Glottometrics 45, 2019, 7-23

Context Specificity of Lemma. Diachronic Analysis

Jan Hůla[1]
Miroslav Kubát[2]
Radek Čech[3]
Xinying Chen[4]
David Číž[5]
Kateřina Pelegrinová[6]
Jiří Milička[7]

Abstract. This study deals with the recently proposed concept of so-called Context Specificity of Lemma (*CSL*). *CSL* is based on the word embedding technique called Word2vec which enables measuring lexical context similarity between lemmas. Specifically, a recently proposed method Closest Context Specificity (*CCS*) is applied to a diachronic analysis of Czech texts. This method expresses how unique is a context within which a given lemma appears. The aim of the paper is to study what kind of semantic features can *CCS* detect and how useful could *CCS* be in a diachronic semantic analysis. The second goal is to observe the relation of *CCS* to frequencies in the corpora.

Keywords. *Word2vec, semantics, diachronic analysis, context specificity.*

1. Introduction

Generally speaking, the semantics of any linguistic unit is a very complex issue which is difficult to study in a quantitative way. Considering the number and the variation of the factors playing a role (especially pragmatic ones), it seems to be nearly impossible to express the meaning of a linguistic unit (in our case a lemma) using quantitative methods. However, very innovative methods based on neural networks approach have recently shown promising results. Namely, Word2vec technique enables measuring semantic similarities between words, where the meaning of a word is given by its context (Mikolov 2013a, 2013b). Čech et al. 2018 proposed a concept of so-called Context Specificity of a Lemma (*CLS*) which measures how unique is the context of a given lemma.

[1] Jan Hula, University of Ostrava, jan.hula21@gmail.com
[2] Miroslav Kubát, University of Ostrava: miroslav.kubat@gmail.com, https://orcid.org/0000-0002-3398-3125, corresponding author, University of Ostrava, Reální 5, Ostrava 701 03, Czech Republic
[3] Radek Čech, University of Ostrava: cechradek@gmail.com
[4] Xinying Chen, University of Ostrava, Xi'an Jiaotong University, cici13306@gmail.com, https://orcid.org/0000-0002-5052-4991
[5] David Číž, University of Ostrava, davidciz95@gmail.com
[6] Kateřina Pelegrinová, University of Ostrava, pelegrinovak@gmail.com
[7] Jiří Milička, Charles University, milicka@centrum.cz, http://orcid.org/0000-0001-8605-1199

A lemma has high context specificity when there are not many other lemmas which appear within a similar context. For instance, function words (synsemantics) like conjunctions or prepositions should have lower context specificity than content words (autosemantics). There is a limited number of function words and they have very low or no lexical meaning. Their role is to express some grammatical function. Therefore, function words should not be very tied to any context in general. Another example could be the difference between highly frequent lemmas with common usage such as *car, house, grass, money* on the one hand; and technical terms such as *atom*, *phoneme*, *molecule*, etc. on the other hand. The technical terms should have a much more specific context in general because their usage is very limited to the specific topics and style. Closest Specificity of Lemma (*CCS*) can detect the context of target lemmas and express the uniqueness of the context. This approach showed very promising preliminary results from synchronic (Kubát et al. 2018) and diachronic (Čech et al. 2018) points of view. This study follows up the recently proposed approach by the application of *CCS* to a diachronic analysis.

Context specificity can be considered as a semantic feature of lemmas which can be measured in a quantitative way and at the same time allows linguistic interpretation. This study is focused on the semantic changes of selected lemmas in Czech journalism during more than 20 years. The main goal of the paper is to discover whether *CCS* is a suitable tool for diachronic semantic analyses of lemmas and test the preliminary conclusion made by authors of this approach (Čech et al. 2018). The lemmas are selected in a qualitative way, i.e. we choose those lemmas where we intuitively expect potential changes in meaning during the analyzed time period. The following step is the linguistic interpretation of obtained data. We, therefore, cannot observe many lemmas, this study is rather focused of deeper insight into the behavior of *CCS* in individually selected cases because we want to understand what kind of semantic feature(s) (if any) the concept of measuring Content specificity can detect.

As the source of data, we use the Czech National Corpus. Specifically, we use one of the largest Czech corpora SYN_V4. This corpus consists of more than 3 billion tokens and covers the Czech language from 1990 to 2014. We can, therefore, analyze more than 20 years of development of the Czech language from the beginning of a democratic state after the so-called Velvet revolution in 1989 when the communistic regime fell.

Since many indicators from quantitative linguistic analyses such as vocabulary richness are influenced by text length (cf. Kubát 2016), we also pay attention to this problem in this study. The relation of Closest Context Specificity (*CCS*) to the relative frequencies in the corpora is tested.

2. Methods

2.1 Word Embeddings

Word Embeddings represent a set of methods which are effective for finding useful representations of textual data which are usually collected in a form that is not suitable for a task at hand. These representations are produced by taking the original representation (with dimensionality equal to the number of distinct words within the corpus) as input and transforming it through series of numerical operations to different representations (usually with much lower dimensionality) which have certain desirable properties. The exact value of the output representation is dependent on the learnable parameters which are found by maximizeing a score function on a concrete task. For word embeddings, the task is usually language modeling where we try to predict the words within the corpus conditioning on the words in its neighborhood. We can use the obtained score to update the parameters of the model in a way

which tries to increase the score. By iterating this process, we are trying to maximize the score and thus to find a better representation for the task. In our case, we want the representation of a word 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 next to it within the corpus). Thus, if two words often appear in the same context, their vector representations should be close to each other.

Such word embeddings are easy to obtain with algorithms such as Word2Vec or GloVe (Mikolov et al. 2013a; Manning et al. 2014). In our work, we are focusing on the Word2Vec algorithm, concretely the Skip-Gram version of it. The algorithm aims to represent a word (in our case the lemma) as a high-dimensional (50–1000) vector which captures co-occurrence statistics between the lemma itself and other lemmas in the small window centered at this lemma. The window acts as a context for the lemma in the center. Intuitively the vector representing the lemma should contain information about the contexts where it appears. Concrete values of these vectors are found by a process which tries to maximize an objective function which measures how well can be every lemma within the window predicted based on the lemma in the center of this window. 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 individual summands (log probabilities) are maximized. The first sum (indexed by t) iterates over all tokens within 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" and such co-occurrence is frequent within the corpus, 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:

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

where $u(o)$ and $v(c)$ are vector representations of lemmas o and c (o for the outer lemma, c for center lemma).

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 vector v("funny") 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 vector u("funny") as its representation. These two vectors are used only to simplify the optimization problem. In 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 assigned to them. Thus, even if the exact values of these vectors are not interpretable, their closeness could be interpreted. For measuring this kind of lexical context similarity 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, when it is close to -1, 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).

