

SlaviCorp 2018, 24–26 September 2018, Charles University, Prague

# Context Specificity of Lemma. Diachronic analysis

Miroslav Kubát, Jan Hůla, Radek Čech, David Číž, Kateřina Pelegrinová



The research is supported by the grant of the University of Ostrava SGS01/UVAFM/2018

# Outline

1. Introduction of the method for measuring context specificity of a lemma
2. Introduction of the dataset
3. Results of a diachronic analysis

# Motivation

- Context specificity of a lemma should measure how unique is the context in which the lemma occurs.
- The lemma should have high context specificity if there are not many other lemmas which appear within the similar context.
- **Example:** *conjunctions* (low context specificity), *atom* (high context specificity)
- How to capture context in which the lemma occurs?
- How to measure uniqueness of this context?

# Word2Vec

The president be elect today by majority of people

- Assigns a vector to every lemma
- The relationship of the assigned vectors reflects the relationship of the lemmas within the corpus
- The cost function: how well can the model predict the most probable lemmas lying within the window centered around a given lemma
- Training data created from the window centered on the lemma “elect”:
  - (elect, president), (elect, be), (elect, today), (elect, by)
- By learning to predict neighbour lemmas, it learns to capture the co-occurrence statistics of lemmas within a corpus

$$J(\theta) = \frac{1}{T} \sum_{t=1}^T \sum_{-m \leq j \leq m, j \neq 0} \log p(w_{t+j} | w_t)$$

# Word2Vec

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

- Parametrization of the probability
- By maximizing this probability, the vectors  $v_c$  the vector  $u_o$  become correlated
- We use the the learned vector as simplified representations of the context in which the given lemma occurs
- We can measure how correlated are the contexts in which two different lemmas occur by measuring how correlated are their assigned vectors
- We use cosine distance between two vectors to measure similarity of contexts in which the given lemmas appear (the values lie in the interval  $[-1,1]$ )

# Context specificity of lemma (CSL)

- CSL measures how unique is the context in which the lemma appears.
- We can compute the similarity of a given lemma to all other lemmas. 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.

**20 closest (most similar) lemmas to the target lemma ATOM [ATOM]**

1	deuterium [deuterium]	0.51		11	antihmota [antimatter]	0.45
2	elektron [electron]	0.50		12	vodík [hydrogen]	0.45
3	částice [particle]	0.48		13	meteoroid [meteorid]	0.43
4	molekula [molecule]	0.48		14	křemík [silicon]	0.43
5	termojaderný [thermonuclear]	0.47		15	cesium [ceasium]	0.43
6	fúzní [fusion]	0.47		16	foton [photon]	0.42
7	termonukleární [thermonuclear]	0.45		17	vodíkový [hydrogen]	0.42
8	neutron [neutron]	0.45		18	helium [helium]	0.42
9	tritium [tritium]	0.45		19	iont [ion]	0.42
10	atomový [atomic]	0.45		20	hélium [helium]	0.41

**20 closest (most similar) lemmas to the target lemma PROTOŽE [BECAUSE]**

1	takže [so]	0.82		11	kdyby [if]	0.71
2	jelikož [because]	0.78		12	navíc [extra]	0.71
3	ale [but]	0.78		13	muset [have to]	0.70
4	proto [therefore]	0.78		14	že [that]	0.70
5	tak [so]	0.77		15	sice [though]	0.70
6	neboť [because]	0.77		16	prý [apparently]	0.69
7	ten [that]	0.74		17	tudíž [hence]	0.69
8	když [when]	0.72		18	pokud [if]	0.69
9	totiž [namely]	0.72		19	být [to be]	0.68
10	samozřejmě [of course]	0.71		20	vůbec [at all]	0.68

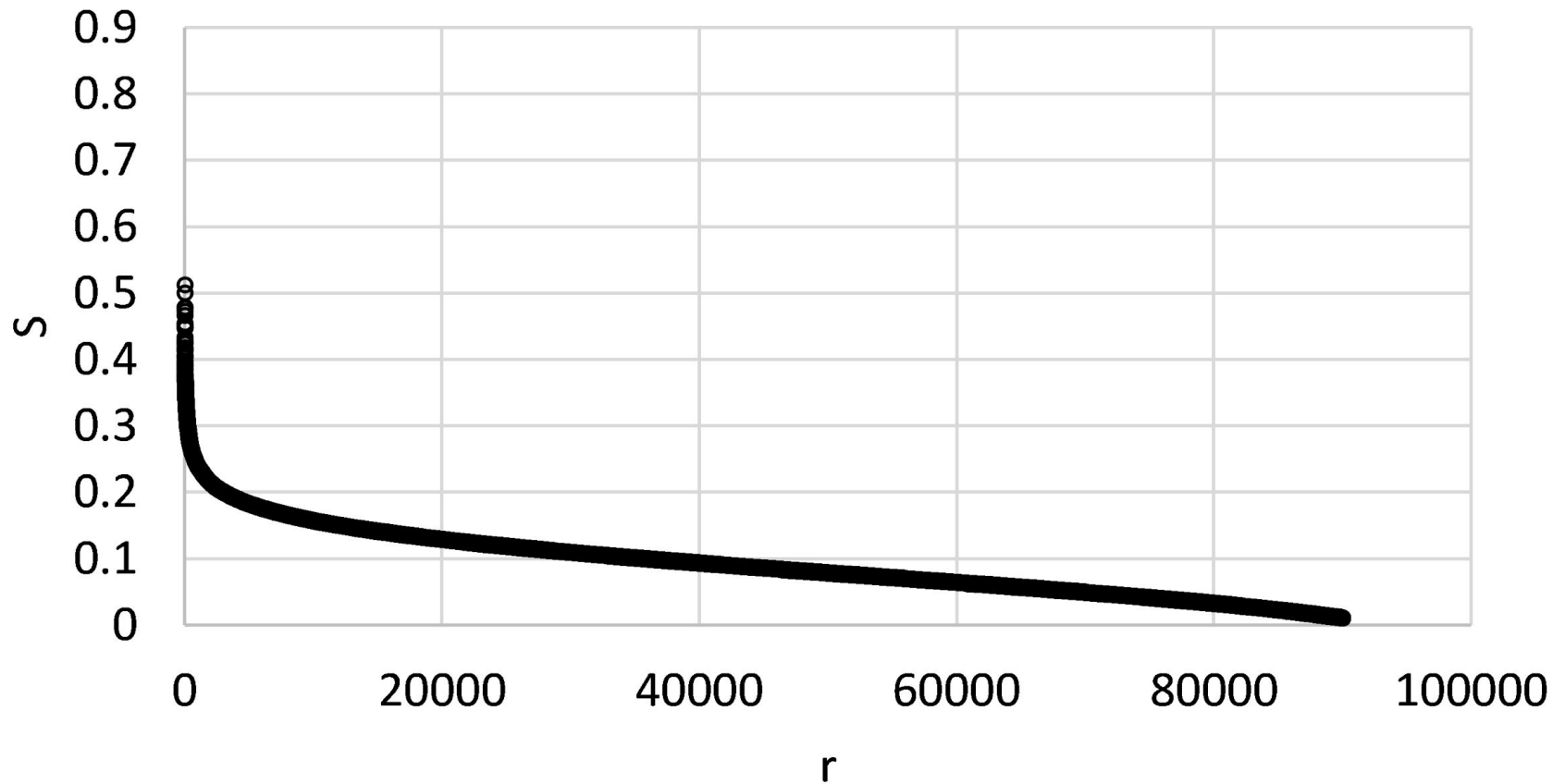
# Closest context specificity (CCS)

