Petr Pořízka

petr.porizka@upol.cz
https://orcid.org/0000-0001-6980-9148
Department of Czech Studies, Faculty of Arts,
Palacký University
Olomouc, Czechia

https://doi.org/10.4467/K7478.47/22.23.17743

The Function of Proper Nouns in Quantitative Analysis of Dramas: A Case Study of Karel Čapek's Plays

Abstract

This paper falls in the field of literary onomastics, and the focus of the study is mainly methodological. However, I will present the first results of a quantitative analysis of Čapek's plays, especially engaging with semantically oriented keywords of the text (*prominent units*). I introduce the possibilities of using proper nouns for the quantitative analysis of a literary text, specifically a drama, within the framework of content analysis. For this purpose, an annotated corpus of Karel Čapek's plays was created, which served as a basis for the analysis and was processed in several variants, the most important being segmentation according to a literary character and the distinction between text (dialogues) and metatext (commentary on the characters' actions, the scene on stage, etc.). The key distinction here is one between proper nouns in dialogues and *proper noun labels*, i.e. speech prefixes found at the beginning of the lines (dialogues) of individual characters. In this study several methods of quantitative linguistics are used, namely KWs analysis (TF*IDF calculation), the extraction of thematic words (TC/STC indexes of thematic concentration method), an exploration of lexical dispersion (specificity score and neutral, overused, underused KWs), and multidimensional scaling (hierarchical clustering, correspondence analysis).

Keywords

drama, keywords, multidimensional scaling, proper nouns, quantitative linguistics

1. Introduction

In this study, we do not focus on the function of proper nouns from a linguistic theoretical or onomastic point of view, but rather try to show the different textual layers in which proper nouns appear in a drama and their use and implications for a quantitative analysis of the text. Thus, we explore the use of proper nouns as structural and segmenting elements in the quantitative analysis of dramas. The main exploratory method is the analysis of keywords (also prominent text units) with the intention of detecting how particular thematic (or topic) words are projected within a particular literary character of the drama.

Within quantitative linguistics, there are various approaches and methods for detecting prominent units of text (PU), the most common of which are topic words (TWs), as a result of measuring the thematic text concentration (TC) (cf. Čech et al., 2013, 2015), and keywords (KWs), based on a comparison of the source and reference corpus (cf. Scott & Tribble, 2006). The choice of a particular method depends both on the type of linguistic analysis and the research goal, and on the text type to be used for analysis. Moreover, the resulting keyword lists are often differentiated, as they depend on settings of various technical parameters, including, among others, the choice of the basic unit of analysis (word form × lemma × other units); the text size, genre, choice of statistical test and reference database, and other configurations. These methods have so far been used mostly to analyse written texts. As shown in an earlier study (Pořízka, 2019), it is very difficult to find a universal method that works across different text genres or language varieties, and the type of text can play an important, if not crucial, role. Methods that are typically applicable to written language show certain deficiencies when used for spoken discourse and other, more appropriate solutions need to be found. One possible reason for this is the different frequency structure of the text, or the changing proportional representation of parts of speech (POS), especially nouns, which are unquestionably the most important category for the analysis of key or thematic words (based on a frequency dictionary or a comparison of frequency lists). For more on the difference between written and spoken form with regard to keyword analysis and thematic words, see Pořízka (2019).

2. Corpus and data processing

For the purposes of this study I have created a corpus of five dramas by Karel Čapek processed in several segmentation variants: (1) all texts together, (2) by drama(s), (3) by a literary character within each drama.

Čapek's drama(s) in the corpus:

- "Loupežník" ('The Outlaw', 1919–20);
- "R.U.R." (1920);
- "Věc Makropulos" ('The Makropulos Affair', 1922);
- "Bílá nemoc" ('The White Disease', 1937);
- "Matka" ('The Mother', 1938).

During the analyses, I also worked with other variants of text structure and processing:

(1) A version of Czech National Corpus (CNC; see also below), where the text and metatext levels are not distinguished: proper names as names of literary characters (PropN-labels or PN-labels) introducing dialogues and other types of metatext are part of the text (not separated).
(2) The version without metatext, where the names of literary characters (PropN-labels) at the beginning of the dialogues (text lines) are removed, as well as the comments and all other parts of the metatext.
(3) The version without PropN-labels introducing the lines and without the irrelevant (unnecessary) parts of the metatext (indicating the structure of the text: act, scene, drop-scene, etc.), but with the relevant metatext left in (commentary on the characters' actions, the scene on stage, etc.).

The technical processing of the texts included in the corpus involved mainly (a) the conversion of character/text encoding (from ANSI set to Unicode UTF-8 version), and (b) text cleaning: removal of all unnecessary metadata (references, editorial notes, bibliographic data on the life and work of K. Čapek, etc.) and, if necessary, metatextual parts of the drama(s), if the database structuring required it – removal of the preface, introductory notes, synopsis, e.g. overview of literary characters (and their brief description) at the beginning of each play.

The segmentation of the texts of Čapek's plays, as already mentioned, was carried out in two main variants (i) according to the specific play, and (ii) according to the literary character(s) (i.e. represented by proper names) in each play.

All texts were subsequently linguistically annotated and converted into a technical XML format.[1] For lemmatization and morphological tagging of the texts, I used the freely available MorphoDiTa tool (ver 1.3; Straka, Straková & Hajič, 2014) with the MorfFlex dictionary (ver. czech-morfflex-pdt-161115), with the "raw lemmas" function activated for lemmas, and the morphological "guesser" activated for unknown words.[2]

3. Analysis methods and tools

I chose TXM (ver. 0.8.1; Heiden, 2010) as the main corpus manager for data mining.[3] For the analysis of Karel Čapek's plays I used several methods (with a more detailed description below) and other software tools, namely:

(1) Keyword extraction (key-words, or key-lemmas) using the TF*IDF method and the KER tool (Keyword extractor; Libovický, 2016);
(2) The topic concentration method and the detection of topic terms (TC and STC indexes) using the QuitaUP tool (Cvrček et al., 2020)
(3) Lexical dispersion of words and their visualization with the statistical tool R (R Core Team, 2021) and the package "qdap" (ver. 2.2.0; Rinker, 2013);
(4) Lexical specificity of KWs (TXM corpus manager; Heiden, 2010);
(5) Multidimensional analysis (dendrograms, cluster analysis) (TXM corpus manager).

