

CORPUS OF DRAMAS BY THE ČAPEK BROTHERS FROM A QUANTITATIVE PERSPECTIVE

PETR POŘÍZKA¹ – JÁN MAČUTEK²

¹Department of Czech Studies, Faculty of Arts, Palacký University, Olomouc,
Czech Republic (ORCID: 0000-0001-6980-9148)

²Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
& Department of Mathematics, Faculty of Natural Sciences and Informatics, Constantine
the Philosopher University, Nitra, Slovakia (ORCID: 0000-0003-1712-4395)

POŘÍZKA, Petr – MAČUTEK, Ján: Corpus of Dramas by the Čapek Brothers From
a Quantitative Perspective. *Journal of Linguistics*, 2025, Vol. 76, No 1, pp. 378 – 389.

Abstract: This text presents the first quantitative analysis of the plays of the Čapek brothers, exploring the linguistic and stylistic differences between their individual and collaborative works. Utilizing computational methods and quantitative approaches, it analyses a corpus of ten plays, focusing on the distribution and proportion of parts of speech in both dialogue and stage directions. The analysis reveals significant stylistic differences: Josef Čapek is characterized by a descriptive language rich in nouns with fewer words overall, while Karel Čapek uses a more dynamic approach with a predominance of verbs. Cluster analysis shows that Josef's dramas form a separate, distinct group when both dialogue and stage directions are considered, with stage directions showing particularly marked differences. Morphological coefficients, including the noun coefficient (Kn) and Busemann coefficient (B), quantitatively confirm these stylistic differences, with Josef's plays showing extreme values that indicate high descriptive saturation, especially in the stage directions. This structural analysis not only provides quantitative evidence of different authorial styles, but also lays a foundation for future research.

Keywords: computational literary studies, drama, CapekDraCor, Karel Čapek, Josef Čapek, quantitative analysis, parts of speech

1 INTRODUCTION AND STATE OF THE ART

Karel Čapek and his brother Josef Čapek are among the key figures of 20th century Czech literature. Both brothers contributed significantly to the development of drama, both in their individual works and in those they wrote together. The dramatic work of the Čapek brothers comprises ten plays (see below for details), which have been extensively discussed in the academic literature and analysed from various perspectives, including theatrical, literary, and linguistic. However, previous scholarly studies have focused on such aspects of their plays as the poetics, compositional or narrative aspects of their work (Sunbee 2011; Novák 2013; Doležel 2014), textual adaptations (e.g. Janáček's libretto in the opera *The Makropulos*

Affair, cf. Křupková 2008), the relationship between the literary version and the film adaptation (e.g. *The White Disease*), translational and theatrical aspects, or reflections on the dramatic work in a broader social and cultural context – the ethical or philosophical aspects of the work and its influence on the literary works of other authors (Janiec-Nyitrai 2012; Droženová 2020). In general, these are analyses and interpretations based on introspection, readerly reception and a traditional structuralist literary scholarly approach (Holý 1984, 2014; Janoušek 1989, 2018).

Rarely, there are also investigations using computer-assisted text analysis or partial quantitative analyses. However, all of them concern only selected plays by Karel Čapek, e.g. using a semi-automatic phrase recognition tool in *R.U.R.* and *The White Disease* (Kováříková and Kopřivová 2012), quantitative analysis of proper nouns in the plays (Pořízka 2023b) or, more often, deal with other genres of his work. The works and contribution of his brother Josef have been largely overlooked, especially in terms of quantitative analysis.

The Karel Čapek Dictionary (*Slovník Karla Čapka*, Čermák 2007b), published by the Institute of the Czech National Corpus (CNC) in 2007 and based on the *capek* corpus (see below), includes the chapter *Statistical Aspects of Karel Čapek's Language, Especially His Lexicon* (Cvrček et al. 2007). It presents statistical data on Čapek's lexicon, parts-of-speech ratio (POS), and lexical richness. However, the authors themselves admit inaccuracies in the quantitative indices used, which are distorted by the length of the text (Cvrček et al. 2007, p. 675).

Previous quantitative analyses of Karel Čapek's works have focused on exploring thematic text concentration, lexical compactness across genres, and the use of selected lexical-statistical indices, such as average token length, verb distance, and vocabulary richness (Davidová et al. 2013; Čech 2015; Kubát 2016; Mačutek et al. 2016). However, these papers consistently exclude the plays of the Čapek brothers, leaving the multilayered textual structure of the plays largely unexplored.