2.2 Context Specificity of Lemma (*CSL*)

The concept of measuring the so-called Context Specificity of Lemma (*CSL*) was recently proposed by Čech et al. (2018). This method measures how unique is the context in which the lemma appears. This approach is based on the fact that we can compute the similarity of a given lemma to all other lemmas using Word2vec technique (Mikolov et al. 2013a). Each lemma is represented by a vector. Both the size and the orientation of the vector express the position of a lemma in a contextual multi-dimensional space. Statistics of these similarities (e.g. mean value) can be used for characterizing the *CSL*. The lower the mean of similarities, the higher the *CSL*.

There are several methods of measuring the context specificity (cf. Čech et al. 2018). The most promising preliminary results in discourse analysis were obtained by Closest Context Specificity (*CCS*). This measurement is based on the average value of the similarities *S* of the 20 closest (most similar) lemmas to the target lemma. The formulas for CCS calculation is as follows:

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

where S = the similarity of the lemma.

It should be mentioned that we modified a bit the originally proposed formula by Čech et al. (2018) which is as follows:

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

We just use a reverse value. The reason for this modification lies in the easier interpretation. Originally, the higher the *CCS*, the less specific the context of the target lemma. After the modification the higher the *CCS*, the more specific the context of the target lemma. We consider the original version quite misleading and therefore we modified it.

For instance, we can illustrate the *CCS* calculation procedure on a lemma “banka” (a bank) based on the data from the subcorpus restricted to the year 2014. First, we need a list of the 20 closest lemmas to the target lemma “banka” (a bank) with the values of similarities S_i. The S_i values express how much similar is the context of a given lemma to the target lemma (see Table 1). Second, we apply the aforementioned formula and gain the resulting value *CCS* = 0.37 (i.e. 1 - the arithmetic mean of the *S* values).

Table 1
20 closest lemmas to the target lemma "banka" (a bank) in the subcorpus 2014

#	lemma	S
1	bankovní (bank - adjective)	0.742
2	LBBW	0.674
3	spořitelna (bank)	0.661
4	Citibank	0.660
5	Equa	0.658
6	Raiffeisenbank	0.654
7	úvěrování (crediting)	0.634
8	kreditní (credit - adjective)	0.631
9	bankéř (banker)	0.628
10	mezibankovní (interbank - adjective)	0.627
11	debetní (debit - adjective)	0.625
12	Hypoteční (mortgage - adjective)	0.625
13	Sberbank	0.622
14	bankovnictví (banking)	0.618
15	Citigroup	0.614
16	Kontokorent (overdraft)	0.613
17	mBank	0.613
18	Barclays	0.613
19	splácený (paid)	0.612
20	úročení (interest)	0.612
CCS		**0.363**

3. Data

Methods based on neural networks require large training data for producing reliable results. Since we analyze the Czech language, we decided to use the Czech National Corpus which is a suitable source for this kind of research. Namely, we work with the corpus SYN_V4. "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). This corpus is not balanced from the point of view of genres or styles. The majority of texts belong to journalism, and smaller parts consist of fiction and non-fiction texts. The structure of the corpus can be seen in Figure 1.

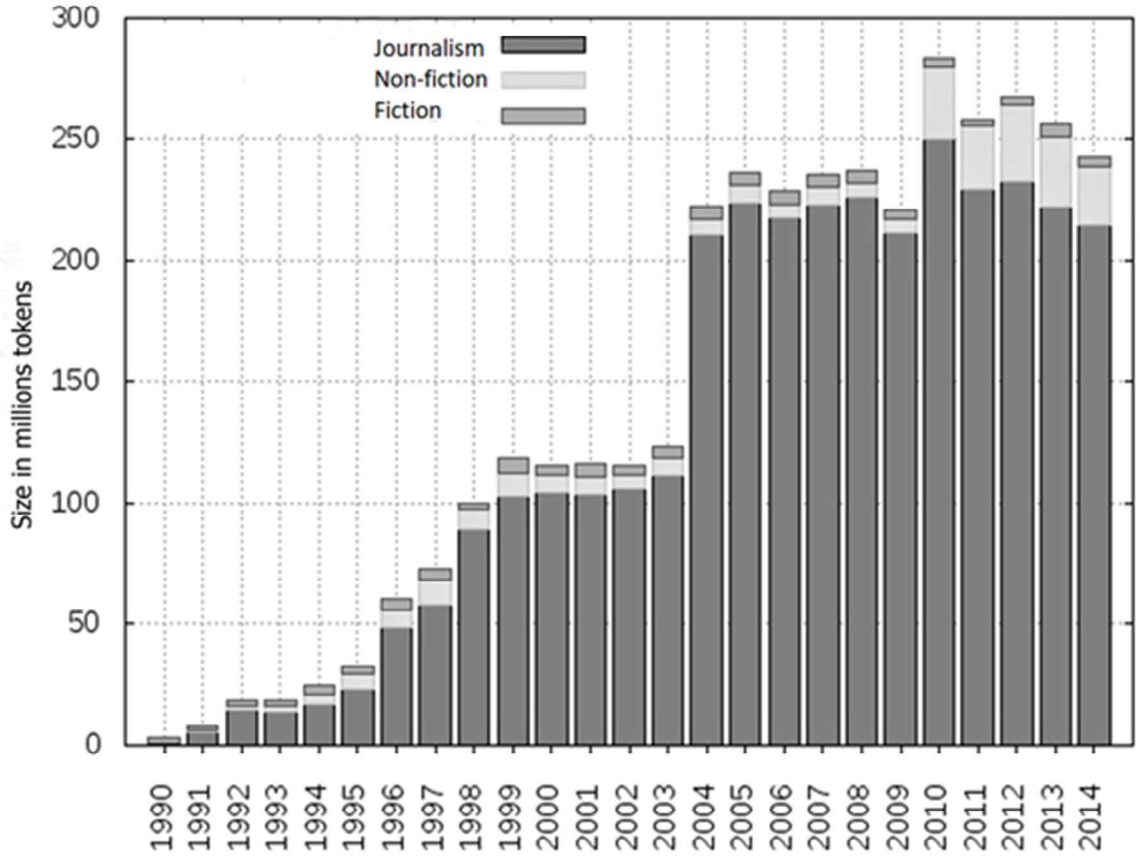

Figure 1 The composition of the corpus SYN_V4

Considering the composition of SYN_V4, we decided to use only journalistic texts due to potentially biased results. 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). Since the goal is to analyze diachronic development of the *CCS*, we divide the data into 19 subcorpora where each represents one year (see Table 2). Only the subcorpus 1990-1996 consists of texts from several years because of the small data sizes (cf. Figure 1).

