Application of Neural Networks in Diachronic and Synchronic Semantic Analysis of Texts

Radek Čech, Miroslav Kubát,
Jan Hůla, Xinying Chen

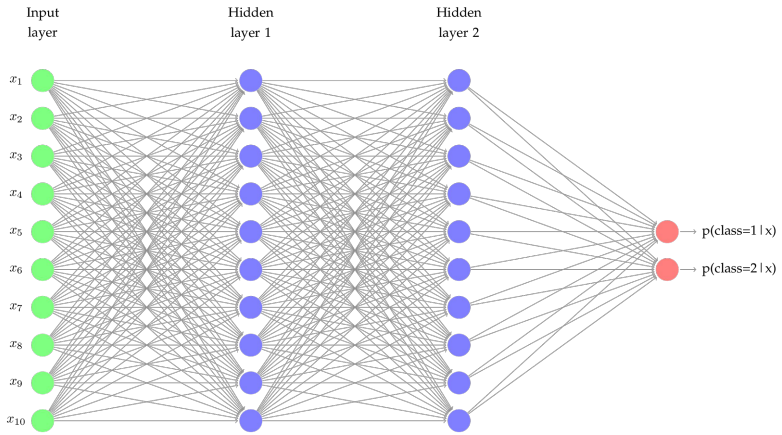
The research is supported by the grant of the University of Ostrava SGS02/UVAFM/2017

Outline

1. Neural Networks
2. Word Embeddings - Word2vec technique
3. Data
4. Context Specificity of Lemma
5. Preliminary Political Discourse Analysis
6. Questions & Comments

Neural networks

- Finding useful representations of data
- Input representations are usually not useful for the task at hand
- Representing words with individual characters is not useful for finding out whether two words have the same meaning
- Representing images with pixels is not useful for finding out what is in the image
- Parameters are found by optimization



Neural networks in linguistics

- In linguistics neural networks are usually used for finding representations which reflect the meaning of words
- Tomáš Mikolov - Efficient Estimation of Word Representations in Vector Space, 2013
- Distributional semantics (1954, Harris, Firth)
- “words that are used and occur in the same contexts tend to have similar meanings”
- "a word is characterized by the company it keeps" - Firth

Word embeddings

- Models the meaning of a word by tracking a contexts where the word appears
- Simplification: count the frequencies of words appearing next to it.

Example (prime minister)

- A prime minister is the head of a cabinet and the leader of the ministers in the executive branch of government
- Collect every instance of the word “prime minister” in the corpus
- Count how many times it appears next to the word “cabinet”, “parliament”, “algebra” and all other words in the corpus to get vector of co-ocurrence statistics
- Words appearing in similar contexts will have similar vectors

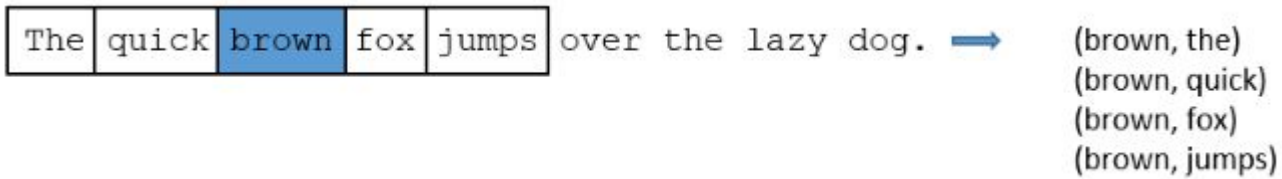
Problems

- **Problem**: vectors are too long
- **Solution**: find best low dimensional approximation (25-1000) with decomposition algorithm (SVD,etc.)
- **Problem**: For corpuses with large vocabularies the decomposition becomes impractical
- **Problem**: It has to be recomputed when adding new texts to the corpus
- **Solution**: Learn the vectors with neural network in online fashion (Word2Vec)

Word2vec

- Learn to predict the words in the window centered around the word
- By learning to predict neighbour words it captures the co-ocurrence statistics
- Learning is done by maximizing the objective function:

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



Word2vec ...

$$p(o|c) = \frac{\exp\left(u_o^T v_c\right)}{\sum_{w=1}^{W} \exp\left(u_w^T v_c\right)}$$

- Parametrization of the probability
- 2 vectors for every word
- By maximizing this probability w.r.t. the vector v_c the vector u_o will get closer to v_c during the optimization

After the optimization

- 1 vector for each word
- Vector arithmetic: [android]-[google]+[apple] ≈ [iphone]
- Useful representation for analogical reasoning and measuring the similarities

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).

