Miroslav Kubát*, Jan Hůla, Xinying Chen, Radek Čech and Jiří Milička

The lexical context in a style analysis: A word embeddings approach

https://doi.org/10.1515/cllt-2018-0003

Abstract: This is a pilot study of usability of Context Specificity measure for stylometric purposes. Specifically, the word embedding Word2vec approach based on measuring lexical context similarity between lemmas is applied to the analysis of texts that belong to different styles. Three types of Czech texts are investigated: fiction, non-fiction, and journalism. Specifically, forty lemmas were observed (10 lemmas each for verbs, nouns, adjectives, and adverbs). The aim of the present study is to introduce a concept of the Context Specificity and to test whether this measurement is sensitive to different styles. The results show that the proposed method Closest Context Specificity (*CCS*) is a corpus size independent method which has a promising potential in analyzing different styles.

Keywords: neural networks, word embedding, word2vec, stylometry, style

1 Introduction

It has been known for decades that a human language behavior is influenced by pragmatic factors (Bublitz and Norrick 2011). There are many methods how to capture and describe this influence in both qualitative and quantitative way (Golato and Golato 2012). Naturally, it is easier (and more unambiguous)

***Corresponding author: Miroslav Kubát,** University of Ostrava, Reální 5, Ostrava, Czech Republic, E-mail: miroslav.kubat@gmail.com
http://orcid.org/0000-0002-3398-3125
Jan Hůla, University of Ostrava, 30. dubna 22, Ostrava, Czech Republic, E-mail: jan.hula21@gmail.com
Xinying Chen, Xi'an Jiaotong University, Xi'an, China, E-mail: cici13306@gmail.com
Radek Čech, University of Ostrava, Reální 5, Ostrava, Czech Republic; Xi'an Jiaotong University, Xi'an, China, E-mail: cechradek@gmail.com
Jiří Milička, Faculty of Arts, Charles University, Institute of the Czech National Corpus, Prague, Czech Republic, E-mail: milicka@centrum.cz
http://orcid.org/0000-0001-8605-1199

to use formal characteristics of language in these kinds of analysis. For instance, phonetic, morphological or syntactic properties are successfully utilized in authorship analysis, stylometry and sociolinguistics (cf. McMenamin 2002; Juola 2006; Grieve 2007; Popescu et al. 2009; Mikros and Perifanos 2013; Kubát 2016). On the other hand, there are some fundamental language characteristics which seemed to be used only in qualitative research because of their fluidity. The outstanding example is, for some scholars, semantics, c.f.

> This coupling between word meanings and innovative thought means that word meanings have an unpredictability that, arguably, makes them incapable of being brought within the purview of empirical scientific theorizing. Fluidity of word sense is not merely a phenomenon arising from changes over centuries in the use of words relating to lofty subjects: it applies to all words all the time. In consequence, trying to produce a rigorous, scientific account of the semantics of human language may be task as futile as chasing the rainbow. (Sampson 2001: 184–185).

However, very innovative methods of language analysis which are connected to its semantic properties have appeared in recent years; namely, approaches based on neural network representations which are also known as word embeddings (Mikolov et al. 2013a; Manning et al. 2014). Profiting from the advancement of computation capacity, we can now quantitatively model the dynamic lexical context feature, which of course includes semantic feature, of "all words" in a text. In linguistics, these methods were applied, among others, for modeling semantic similarities among words as well as semantic changes (Hamilton et al. 2016). In this study, we tend to scrutinize a potentiality of the word embedding model for a style analysis. Specifically, each lemma's context specificity is computed based on the neural network analysis first. Furthermore, differences of particular lemmas' context specificity among styles are observed. We assume that there should be significant distinctions among lexical context characteristics (i.e. the context specificities, in our case) of these lemmas caused by styles, which represents an "aggregate" of pragmatic factors.

Lemmas are considered as the basic units in this research because of the inflectional character of Czech language. Czech is a highly inflected language where different endings express different grammatical categories such as case, number or gender in declension (nouns, adjectives, pronouns, numerals), and person, number or tense in conjugation (verbs). For example, the lemma *pes* 'a dog' has twelve different word forms for indicating its grammatical categories: pes, psa, psovi, psu, pse, psem, psi, psové, psů, psům, psy, psech.

For testing the idea, the measurement, Context Similarity of Lemma, is applied to the Czech National Corpus data. This recently proposed method (Čech et al. 2018) showed convincing results in the lexical context analysis in the Czech political discourse. The present work follows up this research and aims to discover whether the proposed approach could reveal some hidden properties of different styles and can be therefore potentially applied in stylometry. However, it should be emphasized that this research is just the first look at the issue and cannot be therefore considered as a well-established style analysis. It is a pilot study of Context Specificity in linguistic research (differences between styles in this specific case).

The article is organized as follows; first, the neural network approach is briefly introduced. Second, the corpus and the methods are described. Finally, the results are presented and discussed.

2 Neural networks approach in stylometry

Stylometry is predominantly a branch of applied linguistics which can be described as quantitative stylistics. The stylometric research can be traced back to nineteenth century and it is connected to the main domain of this linguistic field – the authorship attribution. The pragmatic goal is to recognize authors of anonymous texts or to solve the cases of questionable authorship of some literary work. The most famous cases are for instance disputed authorship of some Shakespeare's plays or anonymously published *The Federalist Papers* or the novel (so-called roman à clef) *Primary Colors*. The need of finding real authors of such texts pushes scholars to apply quantitative methods in linguistics because of their rigor and efficiency. Besides authorship, stylometric research also focuses on genre analysis which is a more theoretical field (cf. Grieve 2005, Grieve 2007; Juola 2006).

Stylometric methods can be divided into two basic groups: "interpretable" and "uninterpretable". The first group consists of relatively simple indicators such as word length, sentence length, vocabulary richness, activity of text or thematic concentration. These methods are based on traditional linguistic approach and units, can be easily computed and, above all, are well interpretable. Their application is suitable mainly in genre analysis or single author's style description for linguists or literary criticism scholars. The goal is to understand the mechanism of writing various styles or to describe a genre or an author's style. The second group of stylometric

methods consists of more complex computational procedures and it often uses "artificial" units such as n-grams which are not common in traditional linguistics. These methods are usually referred to "black box methods" because they are constructed with respect to a performance on some concrete tasks and not with interpretability in human mind. This approach is commonly predominant in authorship attribution or forensic linguistic studies where the most important criterion is the probability of identifying a real author of a text.