Table 2
The number of lemmas in each year. Years 1990-1996 are merged because of an insufficient amount of data

Year	Number of lemmas
1990-1996	37292
1997	44023
1998	40954
1999	45038
2000	45490
2001	44930
2002	44624
2003	45757
2004	64119
2005	65008

2006	64110
2007	65698
2008	66113
2009	63695
2010	69212
2011	66167
2012	66783
2013	65381
2014	64186

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 *kočka* (a cat) has eleven different word forms for indicating its grammatical categories: *kočka, kočky, koček, kočce, kočkám, kočku, kočko, kočce, kočkách, kočkou, kočkami*. Since we focus on the semantic features of lexical units, lemmas are considered as the basic units in this research.

4. Diachronic Analysis

The goal of this analysis is to apply the recently proposed method called Closest Context Specificity (*CCS*) in diachronic semantic analysis. We select several lemmas from various fields where we expect some semantic changes. This study thus combines qualitative and quantitative approach. First, the lemmas are chosen qualitatively. Second, the lemmas are analyzed quantitatively. Third, the obtained results are qualitatively interpreted. We can then see what kind of semantic feature(s) (if any) could be detected by Context Specificity. It should be emphasized that this work does not have the ambition to make a final conclusion about the concept of Context Specificity of Lemma. However, we can do the first step to better understand this recently proposed method by a deeper look into several qualitatively chosen lemmas.

4.1 Political parties

The first analyzed group of lemmas is devoted to the Czech political parties. We chose traditional parties which continually existed from 1990 to 2014, namely: ODS, ČSSD, KDU-ČSL, KSČM. ODS is a right-wing conservative party. ČSSD is a left-wing labour party. KDU-ČSL is a Christian-democratic political party. KSČM is an extreme left-wing communistic party.

Looking at Figures 2-6, we can see a similar pattern of the four most traditional Czech political parties after 1989. The biggest changes can be seen during the time of the parliament election (1998, 2002, 2006, 2010, 2013). In these years the *CCS* is going down which means that the context of the names of political parties is less specific during elections. The reason for this behavior lies probably in the fact that newspapers focus more on the future agenda of the political parties and try to provide adequate information for voters for the election. The parties are mentioned in journalistic texts on various topics and that is why the context of the names of parties is less unique.

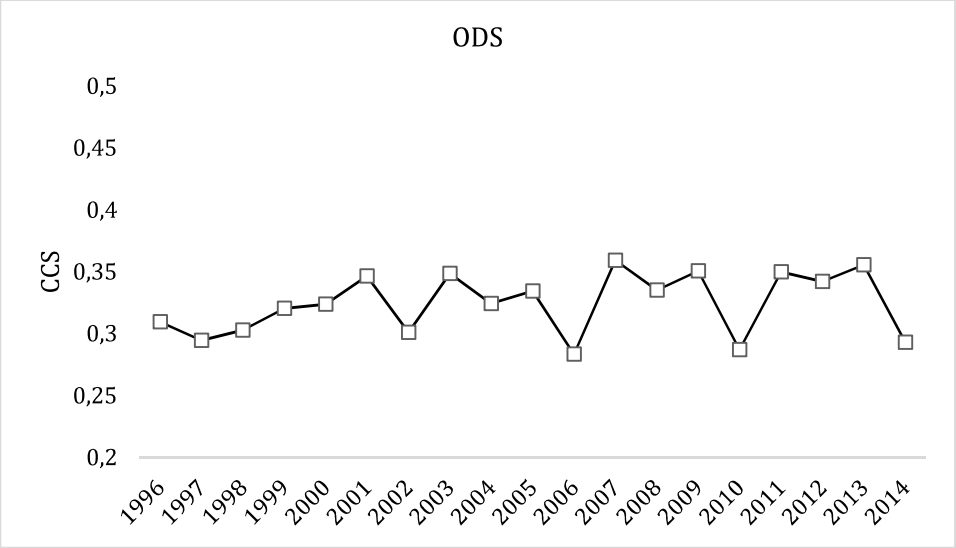

Figure 2 The *CCS* development of lemma "ODS"

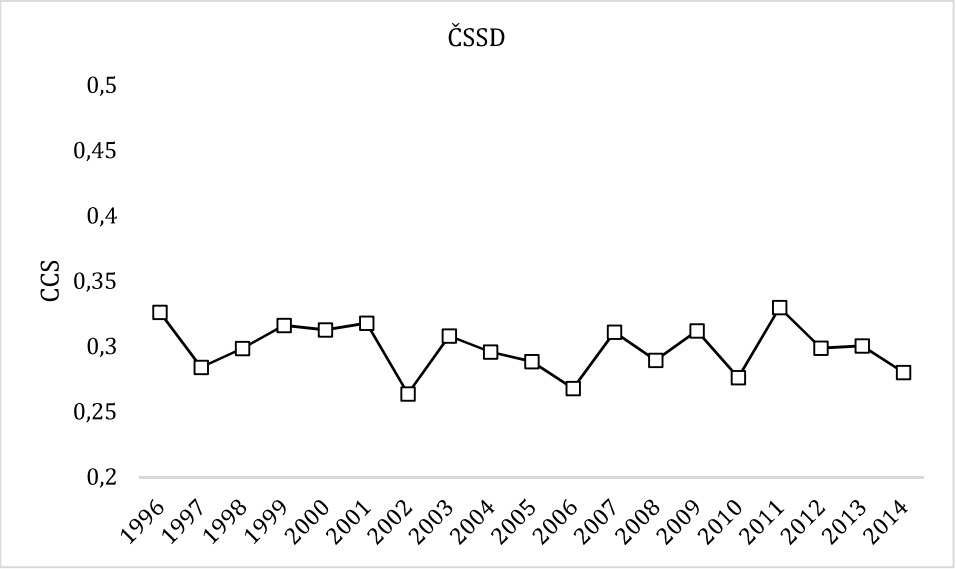

Figure 3 The *CCS* development of lemma "ČSSD"

