„Bohemistyka” 2016, nr 3, ISSN 1642–9893

Richard ZMĚLÍK
Wrocław

Potentiality of Quantitative and Corpus Analysis to Literary Studies – Toward Methodology (The Analysis of Thematic Fields)

Keywords: Jan Čep, thematic concentration of a text, quantitative and corpus linguistics and theory of literature, model of Jan Čep authorial vocabulary

Abstract

The first part of presented study is focused on quantitative and corpus analysis of the first three collections of short stories by the Czech prose writer Jan Čep. Using the definition of so-called thematic field we try to demonstrate how this methodology can be used in the context of literary theory. The study is intended mainly to present the possibilities that contemporary quantitative and corpus analysis offer to the theory of literature. Both disciplines have in fact so far developed without noticeable mutual contacts. Special model of Jan Čep authorial corpus and its analysis should prove potentiality of these primarily linguistic methods also in the theory of literature.

I Introduction (of the First Part)

The present study falls into a relatively new, progressively developing sphere of literary studies which involves the utilisation of quantitative analysis, conducted in the context of modern electronic corpora. Since the study contains longer frequency lists of lemmas and a number of graphs, it was necessary to divide it into two parts rather than publishing it as one long text. In the first part, the present author introduces the core of the research project involving a quantitative-corpus analysis of the prose work of the author Jan Čep, which was conducted between 2012 and 2015 at Palacký University in Olomouc, Czech Republic; moreover, the author presents the chosen methodological approach, that is, an analysis of the thematic concentration of the text, and also prepares the results of this analysis for the subsequent literary analysis and interpretation, which will be included in the second part of the study.

Quantitative linguistics and corpus linguistics are primarily focused on exact study of natural language. These disciplines concentrate namely on objective analysis, taxonomization and interpretation of partial aspects of the language system, its individual structural layers and components. Findings concerning quantitative stratification, distribution and combination of the system components enable them also to view the language in an unbiased way as a system complex of structurally interconnected phenomena. Using exact methods of language analysis is typical for quantitative linguistics that became influential namely in the second half of the 20th century. Before modern language corpuses were compiled, quantitative linguists had often used fictional texts as reliable language sources.[1] Nevertheless, it was the appearance of electronic corpuses that brought about real progress in exact study of language since the corpuses not only offered linguists the possibility to study language with the use of a vast body of texts but also enabled them to pay attention not solely to the language of fiction, but also to different functional areas of national languages. Thus the knowledge of a language becomes more plastic and it reflects the real character of the language more accurately. In spite of all the success and progress in this area of research we cannot expect that any of recent advanced corpuses could reflect all aspects of a national language truly. Nevertheless, it is a fact that advancement of technology is accompanied by new efforts in the field of exact analysis of the linguistic material, by formation and ceaseless improvements of sophisticated language corpuses and also by advancements of methodology. Here a question arises whether methods and procedures developed primarily for linguistic research can be employed meaningfully also in the context of literary theory. We assume that possibilities of modern

[1] For example Laslie Hancock: *Word Index to James Joyce ´s Portrait of the Artist* (1967), Jitka Štindlová: *Konkordanční a frekvenční index k Slezským písním Petra Bezruče* (1969), František Čermák et al.: *Slovník Karla Čapka* (2007), *Slovník Bohumila Hrabala* (2009).

language corpuses and namely parameters of certain statistic methods have an interesting potential for the use in the field of literary theory even if their application will be liable to specific criteria. In other words, procedures and results of quantitative and corpus analysis must be functionally connected to particular methodological aspects of literary theory.[2]

II Goals of research

In this part of study, we intend to demonstrate one such viable approach that pertains measuring of so-called *thematic concentration of a text*. The method, theoretically formulated by Ioan-Iovitz Popescu,[3] was tested on the works of some Czech authors.[4] Nevertheless, we will not only apply this method, i.e. demonstrate the analytical procedure itself; first of all we will ask what possibilities the method offers for theory of literature and what use it can have in the context of this discipline. Such question has not been asked yet. Its significance seems to exceed the confines of the selected partial procedure, that is of measuring of thematic concentration of texts; the question provokes more general reflection over the possibility of using exact methods, so far applied exclusively by modern quantitative and corpus linguistics, in literary theory (and not solely in linguistics), and over the extent to which such approach can enhance research in literary theory.