Applying neural networks approach to stylometry can be traced back to the last decade of twentieth century (cf. Matthews and Merriam 1993). In recent years, the very innovative and successful approaches to language analysis have appeared; namely, methods based on representations which are also known as word embeddings (Mikolov et al. 2013a; Manning et al. 2014). In linguistics, word embeddings were for example used for tracking the semantic evolution of words and to quantitatively tested hypotheses about semantic change (Hamilton et al. 2016). And for this study, neural networks approach is applied to stylometry, specifically to style analysis.

Neural networks represent a set of methods which are effective for finding useful representations of data, which are usually collected in a form that is not ideal or even suitable for a task at hand. For example, words are represented as a sequence of characters which is not a suitable representation for finding out whether two words are used in similar contexts. Neural networks produce useful representations by taking the original representation as an input and transforming them through series of numerical operations to different representations. The exact value of the output representation is dependent on the learnable parameters. Concrete values of these parameters are found by minimizing an error function on a concrete task. We can use the obtained error to update the parameters of the network in a way which tries to lower the error. By iterating this process, we are minimizing the error and thus finding a better representation for the task. In our case, we want the representation to be a good predictor of the contexts in which the word appears (this is measured by how well it can predict the words which appear in specific contexts). Thus, if two words often appear in the same context, their representations should be similar.

Representations with this property are easy to obtain with methods collectively called Word Embeddings (Mikolov et al. 2013a; Manning et al. 2014) where the aim is to represent a word (in our case a lemma) as a multi-dimensional (50–1000) vector. This vector captures co-occurrence

statistics between the lemma itself and other lemmas in the small window centered at the lemma at hand. The window acts as a context for the centered lemma. Intuitively the vector representing the lemma contains information about the contexts. Concrete values of these vectors are found by maximizing an objective function which measures how well every lemma can be predicted based on its neighbor lemmas. This objective function has the following form:

$$J(\theta) = \frac{1}{T}\sum_{t=1}^{T} \sum_{-m \le j \le m, j \neq 0} \log p(w[t+j]|w[t])$$

This function is maximized when the sums of the individual probabilities are maximized. The first sum (indexed by t) iterates over all tokens in the corpus (the number of tokens is T). The second sum (indexed by j) iterates over all tokens in the small window centered at the token with an index t. This window is of length $2m + 1$ (there are m lemmas on every side of the central lemma). Intuitively we want the lemmas inside this window ($w[t + j]$) to be predictable from the central lemma ($w[t]$). For example, when the lemma $w[t]$ is *funny* and the lemma $w[t + 1]$ is *joke*, we want p(joke|funny) to be high so that the lemma *joke* is predictable from lemma *funny*.

This kind of predictability is measured by a function with related vectors as arguments. Concretely, the conditional probabilities in the equation above are estimated by the following function:

$$p(o|c) = \frac{exp\left(u(o)^T . v(c)\right)}{\sum_{w=1}^{W} exp\left(u(w)^T . v(c)\right)}$$

where $u(o)$ and $v(c)$ are vector representations of lemmas o and c.

The first thing to notice is that every lemma is parametrized by a set of two vectors (u and v). One vector (v) is used when the lemma appears in the center of the window and the second vector (u) is used when the lemma appears as a context lemma. For example, when the window is centered at the lemma *funny*, then the first vector is used as its representation, but when the window is centered at some other lemma and the lemma *funny* appears in this window as a context word, then we use the second vector as its representation. These two vectors are used only to simplify the optimization problem. At the end, these representations could be averaged or one of them can be discarded. After the optimization, the lemmas which appear in similar

contexts will have similar vectors. Thus, the exact values of these vectors are not interpretable, but similarity of two vectors could be interpreted. For measuring this kind of lexical context similarities between lemmas we use the cosine similarity as suggested by (Levy et al. 2015). We first normalize all vectors to unit length and then the cosine similarity is equivalent to dot product between these normalized vectors. Therefore, when the vectors point in the same direction, their similarity is 1, when they point in opposite directions their similarity is –1, and when they are orthogonal then their similarity is 0. In other words, if the similarity is close to 1, then the contexts in which these lemmas appear are positively correlated, and when it is close to –1, then they are negatively correlated, and when it is close to 0, then they are uncorrelated. For the concrete details about this optimization procedure see Mikolov et al. (2013b) and Čech et al. (2018).

3 Language data

The neural network approach in linguistic research requires huge data for training. That is why we choose one of the largest corpora of contemporary Czech written language SYN_V4. This corpus is one of so-called SYN series corpora of the Czech National Corpus (Křen et al. 2016). "SYN" refers to "synchronic" and every version consists of texts from all reference synchronic written corpora of the SYN series published up until the given version of the SYN corpus (Hnátková et al. 2014). The size of the SYN_V4 is 3, 626 billion tokens. This corpus consists of three text types: journalism, non-fiction, and fiction. The number of texts is not proportional in this data set; the journalistic texts comprise more than 90 % of the whole corpora (see Figure 1). To avoid a potential bias caused by big differences between the size of the subcorpora in our style analysis, the journalism is arbitrarily limited only to the texts published in 1999, 2006, 2014. The size of the subcorpora used in this study can be seen in Figure 2.

Although the study deals with styles, it is necessary to emphasize that we use the term "style" in accordance with the metadata used in the corpus. Journalism is represented by articles in newspapers and leisure magazines. Non-fiction consists of scientific, professional and popular texts from various fields (humanities, social sciences, natural sciences, technical sciences) and also memoirs, autobiographies or administrative texts. Fiction contains novels, short

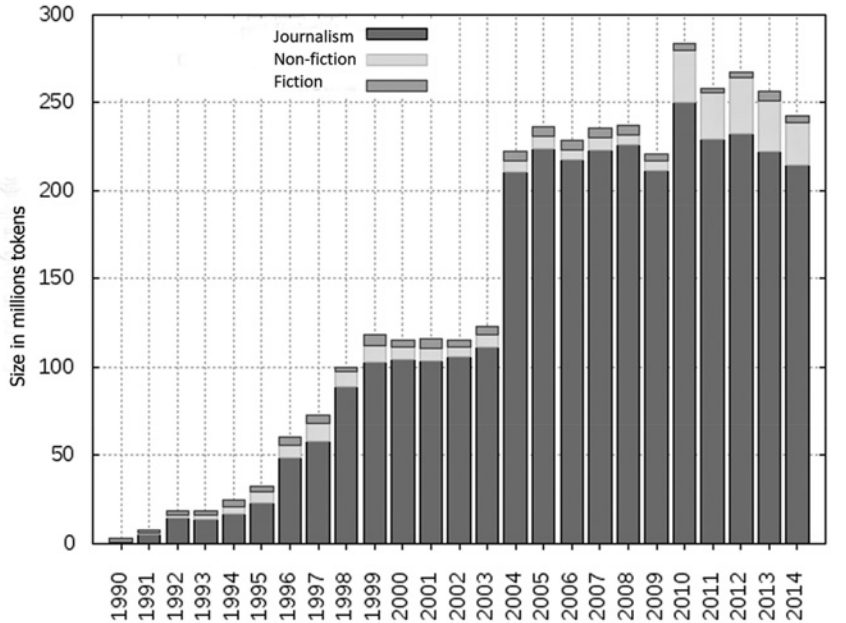

Figure 1: Composition of the corpus SYN version 4.[1]