[1] XML – abbreviation of Extensible Markup Language. More information about this format is available here: https://www.w3.org/XML/

[2] MorphoDiTa – abbreviation of the Morphological Dictionary and Tagger, http://lindat.mff.cuni.cz/services/morphodita/

[3] TXM – abbreviation of textometry, http://textometrie.ens-lyon.fr/?lang=en

In the first phase of the probe, we also used the corpus “capek” (Čermák et al., 2007) from the Czech National Corpus (CNC), specifically the sub-database of plays (for the comparison of wordlists).[4]

4. Čapek’s corpora in CNC

There are currently two databases available in the Czech National Corpus, “capek” and “capek_uplny”, containing Čapek’s complete works. However, the method of processing his plays is not suitable from our point of view and for the purpose of quantitative analysis. First of all, the different levels of the text are not separated, which leaves parts that may affect the quantitative analysis. One example from the corpus is CNC “capek” – KWIC *osoba* ‘person’ (Figure 1):

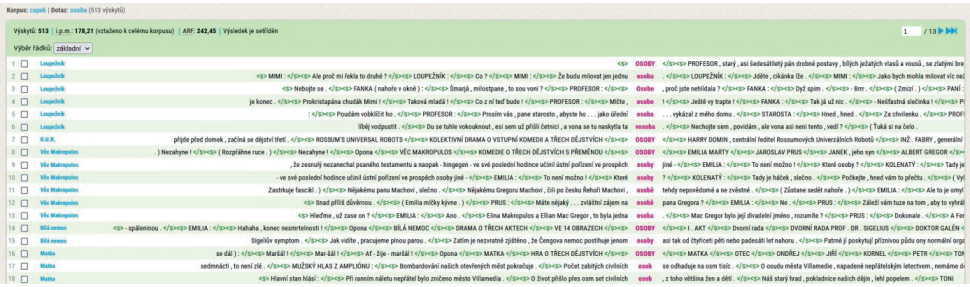

Figure 1. Concordances of KWIC *osoba* ‘person’; demonstrating that text and metatext are not separated

Source: CNC – capek corpus.

The concordance list shows that the metatextual segments are still part of the text – see the following examples:[5]

[4] More information about this corpus is available at: https://wiki.korpus.cz/doku.php/cnk:capek (retrieved May 25, 2023).

[5] All translations by the author.

- e.g., line 1: there is a description of the literary characters of the play: ("Loupežník" 'The Robber') *OSOBY PROFESOR, starý, asi šedesátiletý pán drobné postavy, bílých ježatých vlasů a vousů*... 'PERSONS THE PROFESSOR, an old gentleman of about sixty years of age, of small figure, with white hair and beard...'.
- cf. lines 7, 8, 14 and 16: see again the overview of characters, possibly their characteristics (description) and other metatextual data:
  - title, subtitle, etc.: *ROSSUM'S UNIVERSAL ROBOTS, KOLEKTIVNÍ DRAMA O VSTUPNÍ KOMEDII A TŘECH DĚJSTVÍCH* 'ROSSUM'S UNIVERSAL ROBOTS, A PLAY IN INTRODUCTORY SCENE AND THREE ACTS';
  - terms relating to parts of the text: *opona* 'drop-scene', *I. akt* 'Act I';
  - comments by the author (in round brackets) describing the actions of characters, a particular scene, etc.: *zůstane sedět nahoře* 'he remains seated upstairs', *Emilia mlčky kývne* 'Emilia nods silently';
  - names of literary characters – tags preceding the dialogue of a given character: *LOUPEŽNÍK*:, MIMI:, PROFESSOR:, EMILIA:, PRUS:.

This processing can – and, as we shall see below, does – affect quantitative data, and subsequently the results based on that data (cf. the two versions of the wordlist of nouns below).

5. Specificities of dramatic texts

The texts of plays can be viewed from several perspectives. From a linguistic point of view, it is a borderline (or zone) between written and spoken language (see below for the proportion of POS). In terms of format, it is a mixture of text and metatext. The text is represented by the characters' lines (dialogues) and the metatext includes a foreword/preface, opening remarks, overview/description of literary characters, commentary and other notes like the act, scene, etc.

Proper nouns have a specific role in plays: When analysing literary texts, dramas are very different compared to other forms of fiction (novels, etc.) from this point of view, because in fiction, generally speaking, the naming of a literary character (i.e., the proper noun) is an integral part of the text. In contrast,

in drama it is usually a label for each of the character's lines, which need to be removed for quantitative exploration, as well as part of the text in dialogues.

All these aspects are very important because the way the data is processed influences and ultimately leads to the subsequent results of the quantitative analyses. There may be nuances in terms of the effect on the analysis, but sometimes the differences can be more significant. My assumption is that if one uses proper nouns which denote literary characters in the text as a means of both segmentation and interpretation, the analysis will be more focused, detailed, complex, and accurate.

As seen above, the problem is that the different levels of text in Čapek's plays are not separated in the CNC database. The following table shows two versions of the wordlist of nouns (a sample of the first 30 words): (1) the first unmodified (left), and (2) the second with the PropN-labels removed (right).

	(1) unmodified version (CNC)		(2) modified version (PN-labels removed)	
rank	**lemma**	**AF**	**lemma**	**AF**
1	Helena	590	člověk	398
2	Emilia	493	pan	350
3	doktor	452	robot	190
4	Loupežník	436	slečna	179
5	člověk	418	Mimi	177
6	Mimi	412	rok	162
7	Domin	371	bůh	150
8	pan	352	doktor	150
9	matka	336	maminka	137
10	Gregor	323	Toni	135
11	profesor	266	ruka	134
12	Prus	255	Helena	128
13	rada	247	dítě	117
14	maršál	245	život	113
15	Galén	239	svět	110
16	Toni	234	sto	105
17	robot	229	máma	103

	(1) unmodified version (CNC)		(2) modified version (PN-labels removed)	
18	Petr	227	pán	103
19	otec	211	Gregor	98
20	Alquist	196	věc	95
21	Kornel	196	rada	93
22	paní	190	Galén	84
23	slečna	189	láska	83
24	rok	178	maršál	83
25	ruka	177	pravda	83
26	Krüg	170	válka	83
27	Fanka	160	pauza	79
28	bůh	154	hlava	76
29	Gall	154	Petr	76
30	maminka	145	nemoc	72

Ad 1 (left): Proper nouns are almost everywhere at the top of the list with a very high frequency of occurrence, which appear because all proper noun labels preceding dialogues are part of the actual text. Of course, one should work with proper nouns, or more precisely with their PN-labels, but rather as an element of text segmentation and consequently interpretation of linguistic aspects and relations of literary characters.