We use prosaic works of Jan Čep[5], a Czech author, as the material on which we intend to demonstrate the procedure. For this study we chose only a part of Čep‘s literary texts, namely the first three collections of his short stories: *Dvojí domov* [Double Being] (1926), *Vigilie*

[2] This relation was ushered in by the methodological base of corpus linguistics that had grown from structural foundations.

[3] Professor of Physics at Bucharest University.

[4] Davidová-Glogarová – David – Čech (2013), Davidová-Glogarová, Čech (2013), David – Čech – Davidová-Glogarová – Radková – Šústková (2013).

[5] Jan Čep (1902–1974), a Czech writer of Catholic orientation. In 1948, he emigrated to France where he found permanent residence.

[The Vigil] (1928) a *Zeměžluč* [Centaury] (1931). These books do not represent solely a particular chronological line in the initial phase of Čep‘s literary activity, their mutual relations are much more complex. *Zeměžluč* contains three parts the first two of which are largely rewritten versions of the other two titles.

Table 1. Structure of the first three books by Jan Čep

Double Being (1926)	The Vigil (1928)	Centaury (1931)		
		Double Being	The Vigil	Centaury
The Little House	Parting	The Little House		The Old Man’s Laughter
The Storm	The Astray			The Sorrow of Love
The Death of Shoemaker Nerušil	Rozárka Lukášová		Rozárka Lukášová	The Jolly Funeral
The Revolt	The Goose Herder		The Goose Herder	The Moth
Double Being	Who Will Be Victorious	Double Being	The Diligent Family	The Archaic
Kozlovice	The Green Sparks	Kozlovice		Lucie Laurová
The Conqueror	The Vigil		The Vigil	Albina Drůzová
Little Justine	The Epilogue			On the Way to the Morning Mass
The Purse		The Purse		Phantoms
To the Town		To the Town		
Delusion				
The Quail				
The Elegy		The Elegy		

In case of the first Čep‘s collections of short stories one must not take into consideration only temporal sequence (dates of the first edition of the collections). The situation must be viewed as a holistic phenomenon (event). The author returned to his previous texts, after some time he newly reflected the first two collections and he did not hesitate to include them, in a considerably rearranged form, in a new book titled *Zeměžluč* [Centaury]. The first Čep‘s prosaic titles manifest both sequence and returnability of time. Thus they form a closely linked up whole.

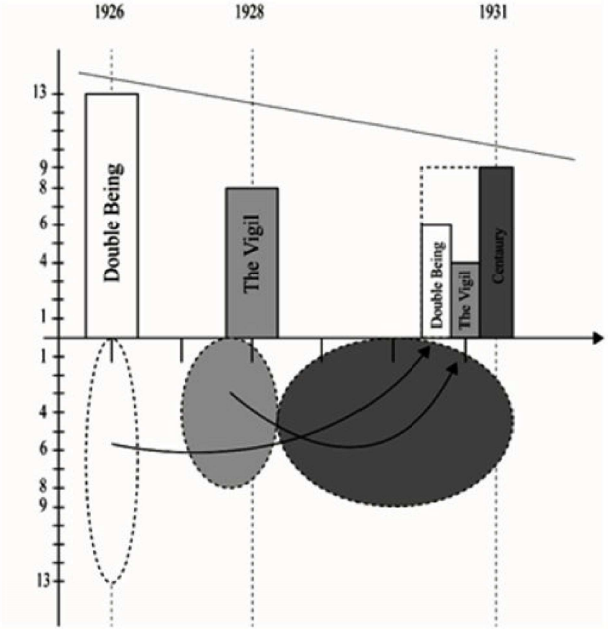


Chart 1. Chronological layout of Čep‘s first collections of short stories. The lower area of the chart shows time intervals of formation of individual collections, so-called relative chronology