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

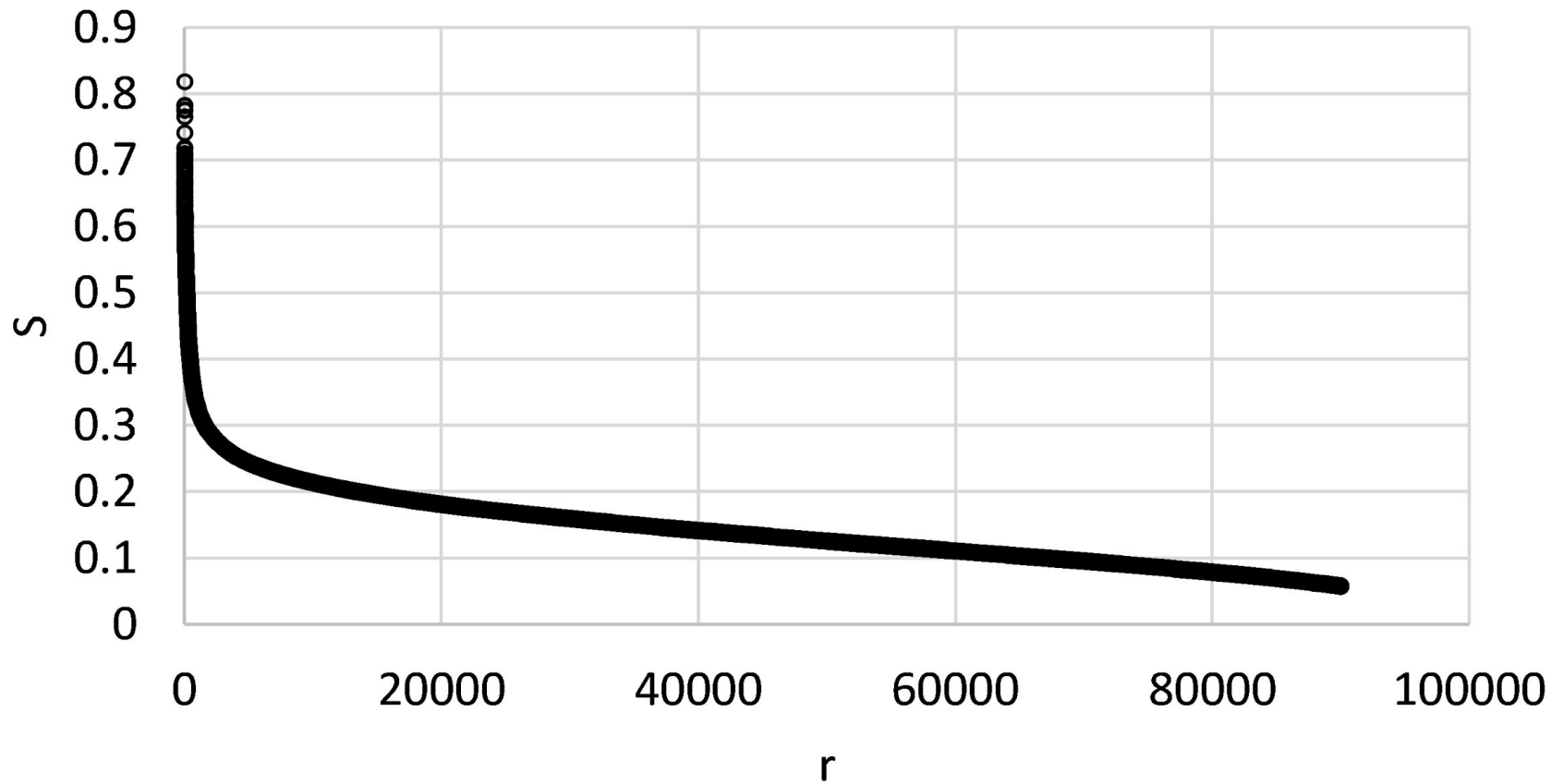
$S$  = the similarity of the lemma

For example, CCS of the lemma “atom” [atom] is 0.55 while CCS of the lemma “protože” [because] is 0.27.

atom [atom]



# protože [because]



# Data

SYNv4 (Czech National Corpus)

Only journalistic texts.

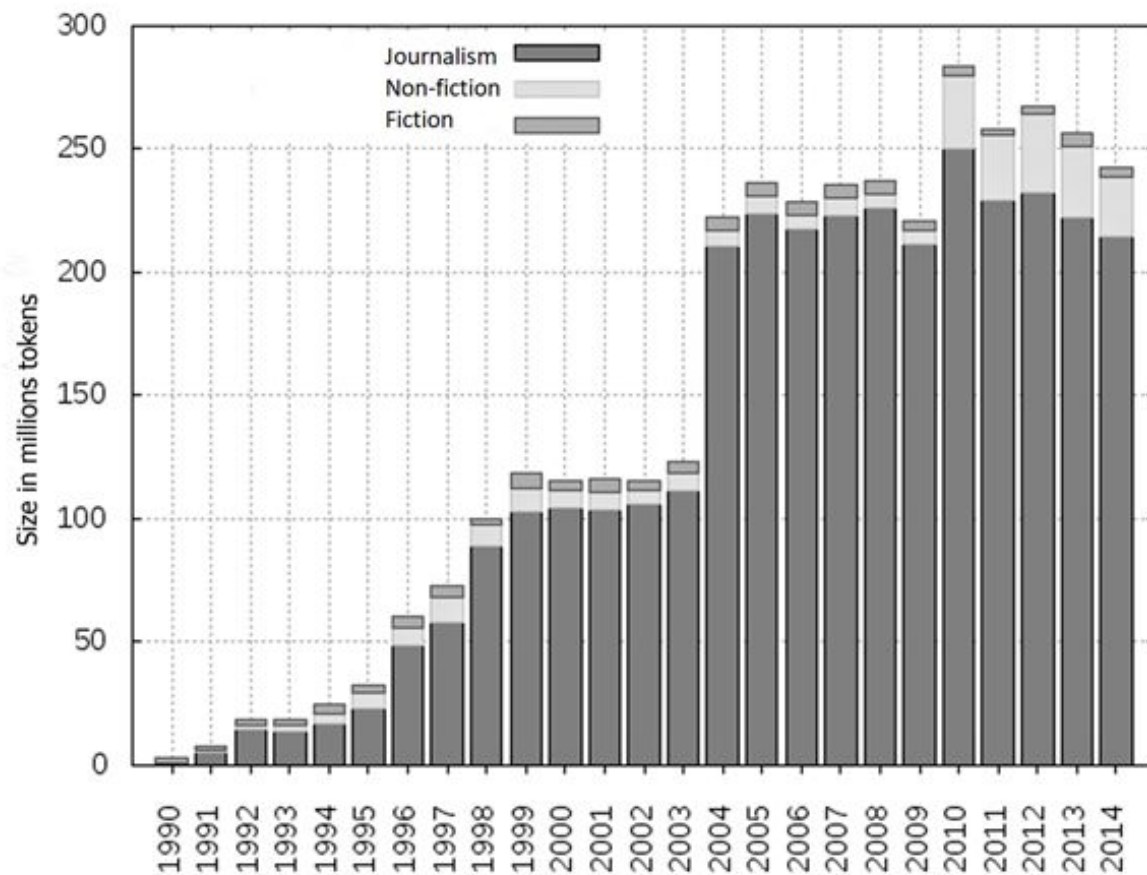
More than 3 billion tokens (3,045,389,630) and more than one hundred thousand types (102,707).

Lemmas are the basic units of this research.

Divided into 19 subcorpora that each represents one year.

Years 1990-1996 are merged because of insufficient amount of data for each year.

# Composition of the corpus SYN version 4



# Diachronic Analysis

The goal is:

- a) to find out whether CCS can detect semantic changes of lemmas from a diachronic point of view,
- b) to observe the relation of CCS to frequencies in the corpora (AF, i.p.m.).

# i.p.m. (instances per million)

A relative frequency measure.