Ad 2 (right): Looking at the modified (second) version of the wordlist with PropN-labels removed, common nouns such as *human*, *robot*, *year*, *god*, *life*, *world*, *thing*, *love*, *truth*, *war*, etc., come to the highest positions in the wordlist and the frequency dictionary thus provides a much more accurate picture of the occurrence of words in the text. It is not distorted by the occurrence of PropN-labels. Proper nouns are still part of the frequency dictionary, but only those occurrences from the characters' dialogues are counted; cf. the differences (the first number indicates the frequency of occurrence, the second number is the ranking):

Helena: 590/01 × 128/12
Emilia: 493/02 × 13/233
Mimi: 412/06 × 177/05
Domin: 371/07 × 30/88
Prus: 255/12 × 55/40

Galén:	239/15	×	84/22
Toni:	234/16	×	135/10
Petr:	227/18	×	76/29
Alquist:	196/20	×	41/59
Kornel:	196/21	×	57/37
Krüg:	170/26	×	54/41
Fanka:	160/27	×	30/89
Gall:	154/29	×	21/134

The number of occurrences has changed significantly (downwards): cf. Helena, Emilia, Mimi, Domin, or Prus. In most cases, proper names move to lower positions in the wordlist after the removal of PN-labels (significantly, e.g., for Emilia from position 2 to 233), but there are also cases where the term moves up in the wordlist even though the frequency of occurrence is reduced (cf. Mimi and Toni). There is no doubt that this indicated difference affects the subsequent analysis of the text as well as the results and interpretation of the quantitative data.

The next step towards a more detailed quantitative content analysis is the division of character's dialogues and metatextual comments; and then the segmentation of the texts into sections according to the individual literary characters of the play, i.e., according to PropN-labels. At this point, it is possible to perform a focused analysis and use these proper nouns to explore and interpret the content of the text more accurately.

6. Drama and written vs. spoken form: differences

As already stated, drama can be viewed as a border zone between written and spoken language. It has a similar textual structure to spoken dialogue in that PN lines are similar to the spoken language speakers' lines; in terms of proportions of POS classes, sometimes a play is closer to spoken, sometimes to written language.

Table 1 presents an overview of frequency structure of open and closed class words. The first and second columns show data from the written and

spoken corpora of the CNC, the third column represents the POS values of all Karel Čapek's dramas without modifications (i.e., including PN-labels), the fourth column represents the same texts after removing PN-labels. The last column shows the differences between the Čapek versions – the effect of removing PN-labels on the quantitative data (frequency in %):

Table 1. Drama and written vs. spoken form: differences of open and closed class words. Numbers represent relative frequency in per cent

POS	CNC written	CNC spoken	Čapek ALL	Čapek minus PN	Čapek DIFF (%)
NOUN	30,53	11,41	28,11	21,68	–6,43
ADJ	11,48	3,50	5,63	5,52	–0,11
PRON	10,48	20,27	18,02	20,16	2,14
NUM	3,17	2,04	1,37	1,23	–0,14
VERB	16,86	20,15	23,28	25,48	2,20
ADV	7,10	12,84	8,08	9,05	0,97
PREP	10,55	5,67	5,97	6,07	0,10
CONJ	7,56	11,48	6,40	7,13	0,73
PART	0,99	8,38	2,30	2,62	0,32
INTJ	0,05	0,42	0,83	1,06	0,23
resp+hes	–	2,15	–	–	
uncompl	–	1,04	–	–	
unknown	1,26	0,65	–	–	

Legend: POS = Part of Speech; CNC = Czech National Corpus; PN = Proper Noun; DIFF = difference; resp+hes = response and hesitation; uncompl = uncompleted words; unknown = expressions not recognized by the tagger.

Source: Czech National Corpus and own work.

The table shows that the distribution of POS classes oscillates between written and spoken forms, demonstrating the following tendencies:

- closer to written language: CONJ, PART;
- closer to spoken language: PRON, PREP;
- "somewhere in between": NOUN, ADJ, NUM, ADV;
- specific class words: VERB, INTJ.

These differences may affect the quantitative analysis according to the nature or type of text, whether and to what extent they shift on the scale from the written towards the spoken form with respect to the frequency structure of the texts. The methods of analysing prominent units are based on frequency dictionaries or their comparison, and especially the ratio of nouns (indicating or denoting substances) in the text is crucial in this respect. Their appearance in spoken texts (SpT) is distinctly lower than in written texts (WrT); dramatic texts (ČdT) are somewhere in between:

approx. 30% WrT vs. 20% ČdT vs. 10% SpT

There is a similarly significant decline in adjectives (approx. written: 11.5% vs. spoken: 3.5% vs. drama: 5.5%). On the other hand, the proportion of verbs increases from written to spoken texts and is even higher in plays than in spoken language (approx. written: 17% vs. spoken: 20% vs. drama: 25.5%).

7. Methods of KWs extraction

In general, one of the methodologically important decisions is the choice of the basic unit. From this point of view, the use of a word form or a lemma can be considered. Not only the type of analysis or the focus of research should be considered, but also, for example, factors such as the typological character of the language in question. It seems that the choice of unit (word form vs. lemma) might be more important for inflectional languages than languages of a non-flective nature, in order to include all types of a given lexeme (paradigm). Consider, for example, all the forms of the noun *člověk* 'human' – one of the most frequent nouns in Čapek's plays:

singular forms: *člověk-Ø*, *člověk-a*, *člověk-u*, *člověk-ovi*, *člověk-em* + *člověče* (sg. vocative)
plural forms: *lid-é*, *lid-í*, *lid-em*, *lid-i*, *lid-ech*, *lid-mi*

In non-lemmatized texts, these 12-word forms would function separately as different word units, each with a lower frequency compared to the overall

frequency of the lemma *člověk* 'human', which may have consequences for quantitative analysis. If the intention is to explore the distribution of lexemes, as in the case of keyword analysis, then it is more appropriate to use lemmas, as was done in this study.

Pořízka (2019) tested the most used methods of prominent words (PWs) extraction on spoken discourse in order to find out if commonly used methods (especially KWs + TC) are "usable/applicable", or if they are somehow "defective" and thus it would be necessary to find another solution. I also tried to characterize the advantages and disadvantages of these particular methods and point out problematic aspects.

I compared the following methods: (i) keywords analysis (KWs – keywords, or key-lemmas), (ii) thematic concentration (TC & STC indexes), and (iii) TF*IDF, which has been newly tested on Czech language data and the most promising. TF*IDF seems to be a good alternative to solve or eliminate the drawbacks of the previous two methods, (i) and (ii) (cf. Pořízka, 2019). Based on the above, we use the TF*IDF method to extract KWs of the selected Čapek's play (see below).

