ABSTRACT

The aim of this article is to present a method for the analysis of the development of thematic characteristics in text called thematic concentration. The method is based on a property of the graph expressing the development of thematic concentration in subsequent chapters. Specifically, the length of the line connecting subsequent points in the graph is used for the analysis. This method allows us to test the differences between texts statistically. To obtain a more comprehensive view on relationships among observed texts, averaging of thematic concentration was used as well.

Radek Čech

The development of thematic concentration of text in Karel Čapek's travel books

1 INTRODUCTION

Thematic concentration is a property of text which can be analysed quantitatively. It represents the degree of intensity with which an author focuses on the topic (or topics) of a text. It can also be interpreted as a manifestation of the author's effort to communicate some topic(s) more intensively than would be expected from 'neutral' language or text behavior. The method of analyzing this property is based on characteristics of rank-frequency distribution of words (or lemmas) of a text. In short, the method assumes that words (or lemmas) which represent the main topic(s) of a text occur more frequently than expected in a thematically neutral text. Consequently, such words (or lemmas) appear in the so-called synsemantic branch in rank-frequency distribution of words (or lemmas), see Figure 1. Further, the method allows us to quantify their 'thematic weight' and, finally, to quantify the thematic concentration of the whole text (see Section 2 for more details, cf. Popescu et al. 2009; Čech et al. 2015). Quantification makes it possible to compare texts, genres, authors, etc. with regard to this property of text.

This method of analysis seems to bring interesting results in text linguistics (Sanada 2013), literary history (Wilson 2009, Davidová Glogarová et al. 2013, Davidová Glogarová and Čech 2013), historical semantics (David et al. 2013), analyses of political speeches (Tuzzi et al. 2010, Čech 2014), and sentiment analysis (Veselovská and Čech 2014). However, the method also has some limits which reduce its use. First of all, it cannot be used properly for 'long' texts; for Czech, it was determined empirically that the method can be applied to texts whose length does not exceed about 6500 tokens (Čech, to appear). There are several ways to overcome this deficiency. For instance, long text can be segmented into smaller parts (e.g., chapters, sections, or arbitrarily long text blocks); the thematic concentration of each part is measured separately and, finally, an arithmetic mean is computed. Averaging is a simple and useful method which, moreover, allows us to test the differences between texts statistically. Another way to solve the problem is an analysis of the sequential development of thematic concentration. In this case, the text is segmented into smaller parts (chapters,

paragraphs, etc.) and the thematic concentration of each part is measured separately. Next, the sequence of values is plotted in a graph and, finally, the graph's characteristics (specifically, the length of the curve expressing the development of values in the text) are used as an indicator. The latter approach has an important advantage over the former, because it takes into account the dynamic aspect of text development.

In this article, both approaches are used for the analysis of five travel books written by Karel Čapek. Books by a single author and in a single genre were deliberately chosen to avoid the influence of authorship and genre factors.

2 SECONDARY THEMATIC CONCENTRATION

The concept of thematic concentration was introduced by Popescu (2007), elaborated further by Popescu et al. (2009), Popescu and Altmann (2011), Čech et al. (2013, 2015), and Čech (to appear). There are several methods of measuring this text property, and each of them has some (dis)advantages (see Čech et al. 2015). For the purposes of this article, a method called *secondary thematic concentration* has been chosen, since it seems to be the best method for a comparison of texts.

The secondary thematic concentration (hereafter *STC*) represents a modification of the original method which is based on two text characteristics: 1) the frequency distribution of words (or lemmas or co-referential units), and 2) the so called *h*-point (cf. Popescu 2007). The *h*-point is defined as a point where frequency equals rank (see Formula 1); in a fuzzy way, it separates the most productive synsemantics from autosemantics in a rank frequency distribution of words or lemmas (for more details see Popescu et al. 2009, p. 17ff, and Figure 1). Specifically,

$$(1) \qquad h = \begin{cases} r_i, & \text{if there is } r_i = f(r_i) \\ \dfrac{f(r_i)r_{i+1} - f(r_{i+1})r_i}{r_{i+1} - r_i + f(r_i) - f(r_{i+1})} & \text{if there is } r \neq f(r) \end{cases}$$

where r_i is a rank and $f(r_i)$ is the respective frequency of this rank, given that r_i is the highest number for which $r_i < f(r_i)$ and r_{i+1} is the lowest number for which $r_{i+1} > f(r_{i+1})$. Thus, if no rank is equal to the respective frequency, one computes the lower part of Formula (1), consisting of neighboring values. Having stated the *h*-point, all autosemantics occurring at lower ranks are considered thematic words because they signal frequent repetition of the given autosemantics.[1] The original thematic concentration *TC* is defined as

