sciendo
DOI 10.2478/jazcas-2019-0075

ON POSSIBILITIES AND METHODS OF ANALYSIS OF THEMATIC EXPRESSIONS IN SPOKEN TEXTS

PETR POŘÍZKA
Faculty of Arts, Palacký University, Olomouc, Czech Republic

POŘÍZKA, Petr: On possibilities and methods of analysis of thematic expressions in spoken texts. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 469 – 480.

Abstract: The treatise focuses on mutual comparison of three methods of detection of prominent text units (prominent in relation to the contents of the text). The methods are: 1) analysis of key words based on comparison of source and referential corpora, 2) thematic concentration and h-point, and 3) the TF*IDF method. We try to thematize their pros and cons and, using the results of the carried out analyses, propose the optimal method for the extraction of thematic words from the spoken texts the frequency structure of which differs distinctly from the frequency structure of written texts.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 INTRODUCTION

Quantitative linguistics disposes of methods that are used to recognize main topic(s) of texts or keywords in the texts. Methods of extraction of these so-called *prominent units* are tested on texts of different genres and they are predominantly used to analyze written texts. This study intends to find out to what extent the selected methods of analysis can be used to extract prominent units in spoken texts. It is well known that in its form spoken language often differs distinctly from written language. From the quantitative viewpoint, the difference is evident even if we compare frequency vocabulary of spoken and written texts. Spoken dialogues have a specific frequency structure and a clearly distinct frequency distribution of individual parts of speech (henceforth POS). This fact can have relevant consequences since these methods of analyzing prominent units are based on word lists (or on the comparison of those lists) and on frequency structure of texts. Let us now see frequency structure of POS in large corpora of written Czech included in the Czech National Corpus (CNC; the column *CNC-written* represents the average values of POS of the SYN line of corpora) and in representative spoken corpora of CNC (the column *CNC-spoken* represents the average values of POS of ORALv1 and ORTOFONv1 corpora). In the table, relative frequency in per cent is stated.[1]

[1] Partial corpora of the SYN line contain approximately 100 million words, ORALv1 includes about 5.5 mil. words and ORTOFONv1 about 1 mil. words.

POS	CNC-written	CNC-spoken
Noun	30.53	11.41
Adj	11.48	3.50
Pron	10.48	20.27
Num	3.17	2.04
Verb	16.86	20.15
Adv	7.10	12.84
Prep	10.55	5.67
Conj	7.56	11.48
Part	0.99	8.38
Interj	0.05	0.42
resp+hes	---	*2.15*
uncomp	---	*1.04*
unknown	*1.26*	*0.65*

Tab. 1. Frequency distribution of parts of speech in written and spoken corpora of CNC. Number represent relative frequency in per cent. Legend: resp+hes = response and hesitation; uncompl = uncompleted words; unknown = expressions not recognized by a tagger

As we can see, the differences are manifested most significantly in the distribution of *nouns*: the frequency of their appearance in spoken texts is distinctly lower than in written texts (approx. 30% written vs. 11% spoken); a similarly distinct decrease is documented in the distribution of *adjectives* (approx. 11.5% vs. 3.5%) and *prepositions* (approx. 10.5% vs. 5.5%). On the other hand, the frequency of *adverbs* (7% vs. 13%) and *particles* (1% vs. 8%) rises.[2] As we will demonstrate, thematic expressions are extracted from nouns, adjectives and verbs. And nouns, as expressions signifying *substances*, are undoubtedly significant for any method the aim of which is to detect prominent text units. On the basis of these differences we intend to find out to what extent the perceptibly lower frequency distribution of nouns (and possibly even other differences) will be manifested in our analyses carried out with the use of selected methods.

2 DATA, METHODS, TOOLS

For our probe we chose two of currently often used methods of extraction of prominent units: 1) *analysis of keywords* and 2) the method of measuring *thematic text concentration*, namely the part of the method in which thematic words are detected. The third method is 3) TF*IDF method (*Term Frequency* vs. *Inverse Document Frequency*), used in semantic analysis of texts. In the text analyses, following freely available software tools were applied: (ad 1) *KWords* [1], (ad 2) *QUITA* [2], and (ad 3) *KER – Keyword Extractor* [3].

[2] Among particles even hesitation and response sounds might be included (the category of *resp+hes* in Table 1); thus their proportional representation would rise by 2% to the final proportion of 10%.