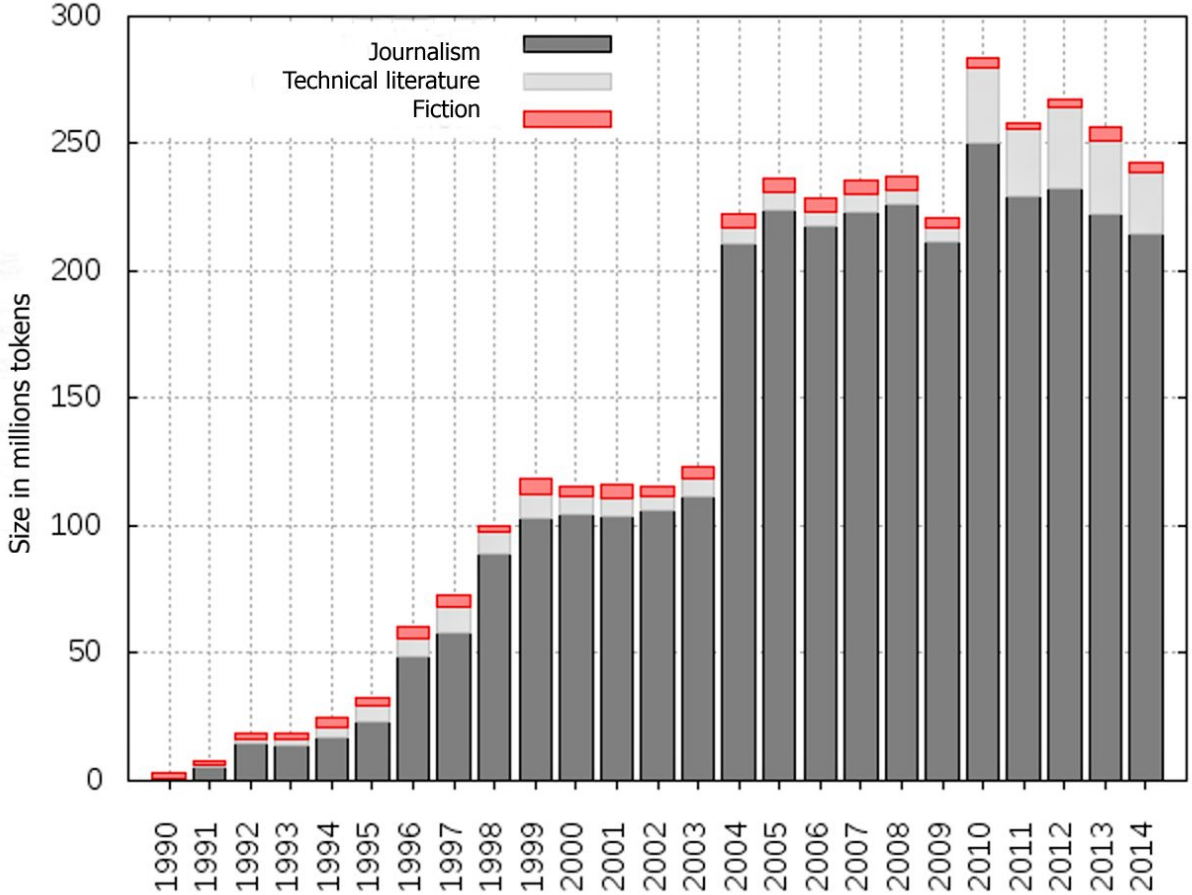
Lemmas are the basic units of this research.

All lemmas with frequency less than 70 were omitted ($f \leq 69$).

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



Context specificity of lemma (CSL)

CSL measures how unique is the context in which the lemma appears in the corpus. if the lemma occurs in many different contexts, it will have low CSL.

The context in which the lemma appears is captured with a vector of co-occurrence statistics which is assigned to every lemma.