TF*IDF (Rajaraman & Ullman, 2011) is an information retrieval technique (a statistical measure) that weighs term frequency (TF) and its inverse document frequency (IDF). It aims to define the "importance" of a keyword(s) within a document. The more often a word occurs in documents, the less relevant it is to the source text as the weight of a term is proportional to its frequency: the smaller the weight, the more common the term. The formula is as follows:

> TF(t) = (freq of term t in a document) / (total number of terms in the document)
> IDF(t) = log_e (total number of documents / number of documents with term t in them)

For keyword analysis via TF*IDF method I used the freely available KER – Keyword Extractor tool with the following settings: basic unit: lemma; TF*IDF threshold level: 0.01; max. number of keywords (key-lemmas): 50 KWs chosen randomly. I am aware of the fact that more analyses will have to be carried out, to test more data in order to find optimal settings of the key parameters.

Here are the analysis results for the different samples. For this purpose, I have chosen the play "Bílá nemoc" ('The White Disease'):

Sample 1 (à la CNC-Capek: text and metatext not separated)

Almost all extracted words are proper names; to be precise the proper name labels of literary characters at the beginning of each character's lines (or other metatextual words, especially remarks and comments on the characters' actions or the stage scene). The negative effect of keeping these names or those descriptive words as text rather than as metatext is clear.

Sample 2 (without PropN-labels at the beginning of each line)

If we remove proper names (as labels of each line) and still keep comments and remarks on the characters' actions and the situation on stage, the result of the extracted words slightly improves, but the list of words still does not convincingly express the theme or topics of the text. Proper names remain, as do words like *pauza* 'pause', *dveře* 'door', *okno* 'window', *opona* 'drop-scene' as situational, contextual words, or commentary.

Sample 3 (without PropN-labels and comments)

There has been another slight improvement with situational words removed, but still topical or thematic words are a minority and appear exceptionally.

Sample 4 (text divided into parts according to individual literary characters)

Only this segmentation allows one to lead the analysis in such a way that one can detect not only much more KWs of texts (not only PN), but also to find out which of them are related to which character.

Look at the key-lemmas of the main characters in the play "Bílá nemoc" ('The White Disease') as an example. Below is a preliminary attempt to categorize these keywords (nouns from the first 50 KWs):

DVORNÍ RADA (COUNSELOR)

DISEASE: *nemoc* 'disease', *klinika* 'clinic', *doktor* 'doctor', *lékař* 'physician', *choroba* 'disease', *malomocenství* 'leper', *tshengi* 'tshengi', *lék* 'cure', *morbus*

'morbus', *morfium* 'morphine', *pacient* 'patient', *nemocný* 'sick', *zápach* 'smell', *nákaza* 'infection', *lepra* 'leprosy', *třináctka* 'room number thirteen'.

ABSTRACT: *člověk* 'human', *národ* 'nation', *veřejnost* 'public', *pravda* 'truth', *věda* 'science', *čas* 'time', *úspěch* 'success', *případ* 'case', *prostředek* 'remedy', *příznak* 'symptom'.

CONCRETE (persons): *asistent* 'assistant', *mládenec* 'bachelor', *klient* 'client', *kamarád* 'friend', *blázen* 'fool', *tchán* 'father-in-law'.

MARŠÁL (MARSHAL)

DISEASE: *malomocný* 'leper', *léčení* 'cure', *čelo* 'forehead', *ruka* 'hand'.

SPECIFIC: *bůh* 'god', *národ* 'nation', *mír* 'peace', *člověk* 'human', *voják* 'soldier', *pravda* 'truth', *válka* 'war', *poslání* 'mission', *právo* 'law', *vůle* 'will', *rozkaz* 'command', *vlast* 'homeland', *hrdina* 'hero', *vojna* 'war', *radost* 'joy'.

ABSTRACT: *podmínka* 'condition', *zpráva* 'message', *výsledek* 'result', *dětinství* 'childishness', *věc* 'thing'.

CONCRETE (persons): *hoch* 'boy', *holčička* 'little girl', *chudák* 'poor guy', *kamarád* 'friend', *mladík* 'young man', *blázen* 'fool'.

GALÉN

DISEASE: *lék* 'cure', *nemoc* 'disease', *malomocný* 'leper', *doktor* 'doctor', *klinika* 'clinic', *lékař* 'physician', *nemocný* 'sick', *malomocenství* 'leprosy', *pacient* 'patient', *zápach* 'smell', *kostižer* 'multiple myeloma', *třináctka* 'room number thirteen'.

SPECIFIC: *člověk* 'human', *národ* 'nation', *vladař* 'ruler', *mír* 'peace', *válka* 'war', *vojna* 'war', *munice* 'ammunition', *yperit* 'yperit', *svinstvo* 'crap'.

ABSTRACT: *prominutí* 'forgiveness', *podmínka* 'condition', *praxe* 'practice', *chvilinka* 'brief moment', *prostředek* 'remedy', *případ* 'case'.

BARON KRÜG

DISEASE: *nemoc* ‘disease’, *malomocný* ‘leper’, *léčení* ‘treatment’, *izolace* ‘isolation’, *nemocný* ‘sick’.

SPECIFIC: *bůh* ‘god’, *člověk* ‘human’, *mír* ‘peace’, *rozkaz* ‘command’, *ultimátum* ‘ultimatum’, *pravda* ‘truth’, *drát* ‘wire’, *tank* ‘tank’, *strach* ‘fear’, *povinnost* ‘duty’, *letadlo* ‘plane’, *stihačka* ‘fighter jet’, *válka* ‘war’, *továrna* ‘factory’.

ABSTRACT: *podmínka* ‘condition’, *prominutí* ‘pardon’, *peníze* ‘money’.

CONCRETE (persons): *děvče* ‘girl’, *drahoušek* ‘darling’.

8. Lexical dispersion and plots

Another suitable method that can help analyse the behaviour of proper nouns in more detail is lexical dispersion, which seeks to reveal the even/uneven distribution of KWs in texts, or to identify specific parts where a particular KW is located. Distribution can be expressed both numerically and graphically. One of the most widely used numerical indexes is the ARF (average reduced frequency), which belongs to the adjusted frequencies (Savický & Hlaváčová, 2002).[6] It is particularly advisable to use a visualisation (such as a dispersion plot), which shows the dispersion of words in the text very clearly. See the three examples (different processing and visualisation options) below that differ in the way the database is structured (in this case the word form is used).[7]

(1) The dispersion is most commonly expressed within the whole, undivided text (Figure 2):

[6] For more information, including the calculation formula, cf. Savický & Hlaváčová (2002), or see the entry ARF in the CNC glossary, retrieved May 25, 2023, from https://wiki.korpus.cz/doku.php/en:pojmy:arf

[7] For this purpose, we have used the R tool and the “qdap” package.

