

The Interpretation of Quantitative and Corpus Analysis of Thematic Fields in First Collections of Short Stories by Jan Čep¹

Keywords: Jan Čep, thematic concentration of a text, quantitative and corpus linguistics and theory of literature, interpretation of quantitative research in literary studies

Abstract

The present paper is a direct continuation of the first part *Potencionality of Quantitative and Corpus Analysis to Literary Studies – Toward Methodology (The Analysis of Thematic Fields)*, which was published in third number of this journal in 2016. In this study we present in detail the exploration of possibilities of interpretation using quantitative and corpus methods and illustrate the way, how specific we can use the results of the named method in the literary interpretation.

I. Assessment and interpretation

As we can observe the frequency analysis is not a mere formal quantification. It naturally has a qualitative dimension which reflects the choice of particular lexis appearing in the area above the *h-point*. Mutual relation of both aspects (quantitative and qualitative) is not coincidental, it is a projection of immanent structural and semantic rules found in a particular text. This fact can be proved by the following test

¹ This is the second part of the study, which first part was published in *Bohemistyka* nr. 3, 2016. Both parts of this studies are basic results of research, which was in detail published in czech written monograph *Kvantitativně-korpusová analýza a literární věda: model a realizace autorského korpusu a slovníku Jana Čepa v kontextu zahraniční* (2015).

in which we attached relative frequency values from three completely different corpuses to frequency values of thematic words from selected Čep subcorpuses. The first corpus (ČAPEK 1) contains selected and processed works of Karel Čapek² written in the 1920s, the second corpus (ČAPEK 2) includes selected works of the same author written in the 1930s. Frequency values were subsequently compared with referential corpuses SYN2010_BEL³ a SYN2010.⁴

Table 1. Compared corpuses

Jan Čep	ČAPEK 1	SYN2010_BEL	SYN2010
Double Being (1926)	The Absolute at Large (1922)		
	Krakatit (1922)		
The Vigil (1928)	Stories from a Pocket and Stories from Another Pocket (1929)		
Jan Čep	ČAPEK 2		
Centaury - part (1931)	Hordubal (1933)		
	Meteor (1934)		
	An Ordinary Life (1934)		

Table 2. Comparative values of relative frequencies for selected corpuses

r	Double Being, 1926			ČAPEK 1	SYN2010_BEL	SYN2010
	lemma	Af	Rf	Rf	Rf	Rf
1	to go	109	4999,31	2798,36	1644,20	1285,68
2	eye	93	4265,47	1504,17	1093,59	569,57
3	head	75	3439,89	1424,00	990,25	559,25

² Karel Čapek (1890–1938), a prominent Czech writer, playwright and journalist.

³ SYN2010_Bel is a corpus of Czech fiction that forms a part of SYN2010. For the structure of the Czech National Corpus, see <http://ucnk.ff.cuni.cz/struktura.php>.

⁴ SYN2010 is a Czech synchronical referential corpus that is being compiled by the team of Czech National Corpus in Prague (<http://ucnk.ff.cuni.cz/struktura.php>).

4	hand	58	2660,18	1973,74	1227,33	679,97
5	mother	53	2430,86	87,81	428,42	269,29
6	window	52	2384,99	545,93	347,38	212,18
7	father	50	2293,26	148,89	466,46	283,49
8	day	50	2293,26	931,51	1088,58	960,28
9	Rudolf	47	2155,67	11,45	20,79	26,61
10	Ludvík	45	2063,94	0,00	23,36	20,75

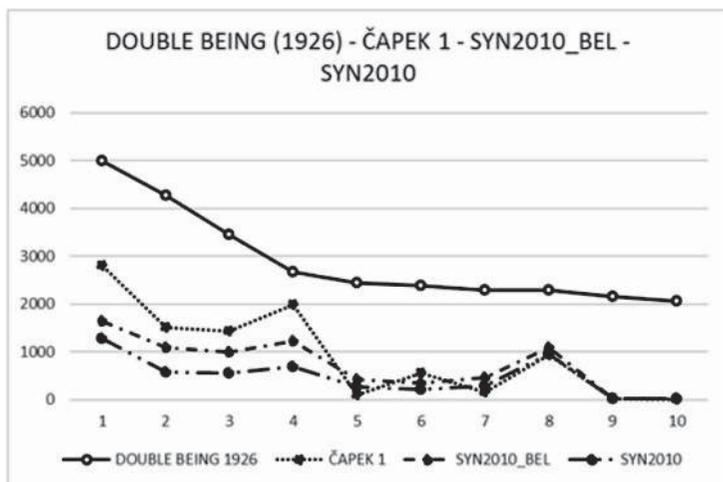


Chart 1. Comparative values of relative frequencies-curves of selected corpuses (the values are in Table 2)