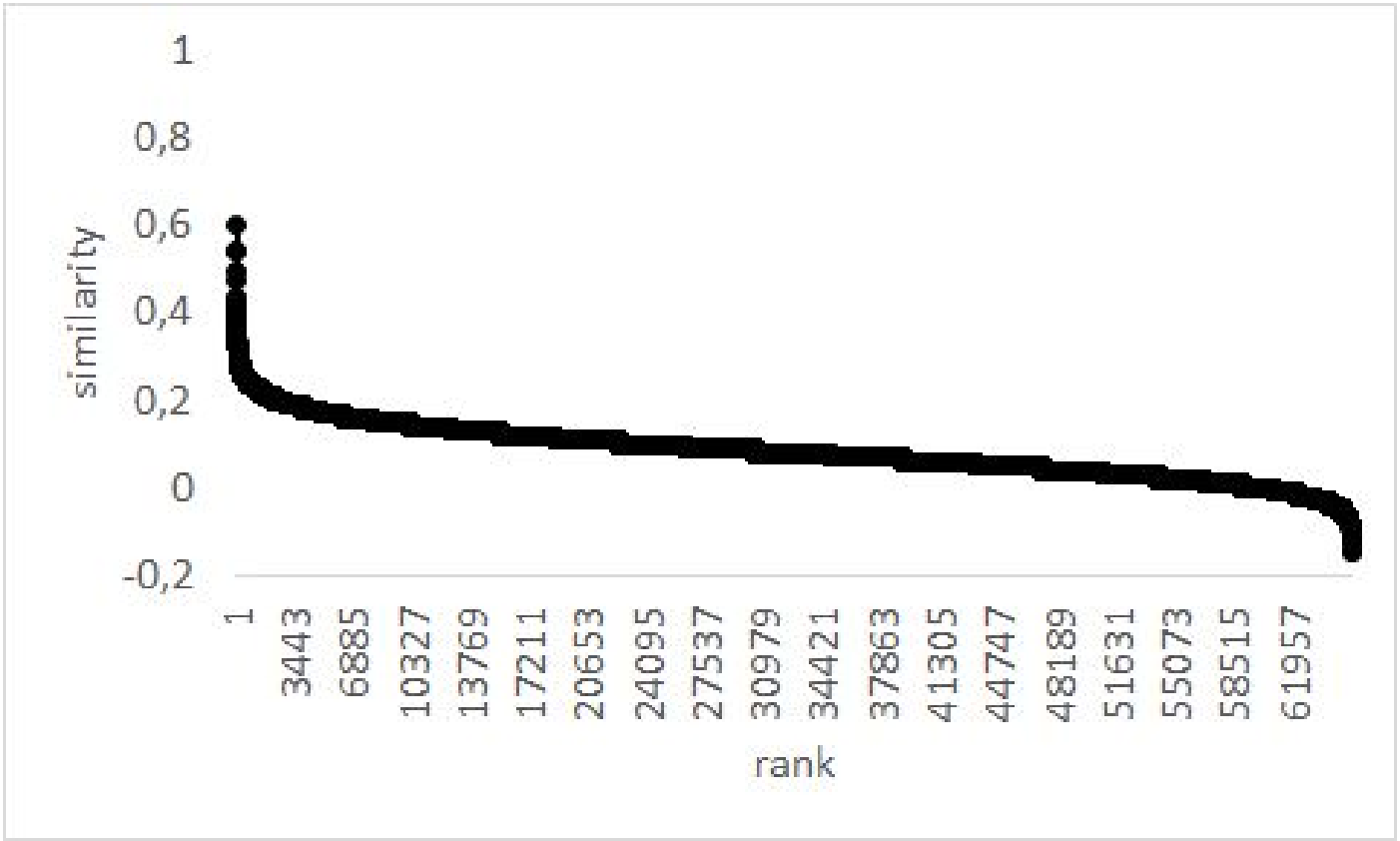
We can compute its similarity to all other lemmas. Statistics of these similarities (e.g., a mean value) can be used for characterizing the CSL. The lower the mean of similarities, the higher the CSL.

For example, CLS of the lemma “atom” (means atom) is 0.0829, while CLS of the lemma “nebo” (means “or”) is 0.1273 in the subcorpus 2013

