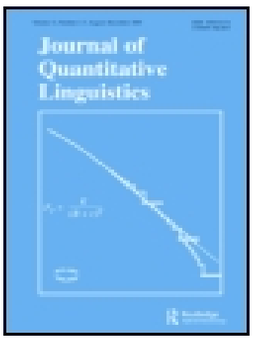

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: http://www.tandfonline.com/loi/njql20

The Development of Context Specificity of Lemma. A Word Embeddings Approach

Radek Čech, Jan Hůla, Miroslav Kubát, Xinying Chen & Jiří Milička

To cite this article: Radek Čech, Jan Hůla, Miroslav Kubát, Xinying Chen & Jiří Milička (2018): The Development of Context Specificity of Lemma. A Word Embeddings Approach, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2018.1491748

To link to this article: https://doi.org/10.1080/09296174.2018.1491748

Published online: 09 Aug 2018.

JOURNAL OF QUANTITATIVE LINGUISTICS
https://doi.org/10.1080/09296174.2018.1491748

The Development of Context Specificity of Lemma. A Word Embeddings Approach

Radek Čech[a], Jan Hůla[b], Miroslav Kubát[a], Xinying Chen[a,c] and Jiří Milička[d]

[a]Department of Czech Language, Faculty of Arts, University of Ostrava, Ostrava, Czech Republic; [b]Institute for Research and Applications of Fuzzy Modeling, University of Ostrava, Ostrava, Czech Republic; [c]School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China; [d]Institute of the Czech National Corpus, Charles University, Prague, Czech Republic

ABSTRACT

The study deals with the application of the neural networks in the linguistic research of word semantics. A new method of measuring context specificity of lemma based on Word Embeddings Word2vec technique is proposed in the first part of the article. Then the method is illustrated in the analysis of the Czech political discourse in the second part. The research is based on the corpus of the Czech journalism consisting of more than 3 billion tokens. The results show that the proposed method is applicable for detecting the semantic development of a lemma and it could have a great potential for linguistic studies if one can apply it with comprehensive explanations.

1. Introduction

A language system undergoes an incessant evolution which is caused by its language usage. Specifically, users of a language influence its properties in two opposite directions: on the one hand, the users keep some stability of the system since the observance of language rules (which can have stochastic form, too) and using 'stable' meanings of words allows them to achieve their needs and goals of communications, on the other hand, a boundless number of specific needs and goals forces users of the language to modify the rules, lexical meanings and even to use new expressions. Needless to say, these modifications or new expressions must be embedded to the existing language system and language users strive to find some equilibrium between old and new language forms as well as old and new functions of given units. To sum up, language represents a dynamic system with continuously ongoing changes.

A description and (in the best case) an explanation of language development have been one of the main goals of linguistics for at least two centuries and there are plenty methods of analyzing this phenomena (Allan,

CONTACT Radek Čech cechradek@gmail.com

2013; Burkhardt, Steger, & Wiegand, 2000). However, very innovative and successful approaches to language analysis have appeared in recent years; namely, methods based on neural network representations which are also known as word embeddings (Pennington, Socher, & Manning 2014; Mikolov, Chen, Corrado, Dean, & Sutskever, 2013a). In linguistics, word embeddings were for example used for tracking the semantic evolution of words and to quantitatively confirm hypotheses about semantic change (Hamilton et al. 2016).

Different from previous studies, in this paper, we use word embeddings to define a concept we call 'context specificity of a lemma'. Context specificity of a lemma measures how unique is the context in which the lemma appears in the corpus. Specifically, if the lemma occurs in many different contexts, it will have low-context specificity. The context in which the lemma appears is captured with a vector of co-occurrence statistics which is assigned to every lemma. In this vector representation, it is possible to measure the similarities among lemmas. To be more specific, it means that for each lemma, we can compute its similarity to all other lemmas. Statistics of these similarities (e.g. a mean value) can be used for characterizing the context specificity of a lemma (hereinafter CSL). The lower the mean of similarities, the higher the CSL (for more details see Section 4). For example, the mean similarities of Czech lemma 'atom' (means 'atom') to all other lemmas is 0.0829, while the equivalent value of lemma 'nebo' (means 'or') is 0.1273 for 2013 subcorpus (for details about corpora, see Section 3). Consequently, 'atom' occurred in more specific contexts than 'nebo'.

This approach allows us a) to measure differences among lemmas with regards to the CLS; b) to measure a development of the CLS, if a diachronic corpus is used. It should be emphasized that the approach enables us to model the development of semantic characteristics which is a more difficult task in a comparison to modeling the development of formal characteristics.

From the more general point of view, the approach seems to be a suitable tool for an operationalization of hypotheses connected to a synergetic model of lexicalization (cf. Köhler, 1993, 2005). In this model, a notion of 'context specificity' denotes so called language-constitutive requirement 'representing the need to form more specific expressions than the ones which are available at a given time' (Köhler, 2005, p. 766). More specifically, this requirement 'can be met in an optimal way if the lexical system under consideration provides expressions which are completely context-specific (i.e. strongly related to particular situations, persons, places etc.) with respect to their applications and meanings. The tendency of a language to lexicalize new words (loan words or neologisms) is a function of the strength of this requirement' (Köhler, 1993, p. 45). In the opposite direction, the process of lexicalization is influenced by the de-specification requirement (sometimes it is called 'context economy') for the cases

where the available expressions are too specific for the current communicative purpose. As a result, equilibrium between too specific and too general expressions emerged in the language system as a balanced consequence of these requirements. In the synergetic linguistics, these two opposite requirements have a decisive impact on the so-called polytextuality of word (the number of different texts in a corpus which contains a given word). Our approach enables us to determine the context specificity as a property of word (or lemma) which can potentially incorporate to the synergetic model of language. Specifically, it could be used to model a relationship between the language-constitutive requirement named 'context specificity' and the context specificity as the property of word (or lemma). However, this aspect is beyond the scope of this study and it will not be discussed in this paper.