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

where $f(1)$ is the highest frequency of word (or lemma) in the text and T is the number of autosemantics with $r < h$; if there are more words (or lemmas) with the same frequency in the rank-frequency distribution, r' can also be represented by the average rank.

1 It should be mentioned that not all autosemantics need be considered to express the thematic properties of the text; for instance, Popescu et al. (2009) use only nouns and their predicates of the first order, i.e. adjectives and verbs. In this paper, this approach is followed.

For reasons presented in Čech et al. (2015), the *h*-point is multiplied by two in the concept of the *STC*, thus, we obtain

$$STC = 2\sum_{r'=1}^{2h} \frac{(2h - r')\, f(r')}{h(2h - 1)\, f(1)} \quad . \tag{3}$$

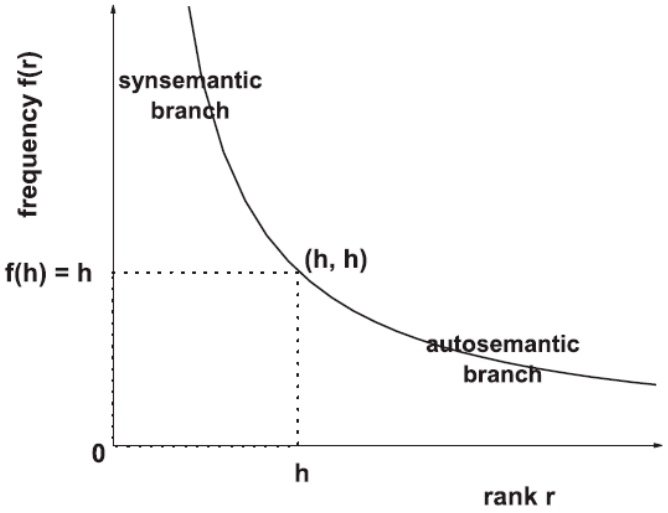

Figure 1:

The usual shape of the frequency distribution of words (or lemmas) in the majority of texts and an illustration of the determination of the h-point (cf. Popescu et al. 2009, p. 17)

3 LANGUAGE MATERIAL AND TEXT PROCESSING

For the analysis, five travel books written by Karel Čapek were chosen. Specifically, these are *Obrázky z Holandska* (Pictures from Holland), *Anglické listy* (Letters from England), *Cesta na sever* (Journey to the North), *Výlet do Španěl* (Trip to Spain), and *Italské listy* (Letters from Italy).[2] It was assumed that there would be no significant differences among these works, because they were written by one author and represent a single (and very specific) genre. If significant differences appear, it could be considered a source of inspiration for literary scholars. However, it should be emphasized that the aim of this article is to present the method, not to perform literary analysis.

Each book was segmented into the given chapters and the *STC* of each chapter was computed separately. The lemma was used as a language unit. Both lemmatisation and computing were performed by the *Quantitative Index Text Analyser QUITA* (2014; for more details, see Kubát et al. 2014). Since the *STC* cannot be properly used for texts with $N < 200$ tokens (Čech, to appear), chapters with less than 200 tokens were not taken into account.

2 For the texts see: *Karel Čapek on-line.* Společný projekt Městské knihovny v Praze, Ústavu Českého národního korpusu FF ÚK, Společnosti bratří Čapků a Památníku Karla Čapka, http://www.mlp.cz/cz/projekty/on-line-projekty/karel-capek/.

Thus, eight chapters out of 132 were excluded. Obviously, exclusion does not represent an ideal procedure. However, due to the small number of excluded chapters, it does not influence the results considerably.

4 RESULTS

The *STC*'s of each chapter are presented in the Appendix. For each book, the arithmetic mean of the *STC* was computed and differences were tested by means of the *u*-test:

$$(4) \qquad u = \frac{|\overline{STC}_1 - \overline{STC}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \ ,$$

where $\overline{STC}$ is the arithmetic mean, s^2/n is the variance of the mean, and n is the number of measurements. The results are presented in Table 1.