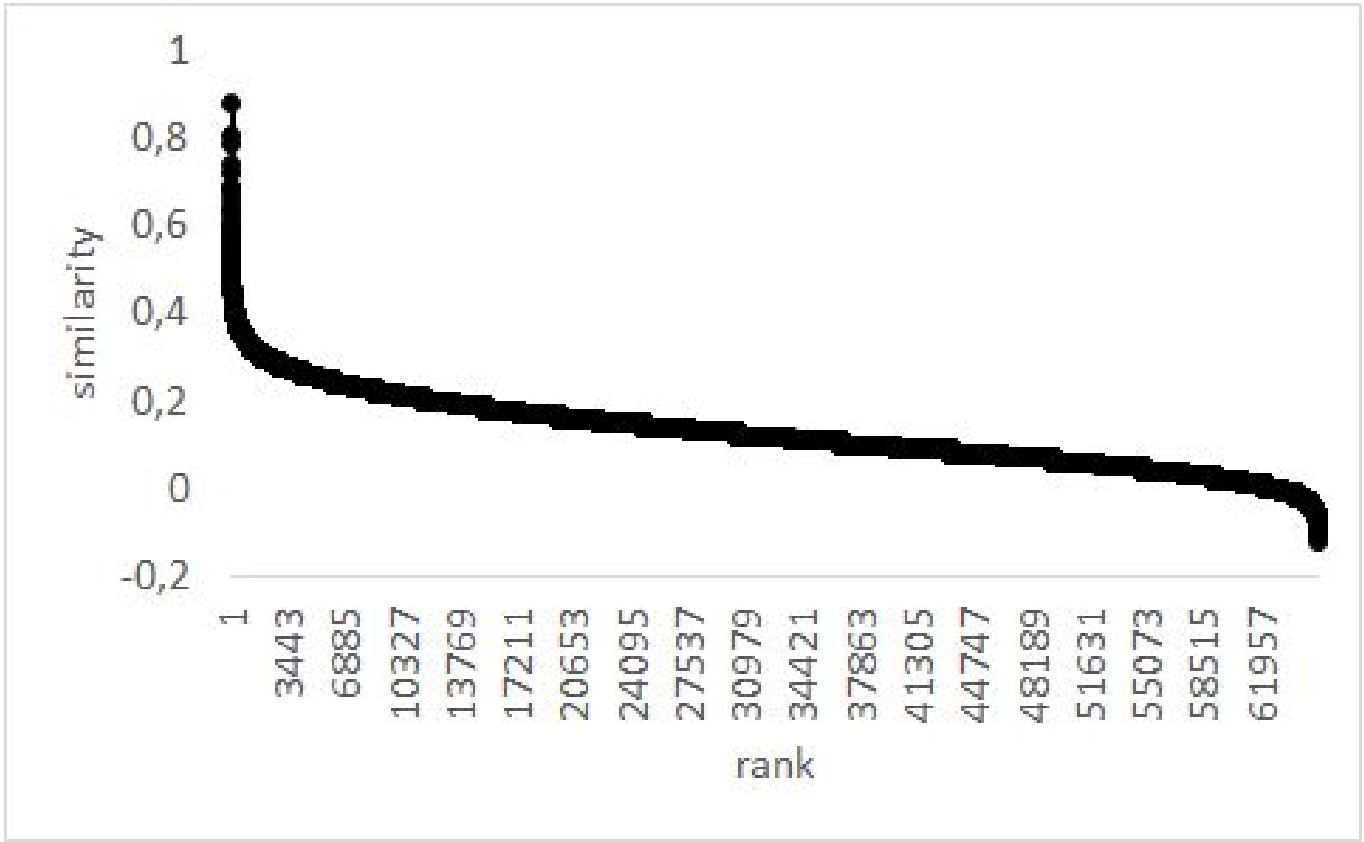
Five the most similar lemmas of the lemma “atom” and “nebo”. *S* assigns the value of similarity from the subcorpus 2013

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

The distribution of distances for the lemma “atom” in the subcorpus 2013.



The distribution of distances for the lemma “nebo” in the subcorpus 2013.



Full context specificity (FCS)

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

S = the similarity of the lemma

n = the number of lemmas in the corpus

Closest context specificity (CCS)