2 CORPUS AND DATA PROCESSING

In terms of the textual structure, plays are multi-layered. This structure includes primarily the character dialogues in a form similar to spoken dialogue, with character labels (proper nouns) preceding each line of text, then structuring words (act, scene, drop-scene), comments (stage, authorial, on the characters' actions), and possibly other sections such as the author's introductory metatextual notes (foreword) and a list of characters (cast list).

There is currently a *capek* corpus (Čermák et al. 2007a) in the Czech National Corpus (CNC), which includes plays by Karel Čapek, but not those by his brother Josef. In addition, this corpus has limitations due to the way the source texts are processed, making it unsuitable for quantitative analyses. The *capek* corpus does not reflect the multi-layered structure of the plays, as the aforementioned textual

and metatextual parts are not separated. This lack of segmentation makes it impossible to analyse the subparts of the plays and can significantly affect subsequent quantitative analyses and their results, as we have shown in a previous study (Pořízka 2023b).

For these reasons, and for the purpose of different types of quantitative analyses, we have recently created a new corpus in two versions, which contains all the plays by the Čapek brothers and reflects the annotation of different (meta)text layers. The first version is made as a standard corpus including also linguistic annotation (lemmatization, morphological tagging) and is available in the *SketchEngine* tool (cf. *Czech Drama Corpus* in *DraCor Drama Corpora*: <https://www.sketchengine.eu/dracor-drama-corpora/>).

The second version of this database called *CapekDraCor* (soon to be publicly available) which we used for this quantitative analysis focusing on the comparison of character dialogues and metatextual comments, was created specifically for the international *DraCor* project (<https://dracor.org/>) and its tools.

The *CapekDraCor* corpus used in the analysis consists of the following texts:

- plays by Karel Čapek: *Loupežník* ('The Outlaw', 1920); *R.U.R.* (1920); *Věc Makropulos* ('The Makropulos Affair', 1922); *Bílá nemoc* ('The White Disease', 1937); *Matka* ('The Mother', 1938);
- plays by Josef Čapek: *Země mnoha jmen* ('The Land of Many Names', 1923);
- plays written together by the Čapek brothers: *Lásky hra osudná* ('The Fateful Game of Love', 1910); *Ze života hmyzu* ('The Insect Play', 1921); *Adam Stvořitel* ('Adam the Creator', 1927).

The data are processed in a standardized format based on XML and general TEI guidelines for processing drama, with a defined basic drama tagset. A more detailed description of the text processing, information about the TEI-XML format and other technical aspects including illustrative examples can be found in (Pořízka 2023a).

3 QUANTITATIVE ANALYSIS AND INTERPRETATION

3.1 Basic word ratios

The two brothers are known for their different approaches to the language of their plays: while Karel Čapek used contemporary language and a more colloquial style, Josef Čapek used a bookish, even archaic style. Because of these differences, we have divided their dramas into three groups: (1) dramas by Karel Čapek, (2) dramas by Josef Čapek, and (3) dramas written by the two brothers together in order to compare these collections in terms of the composition and structure of the plays (authorial style). Using the TEI/XML data format, each play was also divided into two subgroups: (1) character dialogues (the drama itself) and (1) stage directions.

For the purposes of this quantitative analysis, additional metatextual sections were excluded, i.e. structuring words (act, scene, drop-scene), cast list, preface, and character labels (proper names) preceding individual lines of dialogue.

drama	author(s)	word proportions in the stage directions
Bílá nemoc	Karel	0.094
Loupežník	Karel	0.107
Matka	Karel	0.098
R.U.R.	Karel	0.120
Věc Makropulos	Karel	0.089
Adam stvořitel	co-authored	0.098
Lásky hra osudná	co-authored	0.093
Ze života hmyzu	co-authored	0.112
Gassirova loutna	Josef	0.046
Země mnoha jmen	Josef	0.061

Tab. 1. Word ratios in the stage directions of the Čapek brothers’ plays

The basic word ratios or proportions (see Tab. 1) in the stage directions divide the ten plays into two groups. Two of Josef Čapek’s plays (*Gassirova loutna*, *Země mnoha jmen*) clearly have the lowest proportions, while the three co-authored dramas do not differ from Karel Čapek’s dramas in this respect.