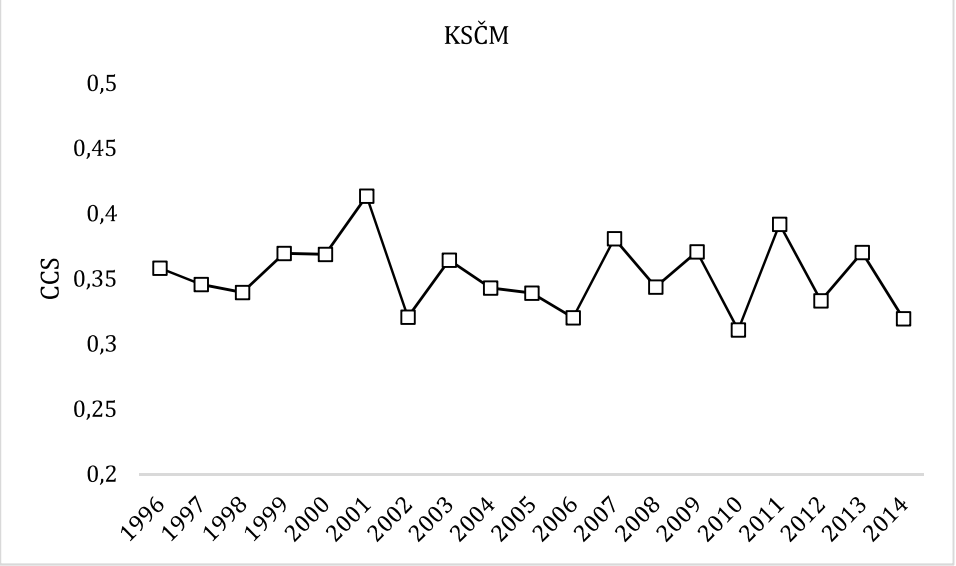

Figure 4 The *CCS* development of lemma "KSČM"

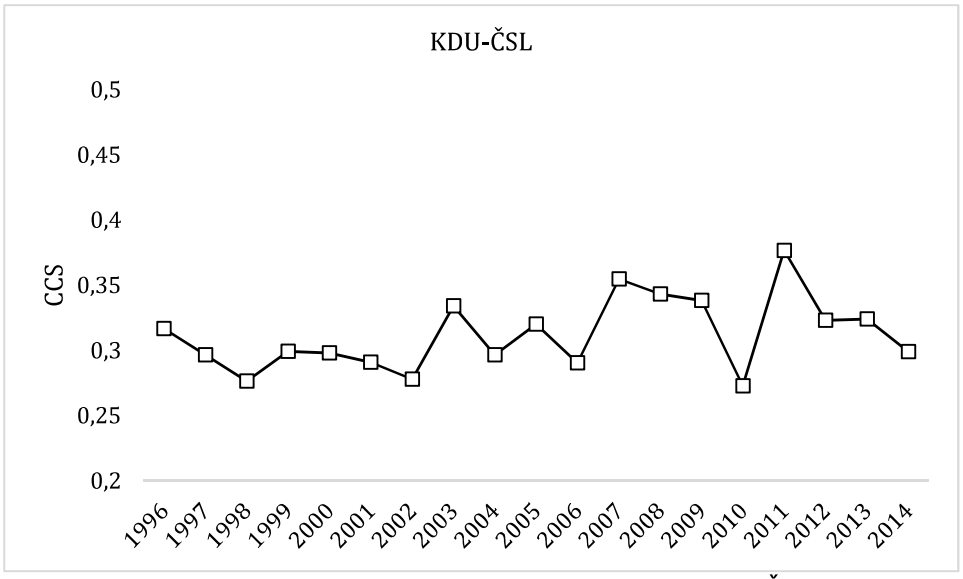

Figure 5 The *CCS* development of lemma "KDU-ČSL"

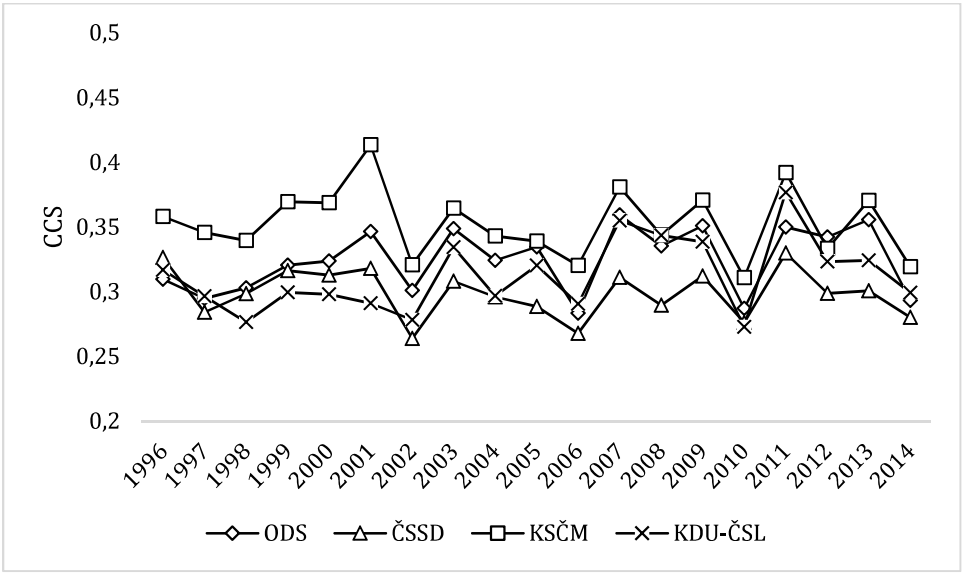

Figure 6 The *CCS* development of four traditional Czech political parties

4.2 Kraj, hejtman

In 2000, the new self-governing units were established in the Czech Republic. The name of this unit is "kraj". This word has several meanings. First, it can mean the place where something, especially surface, ends (an edge). Second, it can be used for referring to some geographical area. The last meaning is the regional unit. It should be mentioned that "kraj" also used to be a self-governing unit before 1989 with different borders and a different administration. Nowadays, the head of "kraj" is "hejtman". "Hejtman" has been used several times during the Czech history in more or less similar meanings. Thus, the usage of this lemma in newspapers in the nineties could refer to the historical meaning or to a discussion about planning new regional units. We can see in Figure 7 that the *CCS* is quite clearly reflecting the mentioned changes. The context specificity has a descending development which changes in 2000 into a rather straight curve. As we mentioned before, in the early nineties, the lemmas

"kraj" and "hejtman" had very specific meaning referring to the history. Since 2000, the context of both lemmas is generally less specific because they are appearing in newspapers in a wide range of various topics.