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

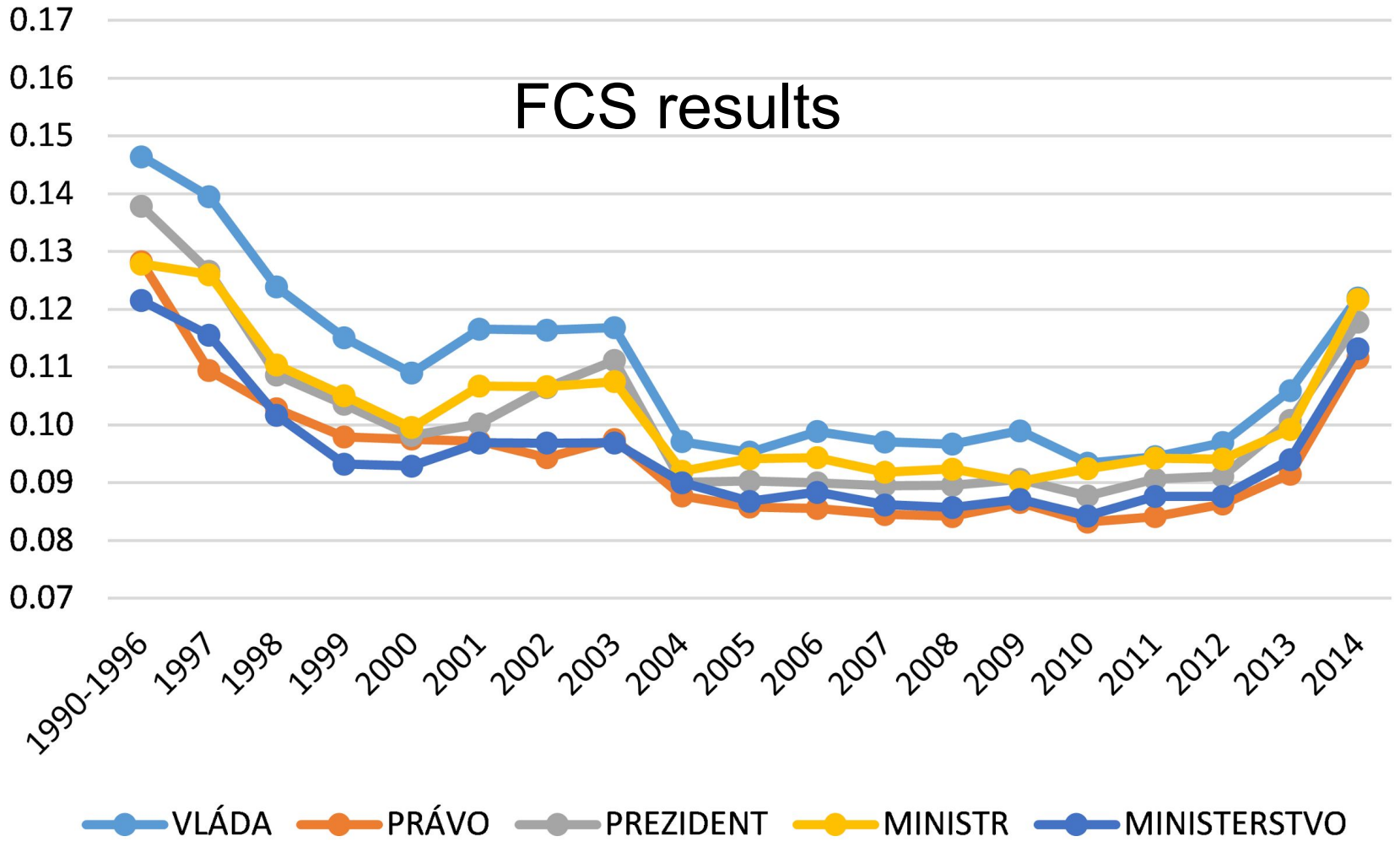
S = the similarity of the lemma

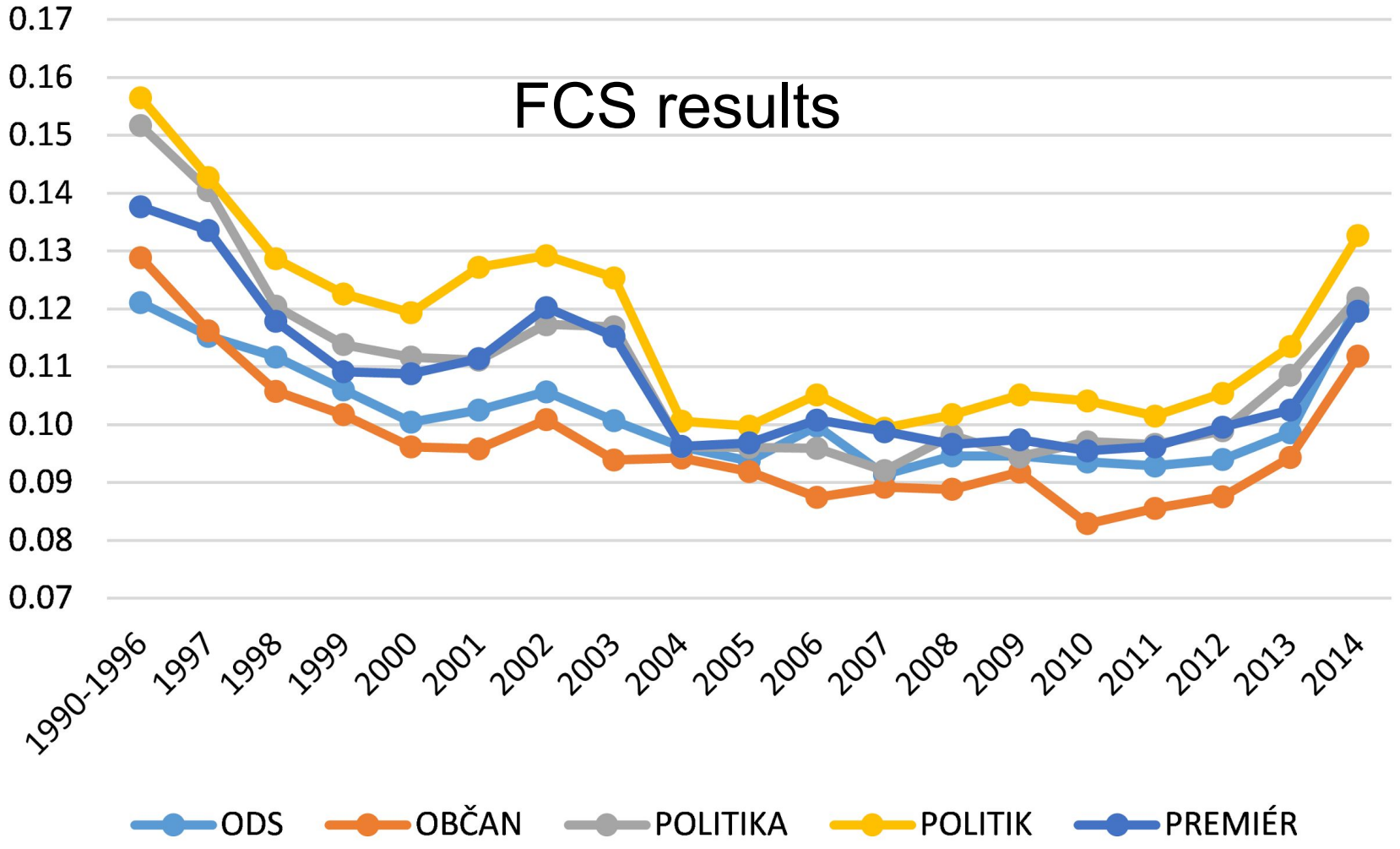
$$FCS_{atom} = \frac{5421.36}{65382} = 0.0829$$