Table 3. Comparative values of relative frequencies for selected corpuses

The Vigil, 1928				ČAPEK 1	SYN2010_BEL	SYN2010
r	lemma	Af	Rf	Rf	Rf	Rf
1	to go	74	3247,32	2798,36	1644,20	1285,68
2	Rozárka	68	2984,03	0,00	0,54	0,42
3	eye	66	2896,26	1504,17	1093,59	569,57

4	road	61	2676,85	458,12	775,99	589,16
5	face	53	2325,79	576,47	526,42	281,85
6	while	49	2150,25	759,72	820,20	481,40
7	man	48	2106,37	3363,38	1775,62	1827,81

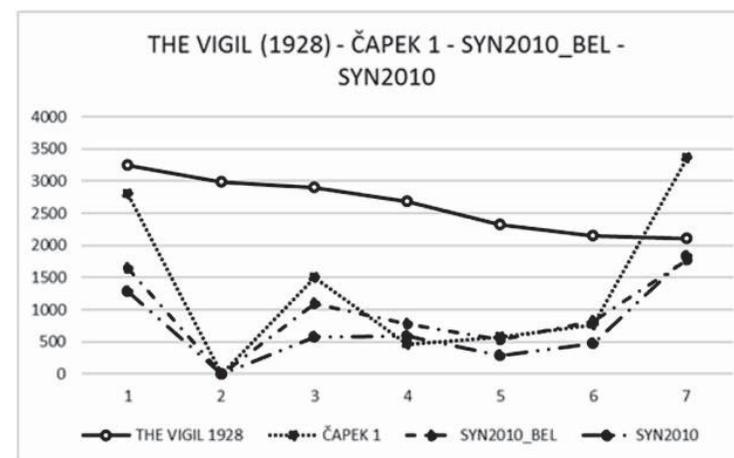


Chart 2. Comparative values of relative frequencies-curves of selected corpuses (the values are in Table 3)

Table 4. Comparative values of relative frequencies for selected corpuses

Centaury - part 1931				ČAPEK 2	SYN2010_BEL	SYN2010
r	lemma	Af	Rf	Rf	Rf	Rf
1	eye	103	4464,29	1818,12	1093,59	569,57
2	face	65	2817,27	533,95	526,42	281,85
3	head	60	2600,55	1709,98	990,25	559,25
4	hand	57	2470,53	2095,23	1227,33	679,97
5	heart	54	2340,50	513,67	285,05	171,77
6	feel	52	2253,81	675,88	482,63	296,56

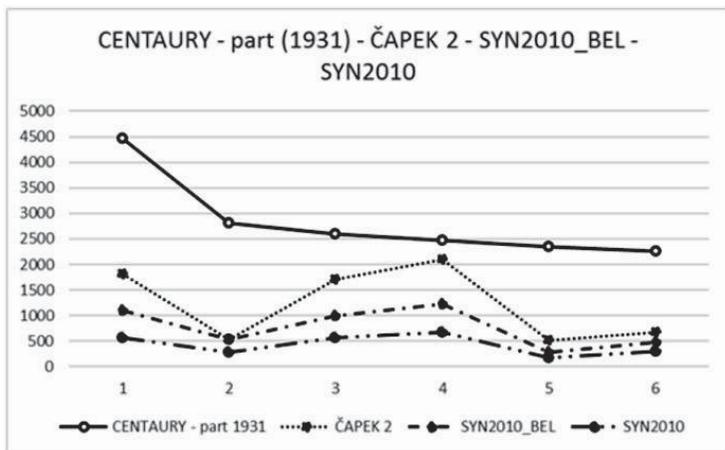


Chart 3. Comparative values of relative frequencies-curves of selected corpora (the values are in Table 4)

Comparison with thematic fields in another authorial (sub)corpus can serve as further evidence of unique evidence of symptomatic character of thematic words and their frequency course in subcorpora of Čep's works. We compiled a frequency list of lemmas that are found above the *h*-point in the corpus of *Povídky z jedné a druhé kapsy* [Stories from a Pocket and Stories from Another Pocket] (1929) by Karel Čapek to serve as an example. Even cursory collation of the cited criteria, i.e. the choice of lexis and the frequency course, reveals evident differences between the two authors.

Table 5. Frequency list of autosemantic lexemes in the (sub)corpus of *Povídky z jedné a druhé kapsy* [Stories from a Pocket and Stories from Another Pocket] (1929) by Karel Čapek

<i>r</i>	lemma	<i>Af</i>	<i>Rf</i>	<i>r</i>	lemma	<i>Af</i>	<i>Rf</i>
1	mister	1237	10924,18	19	to think	157	1386,50
2	to have	898	7930,41	20	whole	144	1271,69
3	you (pl.)	748	6605,73	21	case	134	1183,38