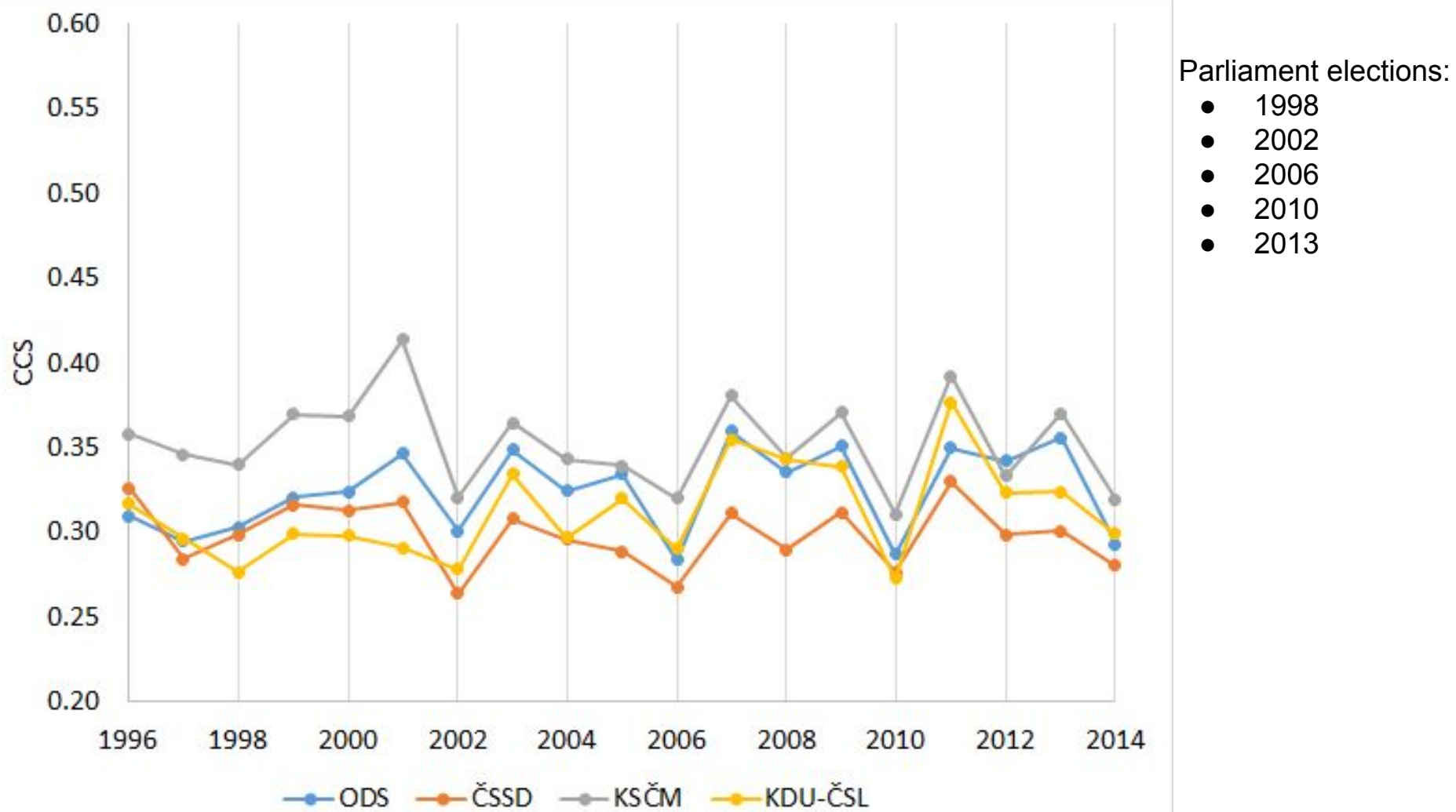
The average number of occurrences of the unit in a hypothetical corpus with the size of 1 million words.

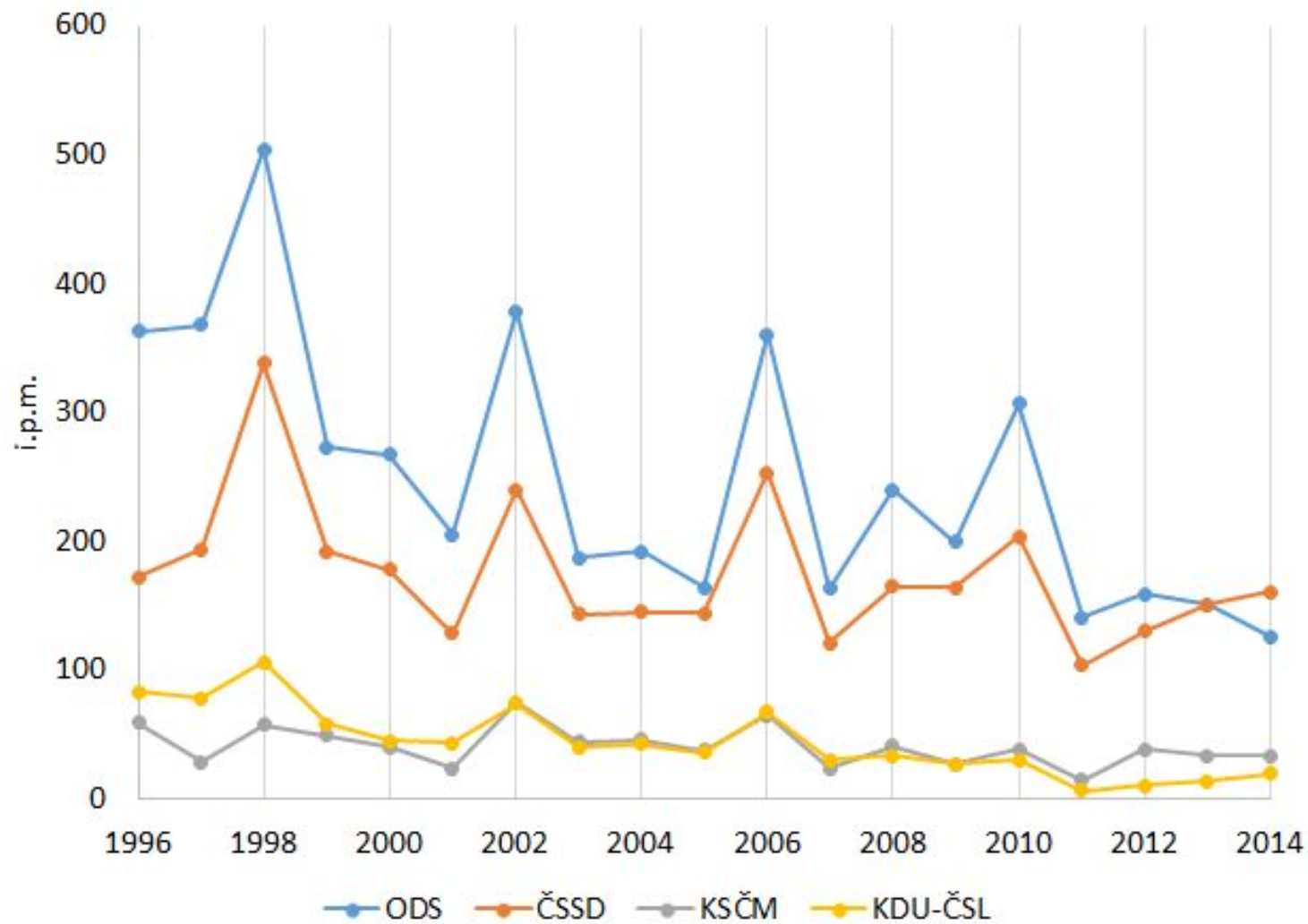
# Analyzed lemmas

We selected several words from various fields where we intuitively expect some semantic changes.