3.2 Parts of speech proportions

The texts were linguistically annotated (lemmatization and morphological tagging via the *MorphoDiTa* tool (Straková et al. 2014)), using the new *LexaMorf* tool (Pořízka 2025), and the frequency distributions of word classes (parts of speech, hereafter POS) were calculated for both the stage directions and the characters’ dialogues of each play. The individual POS categories correspond to the standard classification in Czech, e.g. according to the so-called *Academic Grammar of Czech (Mluvnice češtiny 2)*. The results are shown in the following tables. Since the plays differ in length (as measured by the number of tokens), we relativize the proportions. Absolute frequencies and percentages of parts of speech can be found in the Tab. 2 – Tab. 5:

POS	Loupežník		R.U.R.		Věc Makropulos		Bílá nemoc		Matka	
	rfp	f	rfp	f	rfp	f	rfp	f	rfp	f
adjectives	4.21	543	6.31	927	5.45	745	6.67	935	4.78	729
adverbs	9.79	1262	8.86	1302	9.32	1274	8.83	1238	9.92	1505

conjunctions	7.29	940	6.91	1015	7.47	1021	7.76	1088	7.33	1112
interjections	1.91	246	1.59	233	1.16	159	0.39	55	0.54	82
nouns	17.29	2228	23.10	3393	20.99	2870	22.15	3105	18.34	2784
numerals	0.52	67	1.37	201	1.78	243	1.30	182	0.99	150
particles	2.72	351	2.38	350	2.81	384	3.60	505	2.51	381
prepositions	5.17	666	5.49	806	5.38	736	5.86	822	5.63	854
pronouns	23.30	3003	18.36	2697	20.86	2852	19.68	2760	23.56	3576
verbs	27.78	3580	25.64	3766	24.79	3389	23.76	3331	26.40	4006

Tab. 2. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the dialogues of Karel Čapek's plays

POS	Lásky hra osudná		Ze života hmyzu		Adam stvořitel		Země mnoha jmen		Gassirova loutna	
	rfp	f	rfp	f	rfp	f	rfp	f	rfp	f
adjectives	6.67	453	7.08	777	5.93	980	8.23	916	6.69	280
adverbs	8.93	606	8.66	950	9.00	1487	9.88	1100	8.53	357
conjunctions	9.31	632	6.02	660	8.16	1347	8.37	932	8.96	375
interjections	0.85	58	3.12	342	1.21	200	1.25	139	0.93	39
nouns	22.55	1531	22.51	2470	17.41	2876	22.87	2545	24.74	1036
numerals	0.71	48	2.17	238	0.84	138	0.78	87	1.03	43
particles	2.28	155	2.77	304	2.71	448	2.16	240	2.34	98
prepositions	5.85	397	4.86	533	4.44	734	5.91	658	6.11	256
pronouns	18.59	1262	19.49	2138	23.04	3805	18.57	2067	15.64	655
verbs	24.26	1647	23.33	2559	27.25	4500	21.98	2446	25.03	1048

Tab. 3. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the dialogues of plays by the Čapek brothers and Josef Čapek

POS	Loupežník		R.U.R.		Věc Makropulos		Bílá nemoc		Matka	
	rfp	f	rfp	f	rfp	f	rfp	f	rfp	f
adjectives	2.58	41	6.68	124	5.01	69	5.61	85	7.18	124
adverbs	5.16	82	5.77	107	6.03	83	4.36	66	4.87	84
conjunctions	5.03	80	3.77	70	3.78	52	3.63	55	6.2	107
interjections	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0
nouns	34.03	541	33.19	616	32.17	443	36.53	553	33.14	572
numerals	0.57	9	0.92	17	0.36	5	1.65	25	1.39	24
particles	0.31	5	0	0	0.15	2	1.19	18	0.17	3

prepositions	16.29	259	13.58	252	14.16	195	13.21	200	14.31	247
pronouns	8.74	139	9.21	171	9.08	125	8.78	133	9.97	172
verbs	27.30	434	26.89	499	29.27	403	25.03	379	22.77	393

Tab. 4. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the stage directions of Karel Čapek's plays