There are two aims of this study: (1) to present the new method, (2) to illustrate how this method can be used in an analysis of a political discourse.

2. Neural Networks in Language Analysis

This part presents an intuition behind neural network models. Neural networks represent a set of methods which are very effective for finding useful representations of data. Data are usually collected in a form which is not 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 have similar meaning. As another example, working directly with image pixels is not very effective for verifying whether two images contain the same objects because when the objects are viewed from different angles and in different lighting conditions, the values of their pixels have almost nothing in common. 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 of the network. Concrete values of these parameters are found by minimizing an error function on a concrete task. For example, when we want to identify whether two images contain the same objects, we could measure the error by measuring the distance between the two output representations from the network. If the images actually contain the same objects but the distance between the output representations is very high, then the error would be also very high and vice versa. We can use the error to update the parameters of the network in a way which makes the error lower. By iterating this process, we are minimizing the error and thus finding a useful representation for the task.

In this work, we want to measure context specificity of lemmas and therefore we need to represent a lemma in a representation which captures

the context in which the lemma appears. Representations with this property are easy to obtain with methods collectively called Word Embeddings (Manning et al., 2014; Mikolov et al., 2013a) where the aim is to represent a word (in our case the lemma) as a multi-dimensional (50–1000) vector. This vector captures co-occurence statistics between the lemma itself and other lemmas in the small window centered at the lemma at hand. For example, if we want to obtain the word embedding for the lemma 'president', we need to capture how frequently this lemma appears close to the other lemmas in the corpus. Naively, we could collect every instance of the lemma 'president' in the corpus and count how many times each other lemma appears in the window centered at the lemma 'president' (the window could contain two lemmas on right and two lemmas on left for example, see Figure 1).

These frequencies would constitute the lemma embedding vector of the lemma 'president'. Lemmas like 'state', 'election' and 'minister' would occur many times in this window, whereas lemmas like 'algebra', 'phoneme' or 'atom' would be rather infrequent in this context. Lemmas which appear in similar contexts would have similar lemma embedding vectors.

One problem with this naive representation is that the vectors have enormous dimensionality (the vectors have as many components as there are unique lemmas in the corpus). This problem could be solved by finding a low-dimensional subspace with decomposition algorithms from linear algebra (SVD) which approximates the original vectors. Nevertheless, from a computational standpoint (computer memory requirements), decomposition algorithms are practical only when the corpus does not contain too many unique words. When we want to obtain low dimensional representations for large corpus, neural networks are usually methods of choice, because they can learn the representation in online fashion and thus are not affected by memory requirements. The learning is done by maximizing the following objective function:

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

This objective function is maximized when the individual probabilities inside the sums are maximized. The first sum iterates over all tokens in

Figure 1. The window centered around the lemma 'elect'. It contains two lemmas on the left and two lemmas on the right. The outside lemmas (in light grey) would be considered when counting the frequencies for the lemma 'elect'.

the corpus and for every token t the second sum iterates over all tokens in the small window centered at the token t. This window is of length 2m+ 1. Intuitively, we want that the lemmas inside this window would be predictable from the central lemma in this window. For example, when the lemma w[t] is 'president' and the lemma w[t + 1] is 'election', we want p(election| president) to be high, so that the lemma 'election' is predictable from lemma 'president'.

The conditional probabilities in the equation 1 are parametrized by a neural network and the parametrization has a following form:

$$p(o|c) = \frac{exp\left(u(o)^T \cdot v(c)\right)}{\sum_{w=1}^{W} exp\left(u(w)^T \cdot v(c)\right)} \tag{2}$$

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 an outside lemma. For example, when the window is centered at the lemma 'president', then the first vector is used as its representation, but when the window is centered at some other lemma and the lemma 'president' appears in this window, then we use the second vector as its representation. This is only to simplify the optimization problem and at the end these representations could be averaged or one of them can be discarded.

The optimization is done by randomly sampling pairs of words which appear in the same windows and then computing the probabilities with the equation 2. The parameters of the network (the vectors themselves) are then adjusted in order to maximize these probabilities.

One problem to notice in equation 2 is that the sum in the denominator is over all unique lemmas in the corpus and therefore it is very expensive to compute. In practice this sum is approximated with method called Negative Sampling which computes the sum using only the lemmas in the numerator plus few other lemmas which are sampled randomly from a distribution which takes a frequency of words into account (Mikolov, Chen, Corrado, & Dean, 2013b). Also, in order to diminish the effects of very frequent lemmas like 'the', 'is' etc., subsampling is used in practice, so that lemma are removed from the window with a probability which is proportional to their frequencies.