4	to say (perf.)	700	6181,83	22	to come	134	1183,38
5	to know	568	5016,12	23	to beg	132	1165,72
6	man	536	4733,52	24	old	131	1156,89
7	to go	410	3620,79	25	year	122	1077,41
8	we	274	2419,75	26	doctor	122	1077,41
9	they	247	2181,30	27	to understand	122	1077,41
10	to see	247	2181,30	28	no	120	1059,74
11	madam	240	2119,49	29	hand	118	1042,08
12	to give	227	2004,68	30	to begin	117	1033,25
13	to say (im-perf.)	220	1942,86	31	day	116	1024,42
14	to tell	203	1792,73	32	to get	115	1015,59
15	you (sg.)	198	1748,58	33	to say (literary)	113	997,92
16	to do	181	1598,45	34	to do	113	997,92
17	thing	170	1501,30	35	to look	104	918,44
18	stand	165	1457,15	36	second	102	900,78

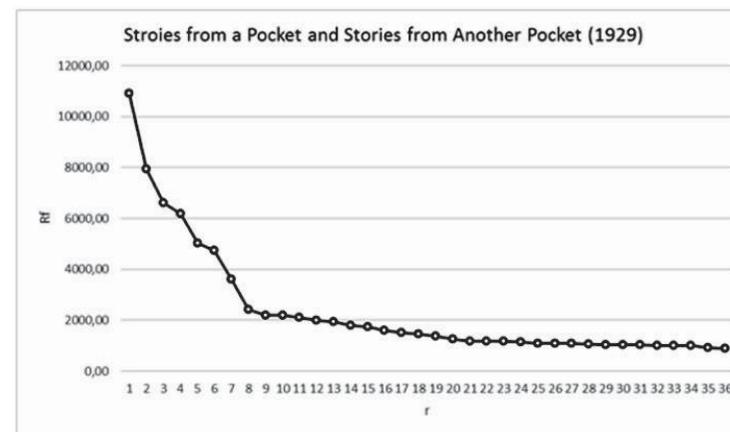


Chart 4. The relative frequencies-curve of selected corpus (the values are in Table 5)

Stated values and comparison suggest that so-called thematic words are unequivocally symptomatic not only due to their frequency that is, compared to their distribution in other corpuses, noticeably higher⁵, but also due to the choice of words, or so-called word type. This fact indicates that in a sufficiently large and presentable corpus, thematic words that belong to their (particular) corpus are distributed in a relatively analogous way, while outside the corpus, in which they reach symptomatic values, they have neutral distribution.⁶ In both cases the occurrence of these words (of *lemmas*⁷, in our case) reflects their structural position in the corresponding corpus. In other words, their frequency distribution and qualitative choice in Jan Čep (sub)corpuses display structural intention of the artistic text, or numerous structural factors that participate in building and formation of a fictional narrative and its artistic testimony. This sphere includes, besides other things, the issues of a narrator or composition, as well as the issues of wider ideological context of author, era, etc. The findings we have obtained so far enable us compare individual areas of occurrence of so-called thematic words in a meaningful way and interpret their frequency (qualitative) and quantitative aspects in detail.

⁵ None of the obtained *Rf* values in the Čep (sub)corpuses drops below 2000 *Rf*, while in other corpuses only very few lexemes can be found below this limit. Specifically they are the lemmas of *jít* [to go] and *člověk* [man] (in ČAPEK 1 corpus) and the lemma of *ruka* [hand] (in ČAPEK 2 corpus) that have significant frequency values which clearly reflects the fact that the lexemes themselves play an important role and are significant semantically in the prosaic works by Karel Čapek (higher positions of the lemmas *jít* [to go] and *člověk* [man] see in Table 9).

⁶ Curves ČAPEK 1 and ČAPEK 2 clearly show certain differences. They are caused by the fact that corpuses ČAPEK 1 and ČAPEK 2, unlike corpuses SYN2010_BEL and SYN2010, are not generally referential. Nevertheless, the tendency towards analogous course with the curves SYN2010_BEL and SYN2010 is evident.

⁷ It is certainly possible to include even inflected word forms.

Frequencies (*Rf*) in the first three observed Jan Čep corpuses evince different courses. Differences pertain namely the first three positions.

Table 6. Reference table and diagram of the first three Čep (sub)corpuses

	Double Being, 1926 (<i>Rf</i>)	The Vigil, 1928 (<i>Rf</i>)	Centaury - part, 1931 (<i>Rf</i>)
1	4999,31	3247,32	4464,29
2	4265,47	2984,03	2817,27
3	3439,89	2896,26	2600,55
4	2660,18	2676,85	2470,53
5	2430,86	2325,79	2340,50
6	2384,99	2150,25	2253,81
7	2293,26	2106,37	
8	2293,26		
9	2155,67		
10	2063,94		