For instance:

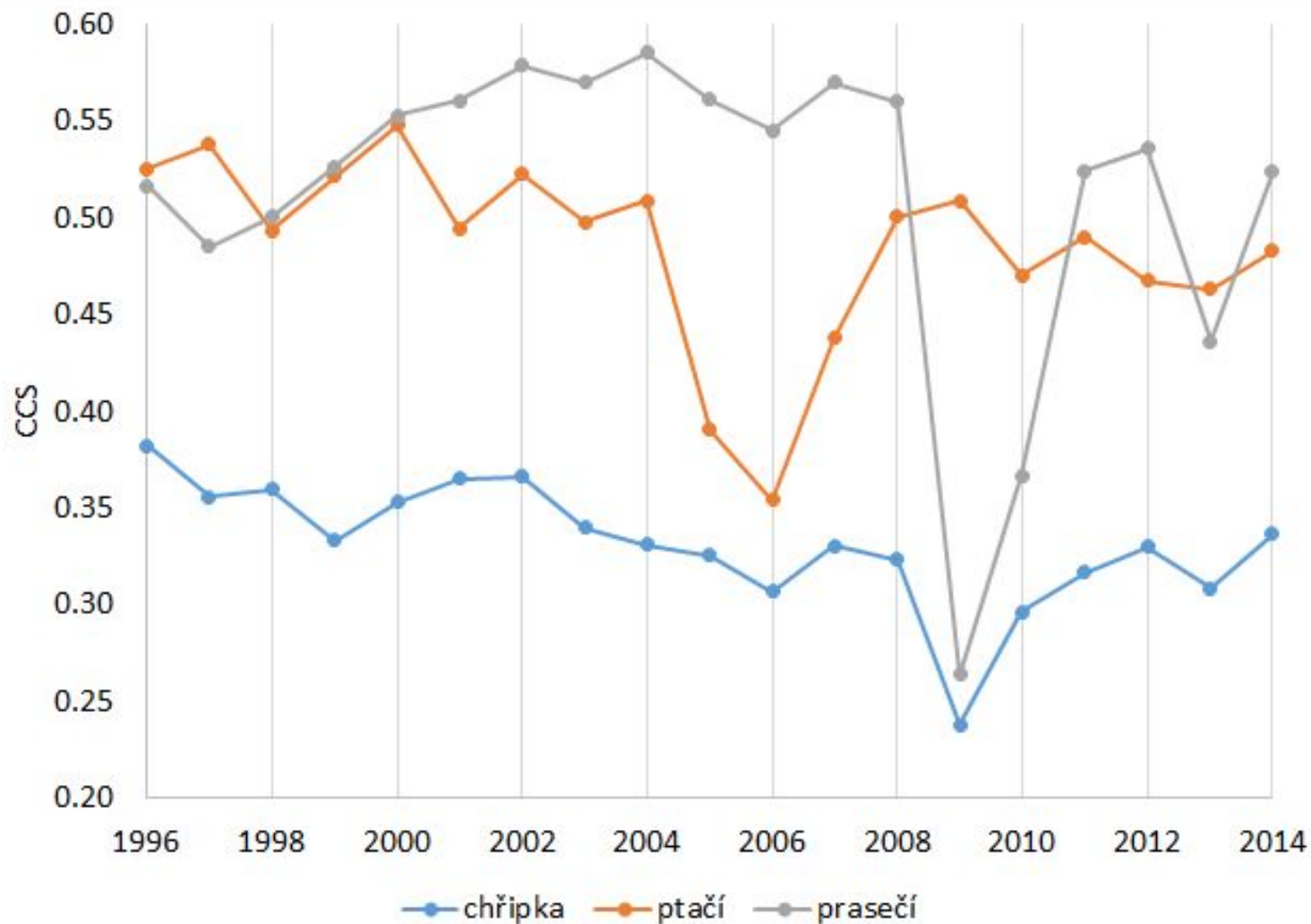
- Political discourse (volby, komunistický, EU, ....)
- Neologisms (web, chat, dealer, show, ...)
- Pople (Zeman, Klaus, Jágr,...)
- Diseases (chřipka, AIDS, alzheimer, ...)
- Countries (Rusko, Čína, USA, Korea, ...)





Parliament elections:

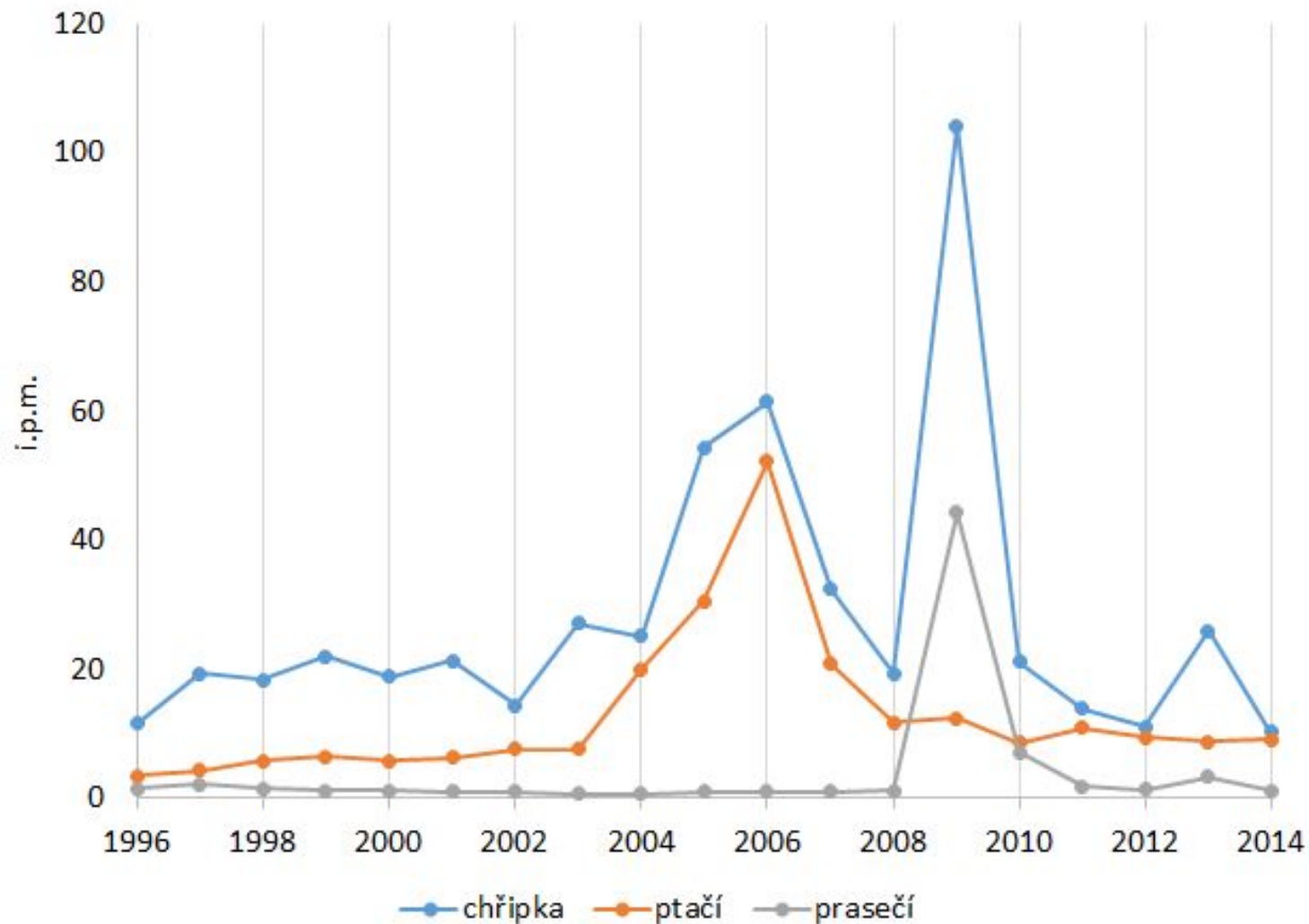
- 1998
- 2002
- 2006
- 2010
- 2013



chřipka = a flu  
(noun)

ptačí = bird  
(adjective)