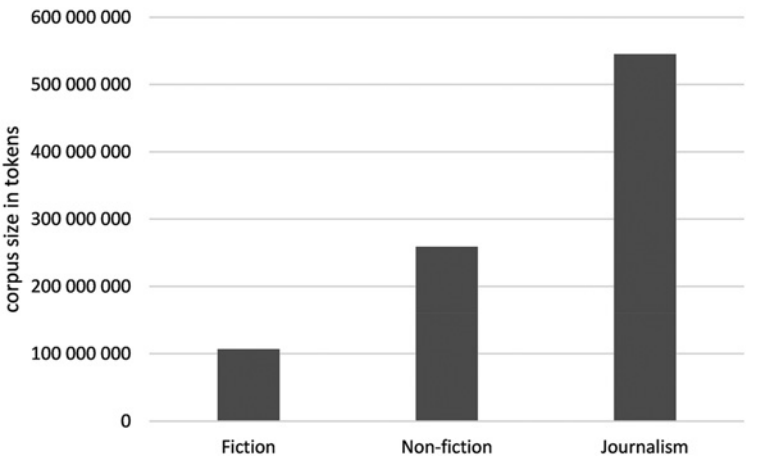

Figure 2: Size of the subcorpora.

1 https://wiki.korpus.cz/doku.php/cnk:syn:verze4

stories, poetry, and drama.[2] These three types seem to clearly belong to quite different style groups, therefore it is suitable for our study. Since the aim of this research is to test whether the presented method can be potentially applied to stylometric analyses, the data can be considered as a sufficient source. A deeper look into the problematic issue of "style" definition is beyond the scope of this study.

Lemmas are considered as the basic units in this research because of the inflectional character of Czech language. In order to avoid a bias caused by low frequencies, all lemmas with frequency less than 70 are omitted ($f \leq 69$).

4 Method

We chose the Context Specificity of Lemma (hereinafter *CSL*) as a method for the semantic analysis of styles (Čech et al. 2018). This method measures the similarity of the context of lemmas. More specifically, each lemma is represented by a vector. Both a size and orientation of the vector express the position of a lemma in a contextual multi-dimensional space. Thus, it is possible to measure lexical context similarities among lemmas. In case there are two lemmas which appear in the very same context in the corpus, these vectors would be identical. The similarity of these two lemmas would be therefore 1. However, this example is artificial and very unlikely to happen in reality.

We decide to apply the Closest Context Specificity (hereinafter *CCS*) to our analysis (cf. Čech et al. 2018). *CCS* measures the average value of the similarities of the 20 closest lemmas of the target lemma. The *CCS* is defined as follows:

$$CCS = \frac{\sum_{i=1}^{20} S_i}{20}$$

where S represents the similarity of the lemma.

For instance, in Table 1 we can see the 20 most similar (in terms of *CSL*) lemmas to the two target lemmas *atom* "atom" and *protože* "because". Since *protože* is a function word, it is rather clear that there are many lemmas in the similar context and thus the context specificity is quite low (and the similarity of the closest lemmas is considerably high). On the other hand, the lemma

2 Detailed information about the structure of the corpus can be found on https://wiki.korpus.cz/doku.php/en:cnk:syn:verze4

Table 1: Example of the 20 most similar lemmas to the two target lemmas *atom* and *protože*.

ATOM "ATOM"	*S*	*PROTOŽE* "BECAUSE"	*S*
neutron "neutron"	0.56	*takže* "so"	0.81
molekula "molecule"	0.49	*ale* "but"	0.78
termojaderný "thermonuclear"	0.46	*tak* "so"	0.77
elektron "electron"	0.45	*ten* "that"	0.74
izotop "isotope"	0.45	*když* "when"	0.73
štěpení "fission"	0.45	*proto* "therefore"	0.72
vodík "hydrogen"	0.44	*jelikož* "because"	0.71
křemík "silicon"	0.44	*prý* "apparently"	0.7
neutrino "neutrino"	0.44	*že* "that"	0.7
částice "particle"	0.43	*být* "to be"	0.69
helium "helium"	0.42	*navíc* "extra"	0.69
iter "iter"	0.41	*já* "I"	0.69
nanočástice "nanoparticle"	0.41	*totiž* "namely"	0.69
iont "ion"	0.4	*vůbec* "at all"	0.69
štěpný "fissionable"	0.4	*opravdu* "really"	0.68
atomový "atomic"	0.4	*dost* "enough"	0.68
štěpit "to do fission"	0.39	*neboť* "because"	0.68
jaderný "nuclear"	0.39	*moc* "power"	0.67
vodíkový "hydrogen"	0.38	*zase* "again"	0.67
mikrovlna "microwave"	0.38	*muset* "have to"	0.67

atom is a technical term, the context specificity of this lemma is therefore high due to its uniqueness. The resulting values of *CCS* of these two lemmas are the average *S* values of the 20 most similar lemmas: $CCS_{atom} = 0.4295$, $CCS_{protože} = 0.708$.

Since the goal of this study is to discover whether *CCS* can reveal some lexical context properties in different styles, we decide to use only autosemantic parts of speech (verbs, nouns, adjectives, adverbs). Namely, 10 most frequent lemmas of each part of speech are analyzed. **Verbs**[3]: *říci* "to say", *dostat* "to get", *dát* "to give", *hrát* "to play", *uvést* "to state", *přijít* "to come", *říkat* "to say", *vědět* "to know", *získat* "to gain", *čekat* "to wait". **Nouns**: *rok* "a year", *člověk* "a human", *město* "a town", *místo* "a place", *koruna* "a crown", *den* "a day", *dítě* "a child", *hodina* "an hour", *doba* "a time", *zápas* "a match".