The analyzed data were formed by 20 spoken texts randomly selected from the so-called *Olomouc spoken corpus* (henceforth OSC) [4]. We used orthographically normalized/standardized versions of transcripts that were further purified in order to suit our intentions. We removed all their parts that could affect textual analysis: particularly marks of individual speakers (before all lines) and all meta-textual marks and commentaries. Individual transcripts contained between 2,300 and 4,500 words (the average of 3,135 words in a transcript); the overall size of the dataset was 62,694 words.

The transcripts were subsequently lemmatized for *KWords* and *KER* with the use of *MorphoDiTa*, a morphological analyzer and tagger [5]. While working with QUITA we used a morphological analyzer *Majka* [6].

Quantitative analysis of so-called *keywords* (further on also KWs) ([7], [8]), based on the comparison of the source text (SourceC) with so-called referential corpus (RefC) is certainly one of the most commonly used methods of content analysis of texts. For keywords we take the words the frequency of which is remarkably higher in the SourceC than in the RefC. Nevertheless, the choice of the RefC influences even the overall result of the analysis and it is therefore recommended to choose textually neutral databases that reflect common language usage. For the detection of statistic relevance of differences two statistical tests are used: *log-likelihood* and *chi-squared test*. Even this exact method has its difficulties that have to be faced, namely with respect to appropriate combination of computing parameters. It is primarily necessary to set *the level of statistic significance* of the test (most frequently to 0.05, 0.01, 0.001, or even more) and sometimes other parameters (see below Sec. 3.2). It is also possible to apply so-called *stop-lists* on the text; by stop-lists we mean the lists of words or word groups that are *a priori* excluded from the analysis of KWs. Among problematic aspects of this kind of analysis belongs the fact that the analysis produces quite large lists of detected KWs (sometimes containing hundreds or even more words) that have to be in some way reduced in order to be used in subsequent analysis and interpretation of the text. Such reductions are often arbitrary, based on some *ad hoc* criteria: most frequently only the group of the initial 20, 50 or 100 words is taken from the list of all detected KWs and applied in the interpretation. That is why even the position of a certain keyword in the final list is important, the position reflecting a simple principle: the higher in the list the KW appears, the more relevant it is for the contents and topic(s) of the text. In this way KWs are hierarchized; KWs can certainly be sorted out according to the coefficient of the main statistical test. We can also use any index reflecting the relevance of different distribution of the word in the SourceC and in the RefC, or the index applied in order to neutralize the different sizes of source and referential corpora. For example, in the latest version (3.5.8) of a concordance tool *AntConc* [9] 10 indexes of this kind are implemented.[3]

[3] Compare individual indexes in the menu of *Keyword Effect Size Measure*.

The above stated characteristics of *keyword analysis* show that this is a relatively demanding procedure during which one must set many parameters that affect the process and the resulting list of KWs. Researchers therefore look for other ways and methods leading to the revelation of main topics of texts. Recently, namely the analysis of thematic words has been tested and developed that utilizes measuring of *thematic text concentration* (further on also TC) [10]. The method is based on simple extraction of thematic words (TW) from a word list; to detect thematic words one needs no external database nor further mathematical modeling of the text that would prefer certain words to others and modify their position in the word list. The method considers as thematic the words that occupy the positions above so-called *h*-point in the word list, while the *h*-point is defined as a position in which the rank of the word equals the frequency of the word.[4] The *h*-point concurrently represents an indistinct borderline between autosemantic and synsemantic POS: all autosemantic expressions, with the exception of adverbs and certain verbs (see below) that appear above the *h*-point are subsequently considered as main topics of the text.

Nevertheless, in practice the use of the method often results in empty TW sets. The texts with an empty set of thematic words are subsequently considered as thematically neutral while the texts in which one detects TWs are thematically determined. In order to eliminate the cases of empty TW sets, the so-called STC (secondary thematic concentration) was implemented in the method which means that the TC value is multiplied by 2 in order to shift the *h*-point lower in the word list and to increase the chance of finding some prominent units. We consider this solution as rather problematic since it is quite arbitrary and it leaves without explanation why TC values are multiplied by 2 and not by other numbers. But there is also a question: Isn't the choice of *h*-point arbitrary in itself?

The choice of an elementary text unit is methodologically relevant as well. Shall it be the word form, a lemma, or even other unit? It is common to take a *text form* as the elementary unit of the analysis but it is evident (from previous analyses) that in case of a strongly inflective Czech *lemma* is definitely a more appropriate choice since it represents all text forms of a lexeme.