The individual probabilities are maximized when the numerator is maximized and the sum in the denominator is minimized. This happens when the inner product in the exponent is maximized for lemmas which often appear next to each other and is minimized for lemmas which do not. At the beginning we initialize these vectors randomly and during

optimization the vectors of lemmas which appear often in same contexts become very correlated (their inner product becomes high) and vectors of lemmas which do not appear frequently in same contexts become decorrelated (their inner product is close to zero). At the end of the optimization, we obtain the vectors which capture the co-occurrence statistics with other lemmas and we can use them to measure the similarity of contexts in which these lemmas appear. For measuring the similarities between lemmas we use the cosine similarity as suggested by literature (Levy, Goldberg, & Dagan 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, 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 learning procedure see (Mikolov et al., 2013b).

3. Language Material

Neural networks need huge training data sets to be capable of producing reliable results. We therefore decided to use the largest Czech text database – Czech National Corpus.[1] Particularly the fourth version (SYN_V4) of so called SYN series corpora was chosen (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, Křen, Procházka, & Skoumalová, 2014). The size of the SYN_V4 is given by the sum, which makes 3,626 billion tokens. The SYN corpus is not representative; the dominant component is journalism, which is the result of the predominance of journalistic corpora SYN2006PUB, SYN2009PUB, SYN2013PUB and the journalistic component from the years 2010–2014. Beside journalism there are other two text types: fiction and technical literature.[2] The structure of the corpus is presented in the Figure 2. Given that this study is focused on political discourse, only journalistic texts were selected for the analysis. The final corpus of our study consists of more than 3 billion tokens (3,045,389,630) and more than one hundred thousand types (102,707).

Considering the fact that Czech as any Slavonic language has rich inflections and the analysis is focused on lexical units, we decided to use a lemmatized corpus and therefore, lemmas are the basic units of this research. In order to avoid a bias caused by low frequencies, all lemmas with frequency less than 70

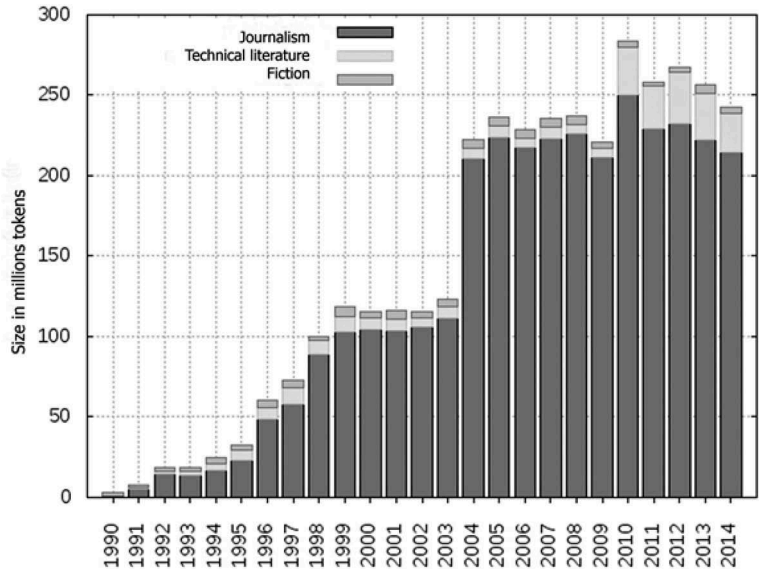

Figure 2. Composition of the corpus SYN version 4[3].

were omitted ($f \leq 69$). Since the goal is to analyze diachronic development of the CSL, we divide the data into 19 subcorpora that each represents one year (see Table 1). Only the subcorpus 1990–1996 consists of texts from several years because of the small data sizes (cf. Figure 2).

Table 1. Number of lemmas in each year. Years 1990–1996 are merged because of insufficient amount of data for each year.

Year	Number of lemmas
1990–1996	37,292
1997	44,023
1998	40,954
1999	45,038
2000	45,490
2001	44,930
2002	44,624
2003	45,757
2004	64,119
2005	65,008
2006	64,110
2007	65,698
2008	66,113
2009	63,695
2010	69,212
2011	66,167
2012	66,783
2013	65,381
2014	64,186

Table 2. Five the most similar lemmas of the lemma 'atom' and 'nebo'. S assigns the value of similarity from the corpus 2013.

target lemma = atom [atom]		target lemma = nebo [or]	
lemma	*S*	lemma	*S*
neutron [neutron]	0.6	či [or]	0.88
molekula [molecule]	0.54	třeba [need]	0.81
elektron [electron]	0.54	anebo [or]	0.79
částice [particle]	0.5	například [for example]	0.74
LHC [LHC]	0.48	i [and]	0.72

4. Context Specificity

When working with word embedding methods, each lemma is represented by a vector. A size and orientation of a vector express the position of a lemma in a semantic multi-dimensional space. Therefore, it is possible to measure similarities among lemmas (see Section 2). If, in an ideal case, there are two lemmas which occur in the identical contexts in the whole corpus, the size and orientation of these two vectors would be identical and, thus, the distance between these two lemmas equals to zero or, reversely, the similarity between them equals one. In the reality, each lemma occurs in different contexts, consequently, they are represented by different vectors which enable us to compute similarities among them. It seems rather obvious that some lemmas are used in more specific contexts than the others. Lemmas that are used in more specific contexts should be more distant to all other lemmas in comparison to lemmas that are used in less specific contexts. As an example, there are five most similar lemmas of the lemma 'atom' and 'nebo' in Table 2.

For the calculation of the context specificity we propose two methods. First, all similarities of the target lemma are taken into account and the average similarity is computed as equation 3:

$$FCS = \frac{\sum_{i=1}^{n} S_i}{n} \tag{3}$$

Where, S is the similarity of the lemma and n is the number of lemmas in the corpus; this method we call the full context specificity (FCS).