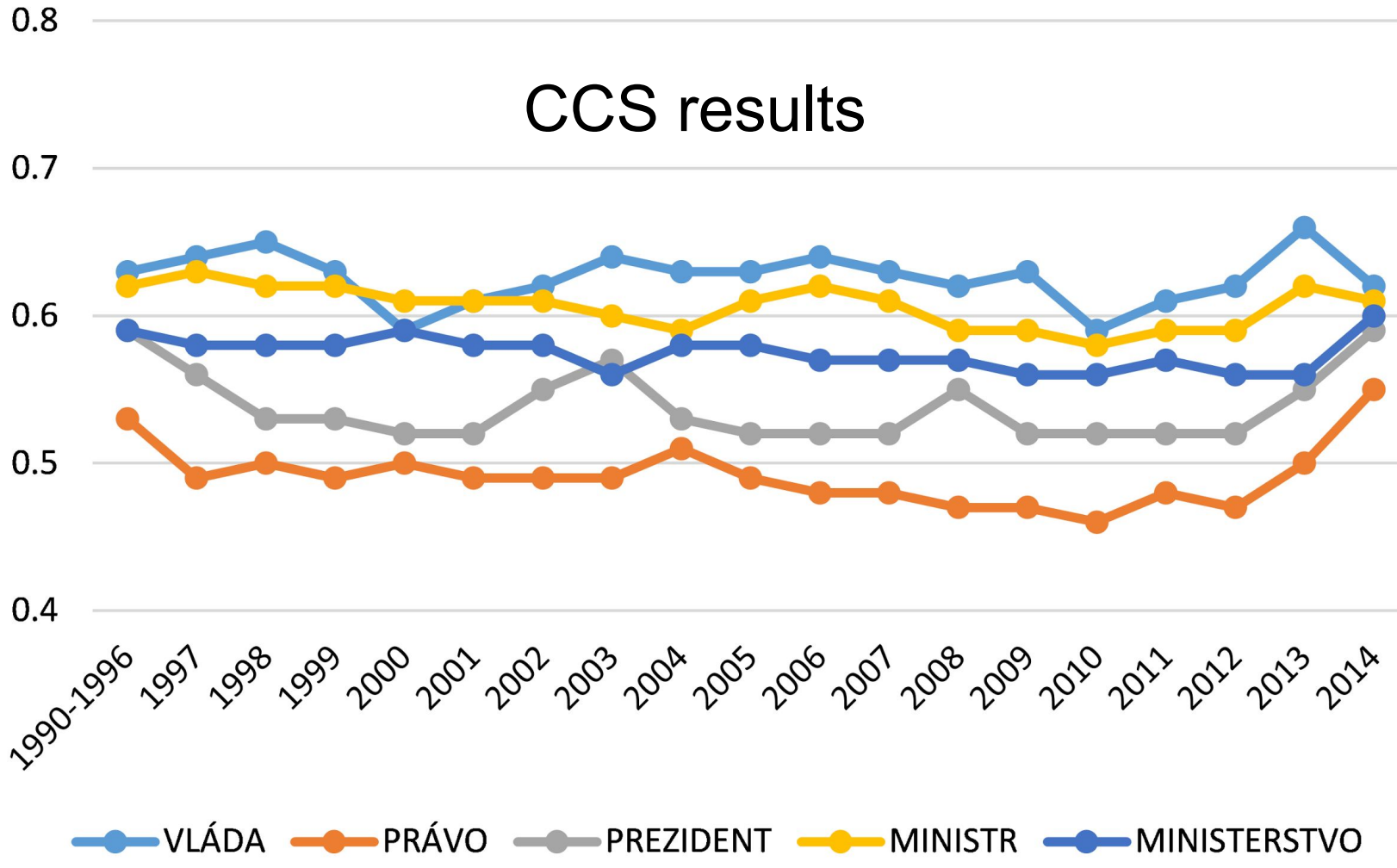
$$FCS_{nebo} = \frac{8324.27}{65382} = 0.1273$$