POS	Lásky hra osudná		Ze života hmyzu		Adam stvořitel		Země mnoha jmen		Gassirova loutna	
	rfp	f	rfp	f	rfp	f	rfp	f	rfp	f
adjectives	4.33	28	6.23	90	7.52	141	9.49	69	8.00	16
adverbs	9.43	61	4.22	61	5.55	104	5.91	43	2.50	5
conjunctions	5.87	38	4.15	60	4.54	85	3.58	26	5.00	10
interjections	0.00	0	0.14	2	0.00	0	0.00	0	0.00	0
nouns	34.47	223	26.92	389	31.27	586	43.74	318	45.00	90
numerals	1.08	7	1.59	23	1.81	34	2.2	16	5.50	11
particles	0.15	1	0.07	1	0.16	3	0	0	0	0
prepositions	10.66	69	13.84	200	12.33	231	9.9	72	10.50	21
pronouns	10.36	67	11.56	167	11.63	218	8.25	60	3.50	7
verbs	23.65	153	31.28	452	25.19	472	16.92	123	20.00	40

Tab. 5. Relative frequencies in percentage (rfp) and frequency (f) of parts of speech in the stage directions of plays by the Čapek brothers and Josef Čapek

Even in the case of the word classes, focusing on the proportions of the parts of speech, we find the same pattern, i.e. works divided into two groups: (1) on the one hand, two plays by Josef, (2) on the other hand, other dramas (plays by brother Karel and co-authored plays). Note: There were four words for which the software did not determine the POS value; these words were not considered.

3.3 Cluster analysis

We then performed a cluster analysis to look more closely at the relationships between the POS frequency distributions. Each drama is represented by a 20-dimensional vector whose coordinates represent the percentage of parts of speech in the dialogues (the first ten coordinates) and in the stage directions (the next ten coordinates). The results of the hierarchical cluster analysis (run in R with default settings) are shown in Fig. 1.

This clear-cut and unambiguous result follows from the stage directions. If we consider only the proportions of parts of speech in the stage directions (i.e. each drama is represented by a 10-dimensional vector), we obtain the same clusters as in

Fig. 1. However, if we consider only the dialogue of the characters, we do not find any significant differences or meaningful distinction between the dramas.

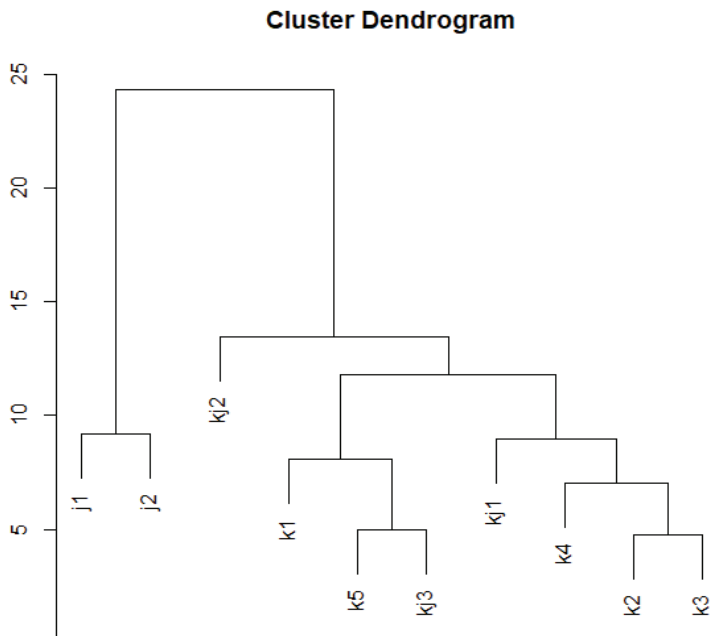


Fig. 1. Hierarchical cluster analysis of POS frequency distribution in the plays of the Čapek brothers

3.4 Morphological coefficients

Differences in the typology of the drama vocabulary can also be characterised using relatively simple statistical indicators introduced in the 1970s by the prominent Czech quantitative linguist Marie Těšitelová. She worked with nominal, verbal and neutral word-groups (Těšitelová 1974, p. 85nn). In particular, she measured the mutual proportionality of the so-called dominant components of the nominal and verbal groups and showed in her analyses that they can be used for individual characteristics of lexical styles, stylistic genres, etc. (Těšitelová 1974, p. 179). She introduced four basic indicators of morphological statistics (Těšitelová 1987, p. 89nn), which we will also use to interpret our data:

- Nominality coefficient: $K_n = N / V$ (ratio of nouns to verbs)
- Coefficient of noun development: $K_{rn} = A / N$ (ratio of adjectives to nouns)
- Coefficient of verb development: $K_{rv} = D / V$ (ratio of adverbs to verbs)
- Busemann coefficient: $B = A / V$ (ratio of adjectives to verbs).