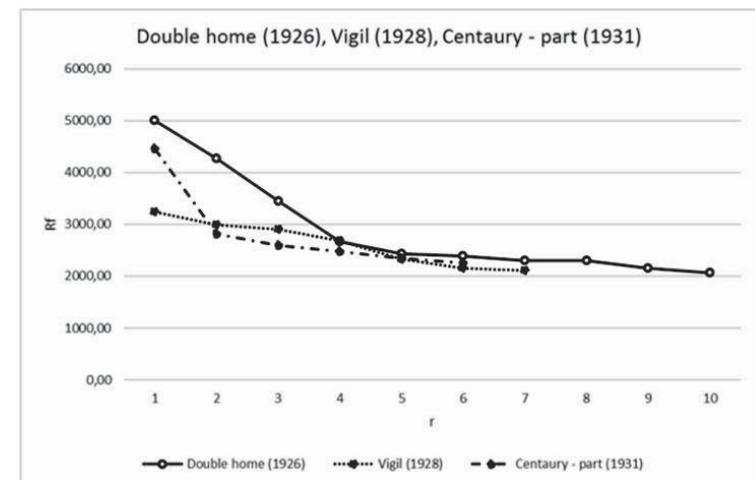


Chart 5. Comparative values of relative frequencies-curves of the first three Čep's novels (the values are in Table 6)

In the corpus *Dvoji domov* [Double Being] 1926 the first three lemmas (*jít* [to go], *oko* [eye], *hlava* [head]) are dominant, while in the corpus *Vigilie* [The Vigil] 1928 the diagram course between thematic words is gradual which is reflected in a higher DTF value (see Změlík 2016, p. 217).. In the corpus *Zeměžluč – oddíl* [Centaur – part] 1931, only one lemma, *oko* [eye], has a symptomatic frequency. The merger of the first three tables of relative frequencies in the above mentioned corpora shows differences in the course of *Rf* curves. The divergences pertain the area between 1 and 3 while from 4 on the courses are analogous. Differences are notable namely between courses of frequency values in the corpus *Dvoji domov* [Double Being] 1926 on one hand and the corpora *Vigilie* [The Vigil] 1928 and *Zeměžluč – oddíl* [Centaur – part] 1931 on the other hand. The mentioned differences show a tendency towards weakening of originally highly symptomatic lemmas (*oko* [eye], *jít* [to go], *hlava* [head]).⁸ The weakening causes a structural change of relations between thematic words and is also reflected in the choice of word types. For example the lemma *jít* [to go] is removed from the area above the *h-point* in the corpus *Zeměžluč – oddíl* [Centaur – part] 1931. In the (sub)corpus *Vigilie* [The Vigil] 1928, this lemma is found above the *h-point*, as well as in the (sub)corpus *Dvoji domov* [Double Being] 1926 where it reaches the first position, but here we record a substantial difference in frequency that makes 53.95%. The situation is reflected in a corresponding DTF value. These aspects signalize a change in artistic strategy which refers to textual (artistic) intention. Here a space for individual theoretical analysis arises that will try to find out to which extent the change is reflected in the categories of a narrator, composition, etc., or to which extent the change is determined by these categories.⁹

Besides frequency (a quantitative criterion), also the type of lemmas (qualitative criterion) occurring above the *h-point* is remarkable.

⁸ The lemma *oko* [eye] reaches high values again in the corpus *Zeměžluč – oddíl* [Centaur – part] 1931.

⁹ The quantitative-corpus analysis in itself cannot substitute theoretical literary analysis.

In the observed thematic fields of Čep (sub) corpora such lemmas occur that constitute semantic groups (due to the character of the sememe).

Table 7. Taxonomy of semantic classes of lemmas above the *h-point*. Percentage in the column stand for proportional differences between *Rf* of lemmas, percentage in lines mark out differences in quantity representation of lemmas in a semantic class

		I'	II	III	IV	V	VI	VII	%
Double Being (1926)	1	to go							100,00
	2		eye						85,32
	3		head						68,81
	4		hand						53,21
	5		mother						48,62
	6			win- dow					47,71
	7		father						45,87
	8				day				45,87
	9		Rudolf						43,12
	10		Ludvík						41,28
%		10,00	70,00	10,00	10,00				
The Vigil (1928)	1	to go							100,00
	2		Rozárka						91,89
	3		eye						89,19
	4					road			82,43
	5		face						71,62
	6				while				66,22
	7		man						64,86
%		14,29	57,14	0	14,29	14,29			

Centaury - part (1931)	1		eye						100,00
	2		face						63,11
	3		head						58,25
	4		hand						55,34
	5		heart						52,43
	6						to feel		50,49
%		0,00	83,33	0,00	0,00	0,00	16,67		
Centaury - book (1931)	1		eye						100,00
	2		face						63,28
	3		hand						60,45
	4		head						60,45
	5					road			50,85
	6		man						48,59
	7						to know		48,59
	8						to feel		48,02
	9				day				46,33
	10		mother						45,76
	11				while				44,63
	12		heart						42,37
	13		old						41,24
	14		Rozárka						38,42
	15							world	36,72
%		0,00	60,00	0,00	13,33	6,67	13,33	6,67	

Note:

I – Motion,
 II – Subject,
 III – Object,
 IV – Time,

V – Space I,
 VI – Properties of a subject,
 VII – Space II.