3 Only autosemantic (lexical) verbs were chosen, the function words including auxiliary verbs were omitted in this study.

Adjectives: *velký* "big", *nový* "new", *další* "next", *český* "Czech", *dobrý* "good", *celý* "whole", *poslední* "last", *jiný* "other", *domácí* "domestic", *vysoký* "high". **Adverbs**: *ještě* "still", *hodně* "much", *dnes* "today", *včera* "yesterday", *dobře* "well", *zatím* "yet", *velmi* "very", *právě* "just", *nyní* "now", *letos* "this year".

As can be seen in Table 1, one can expect quite clear differences between the synsemantic (function) and autosemantic (content) lemmas due to the fact that the autosemantic lemmas have naturally more unique context than synsemantic lemmas in general. The most unique context is then expected in case of technical or scientific terms such as *molecule, neuron* or *phoneme*. The function lemmas are a part of grammar and there is a limited number of them. We must, therefore, use them in most contexts. However, it is not so obvious whether there are general differences between the autosemantic parts of speech. The question is whether the verbs, nouns, adjectives, and adverbs are differently sensitive to the Context Specificity in principle. If so, then which ones have more unique context and why. Although the profound investigation of this question cannot be carried out in the present study, we still try to do the preliminary exploration on this issue. We compute the average *CCS* values of all analyzed lemmas (see above) and compare the results (see Figure 3).

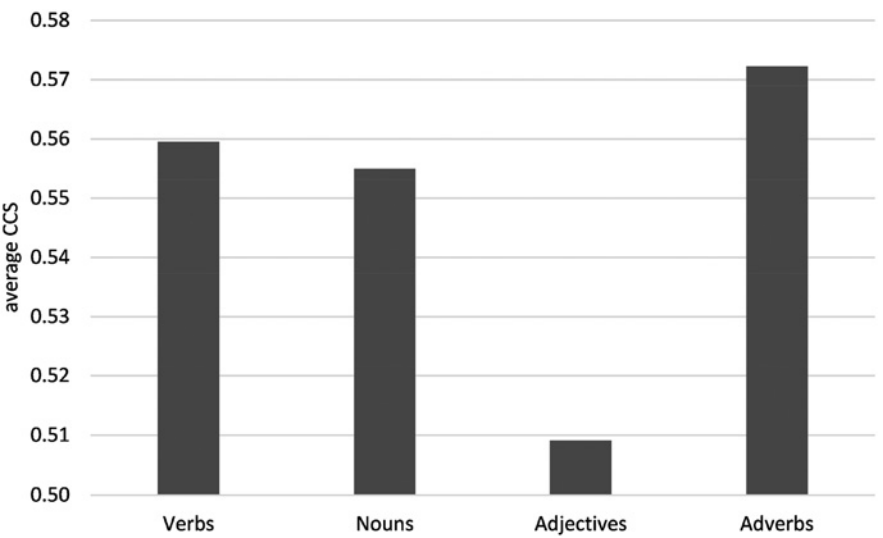

Figure 3: The *CCS* average values of chosen lemmas in four parts of speech.

5 Results

The resulting values of *CCS* in different styles can be seen as follows: verbs in Table 2 and Figure 4, nouns in Table 3 and Figure 5, adjectives in Table 4 and Figure 6, adverbs in Table 5 and Figure 7.

Table 2: *CCS* of 10 most frequent autosemantic (lexical) verbs in styles.

	Fiction	Non-fiction	Journalism
říci "to say"	0.67	0.63	0.72
dostat "to get"	0.46	0.51	0.52
dát "to give"	0.47	0.54	0.54
hrát "to play"	0.55	0.50	0.57
uvést "to state"	0.43	0.53	0.70
přijít "to come"	0.53	0.53	0.50
říkat "to say"	0.61	0.62	0.69
vědět "to know"	0.62	0.66	0.66
získat "to gain"	0.48	0.52	0.52
čekat "to wait"	0.47	0.53	0.50

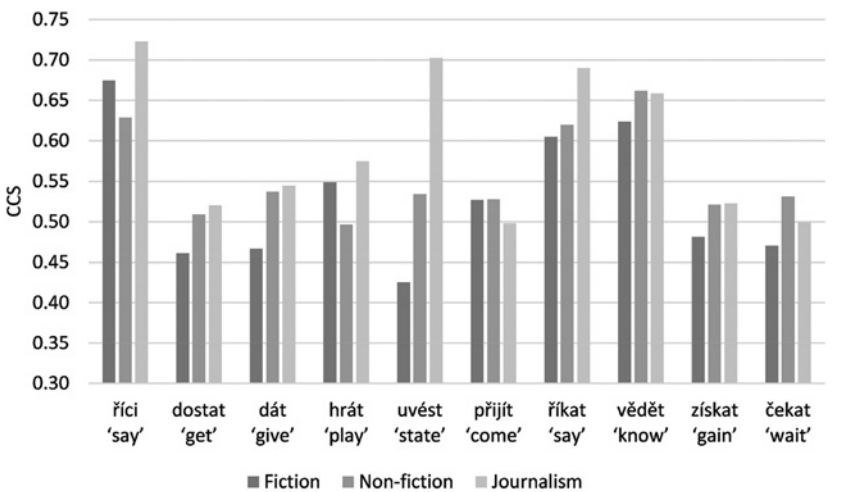

Figure 4: *CCS* of 10 most frequent autosemantic (lexical) verbs in styles.

Table 3: *CCS* of 10 most frequent nouns in styles.

	Fiction	Non-fiction	Journalism
rok "a year"	0.58	0.61	0.58
člověk "a human"	0.48	0.54	0.51
město "a town"	0.56	0.58	0.52
místo "a place"	0.43	0.42	0.44
koruna "a crown"	0.42	0.62	0.64
den "a day"	0.60	0.58	0.61
dítě "a child"	0.59	0.65	0.62
hodina "an hour"	0.57	0.56	0.64
doba "a time"	0.50	0.51	0.52
zápas "a match"	0.52	0.55	0.72

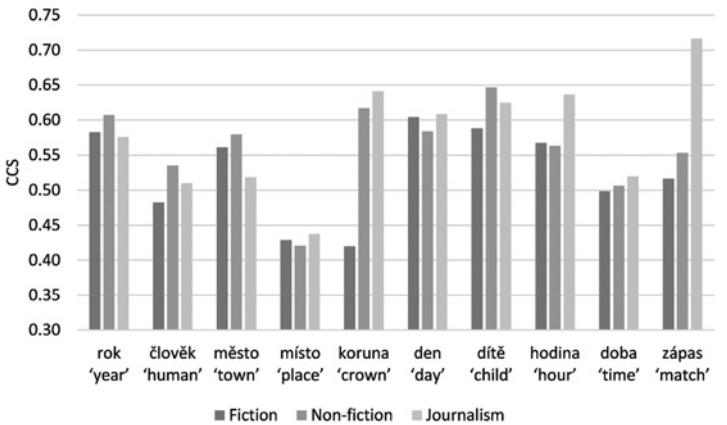

Figure 5: *CCS* of 10 most frequent nouns in styles.