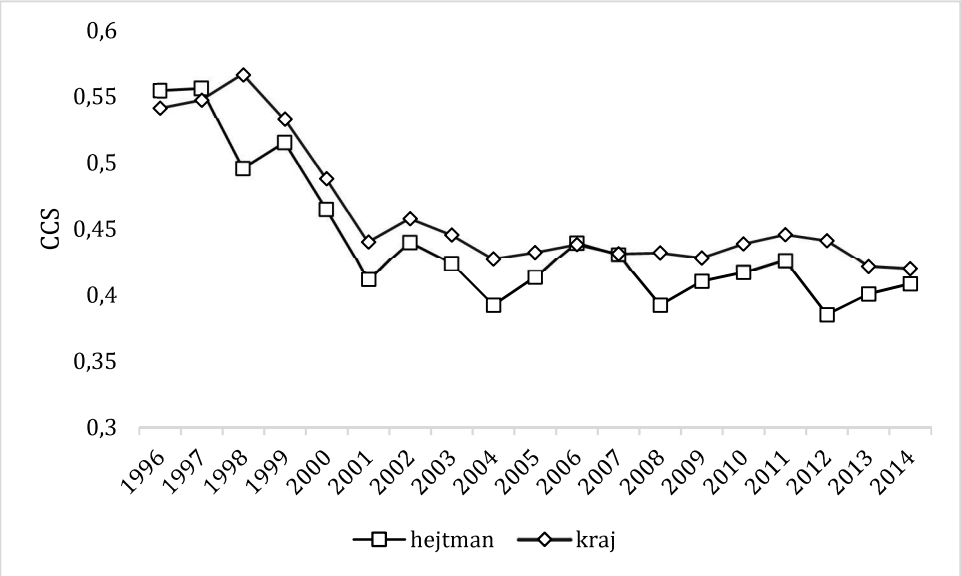

Figure 7 The *CCS* development of lemmas "hejtman" and "kraj"

The change of the meaning of the lemma "hejtman" can be also illustrated by closest lemmas at the beginning of the nineties and in 2014. In 1996, there are only those lemmas connected to the history. There are for example several lemmas referring to various administrative positions in the history of Czech lands such as "komoří", „hofmistr", „purkrabí", „místodržící", „maršálek", „falckrabě". Others are names of some historically important persons such as Pröll, Dietrichštejn, Radecký, Pühringer, Piccolomini. On the other hand, the majority of closest lemmas in 2014 belongs to the surnames of current hejtmans.

4.3 EU, NATO

The Czech Republic joined the North Atlantic Treaty Organization (NATO) in 1999 and European Union (EU) in 2004. These memberships, especially EU membership, has necessarily influenced the political agenda and content of newspapers as well. One could expect that the usage of the names of aforementioned institutions (EU, NATO) changed in a similar way like in the case of "kraj".

If we look at the resulting values in Figure 8, the development is rather the opposite. In the case of both lemmas (EU, NATO) can be seen an increasing tendency of *CCS* which is contradictory to the situation of "kraj" where the new usage of this lemma caused lower context specificity. The tendency could be interpreted in the following way. Both memberships (NATO and EU) were widely discussed before the entrance to these organizations. The newspapers informed readers about all pros and cons in general. Thus, the context was rather less specific. After joining, the news about both organizations refer to some current issues. We can see in Figure 8 that NATO has generally more unique context than EU. It is quite obvious that EU is mentioned in Czech newspapers much more frequently than NATO because the European Union has a higher influence on the daily life of people. NATO is usually mentioned in the news in connection to some NATO summits or some conflicts. The range of potential topics of EU is much wider.

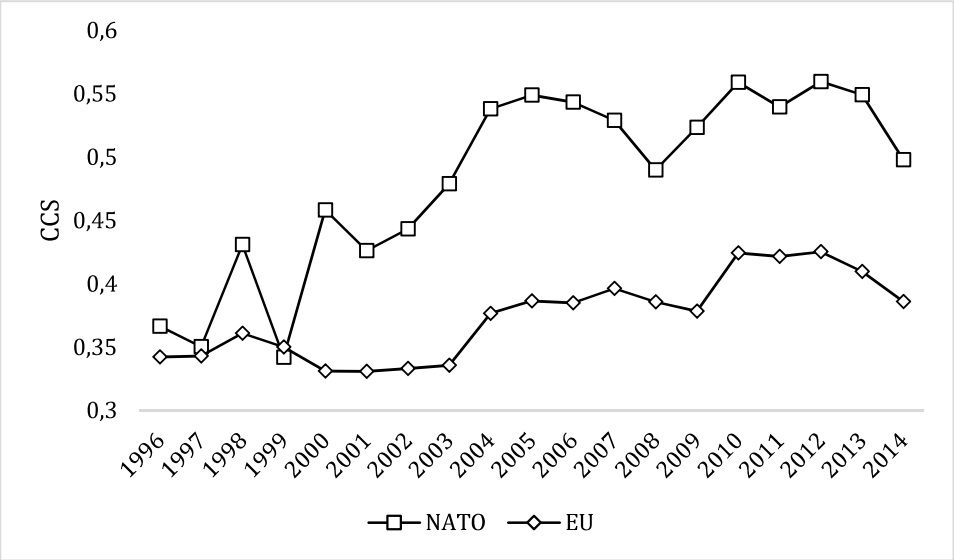

Figure 8 The *CCS* development of lemmas "NATO" and "EU"

4.4 Politicians

Another field where some semantic changes could be expected are names of famous politicians. Since we can detect changes over 20 years, we can see how *CCS* reacts to changes of politician's carriers from a long perspective. We decided to analyze the development of *CCS* of the last three Czech presidents. These politicians can be considered the most famous and influencing Czech politics. The first one, Václav Havel, was a writer, a dissident and the first Czech democratic president from 1993 to 2003. Václav Klaus is a former economist and politician who served as the second President of the Czech Republic from 2003 to 2013, and as the first Prime Minister of the newly independent Czech Republic from 1993 to 1998. Klaus was also the principal co-founder of the Civic Democratic Party (ODS), a Czech free-market Eurosceptic political party. Miloš Zeman is the current Czech president since 2013. He is the first directly elected president in Czech history. He previously served as the Prime Minister of the Czech Republic from 1998 to 2002. For many years, Zeman was also a leader of the Czech Social Democratic Party.

We can see two clear breaking points in the development of *CCS* of Havel in 2003 and 2011 in Figure 9. In 2003, Havel left the office after his second term as Czech president. The context specificity is noticeably higher in the following years. This can be explained by the fact that Havel left politics and the range of topics he was mentioned was therefore much more narrow. Havel died in 2011 and that is why he was often mentioned in newspapers in that year.

Figure 9 The *CCS* development of "Havel"

There are no such dramatic changes in *CCS* development of Klaus as in case of Havel or Zeman (see Figures 10 and 12). The reason lies in the fact that there were no big changes in his political carrier. Klaus entered Czechoslovak politics during the Velvet Revolution in 1989 and became Czechoslovakia's Minister of Finance in the same year. He served as the Prime Minister from 1992 to 1998. In 2003, he was elected as the President of the Czech Republic. Klaus has a rather stable political career where he step by step served several high positions like Minister of Finance, Prime Minister and President. Moreover, he was a leader of one of the most powerful Czech political party (ODS) from 1991-2002. He left the high politics when his presidential office ended in 2013.