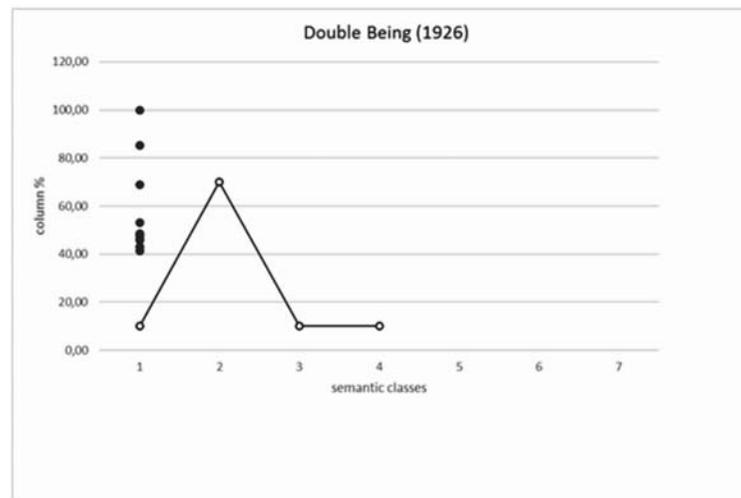


Chart 6. Comparative values (percent) of semantic classes and frequency percent (the values are in Table 7)

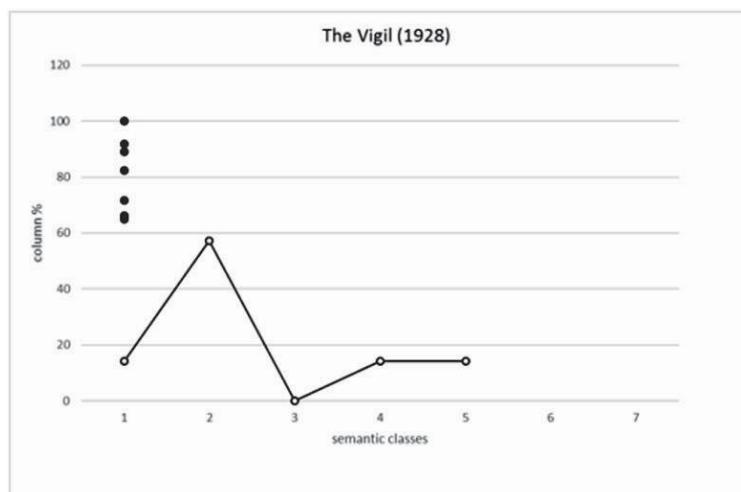


Chart 7. Comparative values (percent) of semantic classes and frequency percent (the values are in Table 7)

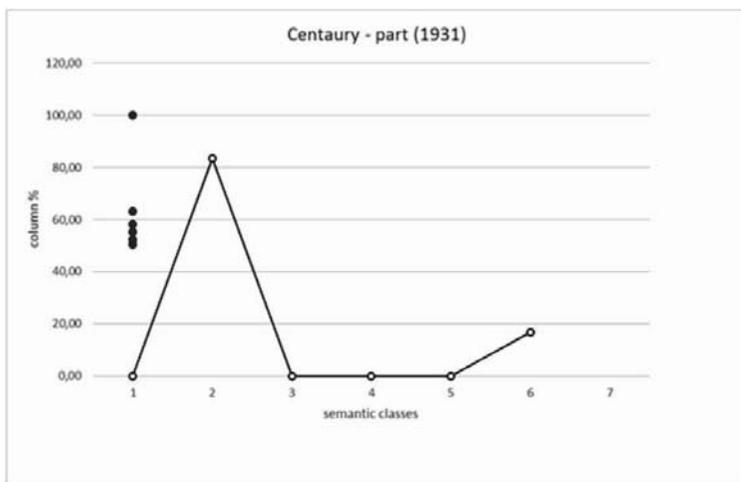


Chart 8. Comparative values (percent) of semantic classes and frequency percent (the values are in Table 7)

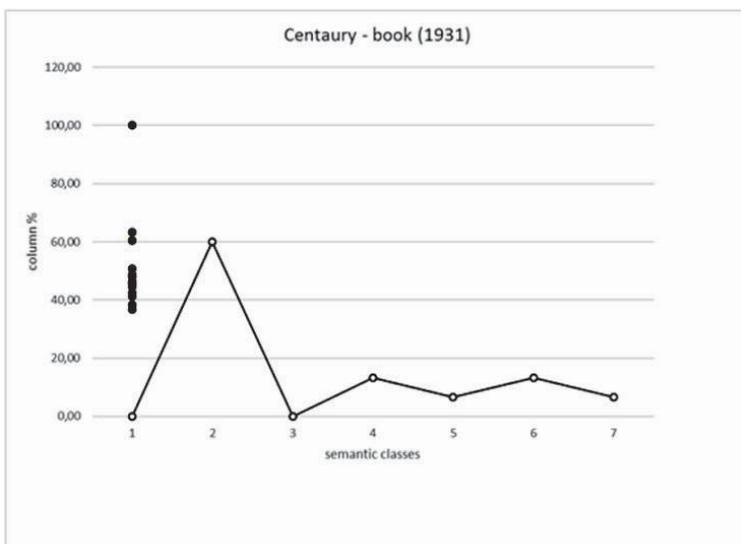


Chart 9. Comparative values (percent) of semantic classes and frequency percent (the values are in Table 7)