Table 4: *CCS* of 10 most frequent adjectives in styles.

	Fiction	Non-fiction	Journalism
velký "big"	0.50	0.52	0.53
nový "new"	0.43	0.53	0.47
další "next"	0.50	0.56	0.59
český "czech"	0.54	0.59	0.52
dobrý "good"	0.51	0.56	0.54
celý "whole"	0.42	0.47	0.54
poslední "last"	0.48	0.52	0.58
jiný "other"	0.47	0.57	0.58
domácí "domestic"	0.35	0.43	0.55
vysoký "high"	0.45	0.54	0.44

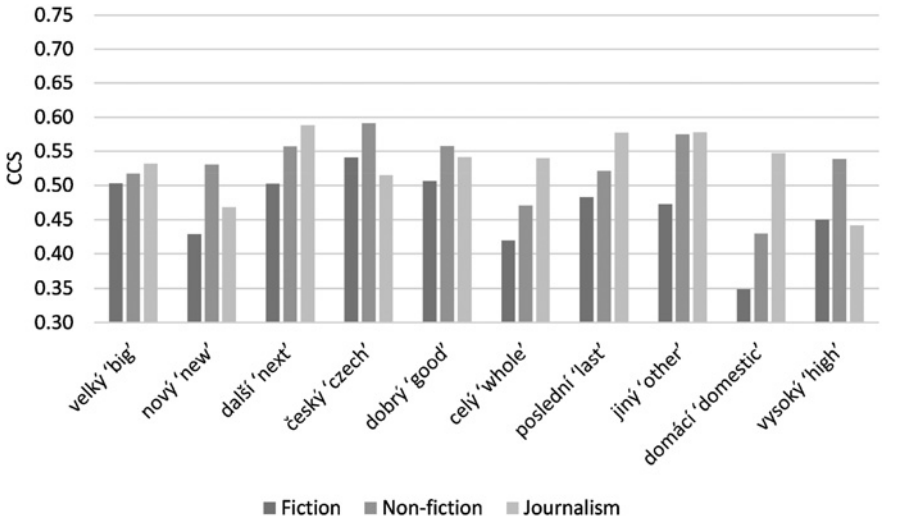

Figure 6: *CCS* of 10 most frequent adjectives in styles.

Table 5: *CCS* of 10 most frequent adverbs in styles.

	Fiction	Non-fiction	Journalism
ještě "still"	0.55	0.58	0.66
hodně "much"	0.51	0.58	0.61
dnes "today"	0.52	0.57	0.52
včera "yesterday"	0.54	0.55	0.65
dobře "well"	0.52	0.60	0.61
zatím "yet"	0.50	0.57	0.56
velmi "very"	0.59	0.63	0.64
právě "just"	0.50	0.58	0.59
nyní "now"	0.52	0.56	0.58
letos "this year"	0.47	0.69	0.61

We can see from Tables 2–5 and Figures 3–6 that *CCS* values of some lemmas such as *rok* "a year", *doba* "a time", *velký* "big" are rather stable and independent to styles, whereas the majority of lemmas (e.g. *uvést* "to state", *koruna* "a crown", *zápas* "a match", *nový* "new", *celý* "whole", *domácí* "domestic") display quite big differences.

The obtained results showed that some lemmas are more sensitive to the styles than others. To interpret the obtained values, we need to find out which

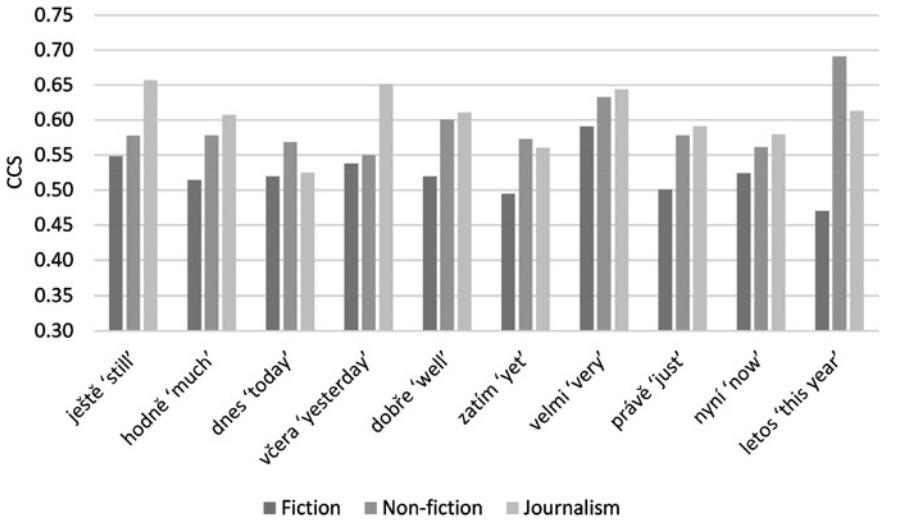

Figure 7: *CCS* of 10 most frequent adverbs in styles.

lemmas are most sensitive to various styles and which lemmas are rather stable. We compute, therefore, the arithmetic mean of the differences of *CCS* of each lemma between styles. It is a simple but sufficient, intuitive measure which we believe is suitable for a pilot study. Of course, other measures are not excluded and can be introduced and compared in future studies. For instance, the calculation process for the lemma *říci* "to say" is as follows:

$$\frac{\left|CCS\ \check{r}\acute{i}ci_{fiction} - CCS\,\check{r}\,\acute{i}ci_{non-fiction}\right| + \left|CCS\,\check{r}\,\acute{i}ci_{fiction} - CCS\,\check{r}\,\acute{i}ci_{journalism}\right| + \left|CCS\,\check{r}\,\acute{i}ci_{non-fiction} - CCS\,\check{r}\,\acute{i}ci_{journalism}\right|}{3}$$

$$\frac{|0.67-0.63| + |0.67-0.72| + |0.63-0.72|}{3} = \frac{0.04+0.05+0.09}{3} = 0.06$$

This measurement enables us to compare the *CCS* sensitivity of lemmas to styles more accurately. The obtained results are shown in Figure 8 in a descending order.