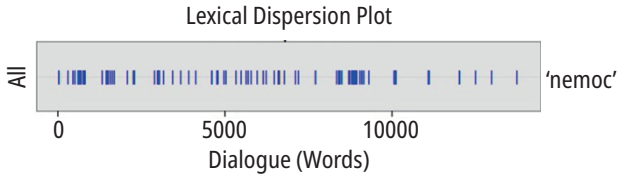

Figure 2. R&qdap: dispersion plot of a keyword *nemoc* 'disease' in Karel Čapek's "Bílá nemoc" ('The White Disease')

Source: own work.

(2) But for plays it is undoubtedly more useful to divide the text into at least major sections according to the particular acts. This is because it provides a more precise idea of how a given word occurs in which part (Figure 3):

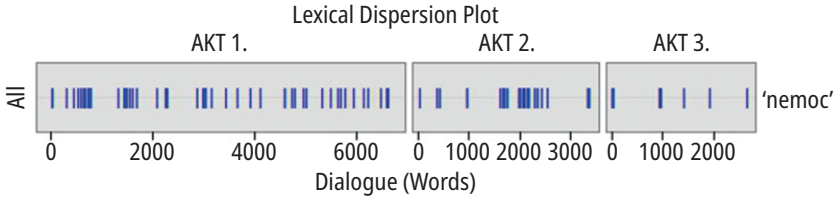

Figure 3. R&qdap: dispersion plot of a keyword *nemoc* 'disease' in Karel Čapek's "Bílá nemoc" ('The White Disease'); divided by parts (Acts 1–3)

Source: own work.

(3) It seems most appropriate to use a division by individual literary characters in the analysis of plays. This way reveals the best idea (image) of the behaviour of KWs under study (Figure 4):

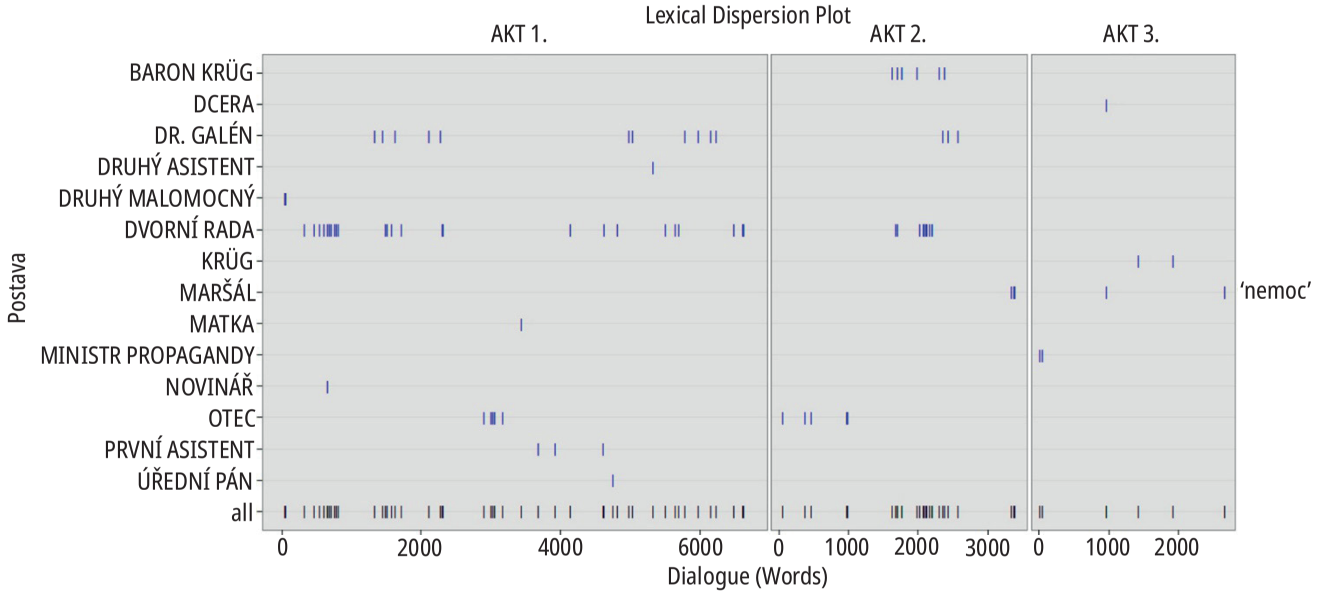

Figure 4. R&qdap: dispersion plot of a keyword *nemoc* 'disease' in Karel Čapek's "Bílá nemoc" ('The White Disease'); divided by parts (Acts 1–3) and literary characters

Source: own work.

Here I have tried at least to introduce lexical dispersion as an important tool for exploring proper nouns in plays and to outline the possibilities and importance of this method, which should be part of any more comprehensive analysis of KWs. The more the text is divided into smaller parts, the better and more detailed is the detection of the behaviour of KWs in those segments, and also in the whole text.

However, other questions arise in this context. Is the lemma a sufficient unit to explore KWs and their lexical dispersion? Is there a sequence or a bunch of words (semantically) related to the topic, or more precisely, to the one linguistic meaning? Shouldn't one consider a higher unit with a wider range of words?

If so, what kind of unit should it be? An aggregate? (And what type?) The key aspect should probably be the meaning. The semantic point of view could lead to a unit that could, for example, be named as a synonymous aggregate (synonymic aggregation). Compare below the lexemes that relate to the main theme of the play, *disease*:

- *nemoc* 'disease'
- *bílá_nemoc* 'white_disease'
- *choroba* 'disease'
- *Čengova_choroba* 'Cheng's_disease'
- *Morbus_Tshengi*
- *mor* 'plague'
- *malomocenství* 'leprosy'
- *bílá skvrna/skvrnka* 'white patch'
- *lék* 'medicine'
- *prostředek* 'remedy'
- *léčebný* 'medical'
- *léčebný_prostředek* 'remedy'
- *léčba* 'therapy'
- *léčení* 'treatment'
- *léčit* 'cure'

The same approach applies to the literary characters of the play, such as the main character: Dr. Galén = doktor Dětina (both MWE represents the same character in a literary work), thus they should be labelled and treated in the same way.

9. Multidimensional analysis

Multidimensional analysis (or scaling) is the analysis of objects organized in meaningful hierarchies or related dimensions. This method allows users to observe data from various viewpoints and enables them to spot trends or exceptions in the data. It is a technique for visualizing the relationships among datasets that are similar to each other on many dimensions by reducing them into smaller segments expressing similarities or differences among texts (or its parts) by means of dendrograms (hierarchical clustering), cluster analysis, etc.