The second method is based on the analysis of rank similarities of distances. Specifically, rank similarities of distances display S-shaped curve, see Figures 3 and 4 for lemmas 'atom' and 'nebo' (from the subcorpus 2013). From the graph, it is obvious that there is a small proportion of lemmas which are close to the target lemma (i.e. lemmas with the highest similarity values), then a majority of lemmas lie in a relatively tiny interval of similarity and, finally, there is a small proportion of lemmas with negative values of similarity which means that the contexts in which these lemmas appear are negatively

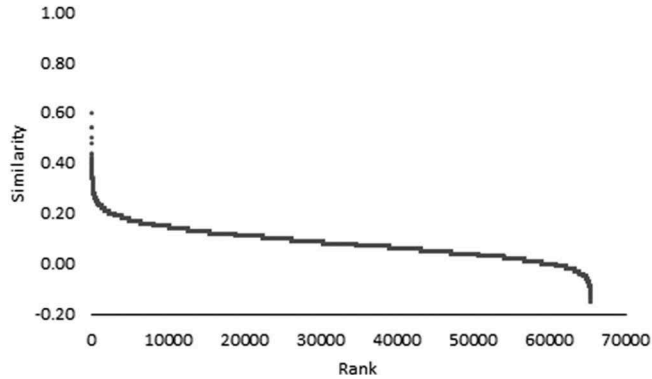

Figure 3. Rank similarities of distances for the lemma 'atom' in the subcorpus 2013.

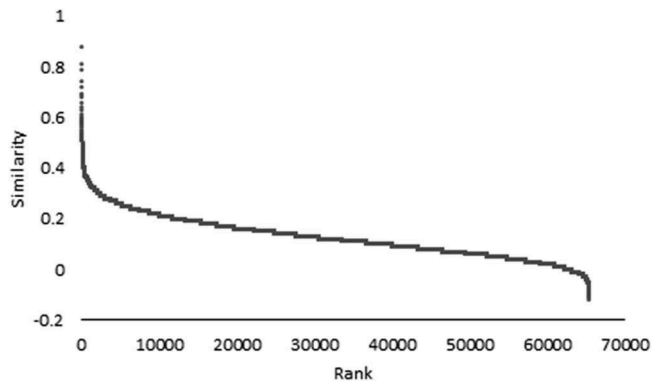

Figure 4. Rank similarities of distances for the lemma 'nebo' in the subcorpus 2013.

correlated (the lemmas which appear in the first context do not appear often in the second context and vice versa). With regard to the concept of CSL, lemmas with highest values of similarities represent the most important units because these lemmas are the closest lemmas of the target lemma in the multidimensional space. Therefore, we decide to compute also the mean value of similarities for the first 20 lemmas, see equation 4, as an alternative way of the calculation of the context specificity:

$$CCS = \frac{\sum_{i=1}^{20} S_i}{20} \tag{4}$$

This method we called the closest context specificity (CCS).

For lemmas in our example, we obtain values as follows:

$$FCS_{atom} = \frac{5421.36}{65382} = 0.0829$$

$$FCS_{nebo} = \frac{8324.27}{65382} = 0.1273$$

$$CCS_{atom} = \frac{8.79}{20} = 0.4395$$

$$CCS_{nebo} = \frac{13.27}{20} = 0.6635$$

These results show that lemma 'atom' display in both measurements higher specificity than the lemma 'nebo'.

5. Results

In this section, we present how the proposed methods (see Section 4) can be used for a modeling of development of the context specificity. We decide to analyze 10 most frequent lemmas representing political discourse from the whole corpus SYNv4. The political discourse is chosen because it can be considered one of the most dynamic genres in a relatively short time span covered by the corpus (c.f., Section 3). Specifically, these lemmas are used for the analysis: VLÁDA [government], PRÁVO [law], PREZIDENT [president], MINISTR [minister], MINISTERSTVO [ministry], ODS [the acronym of the Civic Democratic Party], POLITIKA [politics], POLITIK [politician], and PREMIÉR [prime minister]. The values of FCS and CCS of theses lemmas are presented in Table 3 and Figures 5–8.

Results of FCS display some regularities which must be observed carefully to avoid a misinterpretation. Specifically, there are common tendencies of a development of FCS of all lemmas (see Figures 5 and 6): from the beginning, all values decrease (with exceptions of years 2001–2003), then there is a flat development and, finally, all values increase. It seems very suspicious that all chosen lemmas follow the same tendencies and it opens the question about factors causing this behavior. Among many others, a size effect is one of the most important factor which 'depreciates' many kinds of linguistic analysis (e.g. stylometry, vocabulary richness). Therefore, we decided to observe a relationship between FCS and the size of particular corpora (the size is measured in a number of lemmas, see Table 1). The relationships between these properties are presented in Figure 9. There is an evident tendency to lower values of FCS for larger corpora which corresponds with data presented in Table 1 and Figures 5 and 6 – corpora representing years 1990–2004 are smaller than the others. However, the year 2014 cannot be explained by the size-effect. Closer observation of a training procedure for the word embeddings reveals the reason of this phenomenon. Because the size of the corpus affects the

Table 3. Values of full context specificity (FCS) and closest context specificity (CCS) for ten the most frequent lemmas representing political discourse in journalistic texts.