One of the questions literary theory will ask pertains mutual relations between individual parts of this early phase of Čep‘s literary activities. What made author exclude numerous short stories from the

second edition of *Dvojí domov* [Double Being] and *Vigilie* [The Vigil]? We will surely want to know how individual versions of each book and how different collections of short stories are related, or what the relation between these Čep‘s books and books of other authors who published in the same period of time really is. In this study we will concentrate on the relations between the first editions of the books *Dvojí domov* [Double Being], *Vigilie* [The Vigil] and section *Zeměžluč* [Centaury – part] from the collection of short stories of the same title, the relation between the first and the second edition of *Dvojí domov* [Double Being] and *Vigilie* [The Vigil], and finally on the relations between individual sections of the book titled *Zeměžluč* [Centaury].

Analysis of the above specified relations that we will take into consideration while assessing the overall character of Čep‘s early collections of short stories will be carried out with the use of the data that we obtained from a newly compiled special Jan Čep authorial corpus that currently includes all Čep‘s prosaic works.[6]

We chose a strategy derived from so-called relative chronology that registers relative course of formation of individual collections of shorts stories so that we could interpret obtained statistic and corpus values in a relevant way which in our case primarily means with respect to literary theoretical aspects.

Relative chronology (see chart 2) make us aware of origination and temporal location of individual texts and short-story collections. It displays the situation not in a dot chart (A axis) but in phases (B axis). Unlike A axis, the model on B axis can show us the relations between individual areas of Čep‘s literary production. First of all it enables us notice similarities and differences of text groups that are overlapping and verify given situations in corpus results with the use of statistic calculations and measurements. Chart 2 clearly shows that the state of

[6] In future, a complete Čep authorial corpus will be realized that will include not only Čep‘s fiction, but also his essays, correspondence, diaries and minute publicistic text of numerous genres.

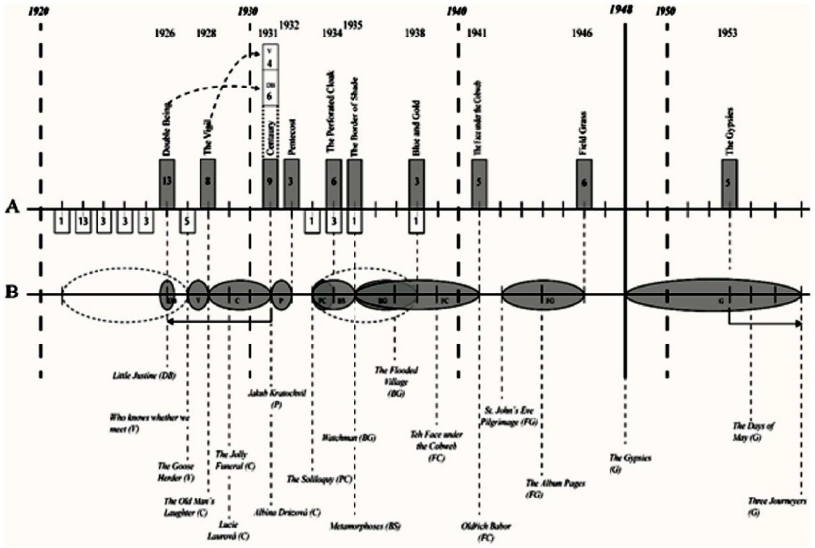


Chart 2. *Overall situation of Čep's prosaic books with relative chronology marked on B axis.*

author's literary production modelled in this way makes it possible to ask numerous partial questions and also to view the overall situation as a non-linear creative process. Our approach includes two aspects of viewing the analysed material:

1. external criterion arising from the temporal characteristics;
2. internal criterion, established by structural relations between the texts.