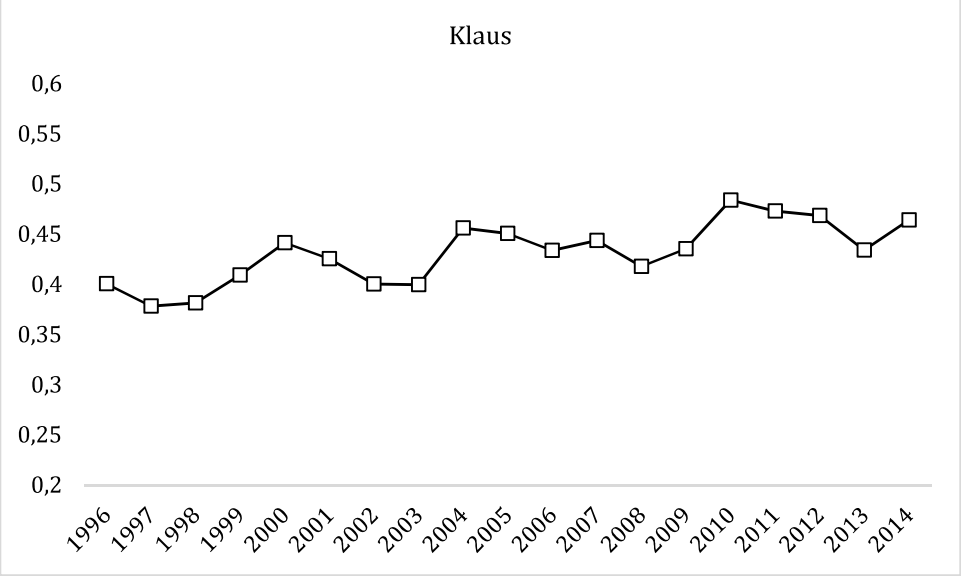

Figure 10 The *CCS* development of "Klaus"

As can be seen in Figure 11, there are two remarkable changes in the development of *CCS* values in 2003, 2013. Zeman left politics after unsuccessful presidential candidacy in 2003. He came back to politics in 2013 when he was elected as the President of the Czech Republic. We can see that the context specificity is considerably higher from 2003 until 2012 than in other years when he was an active politician.

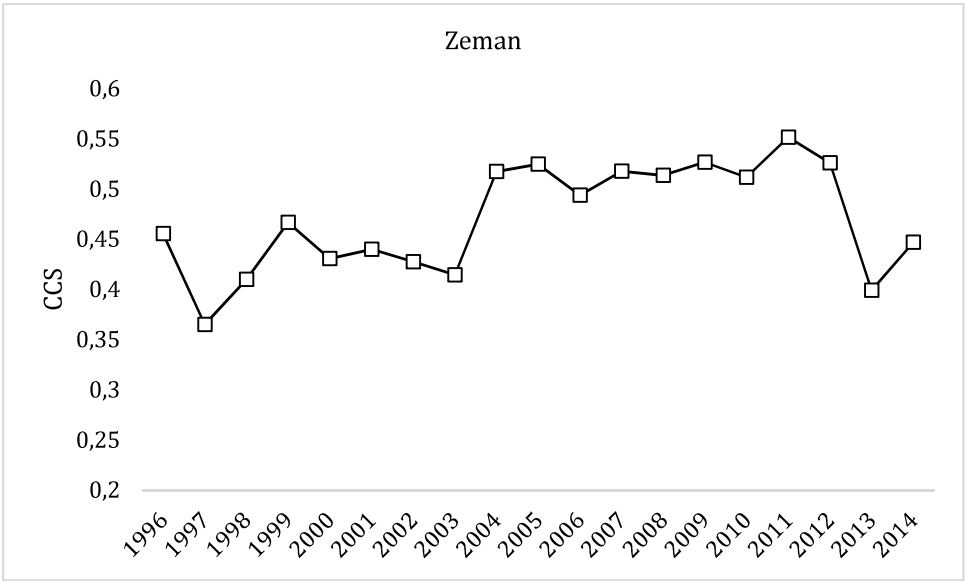

Figure 11 The *CCS* development of "Zeman"

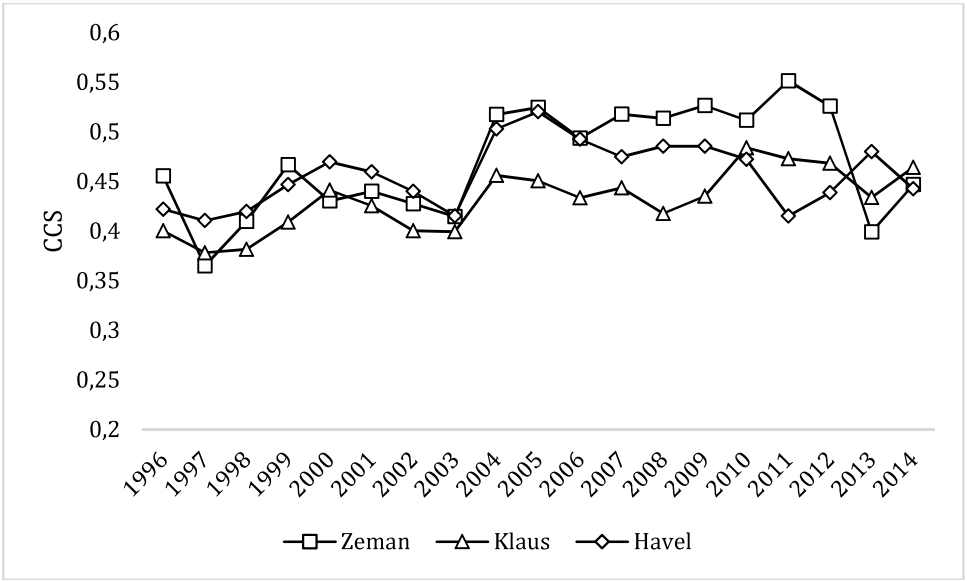

Figure 12 The *CCS* development of lemmas "Zeman", "Klaus" and "Havel"

4.5 Bird and swine flu

There were two epidemics of flu ("chřipka"), bird flu ("ptačí chřipka") and swine flu ("prasečí chřipka") in the last decade. Since these topics were widely reported in newspapers, we can expect semantic changes in the usage of lemmas "chřipka" (flu), "ptačí" (bird - adjective), and "prasečí" (swine - adjective). The years of the occurrence of these diseases are quite clearly detectable in the *CCS* development in Figures 13-16. In the Czech Republic, the bird flu emerged in 2006 and we can see that the *CCS* value drops exactly at that time. The *CCS* value has also a descendant tendency in case of the lemma "chřipka" (the flu).