	VLÁDA		PRÁVO		PREZIDENT		MINISTR		MINISTERSTVO	
year	*FCS*	*CCS*	*FCS*	*CCS*	*FCS*	*CCS*	*FCS*	*CCS*	*FCS*	*CCS*
1990–1996	0.15	0.63	0.13	0.53	0.14	0.59	0.13	0.62	0.12	0.59
1997	0.14	0.64	0.11	0.49	0.13	0.56	0.13	0.63	0.12	0.58
1998	0.12	0.65	0.10	0.50	0.11	0.53	0.11	0.62	0.10	0.58
1999	0.12	0.63	0.10	0.49	0.10	0.53	0.11	0.62	0.09	0.58
2000	0.11	0.59	0.10	0.50	0.10	0.52	0.10	0.61	0.09	0.59
2001	0.12	0.61	0.10	0.49	0.10	0.52	0.11	0.61	0.10	0.58
2002	0.12	0.62	0.09	0.49	0.11	0.55	0.11	0.61	0.10	0.58
2003	0.12	0.64	0.10	0.49	0.11	0.57	0.11	0.60	0.10	0.56
2004	0.10	0.63	0.09	0.51	0.09	0.53	0.09	0.59	0.09	0.58
2005	0.10	0.63	0.09	0.49	0.09	0.52	0.09	0.61	0.09	0.58
2006	0.10	0.64	0.09	0.48	0.09	0.52	0.09	0.62	0.09	0.57
2007	0.10	0.63	0.08	0.48	0.09	0.52	0.09	0.61	0.09	0.57
2008	0.10	0.62	0.08	0.47	0.09	0.55	0.09	0.59	0.09	0.57
2009	0.10	0.63	0.09	0.47	0.09	0.52	0.09	0.59	0.09	0.56
2010	0.09	0.59	0.08	0.46	0.09	0.52	0.09	0.58	0.08	0.56
2011	0.09	0.61	0.08	0.48	0.09	0.52	0.09	0.59	0.09	0.57
2012	0.10	0.62	0.09	0.47	0.09	0.52	0.09	0.59	0.09	0.56
2013	0.11	0.66	0.09	0.50	0.10	0.55	0.10	0.62	0.09	0.56
2014	0.12	0.62	0.11	0.55	0.12	0.59	0.12	0.61	0.11	0.60
	ODS		OBČAN		POLITIKA		POLITIK		PREMIÉR	
year	*FCS*	*CCS*	*FCS*	*CCS*	*FCS*	*CCS*	*FCS*	*CCS*	*FCS*	*CCS*
1990–1996	0.12	0.67	0.13	0.50	0.15	0.57	0.16	0.56	0.14	0.62
1997	0.12	0.69	0.12	0.50	0.14	0.57	0.14	0.58	0.13	0.62
1998	0.11	0.69	0.11	0.49	0.12	0.55	0.13	0.58	0.12	0.61
1999	0.11	0.67	0.10	0.49	0.11	0.54	0.12	0.56	0.11	0.58
2000	0.10	0.66	0.10	0.48	0.11	0.54	0.12	0.56	0.11	0.59
2001	0.10	0.64	0.10	0.48	0.11	0.53	0.13	0.57	0.11	0.59
2002	0.11	0.70	0.10	0.49	0.12	0.54	0.13	0.59	0.12	0.61
2003	0.10	0.65	0.09	0.48	0.12	0.55	0.13	0.57	0.12	0.62
2004	0.10	0.67	0.09	0.52	0.10	0.54	0.10	0.57	0.10	0.62
2005	0.09	0.66	0.09	0.52	0.10	0.53	0.10	0.58	0.10	0.64
2006	0.10	0.72	0.09	0.51	0.10	0.52	0.11	0.57	0.10	0.63
2007	0.09	0.64	0.09	0.53	0.09	0.52	0.10	0.56	0.10	0.62
2008	0.09	0.66	0.09	0.51	0.10	0.54	0.10	0.56	0.10	0.61
2009	0.09	0.64	0.09	0.51	0.09	0.52	0.11	0.58	0.10	0.61
2010	0.09	0.71	0.08	0.51	0.10	0.53	0.10	0.56	0.10	0.59
2011	0.09	0.65	0.09	0.52	0.10	0.53	0.10	0.56	0.10	0.58
2012	0.09	0.66	0.09	0.50	0.10	0.54	0.11	0.58	0.10	0.59
2013	0.10	0.64	0.09	0.51	0.11	0.55	0.11	0.59	0.10	0.62
2014	0.12	0.71	0.11	0.52	0.12	0.56	0.13	0.59	0.12	0.61

quality of the word embeddings and the size is positively correlated with a year, we first train the word embeddings for the year 2014 and then initialize the word embeddings for 2013 from these trained vectors. Therefore the training procedure for the year 2013 does not have to start from scratch, the word vectors are more or less in proper relative positions and during training they are only adjusted to fit the distribution of the year 2013. The same is done for all other years, we always initialize the vectors from previous year. We speculate that the high value of FCS in