In the table (Table 7) and in the diagrams (Charts 6–9) above we can observe the distribution of semantic classes (the line with percentage values in the table) and frequency intensity of individual lemmas in percent (column with percentage in the table). In the corpus *Dvoji domov* [Double Being] 1926, the semantic class of +HUM (Subject) dominates – this class is closely related to the category of motion.¹⁰ Categories of subject (the lemma *okno* [window]) and temporality (the lemma *den* [day]) follow. Even though these categories have the same percentage of representation (10%) the meaning of the category of motion is governed by the frequency of the lemma *jít*. In the corpus *Vigilie* [The Vigil] 1928, the class of +HUM (Subject) dominates together with the category of motion, and besides temporality the motivic element of space (the lemma *cesta* [road]) is strengthened. That does not automatically suggest the absence of the spatial aspect in *Dvoji domov* [Double Being] 1926; here the lemma *cesta* [road] reaches values right below the h-point which proves that the motive of space is also significant.¹¹ It seems that the motive was strengthened in *Vigilie* [The Vigil]. A question here arises, to which extent the motive changes and how the changes influence the character of texts included in the collection. In the section *Zeměžluč* [Centaury – part], the category +HUM (Subject) prevails again; it is accompanied by a class of qualities this category is endowed with (the lemma *cítit* [to feel]). In the collection of *Zeměžluč* [Centaury – book], the categories of temporality (lemmas *den* [day], *chvíle* [while]), space (the lemma *cesta* [road]) and qualities of subject (lemmas *vědět* [to know], *cítit* [to feel]) distributed in a relatively uniform way which is also reflected in the frequency course of these lemmas. The cited comparison shows how

¹⁰ We omitted a highly frequented lemma *stát* the meaning of which is confused in the corpus; the meaning of the motion verb prevails, but *stát* can function also as a copula (to become someone/something) or it can signify monetary value. The percentage of representation of this semantic class should in fact be higher.

¹¹ The limit of h-point should not be seen as a strict boundary but rather as a point of orientation that naturally presupposes the occurrence of a semantic zone of lemmas that participate in the theme of the text below the h-point limit.

frequencies in Čep's works are structurally interconnected with the type of lemmas and with semantic classes of the lemmas. The semantic class +HUM (Subject) is dominant; in the course of writing, classes of semantic categories of time, space and qualities of the subject join the mentioned semantic class. In the final collection of *Zeměžluč* [Centaury – book], these categories (with the exception of still dominant group +HUM-Subject) are distributed in a relatively equal way which suggests possible change or modification of creative focus. It is also obvious that individual categories are capable of mutual pervading or hierarchization. From the semantic point of view, the categories of time and space form a higher structural layer in the fictional narrative which is of course functionally related to the category of +HUM (Subject) and the attributes attached to it (motion, qualities, object). Thus we are finally able to model other relations founded on semantic bases (categories) with respect to their prototypical function in literary fiction. Actions of characters in fictional words unfold primarily in space-time frames. In the analysed works (the first three collections of short stories by Jan Čep), these issues clearly form a central theme and problem of fictional worlds.

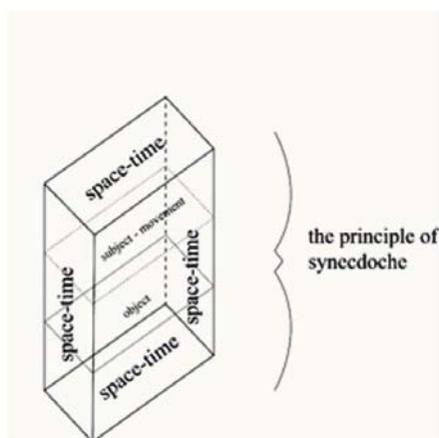


Chart 10. Structure of prototypical semantic categories in the first three Čep (sub)corpus

So far accomplished analyses and different types of systemized generalizations suggest synecdochic character¹³ (though with variable intensity of artistic evocation) as one of structural regularities in the first three collections of Čep short stories. Synecdochic character is also signaled by hyperonymic-hyponymic relations between lemmas. We must, however, ask what looking for such binds between semantic categories implies. First of all they will be a product of textual-narrative intention that projects itself in a specific form of a fictional world in which the principle of synecdoche becomes a vanishing point.