The semantic changes are also very clear when we compare the closest lemmas to "ptačí" in 2006 and other years. For instance, we get following lemmas in 2000: "pták",

(bird), “ptactvo“ (birds species), opeřenec (a bird), “opeřený” (adjective of "opeřenec"), hnízdící (nesting), “voliéra” (aviary), “sýkorka” (a tit), “krahujec” (a sparrowhawk), “krkavec” (a raven), “zoborožec” (a hornbill), “včelojed” (a perninae), “káně” (a buzzard), “ornitolog” (an ornithologist), “nocoviště” (a place for birds for staying overnight), “kroužkování” (bird ringing), “krmítko” (a bird feeder), “poletující” (fliting), “zobák” (a beak), “živočich” (an animal), “zob” (a bird food). We can see that all lemmas are connected to concepts connected to birds such as bird, aviary, ornithologist, etc.

In 2006, when the bird flu emerged in the Czech Republic, we get following closest lemmas to “ptačí” (bird - adjective): “chřipka” (a flu), “H5N1”, “nákaza” (an infection), “virus” (a virus), “pták” (a bird), “ptactvo” (birds - species), “vir” (a virus), “opeřenec” (a bird), “H5”, “nakažený” (infected). “drůbež” (poultry), “uhynulý” (dead), “labuť” (a swan), “nakažení” (an infection), “slintavka” (foot-and-mouth disease), “chřipkový” (flu - adjective), “pandemie” (a pandemic), “ornitolog” (an ornithologist), “H1N1”, “Nořín” (a name of a village where the bird flu emerged). We can see that most of these lemmas are connected to the emerged bird flu. Compare to the aforementioned closest lemmas in 2000, it is clear that the context substantially changed.

The epidemic of the swine flu emerged in the Czech Republic in 2009. This topic was highly reflected in newspapers and that is why the context of lemma “prasečí” (swine - adjective) changed in our corpus. This semantic change also influenced the *CCS* of the lemma “chřipka” (the flu).

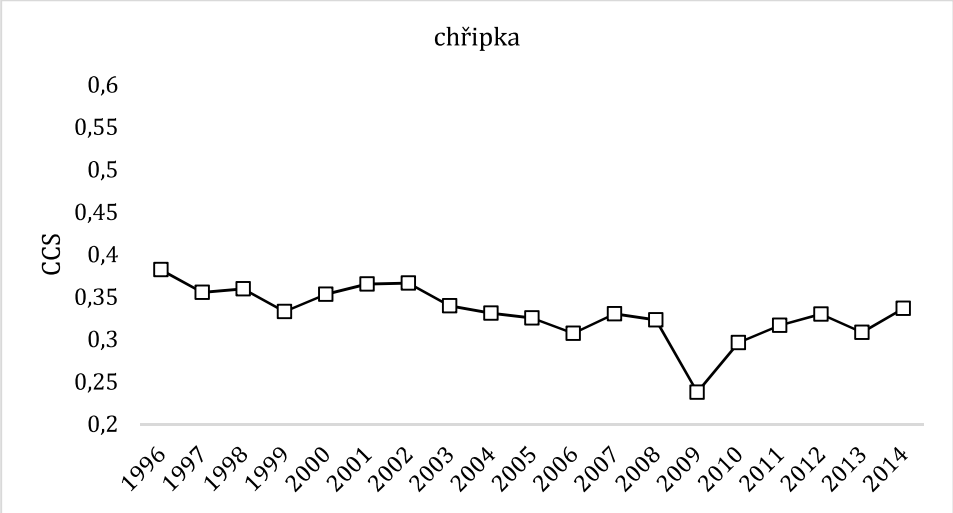

Figure 13 The *CCS* development of a lemma “chřipka” (flu)

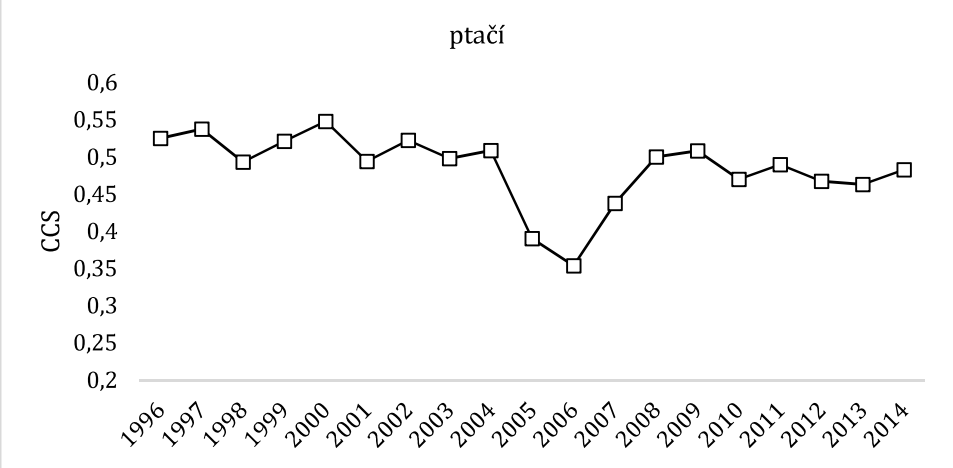

Figure 14 The *CCS* development of a lemma “ptačí” (bird - adjective)

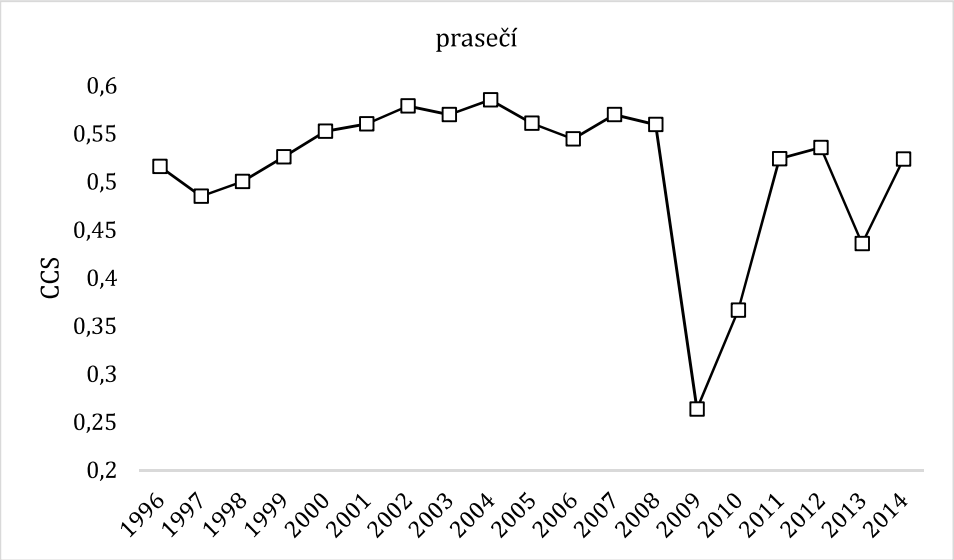

Figure 15 The *CCS* development of a lemma "prasečí" (swine - adjective)