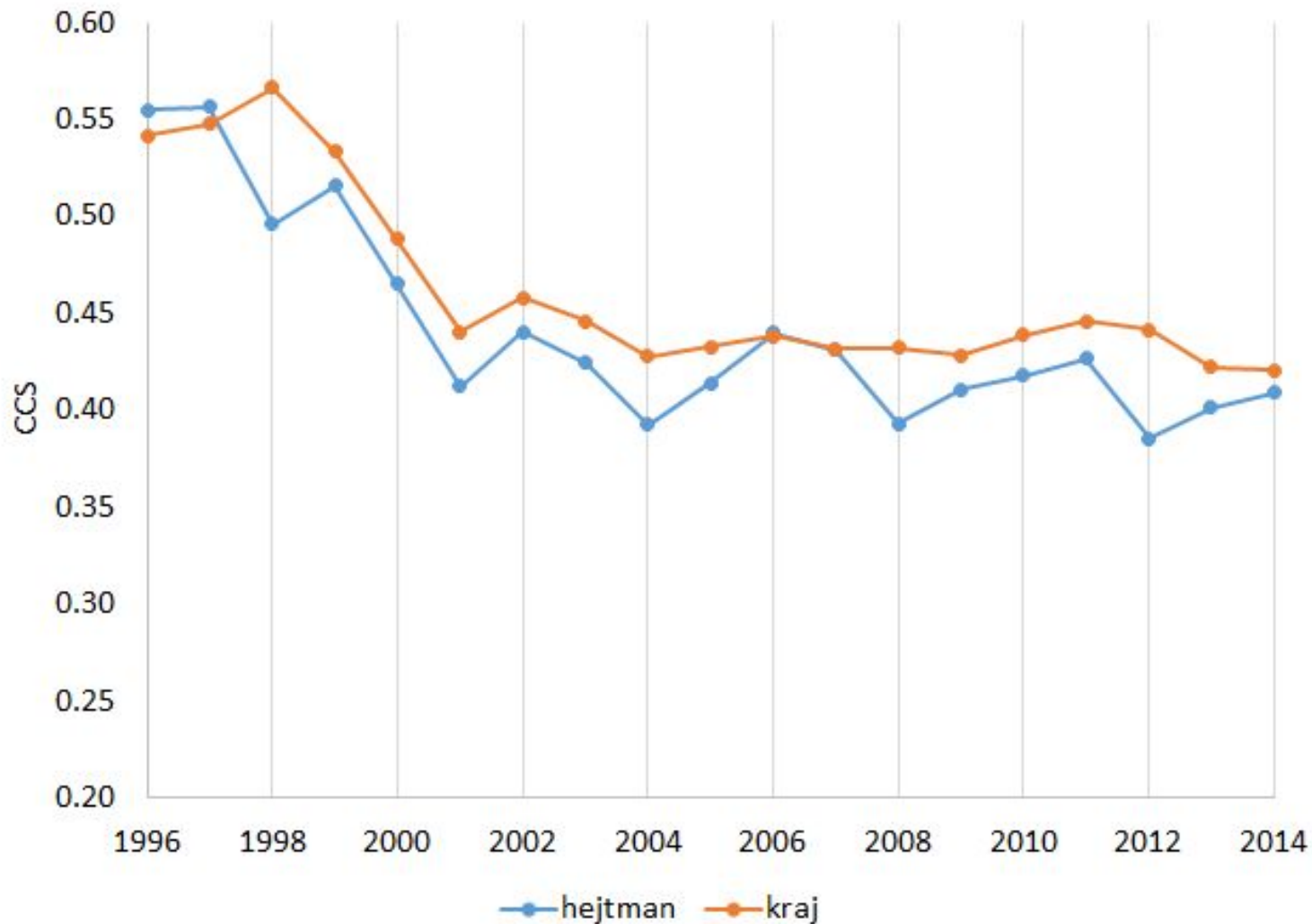
prasečí = swine  
(adjective)



chřipka = a flu  
(noun)

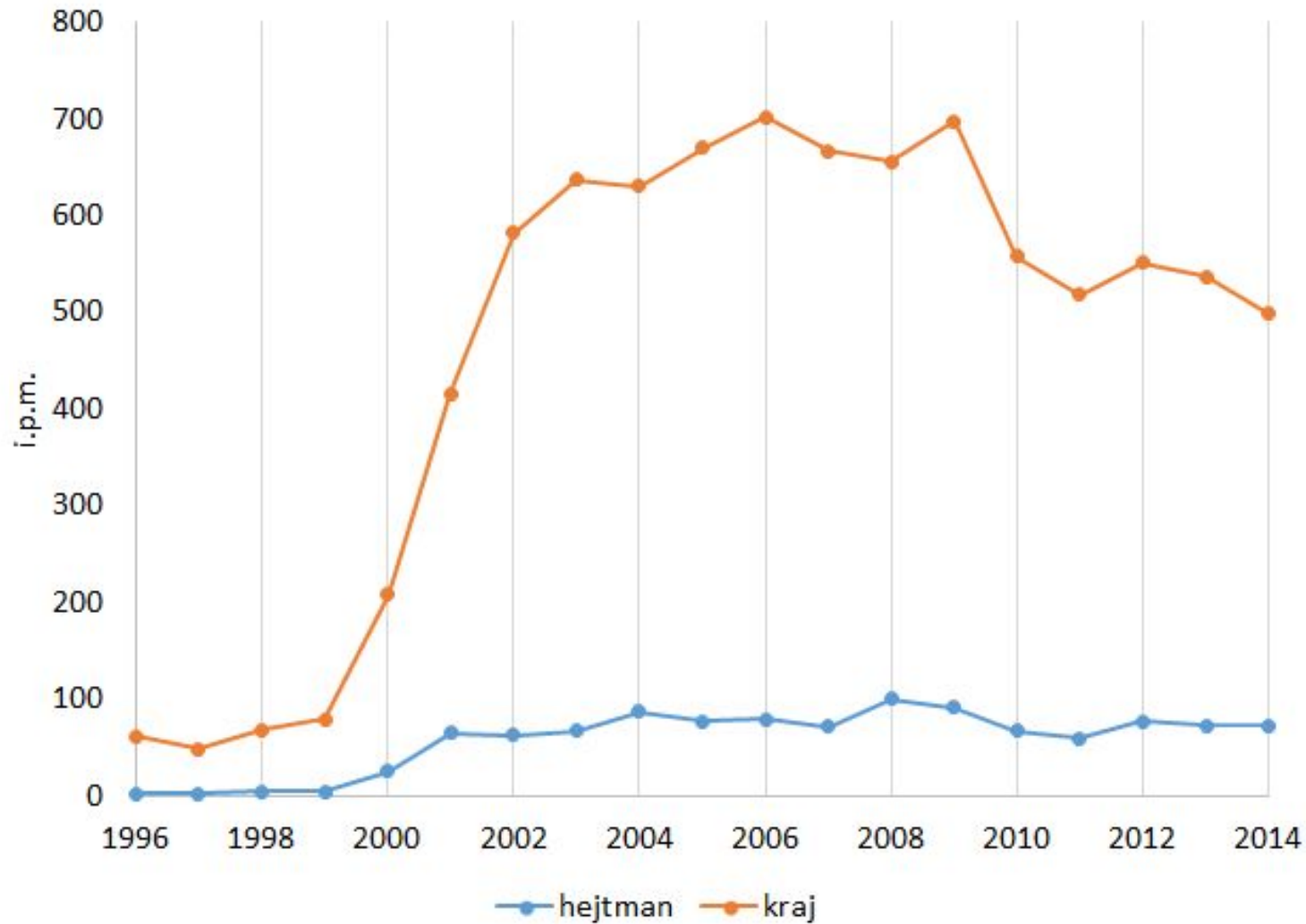
ptačí = bird  
(adjective)

prasečí = swine  
(adjective)