3 ANALYSIS AND INTERPRETATION OF ITS RESULTS[5]

3.1 TC and thematic expressions

Lemma is the elementary unit of our analyses. Besides their lemmatization we annotated the texts even morphologically – we assigned the mark of its affiliation with a particular part of speech to each text unit. In Table 2 below we indicate frequency distribution of individual POS in our specimen of data in comparison with

[4] For comments to the formula and to the calculation of the *h*-point see [10] (pp. 11nn).

[5] Here we will restrict ourselves to interpretational remarks. Complete resulting lists of KWs are available and can be freely downloaded at: `http://corpus.upol.cz/system/files/KWs-lists.zip.`

morphologically annotated spoken CNC corpora. Since we applied two different taggers (see *Sec. 2*) both variants of annotation are presented in the table:

POS	**OSCsample20** *MorphoDiTa*	**OSCsample20** *Majka*	**ORAL v1**	**ORTOFON v1**
Noun	13.59	12.19	11.63	11.18
Adj	4.03	4.02	3.63	3.38
Pron	20.37	20.33	20.86	19.67
Num	1.89	1.81	1.77	2.31
Verb	21.8	21.67	20.46	19.84
Adv	14.37	15.5	12.93	12.74
Prep	5.92	5.96	5.66	5.69
Conj	11.84	11.67	11.51	11.46
Part	5.3	4.51	8.13	8.63
Interj	0.91	0.77	0.43	0.4
resp+hes	---	---	*1.64*	*2.67*
uncomp	---	---	*0.75*	*1.33*
unknown	*0*	*1.62*	*0.6*	*0.7*

Tab. 2. Comparison of frequency distribution of POS in the analyzed specimen of spoken data (*OSCsample20*) and in the spoken CNC corpora. The numeric values signify relative frequency in per cent.

he comparison enables us to suppose that the selected specimen of spoken data can be considered as representative since the frequency distributions of POS correspond with those in much larger databases (OSCsample20: N ≐ 63 thousand; ORTOFON: N 1.03 ≐ million; ORAL: N ≐ 5.5 million words). It is significant, namely with respect to the TC method and its POS limitation of thematic words. We notice certain deviations (for example in the frequency distributions of *particles*, *adverbs* or *nouns*) but they are only minute (avg. 1.5% in case of nouns, 2.1% in case of adverbs, and 3.5% in case of particles) and therefore they cannot affect the analysis of thematic words. The proportional representation of nouns in *OSCsample20* is even slightly higher than in the CNC corpora.

The results of the analysis of thematic words carried out with the use of QUITA tool are presented in Table 3:

DOC	**TWs according to TC**
1	vědět ‘to know’, jít ‘to go’
2	**0**
3	vědět ‘to know’, říkat ‘to say’
4	**0**
5	vědět ‘to know’

6	vědět ‘to know’
7	vědět ‘to know’, **hrát ‘to play’**, dělat ‘to do’
8	**0**
9	říkat ‘to say’, vědět ‘to know’
10	**rok ‘year’**, **fotbal ‘soccer’**
11	vědět ‘to know’, **koupit ‘to buy’**
12	vědět ‘to know’
13	jít ‘to go’
14	vědět ‘to know’, jet ‘to go’, jezdit ‘to go’
15	**0**
16	vědět ‘to know’
17	**0**
18	jet ‘to go’, vědět ‘to know’
19	vědět ‘to know’
20	dobrý ‘good’, vědět ‘to know’

Tab. 3. Results of the analysis of thematic words in spoken texts (OSCsample20). TWs are arranged according to their ranking.

In 5 out of 20 texts, i.e. in 25% of cases, no thematic words were found – they are texts Nos 2, 4, 8, 15 and 17 (mind their absence in Table 3). In all remaining documents only 11 different prominent units were found. We suppose that only some of them can be considered as real thematic words. Particularly they are these: *hrát* ‘to play’*, koupit* ‘to buy’, *rok* ‘year’*, fotbal* ‘soccer’. They are marked by bold print in Table 3 and they were found only in 3 out of 20 texts. Other lexemes rather indicate deviation from semantic (thematic) to pragmatic use (we verified the character of their behavior with the use of concordances in our corpus). It is true namely in case of the verb *vědět* (‘to know’, a verb of mental action, communication) that appeared in 13 out of 15 texts or in the cases of *říkat* (‘to say’, v. dicendi, communication) and *dobrý* (‘good’, an evaluative adjective). As prominent units only two other verbs were detected: *dělat* (‘to do’, v. faciendi), and *jít/jet/jezdit* (‘to go’, v. movendi).