We see in the Figure 8 that there are five extraordinary lemmas in terms of the *CCS* sensitivity to styles, namely *uvést* "to state", *letos* "this year", *koruna* "a crown", *zápas* "a match", *domácí* "domestic". There is an obvious gap between these lemmas and followed decreasing values. The big variability of lemmas *koruna* "a crown", *zápas* "a match" and *domácí* "domestic" is

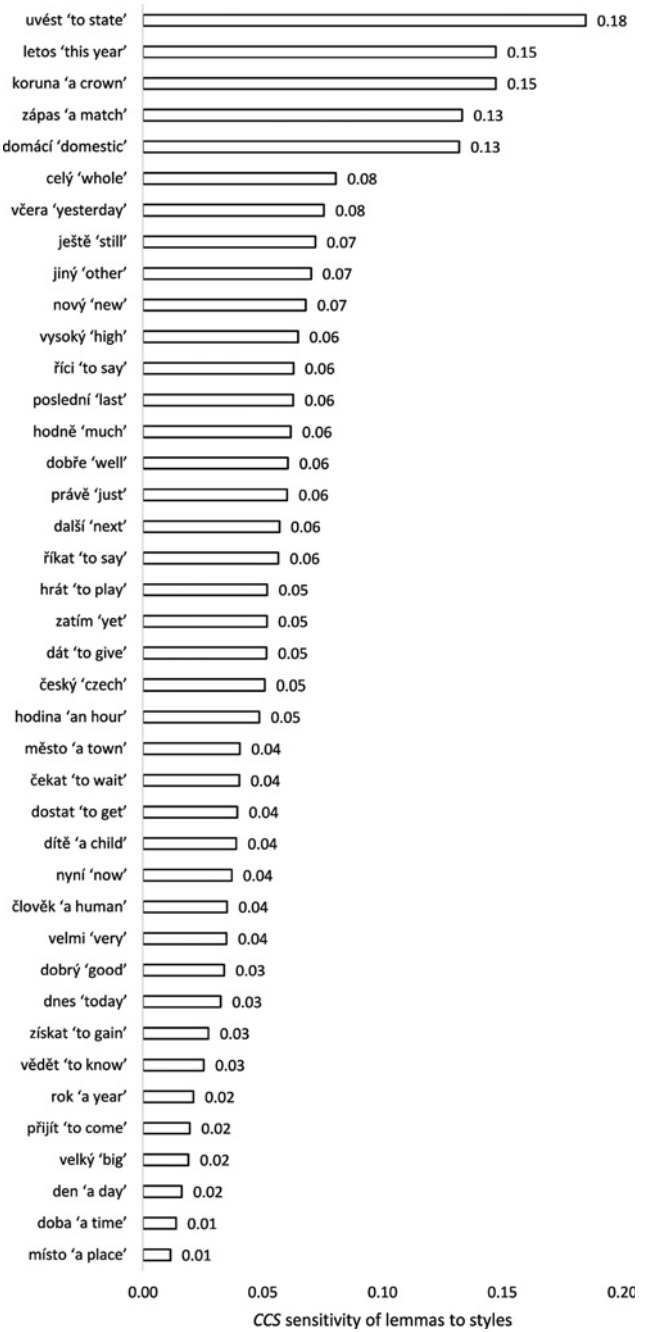

Figure 8: *CCS* sensitivity of lemmas to styles.

probably caused by their specificity in journalism. Koruna "a crown" has several meanings, the most common ones are the Czech currency and a king's crown. However, there are more meanings of this word. The less unique context (therefore the highest *CCS*) is in journalism because *Czech crown* as a currency appears in many different contexts. On the other hand, it should be emphasized that the vector representation can hardly recognize that numbers are just same lexical context. Many news contains some information about money. The explanation of a high variability of *zápas* "a match, a game, a play, a struggle" is also connected to the specific position in journalism. This lemma is highly frequent in sport news but also in news about politics or other issues. It is caused by the wide range of the meanings (including metaphorical ones) of this lemma. The similar situation also exists in the case of lemma *domácí*. This adjective means home, homemade, homegrown, internal or domestic. It is frequently used in all types of news such as politics and economics. The important point is that the lemma *domácí* is highly frequent in sport columns for denoting the domestic teams or players. The number of possible contexts is therefore much higher in journalism. The verb *uvést* "to state" is predominantly used for quoting a citation in journalism. That is why the usage of this lemma in non-journalism texts, especially in fiction, is quite specific and extraordinary. It is very typical to the journalistic or administrative style. Another meaning of *uvést* is "to introduce" which probably appears more frequently in fiction. These facts presumably cause the quite high sensitivity of this lemma to different styles. The adverb *letos* has only one meaning – this year. The reason of the extraordinary position of this lemma lies especially in the low specificity (high *CCS*) in non-fiction which is caused probably by a fact that the closest lemmas (given by the similar context) are rather general expressions such as *letošní* "this year"(adjective), *vloni* "last year"(adverb), *loňský* "last year"(adjective), *2012, 2011, letošek* "this year"(noun), *předloni* "the year before last year"(adverb), *2010, loňsko* "last year"(noun), *příští* "next"(adjective), *2009, 2013, 2008*.

To find out whether there are general differences between parts of speech, we calculate the arithmetic mean of the obtained values in Figure 8. The results can be seen in Figure 9.