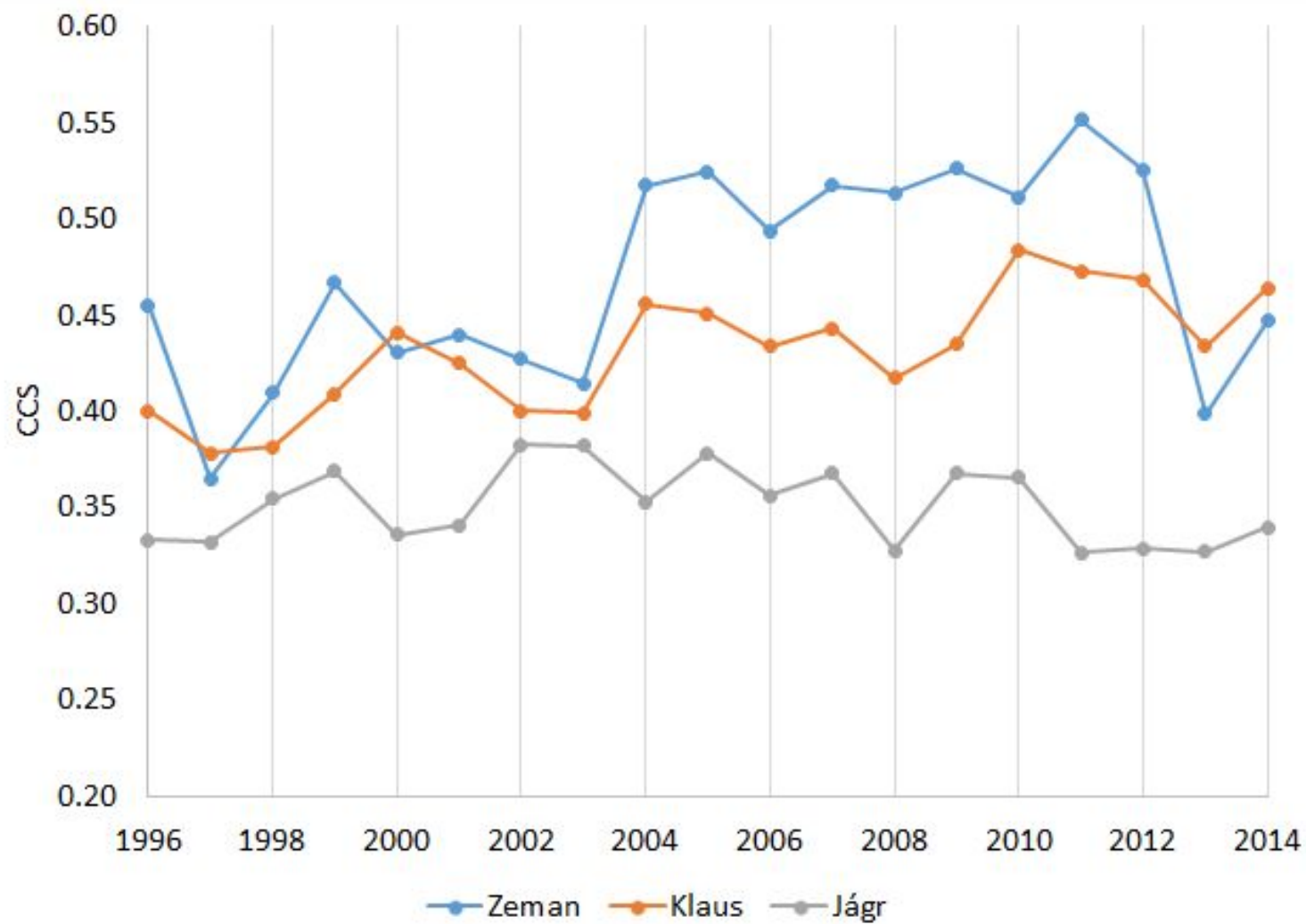
hejtman = a head  
of kraj

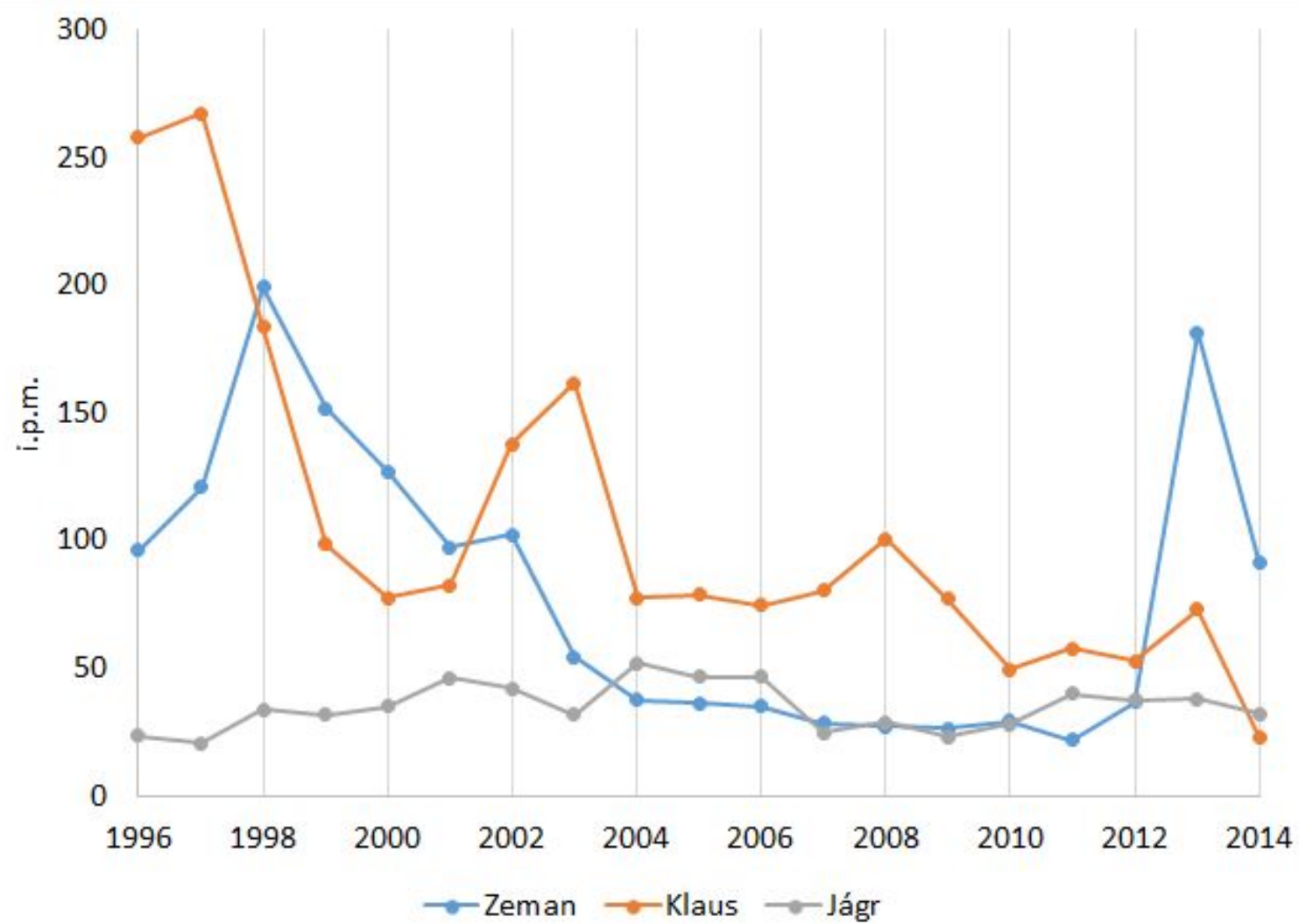
kraj = a region, a  
self-governing unit  
since 2000