In this article, I only present examples of hierarchical cluster analysis and correspondence analysis to illustrate this method without going into more complex interpretations. Multidimensional scaling (MDS) is again based on a comparison of texts' frequency wordlists and can be used to explore the data from various linguistic points of view: i.e., lexical (units: word/lemma/KWs), or grammatical (POS, morphological categories: gender, number, case, mode, person, etc.). MDS analysis was carried out using the TXM & R tool with the following settings:

- method: cosine distance;
- frequency list: first 200 words regardless of POS or other categories;
- basic unit: word form;
- text parts/chunks = all text lines belonging to a particular literary character (PN).

This allows one to perform, for example, this type of exploration (Figure 5).

This graph represents the result of the correspondence analysis and shows the interrelationships between the literary characters from the play "Matka" ('The Mother'), which can be divided into four groups or dimensions. Several characters were grouped into one cluster (1) cf. Toni, *starý pán* 'old man', Petr, *otec* 'father', Kornel, Jiří and Ondra; separate clusters are then formed by (2) the mother ("matka"), (3) the voices from behind the stage ("hlasy z amplionu") and (4) a very different commentary on the characters' actions and the stage scene ("KOMENTAR").

This tendency is confirmed by the subsequent cluster analysis, cf. the dendrogram with the four groups of segments (Figure 6):

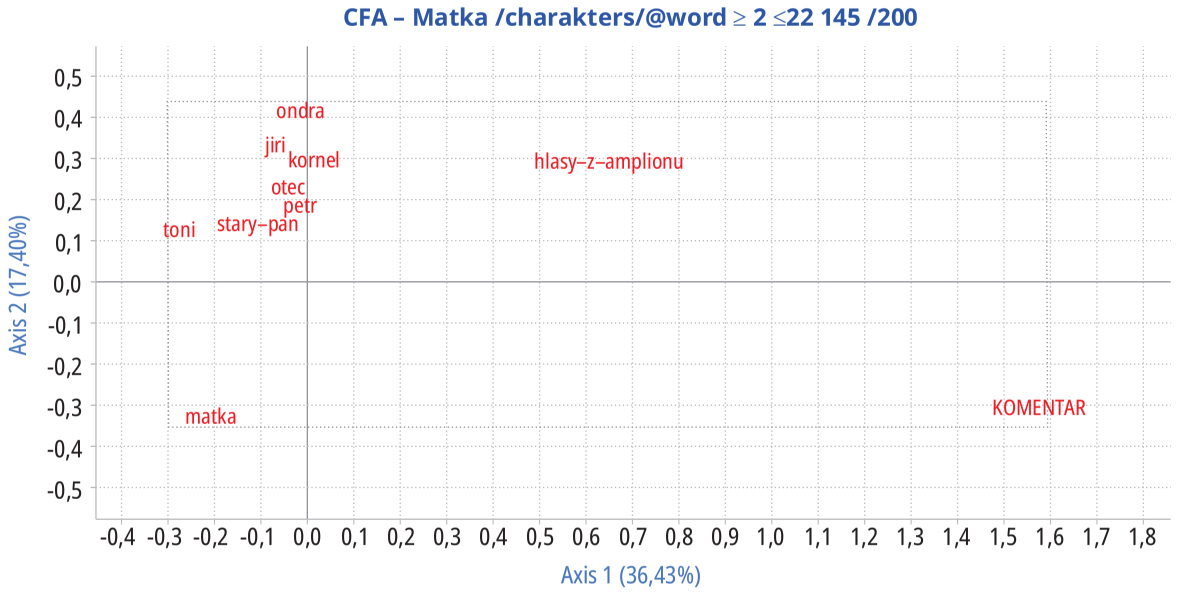

Figure 5. TXM&R – Correspondence Analysis: data – Čapek's "Matka" ('The Mother'); unit – word form

Source: own work.

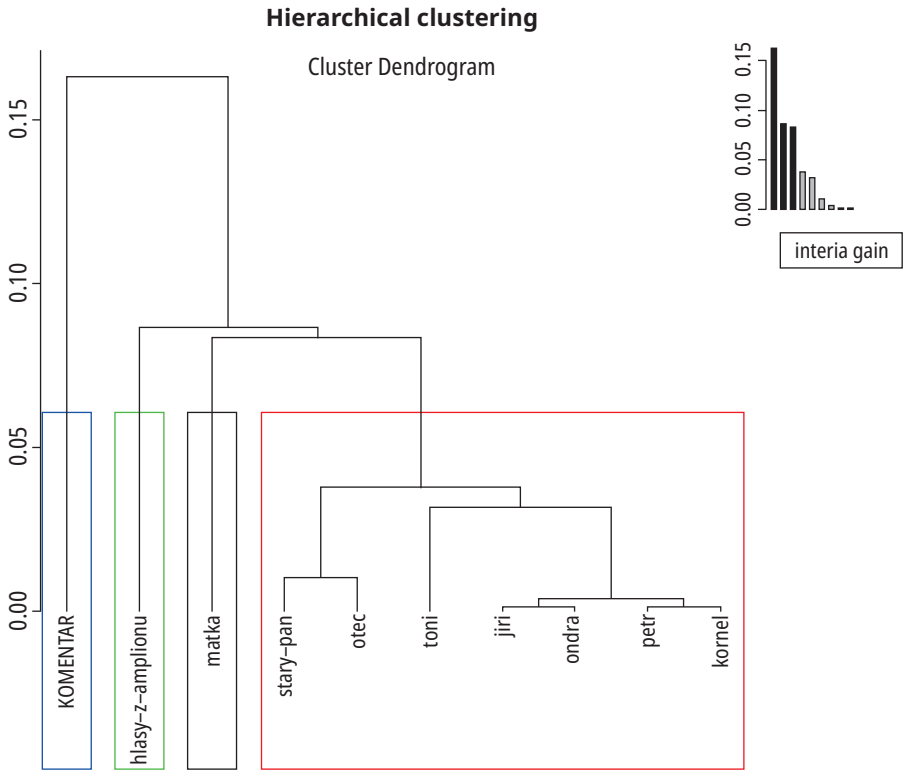

Figure 6. TXM&R – Cluster Analysis: data – Čapek's "Matka" ('The Mother'); unit – word form

Source: own work.

These multidimensional analyses then assist in generating further assumptions and hypotheses to a further and more targeted exploration of the text, helping to detect where to focus the analysis and to discover key factors of difference or similarity between sub-segments.