Busemann’s coefficient is still actively used in quantitative linguistics as an index expressing the activity vs. descriptiveness of a text (Čech et al. 2014, p. 52nn).

We calculated these coefficients for the two structural parts of all the plays under study, i.e. for the characters’ dialogues and the stage directions. All the obtained data are summarized in the following heat-maps (the numbers represent the result of the given coefficients) – see Fig. 2–3.

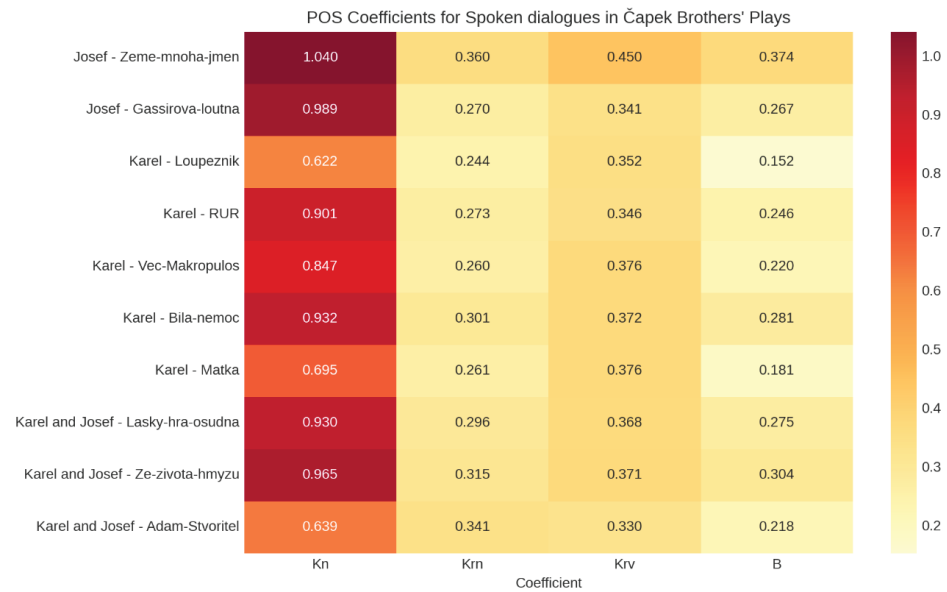


Fig. 2. Heat-map of morphological coefficients of dominant POS for the spoken dialogues in Čapek Brothers’ plays

The difference between spoken dialogue and stage directions

Key findings in spoken dialogue:

- Karel Čapek’s plays and collaborative works generally have consistently lower Kn values, indicating more dynamic, verb-rich and action-oriented dialogue, with relatively more verbs.
- Josef Čapek’s plays have higher Kn and B values, indicating more descriptive, noun-rich dialogue, suggesting more descriptive language relative to action.
- Krn and Krv values are generally higher in spoken dialogue than in stage directions, indicating richer descriptive language of characters, more adjectives to nouns (Krn); characters qualify verbal actions more with adverbs (Krv).