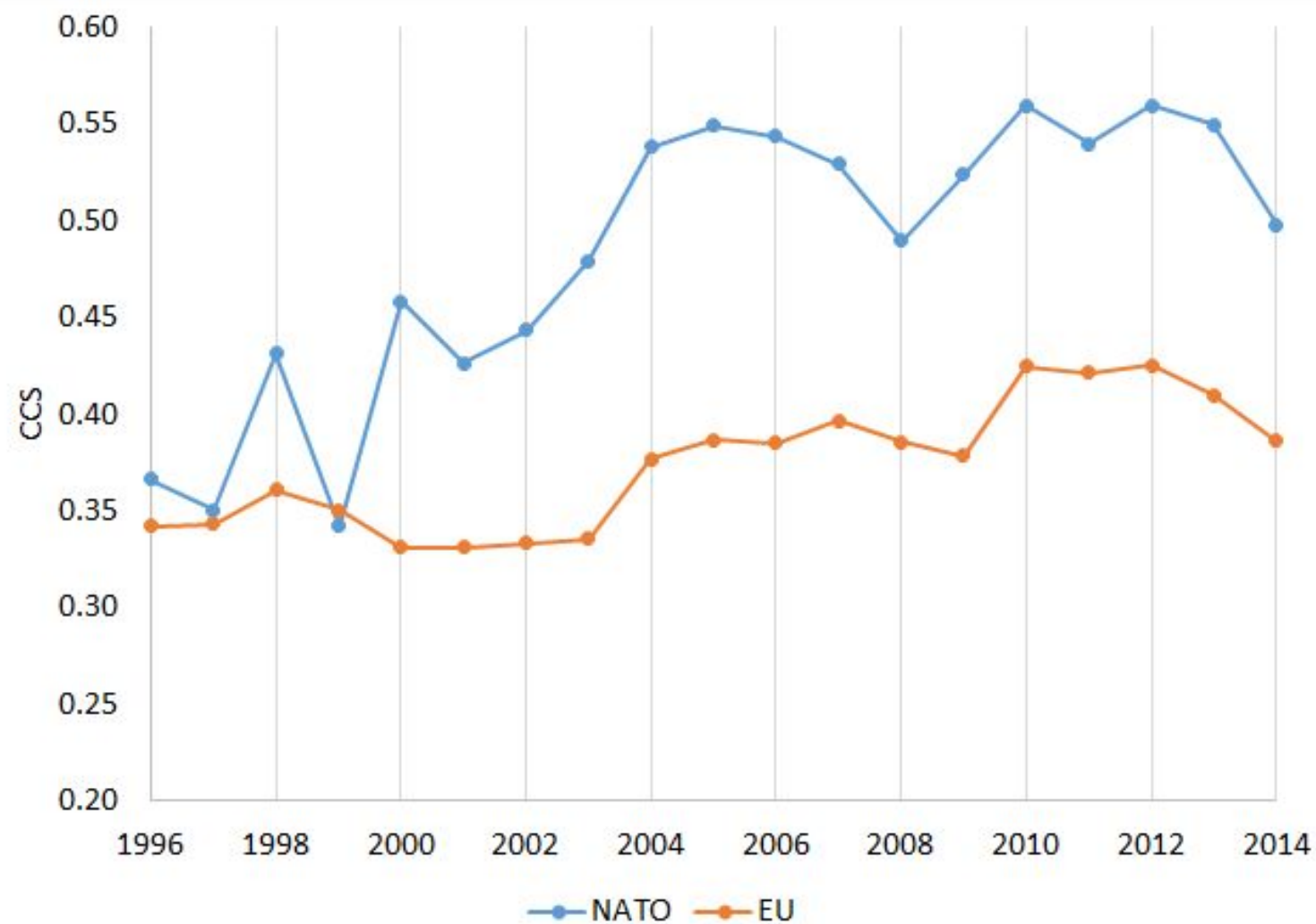


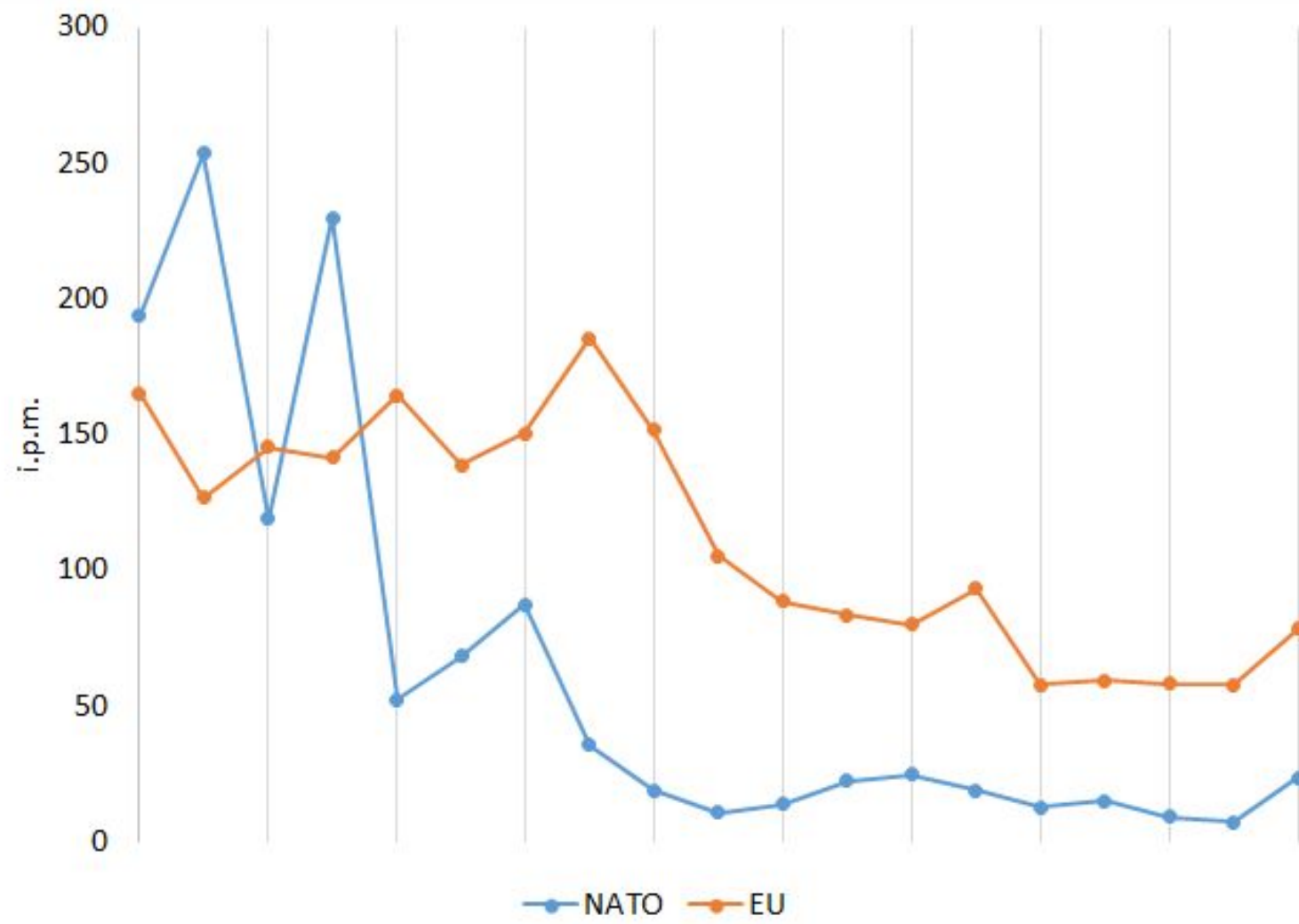
hejtman = a head of kraj

kraj = a region, a self-governing unit since 2000





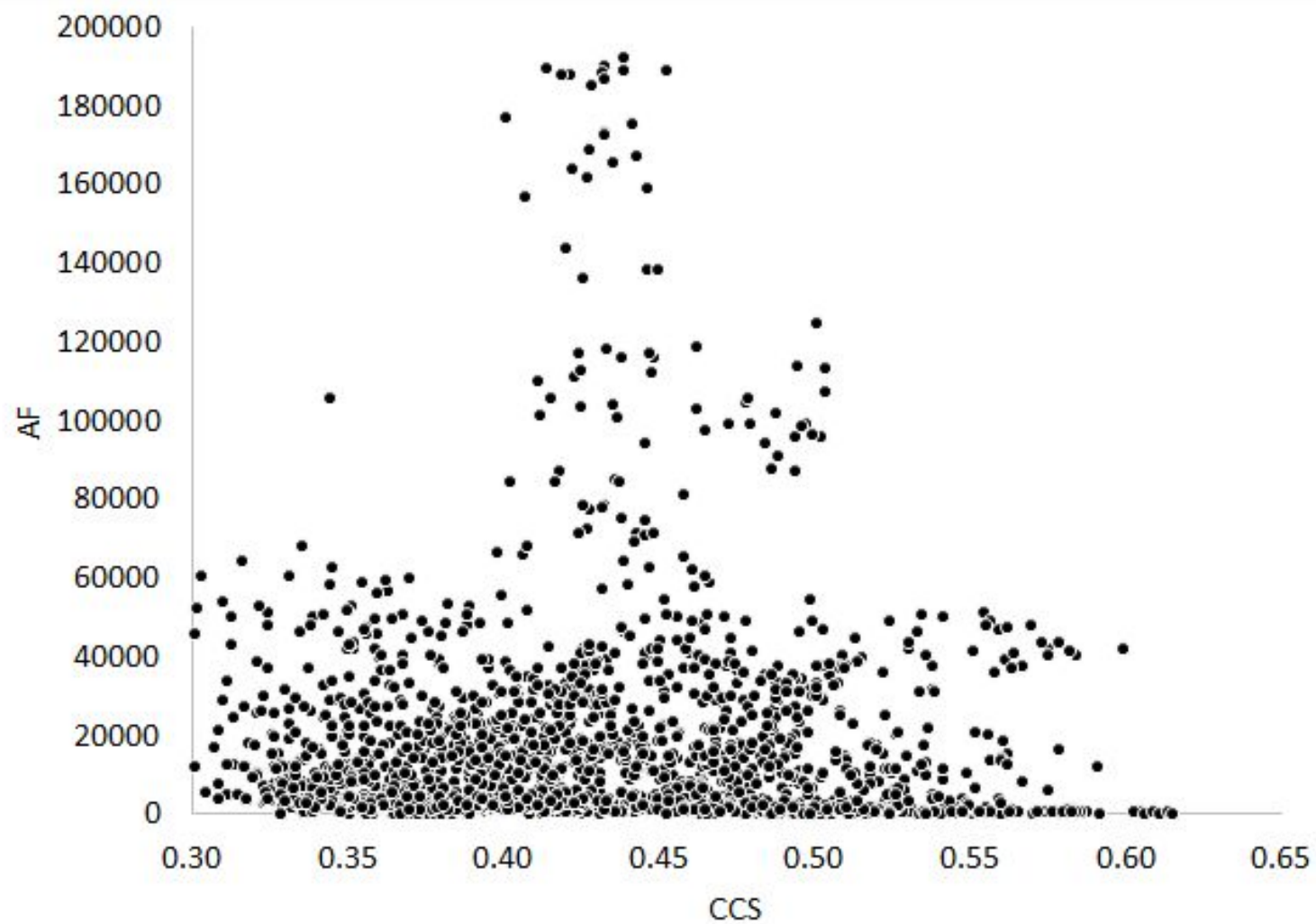


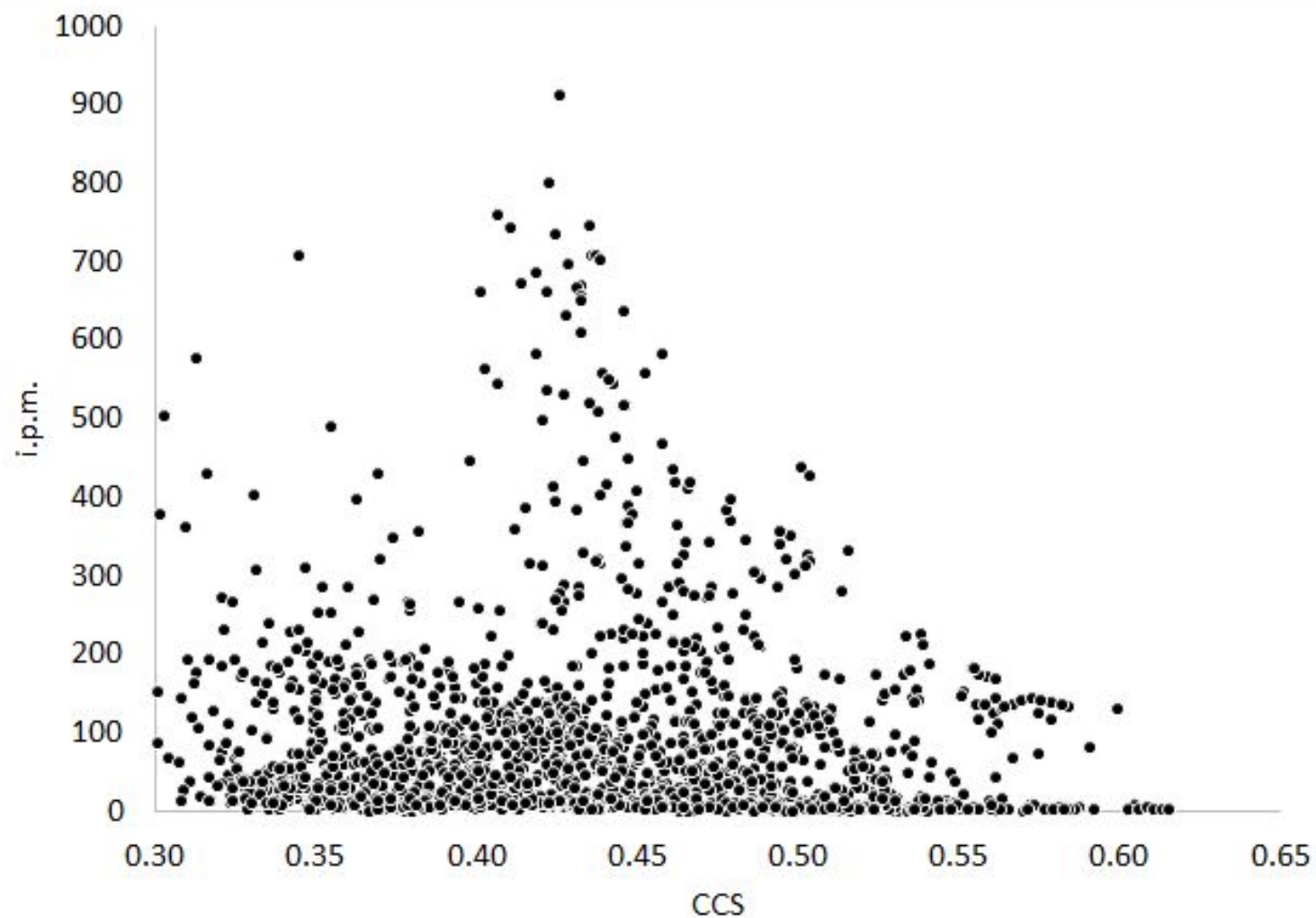


# The relation of CCS to frequencies in the corpopa

AF

i.p.m.





# Conclusion & Discussion

Closest Context Specificity of Lemma (CCS) expresses some kind of semantic feature of lemmas.

CCS can detect semantic changes of lemmas from diachronic point of view.

CCS is not correlated with frequencies of lemmas in a corpus.

The most frequent lemmas tend to reach mean CCS values.

# 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.

Kubát, M., Hůla, J., Číž, D., Pelegrinová, K., Chen, X., Čech, R. Context Specificity of Function and Content word. Conference presentation, QUALICO 2018, Wroclaw.

<https://sgs01-uvafm-2018.webnode.cz/>

<https://sgs02-uvafm-2017.webnode.cz/>

Thank you for your attention!