.5	môcť	{nemôžeme 100}	1
51	40.5	povedať	{povedať 101}	1
52	40.5	ľúbiť	{ľúbim 104}	1
53	40.5	zneistenie	{zneistenie 128}	1
54	40.5	plakať	{plačem 129}	1
55	40.5	smiať sa	{smejem sa 130-131}	1
56	40.5	dúfať	{dúfam 133}	1

57	40.5	uniesť	{unesie 139}	1
58	40.5	ísť	{ide 37}	1

Hence *TC* computed on the basis of hreb-analysis yields more realistic results than the other forms. It is important especially in short texts where the first ranks are occupied by synsemantics and one cannot show formally any concentration.

The three results (Tables 1 to 3) give us the possibility of comparing the rank-frequency sequences. Using the Popescu et al. model (2010) of the rank frequency sequence instead of Zipf's, namely

$$f_r = 1 + \sum_{f \geq 1} a_i^{(-r/b_i)} \tag{5}$$

we obtain very good results in all cases. As can easily be shown, two components of (5), i.e. two exponential expressions on the right hand side are sufficient in all cases. For word-forms we obtain

$$f_r = 1 + 3.3859\exp(-r/11.7152) + 40.9942\exp(-r/0.6054)$$

with determination coefficient $R^2 = 0.97$.

For lemmas we obtain

$$f_r = 1 + 76.9438\exp(-r/0.3873) + 5.7341\exp(-r/9.8111)$$

with $R^2 = 0.97$

and for hrebs we obtain

$$f_r = 1 + 3.6459\exp(-r/10.3217) + 67.2850\exp(-r/1.0465)$$

with $R^2 = 0.99$.

In all cases the F-test yields a probability smaller than 0,00001. The difference between the R^2s is not relevant but the procedure shows that the hreb-analysis is a justified procedure (cf. Altmann 2005). Besides, it shows non-weighted but deterministic associations between the thematic words and other ones. Looking at Table 3 we see that "ja" (*I*) is associated with *reckon, become, be afraid, know, feel, hope, can, love, cry, laugh, be*, demonstrating the mood of the author who is the main subject of the poem.

4. Diffuseness

Diffuseness of a unit is measured as the relative distance between the first and last position of a unit element in text. Needless to say, one can perform the same

computation also with word-forms and lemmas but here the distances will be surely greater. If in the pre-h domain there are no thematic units, the diffuseness has no relevance, or one can say that it is zero because thematic units are only those above the h-point. In texts where there are thematic units, one can compute the diffuseness as

$$D_u = \frac{\sup\langle U_p \rangle - \inf\langle U_p \rangle}{|U|} \tag{6}$$

where |U| is the number of elements of the unit in the text and *sup* and *inf* are the highest and lowest position of the elements of the unit in text respectively.

Let us illustrate the computation of diffuseness using hrebs as units. Contrary to Ziegler and Altmann (2002: 54 ff.), who compute this property for all hrebs of the text, here we restrict ourselves to the thematic hrebs.

As can be seen in Table 3 the hreb "ja" (*I*) begins with the first word (i.e. inf = 1) and ends with the 135th word (i.e. sup = 135) hence

$$D_{ja} = \frac{135-1}{30} = 4.47$$

The other two hrebs have the values

$$D_{ty} = \frac{135-3}{15} = 8.8$$

and

$$D_{my} = \frac{135-33}{5} = 20.4 \quad .$$

The mean diffuseness of the thematic hrebs is then simply the average of these values, i.e.

$$\bar{D}_{thematic} = \frac{1}{K}\sum_{i=1}^{K} D_i \quad , \tag{7}$$

where *K* is the number of thematic hrebs (here 3). For the given text we obtain

$$\bar{D}_{thematic}(Ibaneha) = \frac{4.47+8.8+20.4}{3} = 11.22 \quad .$$

The computation can be extended to all hrebs of the poem having at least two elements, i.e. in Table 3 up to the rank 22. In that case the resulting indicator has the meaning of *referential diffuseness* of the poem. For the above poem we obtain

(4.47 + 8.8 + 18.83 + 20.4 + 20 + 3 + 26.33 + 7 + 19 + 21.5 + 3 + 41 + 2.5 + 11 +1.5 + 2 + 13.5 + 2 + 2 + 1 + 23 + 30)/22 = 341.13/22 = 15.51.