Synecdoche in Čep's works has yet another dimension. It is manifested in author's intention that is realized in radical modification of the first two short story collections and their inclusion in *Zeměžluč* [Centaury – book] (see Table 1). Let us finally try to clarify this aspect also with the use of a quantitative analysis of thematic words. We will specifically compare courses of thematic words in the second editions of *Dvoji domov* [Double Being], *Vigilie* [The Vigil] and the section of *Zeměžluč* [Centaury – part] and subsequently compare the results with the complete anthology of *Zeměžluč* [Centaury – book]. The comparison of frequency courses of thematic words in the first editions of *Dvoji domov* [Double Being], *Vigilie* [The Vigil], and *Zeměžluč* [Centaury – part] has shown how that namely the first positions are notably fluctuating. The following table shows how the relations between individual frequency courses changes in case of the second edition of *Dvoji domov* [Double Being] and *Vigilie* [The Vigil]. Both second editions will be compared with the section of *Zeměžluč* [Centaury – part].

In the table 8, relative frequencies (Rf) of thematic words (in the cells) and of autosemantic lexemes occurring below the h-point (out of cells) are stated. The black horizontal line in the second diagram marks the limit of h-point in corpuses *Dvoji domov* [Double Being] 1931 and *Vigilie* [The Vigil] 1931.

¹³ The significance of the principle of synecdoche in Čep's works has recently been emphasized in the monograph *Dvoji domov Jana Čepa* [Two Homes of Jan Čep] by Tomáš Kubiček (Kubiček 2014, p. 31).

Table 8. Reference table and diagram of Čep (sub)corpus

<i>r</i>	Double Being, 1931 (Rf)	The Vigil, 1931 (Rf)	Centaury - part, 1931 (Rf)	Centaury - book, 1931 (Rf)
1	5174,29	4259,05	4464,29	3816,14
2	3812,64	2881,12	2817,27	2414,73
3	2995,64	2755,86	2600,55	2306,93
4	2859,48	2693,22	2470,53	2306,93
5	2723,31	2442,69	2340,50	1940,41
6	2587,15	2317,42	2253,81	1854,17
7	2450,98	2254,79	1820,39	1854,17
8	2995,64	2192,16	1777,05	1832,61
9	2859,48	2066,89	1777,05	1767,93
10	2723,31	2066,89	1733,70	1746,37
11	2587,15	1941,63	1733,70	1703,25
12	2450,98	2254,79		1617,01
13		2192,16		1573,89
14		2066,89		1466,09
15		2066,89		1401,41
16		1941,63		

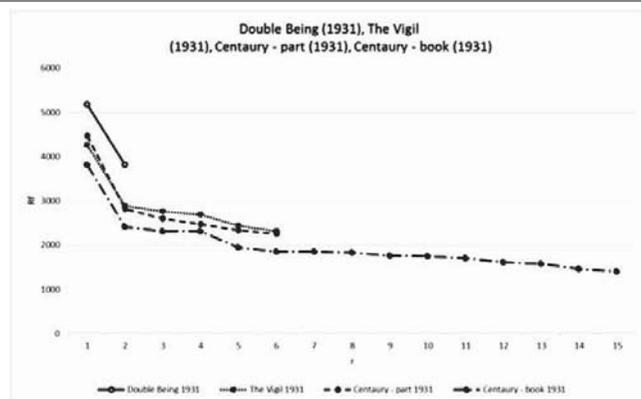


Chart 11. Comparative values of relative frequencies-curves of the second edition of the first two Čep’s novels and Centaury – part (1931) and Centaury – book (1931); the values above the h-points (the values are in Table 6)

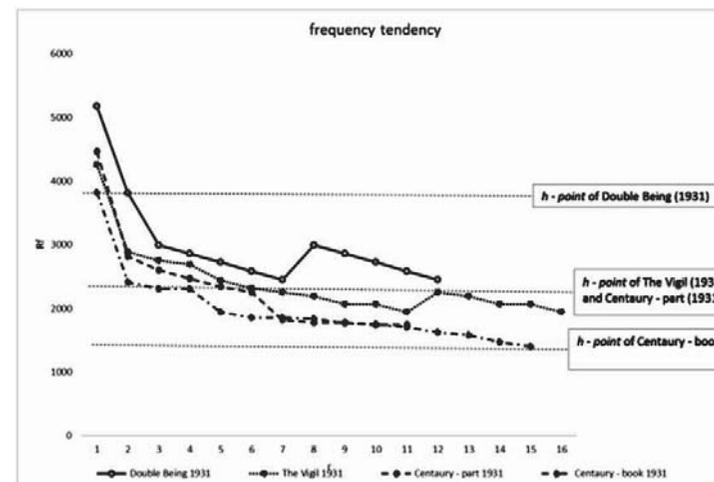


Chart 12. Comparative values of relative frequencies-curves of the second edition of the first two Čep’s novels and Centaury – part (1931) and Centaury – book (1931); the values above and below the h-points (the values are in Table 6)

Frequency courses in thematic fields in the second editions of *Dvoji domov* [Double Being] 1931 and *Vigilie* [The Vigil] 1931 tend to be analogous which is true also in respect to the section of *Zeměžluč* [Centaury – part]. In the second diagram the tendency is even more obvious due the inclusion of autosemantic lexemes below the *h-point*. From this perspective, mutual relations between individual parts of the anthology *Zeměžluč* [Centaury – book] seem more distinctly “harmonized” than the relations between the first editions of *Dvoji domov* [Double Being] and *Vigilie* [The Vigil] and the section of *Zeměžluč* [Centaury – part].