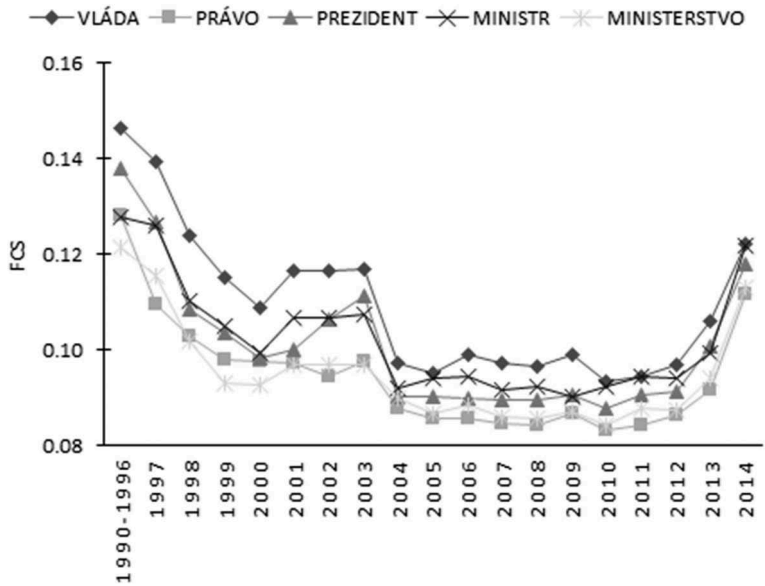

Figure 5. FCS values of selected lemma set A.

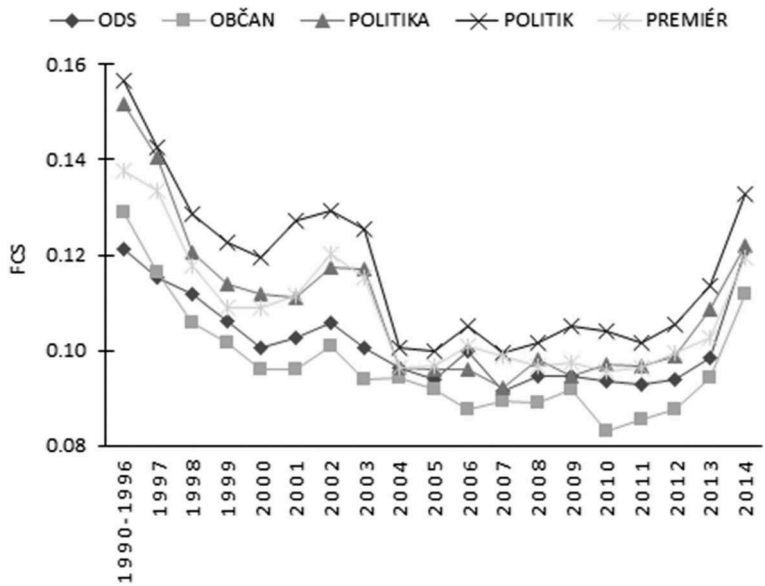

Figure 6. FCS values of selected lemma set B.

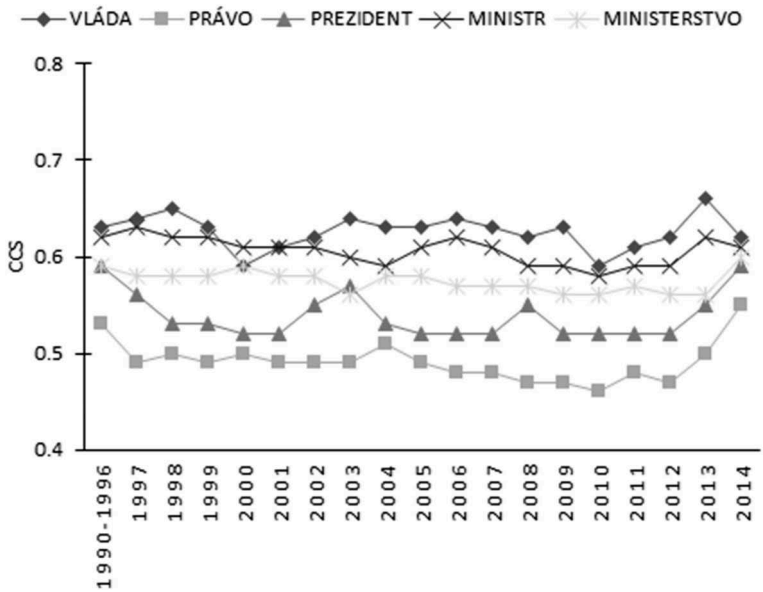

Figure 7. CCS values of selected lemma set A.

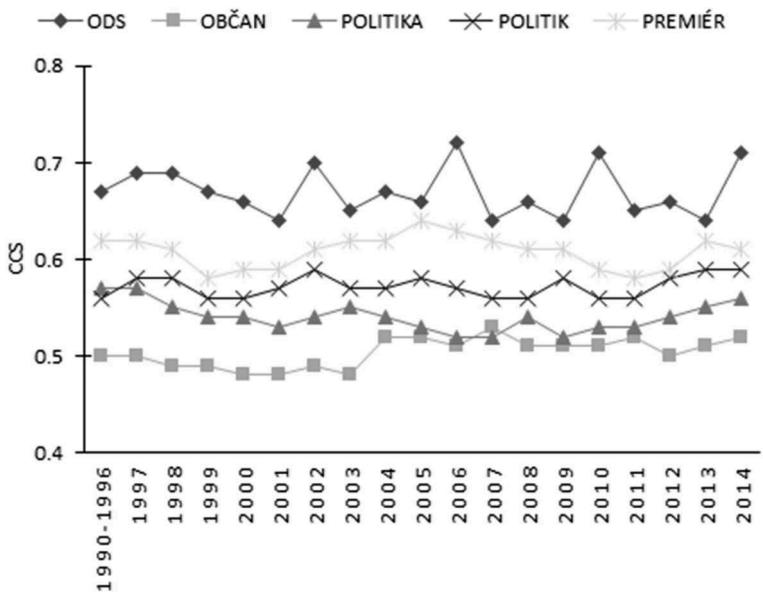

Figure 8. FCS values of selected lemma set B.