This indicator can be interpreted as the mean linear distance between extreme positions of hreb elements. It is something like the memory of the poem, the *mean distance of the recall*. The study of the link between recall and text length, recall and text sort, recall and mean verse line length, etc. is a task that should be scrutinized in the future.

5. Conclusions and further research

The TC can be considered as one of the important properties of the text. Therefore, we assume that a) it should be interrelated to another text properties, especially semantic (e.g., vocabulary richness, repeat rate, text entropy), and b) it should be influenced by pragmatic factors as a genre, style, and even ideology (Čech 2011). Thus, a next step of the analysis of TC should be focused on a testing of hypotheses as follows:

- the lower vocabulary richness of the text, the higher its TC (we assume that the more different words/lemmas/hrebs the author uses, the more themes should be mentioned in the text and, consequently, the text should express lower value of the TC);
- the higher repeat rate of the text, the higher its TC (the idea is clear: a higher repetition of words/lemmas/hrebs should bring higher concentration to the main theme/themes of the text);
- the lower text entropy of the text, the higher its TC (a higher structuring of the text could be accompanied by the higher TC);
- scientific texts should have the higher TC than novels, etc.

The testing of hypotheses of this kind will allow to incorporate the analysis of the TC into more general view on a text properties; specifically, the TC will be able to be interpreted within synergetic linguistics (Köhler 2005). Moreover, we assume that tests of theses hypotheses could also help to reveal which of approaches to the TC, i.e. based on word-forms, lemmas, or hrebs, represents the most appropriate way of text analysis, with regard to the complex functioning of the text.

Acknowledgment
This work has been supported by the Czech Science Foundation, grant no. P406/11/0268 Historical semantics.

References

Altmann, G. (2005). Diversification processes. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds.), Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook, Berlin/New York: de Gruyter, 646–659.

Čech, R. (2011) Thematic Concentration of the Text. The Analysis of Political Speeches of Czechoslovak and Czech presidents (1949—2011). Presentation at *X. Congress of Czech Historians*. Ostrava, 15. 9. 2011.

Hirsch, E. (2005) An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.*, 102, 16569-16572 (2005).

Köhler, R. (2005) Synergetic linguistics. In R. Köhler, G. Altmann & R. G. Piotrowski (Eds.), Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics. An International Handbook, Berlin/New York: de Gruyter, 760–774.

Popescu, I. I. (2007). The ranking by the weight of highly frequent words. In P. Grzybek, R. Köhler, *Exact methods in the study of language and text*. (Berlin - New York, de Gruyter, 2007), 557-567.

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G. (2011). Thematic concentration of texts. In: Kelih, E., Levickij, V., Matskuliak, J. (eds.), *Issues in Quantitative Linguistics 2: 110-116*. Lüdenscheid: RAM.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law – another view. *Quality and Quantity 44(4), 713-731.*

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

Appendix

Iba neha
E. Bachletová

Počítam s tvojím hlasom
tvojou nehou
tvojím slovom
a som prekvapená
ako ľahko sa
stávam závislá
na niečom
tak neuveriteľne
neskutočnom
závratnom
na niečom
čo sa bojím
bližšie označiť
bližšie skúmať
lebo obaja
dávno vieme
že ide o nás

a o veľa.
...-...

Dotýkaš sa ma
slovami
hlasom
perami
a ja cítim
že vo mne prebúdzaš
ženu
lásku
nádej
čakanie
a je mi
tak zvláštne
a neuveriteľne
dobre.
....-....

Tíšiš ma
a ja už cítim
to objatie
v ktoré dúfam
v ktoré dúfame.
...-...

A som s tebou spojená
spätá
uväznená v láske
som zovretá
v tichu, ktoré sa otvorí
keď už nemôžeme povedať
viac
než: ľúbim ťa.
...-...

A neviem čo príde
a neviem či prídeš
a neviem či tu ešte budem
či tu – ešte budeme
a v tom všetkom
zneistení
plačem, smejem sa
a dúfam, že naša láska
toto všetko unesie.