Table 1:

Results of tests comparing the mean *STCs;* bolded values represent significant differences. (u > 1.96, significance level α = 0.05).

	Obrázky z Holandska	*Anglické listy*	*Výlet do Španěl*	*Cesta na sever*	*Italské listy*
mean STC	0,063501	0,031489	0,023527	0,021579	0,017778
Obrázky z Holandska	x				
Anglické listy	1,82	x			
Výlet do Španěl	**2,26**	1,79	x		
Cesta na sever	**2,4**	**2,64**	0,49	x	
Italské listy	**2,6**	**3,15**	1,25	0,98	x

This method reveals the extraordinary position of the book *Obrázky z Holandska*, which has the highest $\overline{STC}$ and the smallest number of non-significant differences. On the other hand, *Výlet do Španěl*, *Cesta na sever*, and *Anglické listy* represent a homogenous group with no significant differences.

The measurement of the sequential development of the *STC* is based on the properties of a graph which expresses the values of the *STC* in subsequent chapters of a book. For an illustration, let us consider two hypothetical texts. In the first text, there are ten chapters with no differences in the *STC*; in the second text, ten subsequent chapters have opposite extreme values of the *STC*, i.e. for the first chapter *STC* = 1, for the second *STC* = 0, etc., see Figure 2.

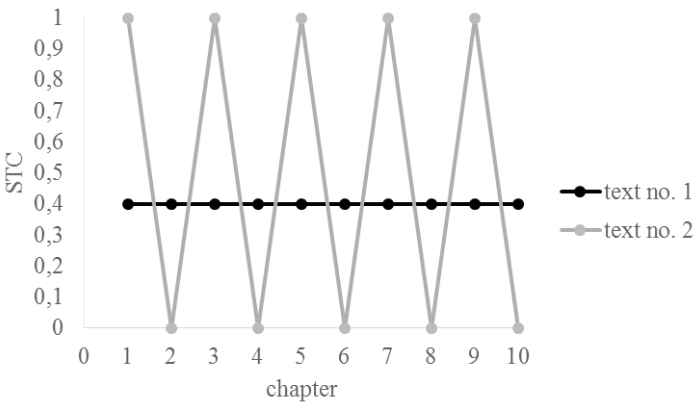

Figure 2:
The development of the *STC* in two hypothetical texts

The length of the line connecting the points representing subsequent values of the *STC* allows us to observe how the *STC* is 'manipulated' by the author in the course of writing.[3] The longer the line (i.e., the more 'jumps' between subsequent points), the more heterogeneous the development, and vice versa.

If i is defined as a value expressing the rank of a chapter, then the distance between subsequent chapters is $(i + 1) - i = 1$. Thus, the length of the line d_i connecting subsequent points can be computed by Pythagoras's theorem

(5) $$d_i = [(STC_i - STC_{i+1})^2 + 1]^{1/2} \ ;$$

see Figure 3.

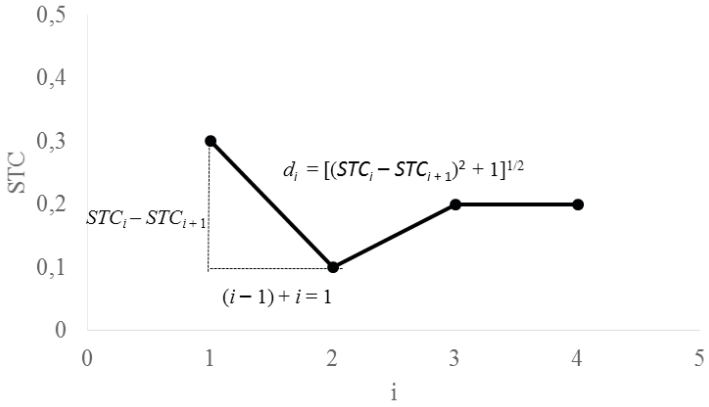

Figure 3:
The method for computing the length of the line

3 The measurement of the length of the line was used by Hřebíček (2000) for the analysis of different text properties.

The total length of line L is a sum of particular distances d_i, i.e.

$$(6) \qquad L = \sum_{i=1}^{N} d_i \ ,$$

where N is the number of lines between points. Since texts usually do not have an equal number of chapters, it is necessary to use the average distance of d_i for appropriate text comparison, specifically

$$(7) \qquad r = \frac{L}{N} \ .$$

For the text with uniform development (hypothetical text 1 in Figure 2), r = 1; for the text with extremely heterogeneous development (hypothetical text no. 2 in Figure 2), r = 1.414214. Thus, the values of r lie in the interval <1; 1.414214>.