A question arises whether it is possible to take the lexeme with noticeably pragmatic use for thematic expression. In spoken texts such lexemes often function as phatic, conative or emotional/expressive words while real thematic words should function as referential units (that signalize the relation to the topic).

The situation slightly improves in case of the STC index (Table 4). Nevertheless, we consider (as we stated above) STC as methodologically problematic. Besides, the authors of TC consider the texts without TWs as thematically neutral and the texts with TWs as thematically determined. A paradoxical situation thus arises in which the same texts with originally empty TW sets suddenly, thanks to STC, become thematically determined.

DOC	TWs according to STC
1	(vědět 'to know'), jít 'to go', dělat 'to do', (říci 'to say'), (říkat 'to say')
2	**baterka 'flashlight'**, dát 'to give', (vědět 'to know'), udělat 'to do', (říkat 'to say'), **třešeň 'cherry'**
3	(vědět 'to know'), (říkat 'to say'), (říci 'to say')
4	(vědět 'to know'), potřebovat 'to need', **lednička 'fridge'**, dát 'to give', udělat 'to do', **koupit 'to buy'**, jet 'to go'
5	(vědět 'to know'), (říkat 'to say'), (myslit 'to think'), jet 'to go', dělat 'to do'
6	(vědět 'to know'), (hezký 'pretty'), (myslit 'to think'), (krásný 'beautiful'), **Krkonoše**
7	(vědět 'to know'), **hrát 'to play'**, dělat 'to do', dát 'to give', **statistika 'statistics'**, (říci 'to say'), **kluk 'boy'**, **zápas 'match'**, jít 'to go', (mhm), (myslit 'to think')
8	(vědět 'to know'), **sval 'muscle'**, jet 'to go', **mozek 'brain'**, (říkat 'to say'), dělat 'to do'
9	(říkat 'to say'), (vědět 'to know'), jít 'to go', (dobrý 'good'), dívat 'watch', **napsat 'to write'**, **psát 'to write'**
10	**rok 'year'**, **fotbal 'soccer'**, **hrát 'to play'**, (myslit 'to think'), jít 'to go', **řada 'row'**, **celý 'all'**, **hráč 'player'**, (říci 'to say')
11	(vědět 'to know'), **koupit 'to buy'**, **libra 'pound'**, jít 'to go', dát 'to give', (myslit 'to think')
12	(vědět 'to know'), vidět 'to see', (myslit 'to think'), (říkat 'to say')
13	jít 'to go', (říkat 'to say'), (vědět 'to know'), **pamatovat 'to remember'**, **dítě 'child'**, chodit 'to go'
14	(vědět 'to know'), jet 'to go', jezdit 'to go', (říkat 'to say'), jít 'to go', psát 'to write', **týden 'week'**, **škola 'school'**, **Honza 'Johnny'**, přijet 'to come', **spát 'to sleep'**
15	(vědět 'to know'), jít 'to go', dělat 'to do', jet 'to go', **člověk 'man'**, (dobrý 'good')
16	(vědět 'to know'), jít 'to go', (říkat 'to say'), (dobrý 'good'), (říci 'to say')
17	**Martin**, (dobrý 'good'), (vědět 'to know')
18	jet 'to go', (vědět 'to know'), jít 'to go', (dobrý 'good'), **Skotsko 'Scotland'**
19	(vědět 'to know'), jít 'to go', přijít 'to come'
20	(dobrý 'good'), (vědět 'know'), **fotka 'photo'**, **jméno 'to go'name**, vidět 'to see', dívat 'to watch'

Tab. 4. Thematic words according to STC (OSCsample20).
TWs are arranged according to their ranking.

This time thematic words were detected in all partial documents of the dataset. Even if the STC caused the growth of detected lexemes they are actually verbs (or adjectives) again, functioning as pragmatic (phatic) words. The words can further be gathered in groups that share the same word-formation base or form pairs in which one verb is imperfective and the other one perfective: *říci–říkat* 'to say'*, dělat–udělat* 'to do'*, psát–napsat* 'to write'*, jít–přijít–chodit* 'to go on foot'*, jet–jezdit* 'to go'. Among adjectives we can find increments with the same meaning and function and belonging to the same category (evaluative words): *dobrý* 'good', *hezký* 'pretty', *krásný* 'beautiful'.