III Methodology

As the scheme suggests, analysis can be carried out not only with regard to the author's complete prosaic work but also in the context of partial phases of its development. Naturally only an all-embracing analysis of all author's works can lead to more general and complex

conclusions. This fact however in no way disqualifies partial tests and analyses that must be carried out in order to complete the final relevant image. At this moment we will perform only a partial but sufficiently demonstrative insight in the authorial corpus that will focus on the first three collections of Čep's short stories. So-called measuring of thematic concentration of a text has already been mentioned as one of the methods suitable for analysis of the corpus material with respect to its potential for literary theoretical research. In our treatise, we will attempt to use the method and the results it leads to in the context of problems primarily inspired by literary theoretical criteria.

To determine thematic concentration of a text, one must first specify so-called *h-point* that that forms a barrier between two parts of quantitatively defined lexicon.

The h-point can be defined as the point at which the straight line between two (usually) neighbouring ranked frequencies intersects the y = x line. Solving two simultaneous equations we obtain the definition

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

In other words, the h-point is that point at which r = f(r). If there is no such point, one takes, if possible two neighbouring f(i) and f(j) such that f(i) r_i and f(j) r_j. Mostly $r_i + 1 = r_j$ (Popescu – Mačutek – Altmann, 2009, p. 24)

Following table represents only a part of the complete frequency list of the Čep's first edition of *Dvojí domov* [Double Being] (1929). The table has three columns in which ranking (*r*), lemma and frequency (*Af*) are stated.[7]

[7] Absolute frequency (*Af*) – total of all forms a particular word.

Table 2. The first 60 lemmas from Čep‘s Druhý domov [Double Being] (1926) arranged in descending order according to absolute frequency (Af)

r	lemma	*Af*	*r*	lemma	*Af*
1	a	1010	31	o	57
2	se	882	32	moci	53
3	být	582	33	matka	53
4	na	361	34	ty	53
5	on	264	35	okno	52
6	v	252	36	stát	52
7	s	205	37	u	51
8	do	193	38	den	50
9	ten	165	39	otec	50
10	ale	137	40	co	50
11	ona	134	41	nad	50
12	z	133	42	před	50
13	k	121	43	Rudolf	47
14	že	115	44	od	46
15	za	112	45	pak	45
16	jít	109	46	Ludvík	45
17	jako	105	47	ještě	45
18	svůj	104	48	tak	44
19	po	98	49	vidět	44
20	už	97	50	já	43
21	oko	93	51	pod	43
22	mít	88	52	muset	43
23	oni	85	53	jeho	43
24	jenž	80	54	cesta	42
25	který	76	55	žena	41
26	hlava	75	56	i	41
27	jak	69	57	Jeník	40
28	když	68	58	člověk	40
29	však	63	59	její	39
30	ruka	58	60	všechen	38

In this case, to specify the *h-point* we can apply the first rule that is valid if there exists an identical correlation between *r* and *Af*. Here thus $h = 45$. In the second corpus of *Vigilie* [The Vigil] (1928), the situation is different.

Table 3. The first 60 lemmas from Čep‘s Vigilie [The Vigil] (1928) arranged in descending order according to absolute frequency (Af)

r	lemma	Af	r	lemma	Af
1	a	1111	31	oko	66
2	se	886	32	cesta	61
3	být	665	33	nad	58
4	v	381	34	její	54
5	na	338	35	tvář	53
6	s	223	36	chvíle	49
7	on	212	37	všecek	49
8	ten	191	38	jeho	49
9	jenž	183	39	člověk	48
10	do	164	40	stát	48
11	z	144	41	až	47
12	že	144	42	pak	47
13	jako	144	43	od	47
14	svůj	136	44	pod	46
15	k	128	45	před	46
16	za	122	46	země	45
17	ona	110	47	cítit	44
18	ale	110	48	však	44
19	už	106	49	co	44
20	který	101	50	moci	43
21	po	90	51	ještě	42
22	tak	88	52	bez	41
23	já	82	53	matka	41
24	mít	77	54	u	41
25	když	77	55	ruka	40
26	jak	75	56	starý	40
27	jít	74	57	den	40
28	oni	72	58	hlava	38
29	tento	69	59	všechen	38
30	Rozárka	68	60	jen	37