In this case, leaving aside the voices from behind the stage ("hlasy z amplionu") and the commentary different in both content and function in the text, MDS analysis suggests that the contrast between the mother and the other main characters (the father and her sons) will be key to a linguistic analysis of the play "Matka" ('The Mother').

9. Lexical specificity of keywords

The lexical specificity of words can be explored through the specificity score (Lafon, 1980).[8] This index is one of the adjusted frequencies that are designed to reflect the dispersion or prominence of KWs in texts. The specificity score expresses their importance in the form of hierarchical lists of frequency distributions; it indicates the tendency for usage of the given word and the degree of its usage: strong(er), weak(er), neutral, i.e. whether the unit is more or less frequent than the average use. This method does not need an external reference database, it is based on a hypergeometric function and the probability of occurrence of the unit (based on observed and expected frequencies). Words that fall into the 'banality zone' are considered neutral. The occurrence of words that exceed this zone is considered significant in a given text segment where they are either overused or deficient.

As a complementary method in this context, I used topic words based on the thematic concentration of the text and its TC/STC indexes. This method is based on the division of the wordlist into two parts determined by the h-point, which is supposed to represent the assumed boundary between autosemantic and synsemantic parts of speech, or, in other words, open vs. closed class words. Words form the open class that get above the h-point are thematic expressions. The h-point is defined as a position in which the rank of the word equals the frequency of the word, the position in which r = f (r), "r" standing for the rank of the unit while "f (r)" signifying its frequency.

Using the TC/STC index, I determined the keyword intersection of all characters from the play "Bílá nemoc" and selected the following words:

> *člověk* 'human', *nemoc* 'disease', *válka* 'war', *národ* 'nation', *mír* 'peace', *lék* 'medicine', *malomocenství* 'leprosy'

I then analysed these words through a specificity score and generated a visualization of their occurrence in the main characters (Figure 7) and in contrast to the other characters and the commentary (Figure 8) as seen below:

[8] For more information about this quantitative index (including the mathematical formula for its calculation), see Lafon (1980, pp. 127–165), and TXM Team (2018, pp. 95ff.).

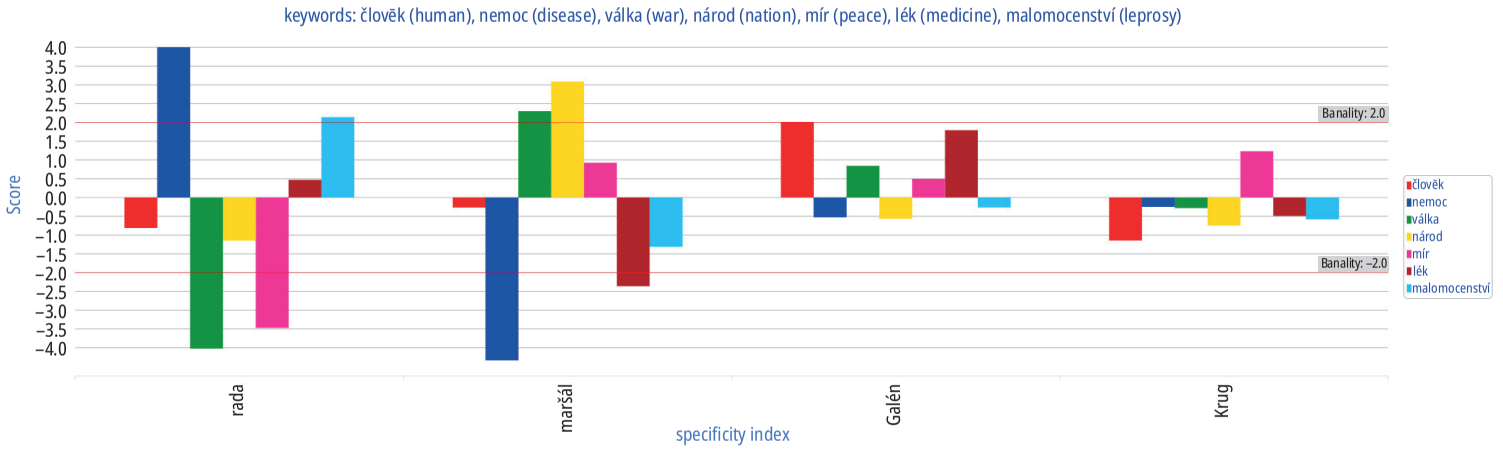

Figure 7. TXM&R – Specificity Score: visualization of the occurrence of selected KWs in the main characters

Source: own work.

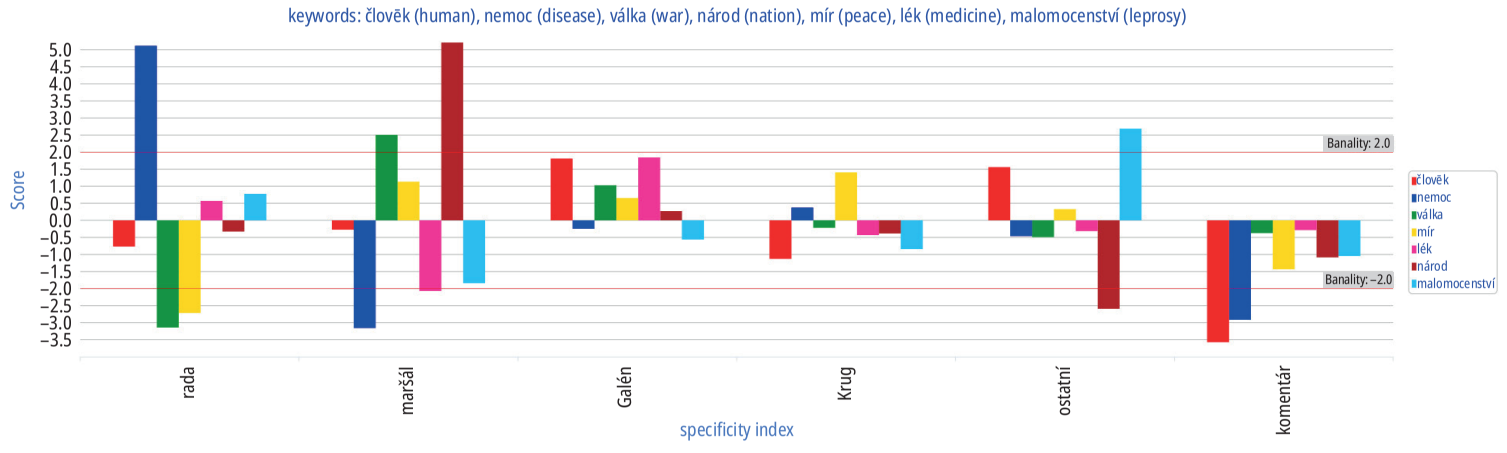

Figure 8. TXM&R – Specificity Score: visualization of the occurrence of selected KWs in the main characters in contrast to the other characters and the commentary

Source: own work.