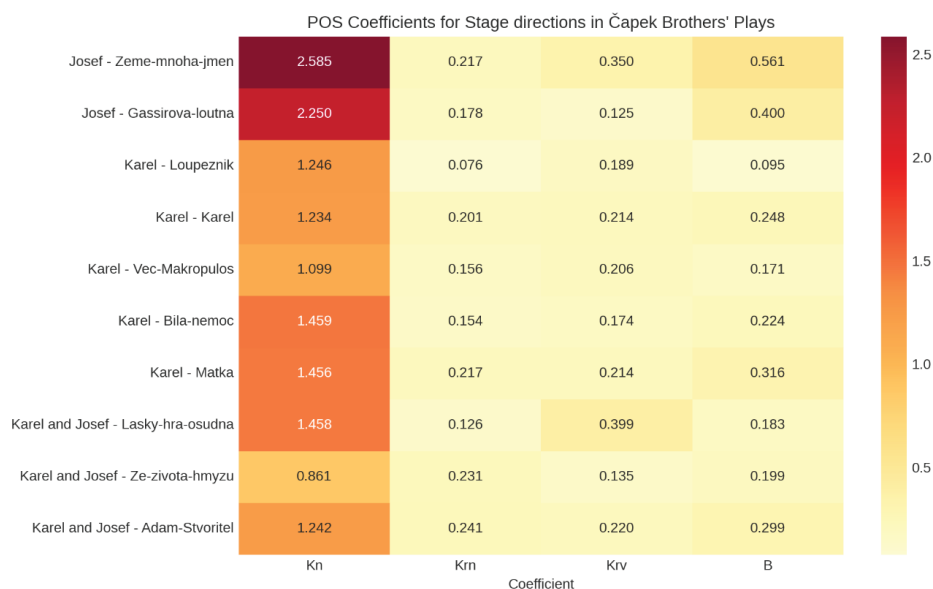


Fig. 3. Heat-map of morphological coefficients of dominant POS for the stage directions in Čapek Brothers' plays

These heat-maps clearly visualize the stylistic differences between the brothers' individual works and their collaborations, as well as the differences between writing styles of dialogue and stage directions.

Key findings for stage directions:

- All stage directions have significantly higher Kn values than spoken dialogue, which is to be expected as they are more descriptive and much richer in nouns relative to verbs.
- Josef Čapek's plays have the highest Kn values in stage directions (are particularly rich in nouns) and have particularly high B values (higher ratio of adjectives to verbs), which confirms his more descriptive style, indicating an emphasis on descriptiveness and detail.

Overall stylistic differences

Karel Čapek's style:

- More action-oriented with more use of verbs than nouns.
- More direct with fewer modifiers (adjectives and adverbs).
- Maintains this style in both spoken dialogue and stage directions, though less pronounced in the latter.

Josef Čapek's style:

- More descriptive with greater use of nouns and adjectives.
- More elaborate with more modifiers.
- Very descriptive, especially in stage directions.

Collaborative works:

- Often and generally intermediate values.
- Stage directions in collaborative works are particularly noun-rich.
- Suggest a blending (or fusion) of the brothers' individual styles.

4 CONCLUSION

Karel and Josef Čapek had completely different styles of writing stage directions. Josef used considerably fewer words. He used more nouns (and adjectives) and fewer verbs than Karel. The stage directions in the co-authored plays are written in the same style (in terms of relative numbers of words and proportions of parts of speech) as those in which Karel is the sole author. The proportions of parts of speech in the spoken dialogue do not indicate one or the other of the two brothers. However, the difference in the stage directions is so clear and strongly pronounced that Josef's dramas form a separate cluster when both the stage directions and the dialogues are taken into account.

Morphological coefficients confirm these differences, and this analysis reveals clear stylistic differences between the Čapek brothers, with Josef preferring a more descriptive language and Karel favouring a more dynamic, action-oriented approach. Their collaborative works often combine these styles. The analysis also shows that there are noticeable differences between the spoken dialogue and stage directions.

In general, it can be said that this comparison shows that Josef Čapek's plays in particular are different from the others – cf. the extreme values of the coefficients K_n and B , which express the high saturation of descriptiveness (especially in the stage directions).

It should be noted that this is the first quantitative analysis focusing on structure that shows some indications of a different authorial style. In the future, we would like to explore the structure of the plays in more detail, by act or by scene, and to look at other aspects and phenomena of the plays of these authors, such as keyword analysis, methods of determining authorship, and characterization of the network of literary characters within each play.

ACKNOWLEDGEMENTS

Supported by the European Commission's Horizon 2020 program (Grant Agreement No. 101004984) through the CLS INFRA project (P. Pořízka) and by VEGA 2/0120/24 (J. Mačutek).