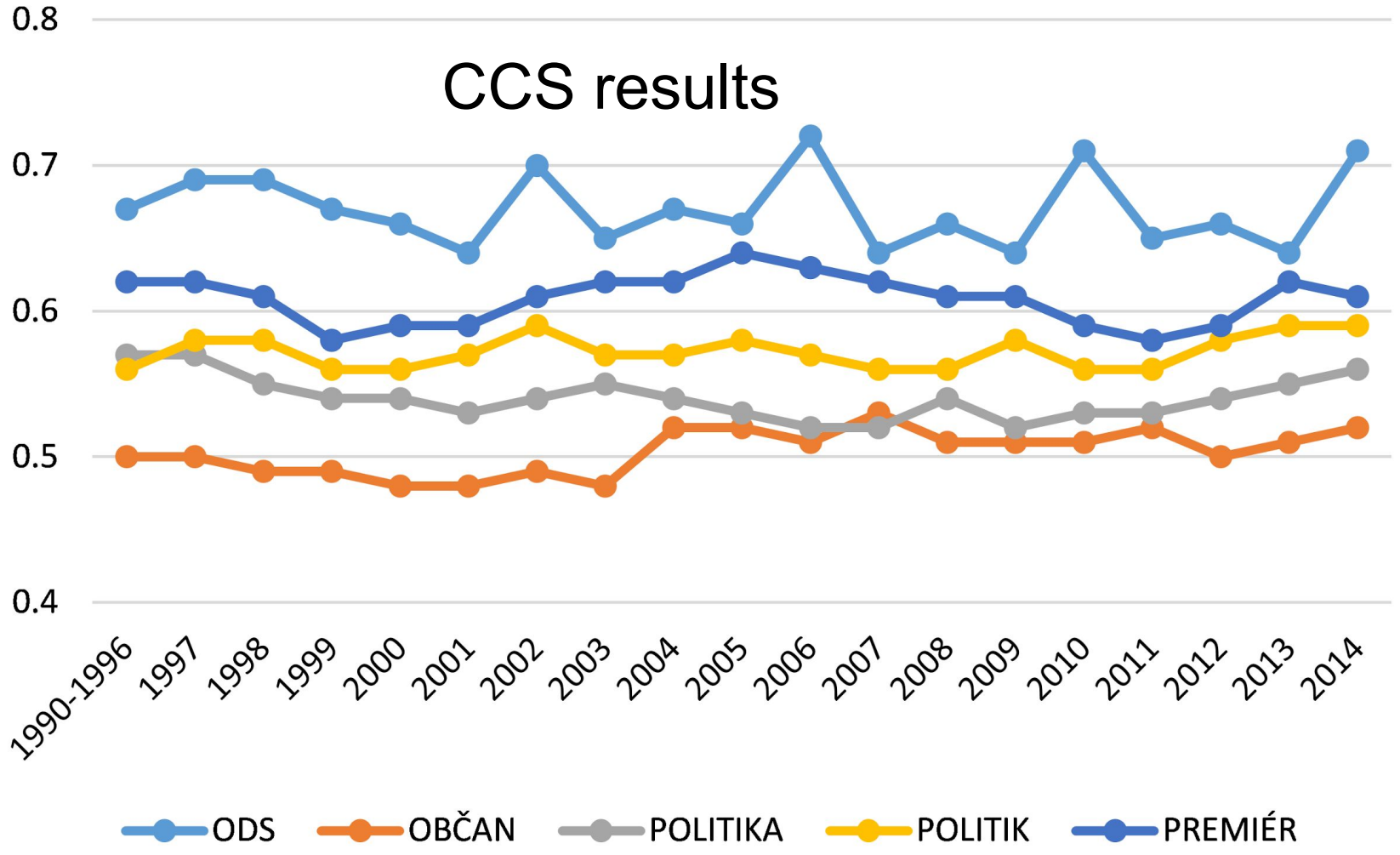
$$CCS_{atom} = \frac{8.79}{20} = 0.4395$$