The d_i of each chapter and r of each book are presented in the Appendix. Development of the *STC* in particular books is presented graphically in Figures 4–5, and the values of r are given in Figure 6.

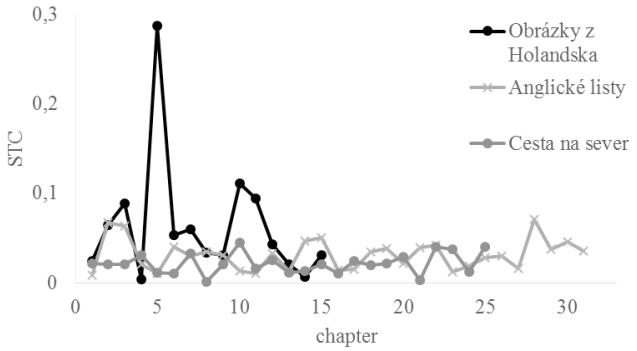

Figure 4:
Development of *STC* in K. Čapek's travel books

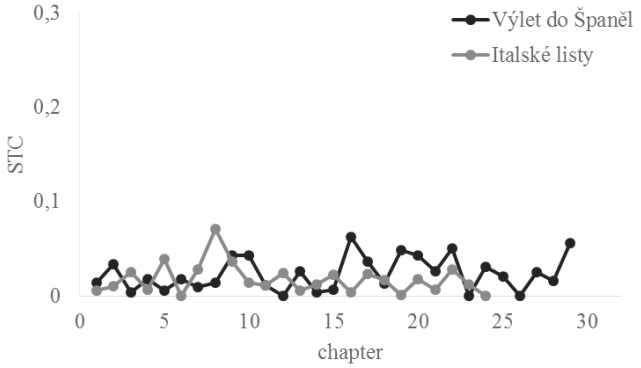

Figure 5:
Development of *STC* in K. Čapek's travel books

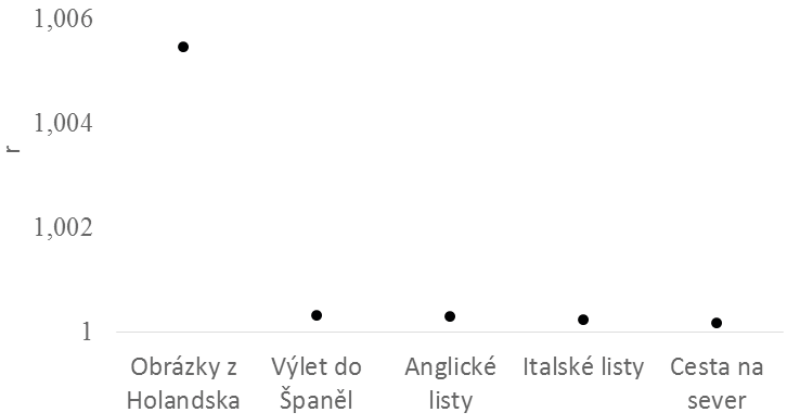

Figure 6:
Average distance of d_i in K. Čapek's travel books

The differences in the development of the *STC* between pairs of books can be tested by means of the Wilcoxon-Mann-Whitney statistical test (Hendl 2004). The results presented in Table 2 confirm the extraordinary position of *Obrázky z Holandska*, which has the highest number of significant differences. On the other hand, *Výlet do Španěl*, *Cesta na sever*, *Italské listy*, and *Anglické listy* represent a homogenous group with no significant differences. Furthermore, there is no significant difference between *Obrázky z Holandska* and *Výlet do Španěl*, which differ significantly in the comparison of the arithmetic mean (see Table 1). It shows that the author can manipulate the development of the *STC* in a similar way in texts that differ significantly with regard to the *STC*. Similarly, *Obrázky z Holandska* and *Anglické listy* reveal that books can have a non-significant difference in the *STC* and significant differences in its development.