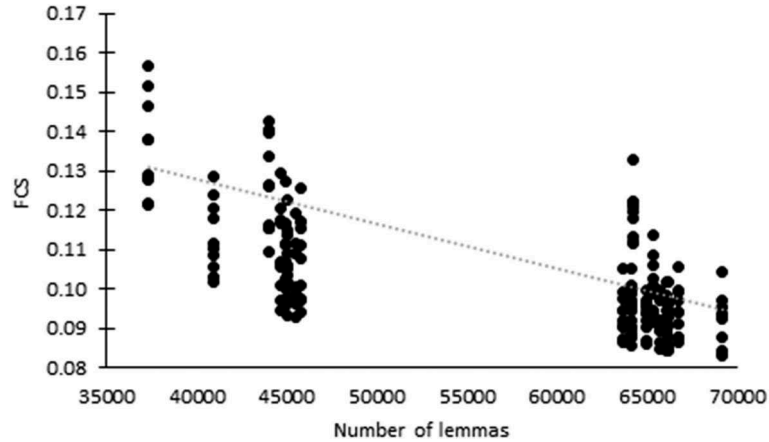

Figure 9. The relation between number of lemmas and their FCS values.

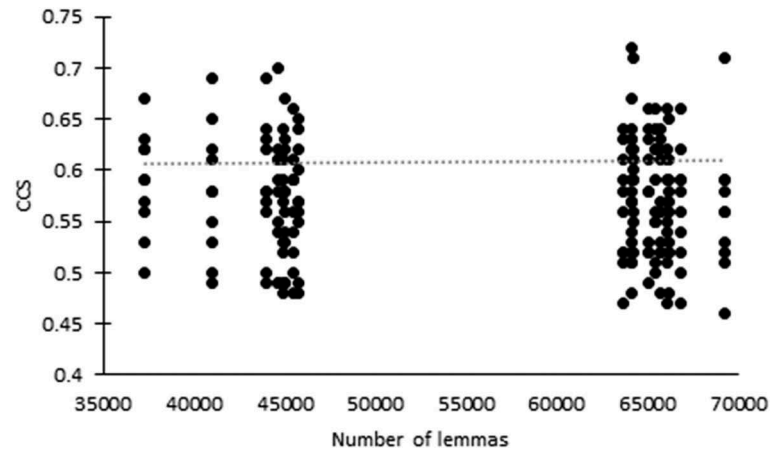

Figure 10. The relation between number of lemmas and their CCS values.

the year 2014 is caused by not having large enough corpus so that the vectors would converge to stable positions even when trained from scratch. Based on these facts, we can conclude that FCS does not seem to be suitable method for the measurement of CSL.

Results of CCS show different picture in comparison to FCS (cf. Figures 7 and 8). First, there is no obvious common tendency for all lemmas. One can see differences of CCS among particular lemmas as well as their different developments – compare relatively flat development of lemma MINISTR and changing development (with four peaks) of lemma ODS. Further, closer observation of the potential impact of the size-effect on CCS reveals no relationship between the size of corpus and

CCS, see Figure 10. Consequently, we adopt CCS as a suitable method for the measurement of CSL. Moreover, from the linguistics point of view, CCS seems to be more acceptable method, too. Specifically, a determination of CSL based on the closest lemmas should reflect its property better than a determination based on all lemmas because a majority of lemmas has little in common to target lemma (as is seen from rank similarities of distances – see Figures 3 and 4).

Differences of development of CCS among particular lemmas are determined by changes of CCS in subsequent years. Specifically, the more changes between subsequent values of CCS, the more dynamic development and vice versa. An index of lemma dynamism D is computed as

$$D_{lemma} = \sum_{year}^{N-1} |CCS_i - CCS_{i+1}| \tag{5}$$

For illustration, D of the lemma ODS is

$$\begin{aligned} D_{ODS} = {} & |0.67 - 0.69| + |0.69 - 0.69| + |0.69 - 0.67| + |0.67 - 0.66| + \\ & |0.66 - 0.64| + |0.64 - 0.70| + |0.70 - 0.65| + |0.65 - 0.67| + |0.67 - 0.66| + \\ & |0.66 - 0.72| + |0.72 - 0.64| + |0.64 - 0.66| + |0.66 - 0.64| + |0.64 - 0.71| + \\ & |0.71 - 0.65| + |0.65 - 0.66| + |0.66 - 0.64| + |0.64 - 0.67| = 0.62 \end{aligned}$$

We repeat the same procedure for the 10 most frequent lemmas, and the results are presented in Table 4 and Figure 11.

As the data shown, lemma ODS has the largest value of D, which means that ODS is the most dynamic lemma that has the strongest volatility. To be more specific, it means the context specificity of ODS is less stable in its diachronic development. If we take a closer look of its development in Figure 8, we can find four peaks. The interesting fact is that these four peaks correspond to election years. Thereby, we conclude that the context specificity of ODS show a different pattern for election years in comparison to others. This result fits one's language intuition well since ODS, as the acronym of the Civic Democratic Party, is obviously more sensitive to elections than other lemmas. One the opposite side, we can find lemmas displaying highly stable development, such as MINISTR, POLITIKA, MINISTERSTVO.