In this case *r* does not equal *Af*. That is why we use an alternative calculation:

$r_i = 45$ $\qquad$ $f(i) = 46$

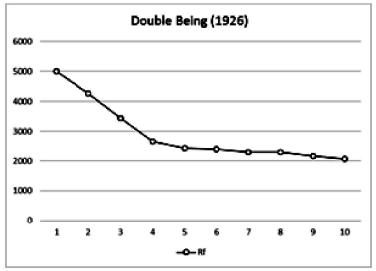
$r_j = 46$ $\qquad$ $f(j) = 45$

$$H = \frac{(46 \times 46) - (45 \times 45)}{(46 \times 45) - (46 \times 45)} = \frac{91}{2} = 45{,}5$$

We proceed analogously also in case of *Zeměžluč* [Centaury – part] (1936) and in case of all this file [Centaury – book]. We can see that above this point synsemantic words generally occur more frequently. In spite of that even some autosemantic words can be found in the same area. The occurrence of these autosemantic words *above* the h-point is – with regard to their natural distribution that is concentrated in the area *below* the h-point – symptomatic. If we compile an individual list of autosemantic words, we will obtain following results.

Table and **chart 4.** Thematic words in Čep's *Dvojí domov* [Double Being], a collection of short stories published in 1926

r	lemma	*Af*	*Rf*
1	to go	109	4999,31
2	eye	93	4265,47
3	head	75	3439,89
4	hand	58	2660,18
5	mother	53	2430,86
6	window	52	2384,99
7	father	50	2293,26
8	day	50	2293,26
9	Rudolf	47	2155,67
10	Ludvik	45	2063,94



Besides absolute frequency (*Af*) the table also hold values of relative frequency (*Rf*) which are necessary for due comparison of frequency of lexical units from corpuses of uneven size. Relative frequency is a standardized value the calculation of which is here determined with respect to parameters of this value standardly used in the National Czech Corpus (ČNK), where the relative frequency of a lexical unit (here a lemma) recounted to a million words (instances per milionem – i.p.m.)

$$Rf = Af \times \frac{1\,000\,000}{N}$$

N in the formula is the total of all lexical units in the corpus. In this specific case, *N* signifies the size of the corpus of Čep's *Dvojí domov* [Double Being] (1929).[8] We will proceed in the same way while compiling lists of thematic words from following corpuses: *Vigilie* [The Vigil] *1928*, *Dvojí domov* [Double Being] *1931*, *Vigilie* [The Vigil] *1931* and *Zeměžluč – part* [Centaury – part] *1931*, and *Zeměžluč – complete collection* [Centaury – book] *1931*. For better orientation we will subsequently transform numeral values of *Rf* in graphs.

Table and **chart 5.** *The Vigil (1928)*

r	lemma	*Af*	*Rf*
1	to go	74	3247,32
2	Rozárka	68	2984,03
3	eye	66	2896,26
4	road	61	2676,85
5	face	53	2325,79
6	while	49	2150,25
7	man	48	2106,37

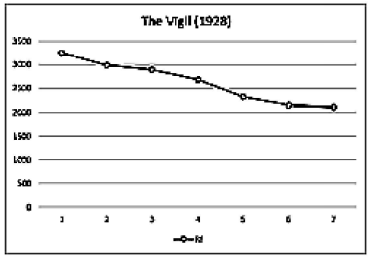
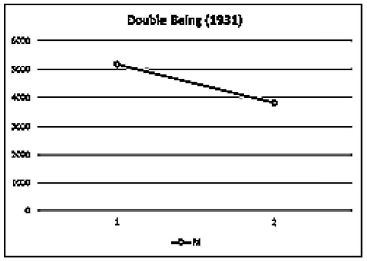


Table and **chart 6.** Double Being (1931)

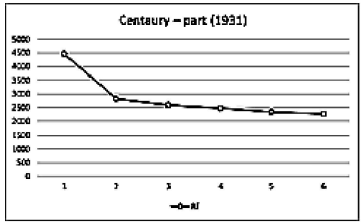
r	lemma	*Af*	*Rf*
1	to go	38	5174,29
2	eye	28	3812,64



[8] The corpus *Dvojí domov 1926* includes 21.803 positions in all. The complete authorial corpus of fictional narratives contains 511.026 positions.