Table 2:
Results of Wilcoxon-Mann-Whitney test; bolded values represent significant differences. (significance level $\alpha = 0.05$).

	Obrázky z Holandska	*Výlet do Španěl*	*Anglické listy*	*Italské listy*	*Cesta na sever*
Obrázky z Holandska	x				
Výlet do Španěl	0,1013	x			
Anglické listy	**0,0135**	0,3369	x		
Italské listy	**0,0284**	0,4603		x	
Cesta na sever	**0,0081**	0,0598	0,6136	0,2417	x

Putting mean *STC* and *r* together in a graph, one obtains a clear picture of the relationships among all books with regard to the observed properties (see Figure 7).

Figure 7:
The values of mean *STC* and *r* in K. Čapek´s travel books.

5 CONCLUSION

The analysis of the development of *STC* shows how to explore a dynamic aspect of the thematic characteristics of text. Generally, the method can be used for many other text properties (such as indexes of vocabulary richness, verb distances, activity of text, etc.) and can also be applied to other text units (such as paragraphs or sentences). From a textological point of view, the results reveal that the impact of both genre and authorship can be 'overcome' by other factors, as is illustrated by the significant differences in the mean *STC* and the development of d_i in the book *Obrázky z Holandska*. This finding can be used by literary scholars.

REFERENCES

Čech, R. (to appear). Tematická koncentrace textu v češtině.

Čech, R. (2014). Language and ideology: Quantitative thematic analysis of New Year speeches given by Czechoslovak and Czech presidents (1949–2011). Quality & Quantity, 48(2), 899–910.

Čech, R., Popescu, I. I., Altmann, G. (2013). Methods of analysis of a thematic concentration of the text. Czech and Slovak Linguistic Review, 4–21.

Čech, R., Garabík, R., Altmann, G. (2015). Testing the thematic concentration of text. Journal of Quantitative Linguistics, 22, 215–232.

David, J., Čech, R., Radková, L., Davidová Glogarová, J., Šústková, H. (2013). Slovo a text v historickém kontextu – perspektivy historickosémantické analýzy jazyka. Brno: Host.

Davidová Glogarová, J., Čech, R. (2013). Tematická koncentrace textu – některé aspekty autorského stylu Ladislava Jehličky. Naše řeč, 96, 234–245.

Davidová Glogarová, J., David, J., Čech, R. (2013). Analýza tematické koncentrace textu – komparace publicistiky Ladislava Jehličky a Karla Čapka. Slovo a slovesnost, 74, 41–54.

Jan Hendl (2004). Přehled statistických metod zpracování dat. Analýza a metaanalýza dat. Praha: Portál.

Hřebíček, L. (2000) Variation in Sequences. Prague: Oriental Institute.

Kubát, M., Matlach, V., Čech, R. (2014). QUITA. Quantitative Index Text Analyzer. Lüdenscheid: RAM-Verlag.

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: P. Grzybek – R. Köhler (Eds.), Exact Methods in the Study of Language and Text. Berlin – New York: Mouton de Gruyter, 555–565.

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

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

QUITA - Quantitative Index Text Analyser (2014). Available at: https://code.google.com/p/oltk/

Sanada, H. (2013). Thematic concentration in Japanese prose. In Obradovic, I., Kelih, E., Köhler, R. Methods and Applications of Quantitative Linguistics. Selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO), Belgrade, Serbia, April 26–29, 2012. Belgrade: University of Belgrade, 130–140.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). Quantitative Analysis of Italian Texts. Lüdenscheid: RAM-Verlag .

Veselovská K., Čech R. (2014). Opinion Target Identification Using Thematic Concentration of the Text. Contributed talk, QUALICO 2014, Olomouc, Czech Republic, May 29 –\ June 1, 2014.

Wilson, A. (2009). Vocabulary richness and thematic concentration in internet fetish fantasies and literary short stories. Glottotheory 2 (2), s. 97–107.