Based on this analysis, I am able to determine not only a list of the most important words for each of the selected characters, but also to express the way in which they occur, divided into three categories: overused, underused, neutral. This method, therefore, enables the expression of the degree of significance of the occurrence. The symbols + and - express the tendency (and degree) of a neutral expression to have a positive or negative index value. The square brackets [ ] indicate words that are on the border of the banality zone, i.e., moving from the neutral field to the zone of significant occurrence (overused, underused).

The lexical specificity of selected KWs in Čapek's "Bílá nemoc" ('The White Disease'):

Strong tendency (above average use; overused):
RADA ('counsellor'): *nemoc* 'disease', *malomocenství* 'leprosy'
MARŠÁL ('marshal'): *válka* 'war', *národ* 'nation'
GALÉN: *člověk* 'human', *lék* 'medicine'
BARON KRÜG: *mír* 'peace'
OSTATNÍ ('others'): *člověk* 'human', *malomocenství* 'leprosy'

Weak tendency (below average use; underused):
RADA ('counsellor'): *válka* 'war', *mír* 'peace'
MARŠÁL ('marshal'): *nemoc* 'disease', *lék* 'medicine', [*malomocenství*] 'leprosy'
GALÉN: NULL
BARON KRÜG: [*člověk*] 'human'
OSTATNÍ ('others'): *národ* 'nation'

Neutral (average use):
RADA ('counsellor'): *člověk* (-), *národ* (--), *lék* (+)
MARŠÁL ('marshal'): *člověk* (-), *mír* (--)
GALÉN: *nemoc* (-), *válka* (++), *národ* (-), *mír* (+), *malomocenství* (-)
BARON KRÜG: *nemoc* (-), *válka* (-), *národ* (--), *lék* (-), *malomocenství* (-)
OSTATNÍ ('others'): *nemoc* (-), *válka* (-), *mír* (+), *lék* (-)

10. Conclusion

This article focused on the role and usage of proper nouns in literary texts, especially in plays, focusing on the possibilities of using these nouns for quantitative linguistic analysis. We tried to discuss content prominent units (thematic words and keywords) in drama within the framework of content analysis, and to show the key role of proper nouns for analysis and interpretation. These nouns appear both as text (in dialogues) and metatext (as labels of these dialogues) and behind all this is the simple but efficient idea to use proper noun labels as a tool for segmentation of the text and for subsequent text interpretation, so that one can analyse parts of the play belonging to a particular literary character. In addition, I also wanted to draw attention to the phenomena that may affect the quantitative analysis and, consequently, the overall results. This study presents "work in progress", testing different methods and tools for later more complex analysis of Čapek's dramas and more generally of plays as a distinct literary type or genre. Drama is a genre that stands, in its linguistic means (phenomena) and frequency structure of the text, on the borderline between written and spoken language.

Akcnowledgements

This research is supported by MEYS, grant IGA_FF_2020_021 "Czech Studies: Literary and Linguistic Overlaps and Interpretations".

Abbreviations

ARF – average reduced frequency
CNC – Czech National Corpus
ČdT – Čapek's dramatic text(s)
freq – frequency
IDF – inversed document frequency
KER – keyword extractor

KWIC – key word in context

KWs – key words

MDS – multidimensional scaling

MorphoDiTa – morphological dictionary and tagger

MWE – multiword expressions

PN-labels – labels of proper names

PropN-labels – labels of proper names

POS – part(s) of speech

PU(s) – prominent unit(s)

PWs – prominent words

SpT – spoken text

STC – secondary thematic concentration

TC – thematic concentration

TF – term frequency

TXM – textometry

TWs – topic words

WrT – written text

XML – extensible markup language

References

Čapek, K. (1994). *Dramata: Loupežník, R.U.R., Věc Makropulos, Bílá nemoc, Matka*. Praha: Český spisovatel.

Čech, R., Popescu, I. I., & Altmann, G. (2013). Methods of analysis of a thematic concentration of the text. *Czech and Slovak Linguistic Review*, *3*(1), 4–21.

Čech, R., Garabík, R., & Altmann, G. (2015). Testing the thematic concentration of text. *Journal of Quantitative Linguistics*, *22*(3), 215–232. https://doi.org/10.1080/09296174.2015.1037157

Čermák, F. et al. (2007). Capek: korpus pouze vlastních textů Karla Čapka. Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz

Cvrček, V., Čech, R., & Kubát, M. (2020). *QuitaUp – a Tool for Quantitative Stylometric Analysis.* Czech National Corpus and University of Ostrava. https://korpus.cz/quitaup/

Heiden, S. (2010). The TXM Platform: Building open-source textual analysis software compatible with the TEI encoding scheme. In R. Otoguro R & K. Ishikawa (Eds.), *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation (PACLIC 24). 4–7 November 2010, Sendai* (pp. 389–398). https://halshs.archives-ouvertes.fr/halshs-00549764/en

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique*, *1*, 127–165. https://doi.org/10.3406/mots.1980.1008

Libovický, J. (2016). *KER – Keyword Extractor*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-1650

Pořízka, P. (2019). On possibilities and methods of analysis of thematic expressions in spoken texts. *Journal of Linguistics / Jazykovedný časopis*, *70*(2), 469–480. https://doi.org/10.2478/jazcas-2019-0075

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rajaraman, A., & Ullman, J. D. (2011). Data mining. In J. Leskovec, A. Rajaraman, & J. D. Ullman (Eds.), *Mining of Massive Datasets* (pp. 1–19). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139058452.002

Rinker T. (2013). *qdap: Quantitative Discourse Analysis Package* (version 2.2.0). University at Buffalo. Buffalo, New York. https://github.com/trinker/qdap

Savický, P., & Hlaváčová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Scott, M., & Tribble, Ch. (2006). *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.22

Straková, J., Straka, M., & Hajič, J. (2014). Open-source tools for morphology, lemmatization, POS tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 13–18). Balitomre, Maryland: ACL. https://aclanthology.org/P14-5003/

TXM Team (2018). *TXM User Manual* (Version 0.7 Alpha). ENS, Lyon & Université de Franche-Comté. Retrieved May 25, 2023, from https://txm.gitpages.huma-num.fr/textometrie/files/documentation/TXM%20Manual%200.7.pdf