II Conclusion

Radical modification of the first two books lead to the publication of the book titled *Zeměžluč* [Centaury – book] 1931 in which short stories from the first collection were included together with the last section that has the same title as the whole anthology. All part of the

newly edited anthology thus form a well balanced composition. We can assume that modifications in the second editions of *Dvojí domov* [Double Being] and *Vigilie* [The Vigil] were carried out in the range of the structural-intentional field of the section *Zeměžluč* [Centaury – part] and that a real reason why Čep modified the original versions of the first two collections does not have to be the same as the one stated by Kubíček in his monograph on Čep: “It is interesting that for the second edition of *Dvojí domov* [Double Being] Čep carefully selects from the stories included in the first edition and finally publishes only seven of them. (R. Z. – In fact Čep only chose six stories of the original collection.) This decision will later be respected by Bedřich Fučík and Mojmír Trávníček (R. Z. – editors of Jan Čep’s works.) while preparing the edition of collected Čep’s works. The reason for exclusion of several stories will always be Čep’s fear that his short stories might be seen as too sentimental or shallow.” (Kubíček 2014, pp. 39–40). Namely a structural-semiotic radius of *Zeměžluč* [Centaury – part] influenced the modification of *Druhý domov* [Double Being] and *Vigilie* [The Vigil] in a significant, or even fundamental way. The alternations of the first two Čep’s books were carried out in accord with the intention of this field since the author intended to join the first collections with the section titled *Zeměžluč* [Centaury – part] in a single volume. In the final part of an autobiographical essay *Sestra úzkost* [The Sister Anxiety] 1975 that was written by Čep in exile, the author remarks: “I did not write all books I could have written and those that I wrote I could have written better. Particularly I have not written **the book** I should have written, the one I will never write.” (Čep 1975, p. 125). Jan Čep was an author who built motivic and thematic invariants. One could even say that his lifelong intention was to write and refine one and the same *book* that he longed for perfect expression and artificial treatment of a central idea that frames the eternal issue of the relation between a man, light, and the transcendent. Čep intended to transform his own concept of the world that is strongly permeated with Christian belief into an artistic form, or to realize reality through artistic form. In Čep’s works, religious belief and literary work are

mutually interconnected. In this context, also the notion of synecdoche receives a new meaning. As a unifying principle, it is projected not only in the fictional world of Čep’s prose, but also on the level of the relation between the author, his work and belief.

Translation by Josef Línek

References

- Čech R. – Glogarová-Davidová J. – David J., 2013, *Analýza tematické koncentrace textu – komparace publicistika Ladislava Jehličky a Karla Čapka*. [The Thematic Concentration Analysis of Text – Ladislav Jehlička’s and Karel Čapek’s journalism comparison.], „Slovo a slovesnost” 74, pp. 41–54.
- Čech R. – Glogarová-Davidová J., 2013, *Tematická koncentrace textu – některé aspekty autorského stylu Ladislava Jehličky*. [The Thematic Concentration of Text – Choose Aspects of Ladislav Jehlička’s authorship.], „Naše řeč” 96, pp. 234–245.
- Čech R. – Popescu I.I. – Altmann G., 2014, *Metody kvantitativní analýzy (nejen) básnických textů*. [The Methods of Quantitative Analysis (not only) in Poetic Texts.], Olomouc: Palacky University.
- Čep J., 1926, *Dvojí domov*. [Double Being.] Prague: Ladislav Kuncíř.
- Čep J., 1928, *Vigilie*. [The Vigil.] Prague: Plejada.
- Čep J., 1931, *Zeměžluč*. [Centaury.] Prague: Publishing House Melantrich.
- The Czech National Corpus – SYN2010. Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Prague 2010. Retrieved from: <http://www.korpus.cz>.
- The Czech National Corpus – SYN2010BEL. Institute of the Czech National Corpus, Faculty of Arts, Charles University in Prague. Prague 2010. Retrieved from: <http://www.korpus.cz>.
- Kubíček T., 2014, *Dvojí domov Jana Čepa*. Brno: Host.
- Změlík R., 2015, *Kvantitativně-korpusová analýza a literární věda: model a realizace autorského korpusu a slovníku Jana Čepa v kontextu zahraniční*. Olomouc: Univerzita Palackého v Olomouci.
- Změlík R., 2016, *Potentiality of Quantitative and Corpus Analysis to Literary Studies – Toward Methodology (The Analysis of Thematic Fields)*. „Bohemistyka”, nr. 3, pp. 203–219.