APPENDIX

Obrázky z Holandska			*Anglické listy*		
Chapter	*STC*	d_i	Chapter	*STC*	d_i
1	0.024270		1	0.008537	
3	0.064849	1.000823	2	0.067470	1.001735
4	0.088523	1.000280	3	0.063351	1.000008
5	0.003551	1.003604	4	0.020768	1.000906
6	0.286859	1.039357	5	0.010136	1.000057
7	0.053571	1.026851	6	0.040486	1.000460
8	0.060000	1.000021	7	0.029841	1.000057
9	0.033381	1.000354	8	0.034398	1.000010
10	0.030909	1.000003	9	0.030725	1.000007
12	0.111183	1.003217	10	0.012963	1.000158
13	0.094180	1.000145	11	0.010802	1.000002
14	0.042735	1.001322	12	0.032789	1.000242
15	0.020524	1.000247	13	0.013763	1.000181
16	0.006588	1.000097	14	0.046470	1.000535
17	0.031390	1.000308	15	0.050926	1.000010
			16	0.014390	1.000667
			17	0.014985	1.000000
			18	0.034712	1.000195
			19	0.038337	1.000007
			20	0.021875	1.000135
			21	0.039668	1.000158
			22	0.041958	1.000003
			27	0.012121	1.000445
			28	0.019290	1.000026
			29	0.028571	1.000043

Obrázky z Holandska			*Anglické listy*		
Chapter	*STC*	d_i	Chapter	*STC*	d_i
			30	0.030100	1.000001
			31	0.016304	1.000095
			32	0.071023	1.001496
			33	0.037546	1.000560
			34	0.045758	1.000034
			35	0.036107	1.000047

Cesta na sever			*Výlet do Španěl*		
Chapter	*STC*	d_i	Chapter	*STC*	d_i
1	0.021289		1	0.013889	
3	0.020443	1.000000	2	0.033597	1.000194
4	0.020618	1.000000	3	0.004141	1.000434
5	0.031315	1.000057	4	0.018131	1.000098
6	0.010997	1.000206	5	0.005445	1.000080
7	0.010681	1.000000	6	0.017756	1.000076
8	0.032645	1.000241	7	0.009470	1.000034
9	0.001443	1.000487	8	0.014323	1.000012
10	0.020949	1.000190	9	0.043132	1.000415
11	0.044664	1.000281	10	0.042810	1.000000
12	0.016306	1.000402	11	0.011111	1.000502
13	0.025174	1.000039	12	0.000000	1.000062
14	0.011512	1.000093	13	0.026684	1.000356
15	0.012940	1.000001	14	0.003676	1.000265
16	0.020624	1.000030	15	0.006808	1.000005
17	0.010401	1.000052	16	0.062915	1.001573
18	0.024370	1.000098	17	0.036667	1.000344

Cesta na sever			*Výlet do Španěl*		
Chapter	*STC*	d_i	Chapter	*STC*	d_i
19	0.019551	1.000012	18	0.012903	1.000282
20	0.021711	1.000002	19	0.048589	1.000637
21	0.029091	1.000027	20	0.043423	1.000013
22	0.003169	1.000336	21	0.026783	1.000138
23	0.040275	1.000688	22	0.050325	1.000277
24	0.037093	1.000005	23	0.000000	1.001266
25	0.012340	1.000306	24	0.030897	1.000477
26	0.039886	1.000379	25	0.020928	1.000050
			26	0.000000	1.000219
			27	0.025511	1.000325
			28	0.016162	1.000044
			29	0.056211	1.000802

Italské listy		
Chapter	*STC*	d_i
1	0.006279	
3	0.010709	1.000010
4	0.025601	1.000111
5	0.007128	1.000171
6	0.038932	1.000506
7	0.000672	1.000732
8	0.028604	1.000390
9	0.070795	1.000890
10	0.036156	1.000600
11	0.013932	1.000247
12	0.011905	1.000002

Italské listy		
Chapter	*STC*	d_i
13	0.024791	1.000083
14	0.005518	1.000186
15	0.012745	1.000026
16	0.022720	1.000050
17	0.003816	1.000179
18	0.023387	1.000191
19	0.016786	1.000022
20	0.001003	1.000125
21	0.017927	1.000143
22	0.006853	1.000061
23	0.028314	1.000230
24	0.012070	1.000132
25	0.000000	1.000073

AFFILIATION

Radek Čech
University of Ostrava
Reální 5
Ostrava 701 03
Czech Republic
cechradek@gmail.com