On the basis of the behaviour of all prominent units in the spoken texts verified with the use of corpus concordances the prominent TC/STC words can be divided in three zones/categories:

1) *non-thematic expressions* with pragmatic function (such as *dobrý* 'good'*, hezký* 'pretty'*, myslet* 'to think', *vědět* 'to know'*, říkat* 'to say') – in Table 4 they are stated in parentheses;
2) *a broad transitional zone of borderline expressions:* namely *verbs* and *adjectives* that can be recognized as both thematic and pragmatic (for example *vidět* 'to see'*, dívat se* 'to watch'*, potřebovat* 'to need'); these expressions appear repeatedly in most analysed texts;
3) *truly thematic expressions* (for example *baterka* 'torch'*, lednička* 'fridge'*, zápas* 'match'*, fotbal* 'soccer'*, škola* 'school' etc.) – they are almost solely *nouns* – in Table 4 they are stated in bold print.

If we sum the results of our analyses up they seem to suggest that, in case of spoken texts, the TC/STC method fails. It may be caused by the fact that spoken texts differ from the written texts significantly: they have a specific frequency structure of the text/vocabulary, they contain many pragmatically used expressions, functioning as phatic, conative or emotive words.

3.2 *KWords* and key-lemmas

We used the *KWords* tool and carried out the analysis with following settings:

- stop-list: pronouns, prepositions, conjunctions, numbers
- methods: *log-likelihood*
- significance level (α): 0.0001
- minimal frequency: 3
- percentage of registered keywords: all significant types
- referential corpus: SYN2015

The list of keywords can be arranged according to DIN that signalizes the relevance of differences in KWs in the SourceC and RefC. We limit the list of lemmas to the units of high and highest prominence (DIN > 95).[6]

[6] For the calculation formula and more detailed description of DIN values see [1].

We analysed 5 documents of the *OSCsample20* set; three of them (Nos 2, 4, and 8) had an empty TC set while in case of the remaining documents (Nos 7 and 11) the set was not empty. Given the extent of the study and the fact that the resulting list of keywords are rather large, we will limit ourselves merely to brief remarks and possible conclusions that follow from our analyses:

- The DIN index functionally and effectively reduces the number of keywords and it also hierarchizes KWs.
- If we limit the list of lemmas to the units of high (DIN: 95–97) and highest (DIN 98–100) prominence, the resulting lists will contain approx. 40 up to 60 words in the texts (cf. Below):

DOC	DIN 95–97	DIN 98–100	DIN 95–100
2	20	36	56
4	16	38	54
7	30	37	67
8	13	43	56
11	29	35	64
MEAN	**21.6**	**37.8**	**59.4**

Tab. 5. The number of keywords in *KWords* tool

- Only very few pragmatically used words appear in the lists: they are following particles (*ano, jo, hm, no, tož*) or interjections (*aha, hele, jé*) and should here be regarded not as prominent units but rather as pragmatically applied words.
- Frequency POS distribution of resulting keywords suggests that the highest positions in the list are actually occupied by thematically significant expressions:

POS	FREQ	FREQ %
Noun	**134**	**44.97**
Verb	**110**	**36.91**
Adj	22	7.38
Part	18	6.04
Adv	6	2.01
Interj	6	2.01
Num	2	0.67
Total	**298**	**100.00**

Tab. 6. Frequency distribution of POS in *KWords* tool

- Unlike in TC, certain adjectives (such as *dětský* ‘childish’, *infekční* ‘infectious’, *levoruký* ‘left-handed’, *etc.*) and verbs (such as *lyžovat* ‘to ski’, *pršet* ‘to rain’, *vyléčit* ‘to rain’, etc.), i.e. the words that can truly be regarded as thematically prominent.

It seems that the *analysis of keywords* is more suitable for the detection of prominent units in spoken texts than the method based on *thematic concentration of texts*.[7]

3.3 *KER* and TF*IDF method

TF*IDF method [11] compares the frequency of the word in the analysed text with the "reversed" frequency of the word in all documents. IDF expresses the "relevance" of the word: the more frequently a particular word appears in the documents the less relevant it is for the analysed text. From mathematical viewpoint the method is relatively simple:

TF(t) = (Number of times term t appears 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 it).