Table and **chart 7.** Cenatury – part (1931)

r	lemma	*Af*	*Rf*
1	eye	103	4464,29
2	face	65	2817,27
3	head	60	2600,55
4	hand	57	2470,53
5	heart	54	2340,50
6	feel	52	2253,81



Note: Here we do not mean the complete collection titled *Zeměžluč* [Centaury – book] but only an independent part [Centaury – part] of the publication with the same title (see Table 1).

Table and **chart 8.** Centaury – book (1931)

r	lemma	*Af*	*Rf*
1	eye	177	3816,14
2	face	112	2414,73
3	hand	107	2306,93
4	head	107	2306,93
5	road	90	1940,41
6	man	86	1854,17
7	know	86	1854,17
8	feel	85	1832,61
9	day	82	1767,93
10	mother	81	1746,37
11	while	79	1703,25
12	heart	75	1617,01
13	old	73	1573,89
14	Rozárka	68	1466,09
15	world	65	1401,41

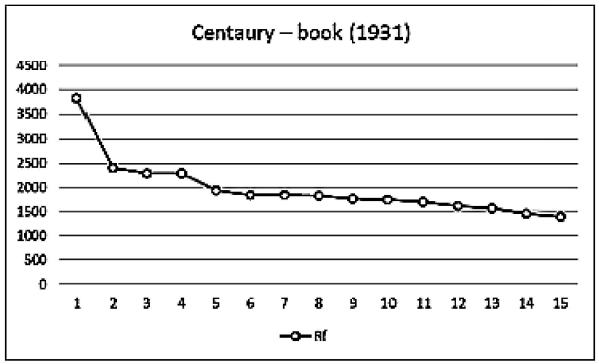
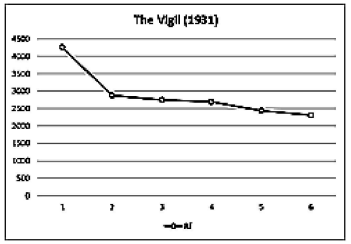


Table and **chart 9.** The Vigil (1931)

r	lemma	*Af*	*Rf*
1	Rozárka	68	4259,05
2	eye	46	2881,12
3	face	44	2755,86
4	to go	43	2693,22
5	mother	39	2442,69
6	hand	37	2317,42



These tables and graphs especially provide us with first comparisons based on the choice of thematic words and their relative frequency (*Rf*). There detailed interpretation will be carried out in due place (see chapter Assessment and interpretation) in the second part of the study. At this moment it is necessary to define main criteria that will interest us further on, the criteria that we will use in the advanced analysis of the thematic field and its elements (lemmas). They are namely these aspects:

1. Number of autosemantic lexemes in the area above the *h-point;*
2. Types of autosemantic lexemes;
3. Mutual semantic relations between them;
4. The range of the thematic field (DTF)

The selected four criteria form elementary parameters which enable us compare distributions of different thematic fields. We will use the possibility not only to compare the first editions of the analysed books, but also to contrast the second editions of *Dvojí domov* [Double Being], *Vigilie* [The Vigil], and the whole collection of *Zeměžluč* [Centaury – book].[9] There are numerous possible combinations to be found within the complete authorial corpus of fictional narratives (see Chart 2). One can for example ask what differences there are between creative periods of 1920s and 1930s or how areas forming significant intersections on the relative chronology axis are related.