6. Conclusion and Discussion

The context specificity of lemma is the concept which can be used for modeling diachronic lexical dynamism. As it is shown in our study, measurements of this property must be thoroughly scrutinized before an application, to avoid misinterpretations of results. Furthermore, we

Table 4. D_{lemma} values of selected lemmas.

lemma	*D*
ODS	0.62
VLÁDA	0.33
PREZIDENT	0.3
PRÁVO	0.26
PREMIÉR	0.21
POLITIK	0.19
OBČAN	0.18
MINISTR	0.17
POLITIKA	0.17
MINISTERSTVO	0.15

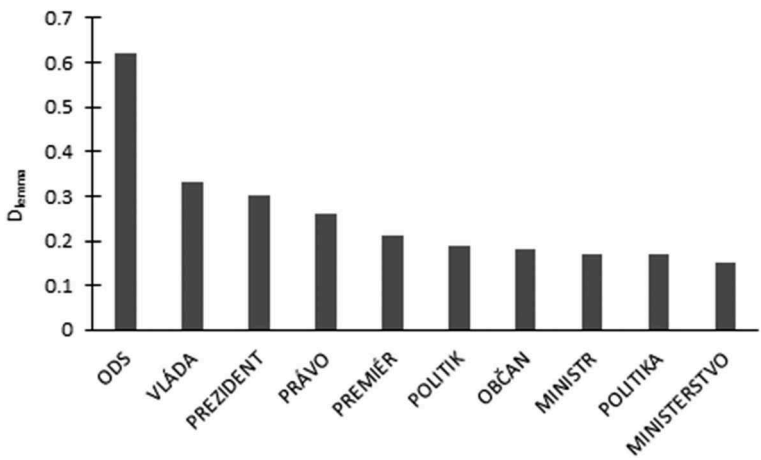

Figure 11. D_{lemma} values of selected lemmas.

found out that the CCS measurement, which is theoretically well interpretable as well as independent of the sample size, enable us both to detect differences among particular lemmas and to analyze the dynamism of lemmas. The preliminary results, presented in this study, seem to be promising and we assume this method can be applied in a broad range of linguistics research, such as critical discourse analysis, content analysis, stylometry, etc. One of the main advantages of applying the word/lemma embedding approach is that it allows quantitative descriptions and comparisons of semantic characteristics of lemmas. However, one should still bear in mind that the computing of word/lemma embedding is very complex – it is based on hundreds or even thousands of parameters which are needed for determination of particular vectors. Needless to say, a linguistic interpretation of such amount of parameters is impossible and, consequently, this method has a character of a black box. On the other hand, if this method is used as a starting point for an

analysis which has clear linguistic interpretation (such as CCS and its dynamic development) and it brings valuable results, its application poses a challenge for linguistic research.

Notes

1. http://korpus.cz.
2. More information about Corpus SYN version 4 can be found on http://wiki.korpus.cz/doku.php/en:cnk:syn:verze4.
3. https://wiki.korpus.cz/doku.php/cnk:syn:verze4.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the University of Ostrava; under Grant SGS02/UVAFM/2017; and Social Science Fund of Shaanxi Province under Grant 2015K001; European Regional Development Fund [CZ.02.1.01/0.0/0.0/16_019/0000734].

References

Allan, K. (Ed.). (2013). *The Oxford handbook of the history of linguistic*. Oxford: Oxford University Press.

Burkhardt, A., Steger, H., & Wiegand, H. E. (Eds.). (2000). History of the language sciences. In *Geschichte der Sprachwissenschaften Histoire des sciences du langage*. Berlin, New York: Walter de Gruyter.

Hamilton, W. L., Jurafsky, D., & Leskovec, J. (2016). Diachronic word embeddings reveal statistical laws of semantic change. CoRR, abs/1605.09096.

Hnátková, M., Křen, M., Procházka, P., & Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of the ninth international conference on Language Resources and Evaluation (LREC'14)* (pp. 160–164). Reykjavík: ELRA

Köhler, R. (1993). Synergetic linguistics. In R. Köhler & B. Rieger (Eds.), *Contributions to quantitative linguistics. Proceedings of the 1st quantitative linguistics conference*, QUALICO, University of Trier, 1991 (pp. 41–51). Dordrecht: Kluwer.

Köhler, R. (2005). Synergetic linguistics. In R. Köhler, G. Altmann, & R. G. Piotrowski (eds), *Quantitative linguistics. An international handbook* (760–774). Berlin, New York: de Gruyter.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., … Zasina, A. (2016). Corpus SYN, version 4 from 16. 9. 2016. Ústav Českého národního korpusu FF UK, Praha 2016. Retrieved from http://www.korpus.cz.

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of The Association for Computational Linguistics*, *3*, (pp. 211–225).

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013b). *Efficient estimation of word representations in vector space* (ICLR Workshop Papers).

Mikolov, T., Chen, K., Corrado, G. S., Dean, J., & Sutskever, I. (2013a). Distributed representations of words and phrases and their compositionality. *Proceedings of Neural Information Processing Systems (NIPS 26)* (pp. 3111–3119).

Pennington, J., Socher, R., & Manning, C.D. (2014) GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, (pp. 25–29) October 2014, (pp. 1532-1543).