The demo version of *KER – Keyword Extractor* has certain limitations. We therefore carried out analyses with following settings:[8]

- TF*IDF threshold level: 0.05
- maximum number of keywords: 25

This setting has turned out as optimal in majority of the analysed texts: the resulting number of KWs is lower than the pre-set maximum limit (5 out of 20 texts reached the maximum limit). Moreover, it turned out that the detected number of KWs does not depend directly on the length of the text. Texts Nos 1, 4, and 14 that reached the maximum limit of 25 KWs do not even contain the average number of words. By the way of contrast, texts Nos 4 and 10 (approx. 2,300 words) and texts Nos 13 and 17 (approx. 3,000 words) are almost equally long. Nevertheless, the numbers of KWs that were found in texts of the same length differ diametrically: doc4: 25 × doc10: 4; doc13: 25 × doc17: 7. Cf. below:

DOC	**TOKENS**	**KWs**
1	2561	**25**
2	3667	14
3	3333	11
4	2358	**25**
5	3183	13
6	2835	15

[7] We should point out that we compared spoken texts (SourceC) with written ones (RefC). Therefore, we would like to examine the possible influence of the reference corpus (different register) by means of further analyses in the future.

[8] For example when set to more than 25 KWs, the application signalizes failure of the database and it stops the whole process.

DOC	TOKENS	KWs
7	3949	20
8	2449	20
9	4537	13
10	2316	4
11	3547	20
12	2306	14
13	3042	**25**
14	3706	**25**
15	2860	11
16	3916	**25**
17	2935	7
18	3874	21
19	2530	21
20	2790	14
MEAN	**3134.70**	**17.15**

Tab. 7. The resulting number of KWs *in KER*

The TF*IDF method generates approximately the same amount of words as analysis of keywords, 340 vs. 300, but the resulting structures of POS differ significantly (compare Tables 6 and 8). TF*IDF actually detects only *nouns* (85%) and *adjectives* (14%), with the exceptions of *hm* (a particle) and *ježiš* (an interjection).

POS	FREQ	FREQ (%)
Noun	**291**	**84.84**
Adj	**49**	**14.29**
Interj	2	0.58
Part	1	0.29
Total	**343**	**100.00**

Tab. 8. Frequency distribution of POS in *KER*

The results document an important characteristic of TF*IDF: the method truly effectively eliminates all phatic expressions, hesitations, responses, and other phenomena that occur in spoken texts very frequently. In the final list, even certain autosemantic POS are missing, particularly verbs and adverbs. Generally we can conclude that the TF*IDF method appears as the most promising; the extracted words can certainly be considered as thematically relevant, their number is not too high and it needs no reduction (necessary if analysis of keywords is applied). During testing we observed that the results were influenced by the length of the analysed text (the setting of elementary parameters was constant, TF*IDF threshold level + max. number of KWs): the longer the text was the less words appeared in the list of Kws.

4 CONCLUSION

The TF*IDF seems to be a good alternative that can solve or eliminate the drawbacks of respective variant methods. Analysis of KWs generates an extensive list of prominent units that needs reduction while the TC method often results in very short or even empty lists of thematic words. We are aware of the fact that more analyses will have to be carried out, testing more extensive materials and various types of texts (prepared vs. unprepared spoken texts) in order to map out the character of the TF*IDF method and to find optimal settings of the key parameters.

References

[1] Cvrček, V., and Vondřička, P. (2013). KWords. Praha. Accessible at: `http://kwords.korpus.cz.`

[2] Matlach, V., Kubát, M., and Čech, R. (2014). QUITA – Quantitative Text Analyzer. Olomouc. Accessible at: `https://code.google.com/archive/p/oltk/.`

[3] Libovický, J. (2016). KER – Keyword Extractor. Praha. Accessible at: `https://lindat.mff.cuni.cz/services/ker/.`

[4] Pořízka, P. (2009). Olomouc Corpus of Spoken Czech: characterization and main features of the project. Linguistik online, 38(2), pages 35–43.

[5] Straka, M., and Straková, J. (2014). MorphoDiTa: Morphological Dictionary and Tagger. Praha. Accessible at: `http://lindat.mff.cuni.cz/services/morphodita/.`

[6] Šmerk, P. (2009). Majka –Morphological Analysis of Czech. Brno. Accessible at: `https://nlp.fi.muni.cz/czech-morphology-analyser/.`

[7] Stubbs, M. (1996). Text and Corpus Analysis. Oxford.

[8] Scott, M., and Tribble, Ch. (2006). Textual Patterns. Key words and Corpus Analysis in Language Education. Amsterdam – Philadelphia.

[9] Anthony, L. (2019). AntConc. Tokyo. Accessible at: `http://www.laurenceanthony.net/software.`

[10] Čech, R. (2016). Tematická koncentrace textu v češtině. Praha.

[11] Rajaraman, A., Ullman, J.D. (2011). Data Mining. In Leskovec, J. et al., Mining of Massive Datasets. pages 1–17.