According to the resulting values in Figure 9, verbs and nouns are less sensitive to styles than adjectives and adverbs in our corpus. It would be very interesting to apply our approach to other styles in different corpora to discover whether there is some general pattern. The explanation for this phenomenon probably lies in the fact that verbs and nouns play central roles within the grammatical structure of sentences. Moreover, almost all languages have these two basic parts of speech. However, beyond these,

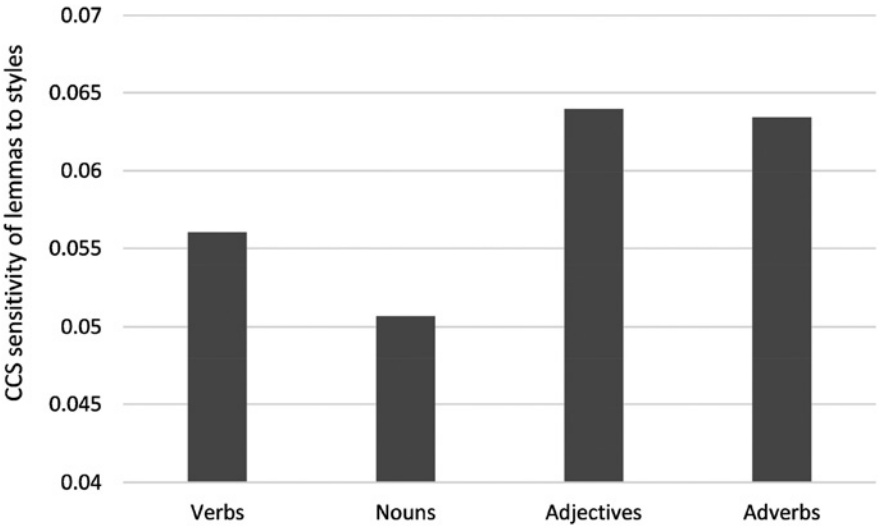

Figure 9: *CCS* sensitivity of lemmas to the styles.

there are significant variations in different languages (Kroeger 2005). Since verbs and nouns are central parts of sentence structure, it is reasonable to expect that (especially in the case of the most frequent lemmas) they are rather stable and not so sensitive to various styles. Adjectives are describing words and the main syntactic role is to qualify nouns and to give more information about nouns. Similarly, adverbs modify verbs or adjectives. They express more detailed information which is more likely connected to a specific style of writing. These parts of speech are therefore more sensitive to different styles.

The crucial question of any frequency-based analyses is whether the results are influenced by the text (corpus) size. This problem is a very common stumbling block of many traditional stylometric indicators such as vocabulary richness (cf. Popescu et al. 2009; Kubát and Milička. 2013; Cvrček and Chlumská 2015; Kubát 2016). One can, therefore, expect that *CCS* as a method based on counting words can be influenced by the corpus size too. Thus, the differences between styles would say nothing about a style. Since we know that the three used corpora have different sizes (see above), the relation between *CCS* and the frequencies of lemmas must be examined. To discover whether *CCS* values are independent on the frequencies, we count both the absolute and relative frequencies of the lemmas. The relation between *CCS* and lemma frequencies can be seen in Figures 10 and 11.

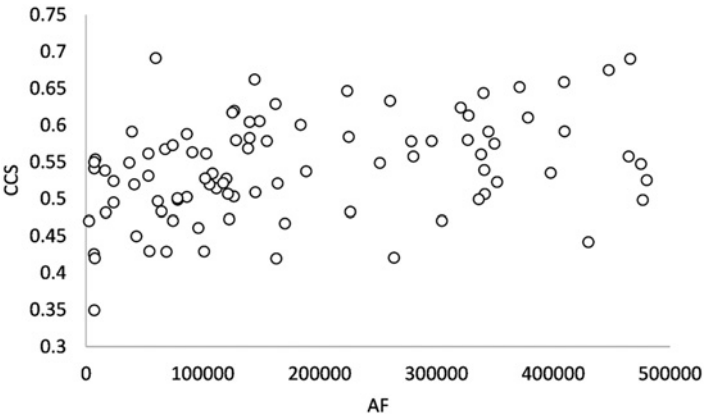

Figure 10: The relation between *CCS* and absolute frequency (*AF*) of analyzed lemmas.

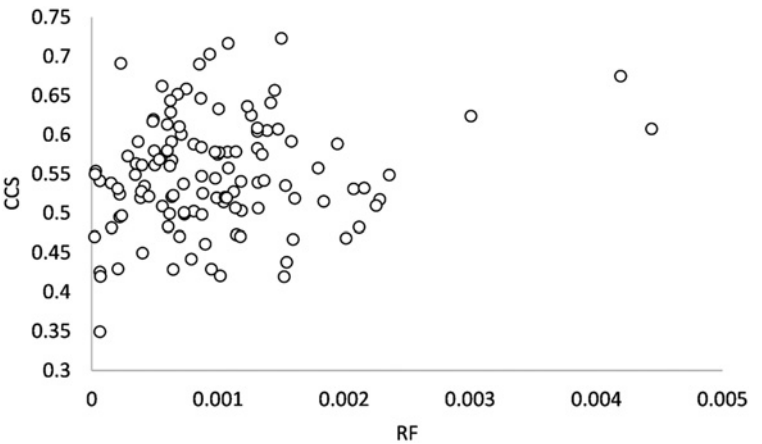

Figure 11: The relation between *CCS* and relative frequency (*RF*) of analyzed lemmas.

As can be seen in Figures 10 and 11, the frequencies (both absolute and relative) seem to have no influence on the *CCS* values. This finding is also supported by the Pearson correlation coefficient (r) where $r = 0.15$ (p-value = 0.1) for the correlation between *AF* and *CCS*; and $r = 0.007$ (p-value = 0.94) for the correlation between *RF* and *CCS*. This is a very important fact for application *CCS* in further text analyses because we can state now that the observed differences between styles are not given just by the different size of the corpora (and thus different frequencies of measured lemmas). We want to emphasize that this is a crucial advantage of *CCS* because even similar methods are correlated with the corpus size (cf. Čech et al. 2018).

6 Conclusion

The word embedding approach based on neural network, namely the Closest Context Specificity (*CCS*) seems to have quite a promising potential in lexical context linguistic analysis. This new method (Čech et al. 2018) showed convincing results in preliminary diachronic analysis of lemmas. Czech political discourse was analyzed in terms of the lexical context changes. This following up research applies *CCS* method to a text style analysis. Specifically, 10 most frequent autosemantic lemmas of chosen parts of speech (verbs, nouns, adjectives, adverbs) were used for the analysis. The obtained results revealed that this approach is an efficient tool for lexical context analysis of lemmas in miscellaneous styles, namely fiction, non-fiction and journalism in this study.

Given that we used four different parts of speech, their *CCS* sensitivity to styles were tested. The results showed that verbs and nouns are less sensitive to various styles than adjectives and adverbs. The explanation for this phenomenon probably lies in the fact that verbs and nouns play central roles within the grammatical structure of sentences. On the other hand, adjectives and adverbs are describing words and express more detailed information which is more likely connected to a specific style of writing. These parts of speech are, therefore, more sensitive to different text styles.