References

- Cvrček, V., Čermák, F., and Křen, M. (2007). Statistické aspekty jazyka Karla Čapka, zvláště jeho lexikonu. In: F. Čermák (ed.): *Slovník Karla Čapka*. Praha: NLN, pp. 671–690.
- Čech, R., Popescu, I.-I., and Altmann, G. (2014). *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého.
- Čech, R. (2015). The development of thematic concentration of text in Karel Čapek's travel books. *Czech and Slovak Linguistic Review*, 2015(1), pp. 8–21.
- Čermák, F., et al. (2007a). *Čapek: korpus pouze vlastních textů Karla Čapka*. Praha: ÚČNK. Accessible at: <http://www.korpus.cz>.
- Čermák, F. (ed.). (2007b). *Slovník Karla Čapka*. Praha: NLN.
- Davidová Glogarová, J., Čech, R., and David, J. (2013). Tematické charakteristiky textu – kvantitativní analýza publicistiky Jaroslava Durycha, Ladislava Jehličky a Karla Čapka. In: J. David – R. Čech – J. Davidová Glogarová – L. Radková – H. Šústková (eds.): *Slovo a text v historickém kontextu – perspektivy historickosémantické analýzy jazyka*. Brno: Host, pp. 62–84.
- Doležel, L. (2014). Dva moderní vypravěči: Karel Čapek a Vladislav Vančura. In L. Doležel: *Narativní způsoby v české literatuře*. Příbram: Pistorius & Olšanská, pp. 97–114.
- Drozenová, W. (2020). Technika, autonomie a etika: ke stému výročí Čapkova dramatu R.U.R. In 100 let R. U. R. *Sborník z konference na Pedagogické fakultě Masarykovy univerzity v Brně* (September 11, 2019). Brno: Masarykova univerzita, pp. 9–16.
- Holý, J. (1984). Funkce jmen postav v dílech Karla Čapka a Vladislava Vančury. *Česká literatura*, 32(5), pp. 459–476.
- Holý, J. (2014). Komentář k edici *Tři hry Karla Čapka*. In: K. Čapek – J. Holý (eds.): *Tři hry*. Brno: Host, pp. 253–279.
- Janiec-Nyitrai, A. (2012). Zrcadlení: Literárněvědné sondy do tvorby Karla Čapka. Nitra: Univerzita Konštantína Filozofa v Nitre, Fakulta stredo európskych štúdií.
- Janoušek, P. (1989). Mezi problémovým dramatem a groteskou (Druhá etapa dramatické tvorby Karla Čapka). *Česká literatura*, 37(3), pp. 193–205.
- Janoušek, P. (2018). Bratři Čapkové: *Lásky hra osudná*. *Theatralia*, 21(1), Supplementum, pp. 33–38.
- Kováříková, D., and Kopřivová, M. (2012). Authorial phraseology: Karel Čapek and Bohumil Hrabal. In: V. Jesenšek – D. Dobrovol'skij (eds.): *Phraseology and culture*. Maribor: Europhras, pp. 522–532.
- Křupková, L. (2008). Proměny Čapkova dramatického textu v Janáčkově opeře *Věc Makropulos*. In: J. Nowak – K. Steimetz (eds.): *Leoš Janáček světový a regionální*. Ostrava: Ostravská univerzita, pp. 25–42.
- Kubát, M. (2016). *Kvantitativní analýza žánrů*. Ostrava: Ostravská univerzita.
- Kořenský, J., et al. (1986). *Mluvnice češtiny 2. Tvarosloví*. Praha: Academia.
- Mačutek, J., Koščová, M., and Čech, R. (2016). Lexical compactness across genres in works by Karel Čapek. In: Mayaffre, D., Poudat, C., Vanni, L., Magri, V., and Follette, P. (eds.): *Statistical Analysis of Textual Data*. Nice: University Nice Sophia Antipolis, pp. 825–832.
- MorphoDiTa: Morphological Dictionary and Tagger (version 1.3) [Software]. Accessible at: <http://lindat.mff.cuni.cz/services/morphodita/>.

- Novák, R. (2013). *Nástin poetiky Karla Čapka*. Ostrava: Ostravská univerzita.
- The R Project for Statistical Computing: R (version 4.2.0) [Software]. Accessible at: <https://www.r-project.org/>.
- Pořízka, P. (2023a). CapekDraCor: A New Contribution to the European Programable Drama Corpora. *Jazykovedný časopis*, 74(1), pp. 244–253.
- Pořízka P. (2023b). The Function of Proper Nouns in Quantitative Analysis of Dramas: A Case Study of Karel Čapek's Plays. In *Onomastics in Interaction With Other Branches of Science*. Vol. 3: General and Applied Onomastics. Literary Onomastics. Chrematonomastics. Reports, pp. 351–379.
- Pořízka, P. (2025). LexaMorf Tool [Software]. Olomouc.
- SketchEngine [Software]. Accessible at: <http://www.sketchengine.eu>.
- Straková, J., Straka, M., and Hajič, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics: System demonstrations*. Baltimore (MD): Association for Computational Linguistics, pp. 13–18.
- Sunbee, Y. (2011). *Znaky, styl a dramata Karla Čapka* [Dissertation]. Praha: Univerzita Karlova.
- Těšitelová, M. (1974). *Otázky lexikální statistiky*. Praha: Academia.