Graphical illustration enables us to read frequency values in a more schematized way. To make their comparison even lucid we introduce yet another value that we call *dispersion of thematic field* (DTF). Its purposed to express the distance between two frequency values (the highest and the lowest) occupied by autosemantic lexemes above the *h-point*. A simple formula serves for the calculation:

$$\mathrm{DTF} = \frac{\mathrm{Rf}\,(y)}{\mathrm{Rf}\,(x)}$$

Rf (y) in the formula is the lowest value of relative frequency of an autosemantic lexeme above the *h-point*, while *Rf (x)* represents the highest value. The higher the value of DTF is, or rather the closer it is to 1, the smaller is the dispersion between individual autosemantic lexemes. The number defines the relation between extreme values of autosemantic lexemes above *h-point* and it specifies the character of the delimited area. If the DTF value drops we can expect uneven relations between thematic words which might, in comparison with their graphical course, indicate symptomatic character of particular lemmas of semantic classes and vice versa.

[9] Contextual characteristic of these lemmas is also important. It must be constantly observed in order not to interpret the meaning incorrectly.

Table 10. DTF values in individual subcorpuses

(sub)corpus	DTF
Double Being (1926)	0,41
The Vigil (1928)	0,65
Centaury - part (1931)	0,50
Double Being (1931)	0,74
The Vigil (1931)	0,54
Centaury - book (1931)	0,37

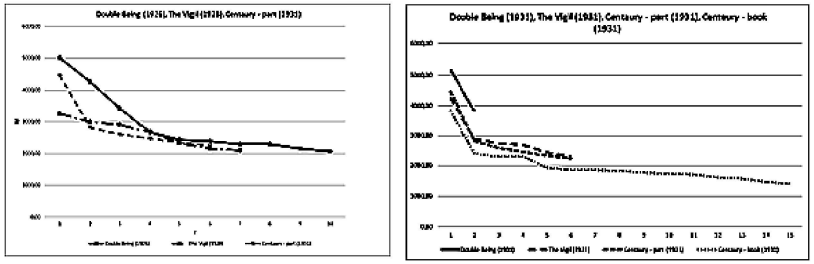


Chart 10. The comparison of two sets of (sub)corpuses

IV Conclusion (of the First Part)

The models of the conducted analysis testify to a change in distribution and stratification of the so-called thematic words (lemmas), which signalises structural changes in the (sub)corpora under study. This is obvious mainly in the comparative display of the first three versions of the short-story collections and a comparison of the second versions of the first two collections with the third collection. As can be observed in chart 3, the revised versions of *Dvojí domov* [Double Being] (1931) and *Vigilie* [The Vigil] (1931) show a change that can be described as a movement towards the stratification paradigm of

Zeměžluč [Centaury – part] (1931). The question remains what motivated this change and what its significance is for the work of Jan Čep. These issues will be addressed in the second part of the study.

Translation by Josef Línek

References

Čech R., Glogarová-Davidová J., David J., 2013, *Analýza tematické koncetrace textu – komparace publicistika Ladislava Jehličky a Karla Čapka* [*The Thematic Concentration Analysis of Text – Ladislav Jehlička´s and Karel Čapek´s journalism comparation*], *Slovo a slovesnost 74*, pp. 41–54.

Čech R., Glogarová-Davidová J., 2013, *Tematická koncetrace textu – některé aspekty autorského stylu Ladislava Jehličky* [*The Thematic Concentration of Text – Choose Aspects of Ladislav Jehlička´s autorship*], *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*.] [w:] 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.

Čermák F., Cvrček V. et al., 2009, *Slovník Bohumila Hrabala*. [*The Dictionary of Bohumil Hrabal*.] Prague: Publishing House Lidové noviny.

Čermák F. et al., 2007, *Slovník Karla Čapka*. [*The Dictionary of Karel Čapek*.] Prague: Publishing House Lidové noviny.

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.

Hancock L., 1967, *Word Index to James Joyces´s Portrait of the Artist*. London – Amsterdam: Southern Illionis University Press.

Popescu I. I., Mačutek J., Altmann G., 2009, *Aspects of Word Frequencies*. Lüdenscheid: RAM-Verlag.