Since many methods in stylometric research are highly dependent on the text length (corpus size), the relation between *CCS* and frequency of lemmas, both absolute and relative values, was examined. The study showed that *CCS* resulting values of lemmas are not proportional to their frequencies in corpora. Thus, we can state that the context specificity measure is not biased by this common problem of many stylometric indicators.

Although this study presents promising results and reveals that *CCS* method seems to be a quite efficient tool for analyzing lexical context differences of lemmas between styles, it should be mentioned that this study is just a first attempt at the application of this method and more analyses on different datasets of different languages must be carried out in the future to verify, modify, or reject our preliminary claims.

Funding: This work was supported by Social Science Fund of Shaanxi State, (Grant Number: 2015K001), Univerzita Karlova v Praze (10.13039/100007397), Progress 4, Ostravská Univerzita v Ostravě (10.13039/501100006704 Grant Number: SGS02/UVAFM/2017).

References

Bublitz, Wolfram & Neal R. Norrick (eds.). 2011. *Foundations of pragmatics*. Berlin: De Gruyter Mouton.

Čech, Radek., Jan Hůla, Miroslav Kubát, Xinying Chen & Jiří Milička. 2018. The development of context specificity of lemma. A word embeddings approach. *Journal of Quantitative Linguistics* https://www.tandfonline.com/doi/abs/10.1080/09296174.2018.1491748 (accessed 28 September 2018).

Cvrček, Václav & Lucie Chlumská. 2015. Simplification in translated Czech: A new approach to type-token ratio. *Russian Linguistics* 39(3). 309–325.

Golato, Andrea & Peter Golato. 2012. Pragmatics research methods. In Carol A. Chapelle (ed.), *The encyclopedia of applied linguistics*. Oxford: Wiley-Blackwell. doi:10.1002/9781405198431.wbeal0946.

Grieve, Jack. 2005. *Quantitative authorship attribution: A history and an evaluation of techniques*. Simon Fraser University MA thesis.

Grieve, Jack. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing* 22(3). 251–270.

Hamilton, William L., Jure Leskovec & Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1489–1501.

Hnátková, Milena, Michal Křen, Pavel Procházka & Hana Skoumalová. 2014. The SYN-series corpora of written Czech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), Reykjavík: ELRA, 160–164

Juola, Patrick. 2006. Authorship attribution. *Foundations and Trends in Information Retrieval* 1(3). 233–334.

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička & Adrian Zasina 2016. *Corpus SYN, version 4 from 16. 9. 2016*. Praha: Ústav Českého národního korpusu FF UK. http://www.korpus.cz.

Kroeger, Paul. 2005. *Analyzing grammar: An introduction*. Cambridge: Cambridge University Press.

Kubát, Miroslav. 2016. *Kvantitativní analýza žánrů*, [Quantitative Analysis of Genres]. Ostrava: University of Ostrava.

Kubát, Miroslav & Jiří Milička. 2013. Vocabulary richness measure in genres. *Journal of Quantitative Linguistics* 20(4). 339–349.

Levy, Omer, Yoav Goldberg & Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics* 3. 211–225.

Manning, Christopher, D. Jeffrey Pennington & Richard Socher. 2014. *Proceedings of the empirical methods in natural language* processing *(EMNLP 2014)*.

Matthews, Robert AJ & Thomas VN Merriam. 1993. Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing* 8(4). 203–209.

McMenamin, Gerald R. 2002. *Forensic linguistics: Advances in forensic stylistics*. Boca Raton: CRC Press.

Mikolov, Tomas, Kai Chen, Greg S. Corrado, Jeff Dean & Ilya Sutskever. 2013a. Distributed representations of words and phrases and their compositionality. *Proceedings of Neural Information Processing Systems (NIPS 26)*, 3111–3119.

Mikolov, Tomas, Kai Chen, Greg S. Corrado, Jeff Dean & Ilya Sutskever. 2013b. Efficient estimation of word representations in vector space. *ICLR Workshop Papers*.
Mikros, George K. & Kostas Perifanos. 2013. Authorship attribution in Greek tweets using multilevel author's n-gram profiles. In E. Hovy, V. Markman, C. H. Martell & D. Uthus (eds.), *Papers from the 2013 AAAI spring symposium "Analyzing Microtext"*, 17–23. Stanford, California. Palo Alto, California: AAAI Press. (25–27March2013).
Popescu, Ioan Iovitz, Gabriel Altmann, Peter Grzybek, Bijapur D. Jayaram, Reinhard Köhler, Viktor Krupa, Ján Mačutek, Regina Pustet, Ludmila Uhlířová & Matummal N. Vidya. 2009. *Word frequency studies*. Berlin, New York: Mouton de Gruyter.
Sampson, Geoffrey. 2001. *Empirical linguistics*. London – New York: Continuum International.

Bionotes

Miroslav Kubát

Miroslav Kubát (born 1984, Ph.D. Palacký University 2015) is an assistant professor in Czech Language at the University of Ostrava (Czech Republic). His research interests focus on quantitative linguistics and stylometry. He specializes in quantitative indices of text analysis, such as vocabulary richness, activity and context specificity.

Jan Hůla

Jan Hůla (born 1985, MgA. Tomas Bata University, 2011) is a PhD student at the Institute for Research and Applications of Fuzzy Modeling, Faculty of Science, University of University. He specializes in Neural Networks and Natural Language Processing; he is also interested in Applied Category Theory and its applications in Linguistics.

Xinying Chen

Xinying Chen (born 1984, Ph.D. Communication University of China, 2012) is a post-doctoral research fellow at the University of Ostrava in the Czech Republic and an associate professor in Linguistics at the Xi'an Jiaotong University in China. Her research interests focus on the empirical syntactical analysis of linguistic units in spoken and written communication. She is also interested in applying interdisciplinary methods, such as social network analysis or statistical clustering algorithms, to quantitative analysis of synchronic and diachronic texts.

Radek Čech

Radek Čech (born 1974, Ph.D. Palacký University 2005) is an associate professor in Czech Language at the University of Ostrava (Czech Republic). His research interests focus on quantitative text analysis and quantitative syntax (valency, syntactic complex networks). He is also interested in the application of quantitative methods to historical linguistics (word ordering of enclitics, stylometry).

Jiří Milička

Jiří Milička (born 1986, PhD Charles University, 2016) is a research associate at the Department of Comparative Linguistics and the Institute of the Czech National Corpus, Faculty of Arts, Charles University. He specializes in quantitative and corpus linguistics and Arabic language; he also develops applications for linguistic research.