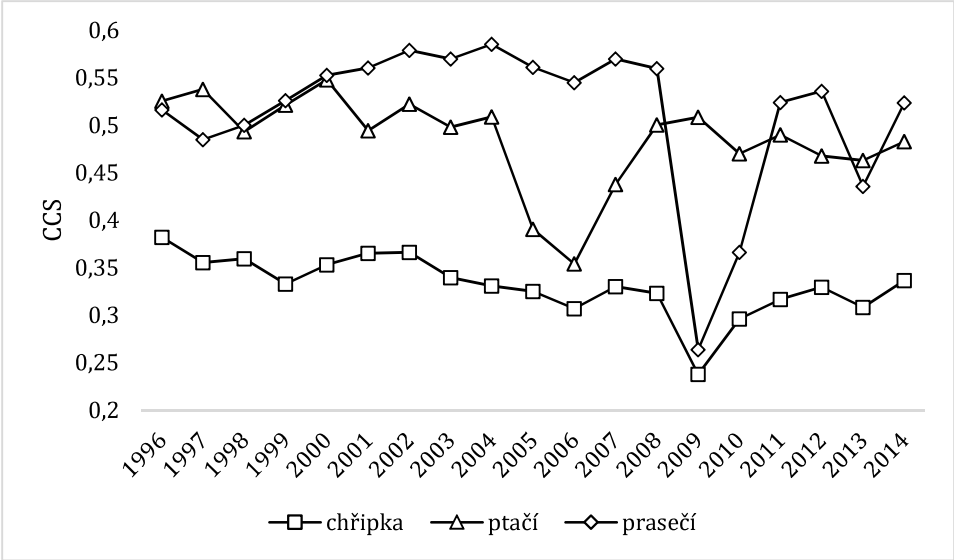

Figure 16 The *CCS* development of lemmas "chřipka", "ptačí" and "prasečí"

4.6 The relation of *CCS* to frequencies in the corpora

One of the most common obstacles of any quantitative linguistic analysis is the relation of a measured feature to frequencies in the analyzed corpus. Linguists have been dealing with this problem since they started to apply statistics to language data. The well-known case is measuring so-called vocabulary richness which is one of the common methods in quantitative linguistics, especially stylometry (cf. Kubát 2016). Given that we work with lemmas with different frequencies, we test the correlation between the obtained *CCS* values and the frequency of lemmas in the corpus. Since the analyzed subcorpora do not have the same size, the relative frequencies are used instead of the absolute frequencies. Namely, we apply the i.p.m. (instances per million) which is the average number of occurrences of the lemma in a hypothetical corpus with the size of 1 million words. We apply the Pearson correlation coefficient with the result r = -0.23. Pearson Coefficient of determination R^2 = 0.05. The correlation is visualized in Figure 17. We can see that generally *CCS* is not strongly influenced by frequencies.

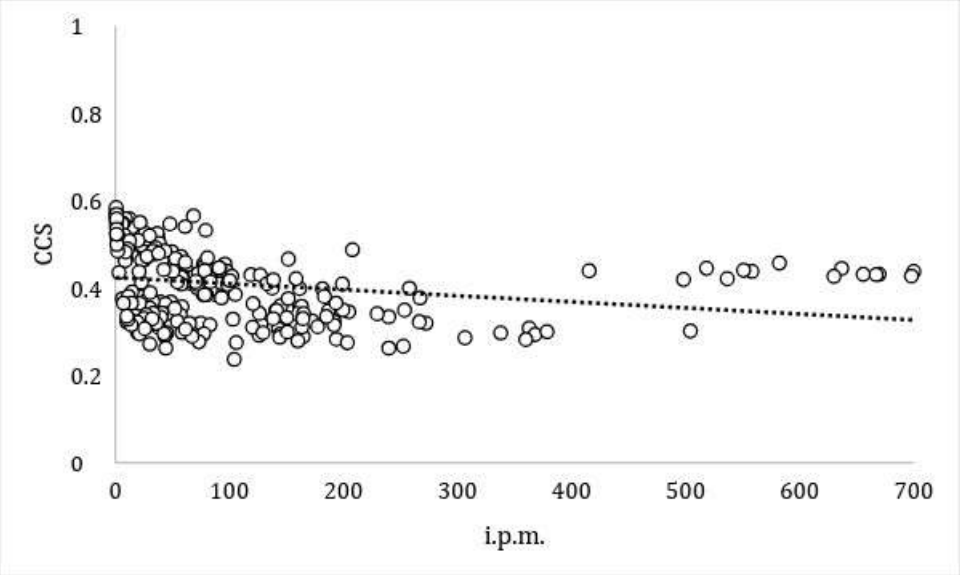

Figure 17. The correlation between *CCS* and relative frequencies (i.p.m.)

5. Conclusion

Closest Context Specificity of Lemma (*CCS*) expresses a kind of semantic feature of lemmas. The measurement is sensitive enough to study changes even in a relatively short time (several years). The behavior of the measured *CCS* development of the analyzed lemmas seems to be quite predictable and interpretable from a qualitative linguistic point of view. We tested the relation between *CCS* and frequencies of lemmas in the corpus. The results of Pearson correlation coefficient show that there is no strong correlation (r = -0.23, R^2 = 0.05).

We can state that the obtained results of this study support the preliminary conclusions given by the authors of the concept Context Specificity of Lemma (Čech et al. 2018, Kubát et al. 2018). This approach therefore seems to be promising tool for lexical semantic analyses. Since it is generally very problematic to study semantics in linguistics by quantitative methods, this method based on Word2vec technique could have a great potential in future research. The important advantage of this approach lies in the fact that even though it is based on neural networks (which are "black box" models), this concept of measuring the uniqueness of the context of the lemma allows linguistic interpretation.

Needless to say, this study is just one attempt to better understand the recently proposed method. More data must be analyzed to support or reject our conclusions based on the obtained findings in this study.

Acknowledgments

This work is supported by the University of Ostrava (No. SGS01/UVAFM/2018), the European Union & Ministry of Education of the Czech Republic (No. CZ.02.2.69/0.0/0.0/16_027/0008472), and the European Regional Development Fund (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

REFERENCES

Čech, R., Hůla, J., Kubát, M., Chen, X., Milička, J. (2018). The Development of Context Specificity of Lemma. A Word Embeddings Approach, *Journal of Quantitative Linguistics,* DOI: 10.1080/09296174.2018.1491748.

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

Kubát, M. (2016). *Kvantitativní analýza žánrů*. University of Ostrava.

Kubát, M., Hůla, J., Chen, X., Milička, J., Čech, R. (2018). The lexical context in a style analysis: A word embeddings approach. *Corpus Linguistics and Linguistic Theory*, DOI:10.1515/cllt-2018-0003

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.

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).