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

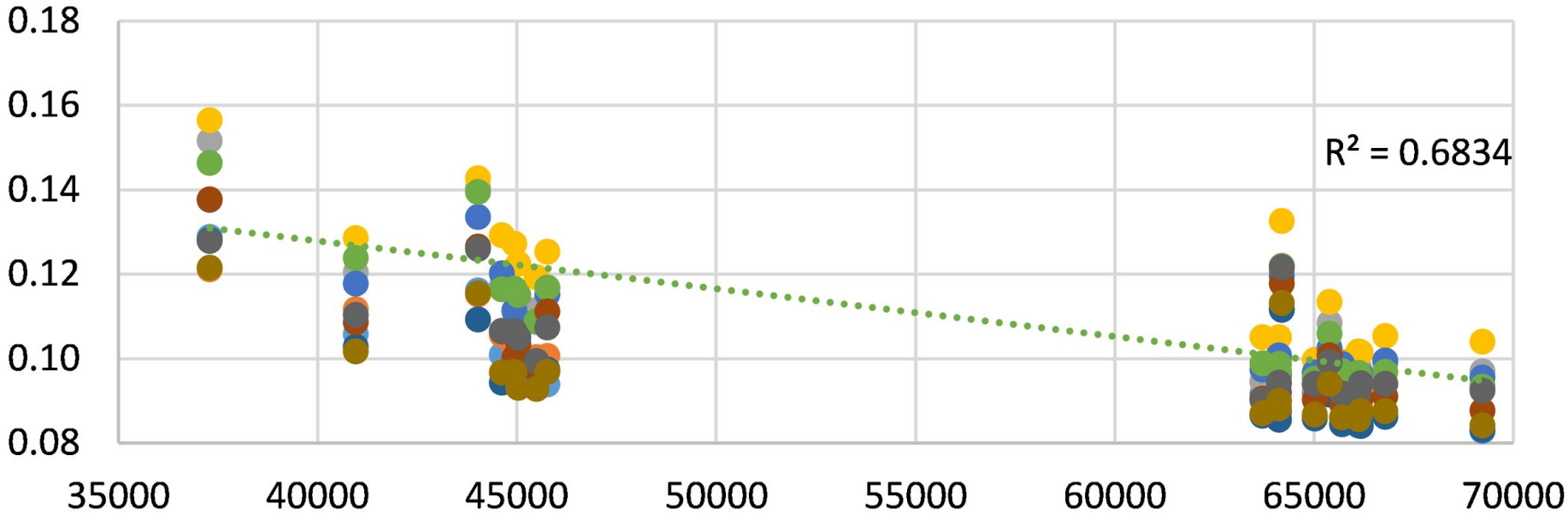

FCS results
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
1990-1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
VLÁDA
PRÁVO
PREZIDENT
MINISTR
MINISTERSTVO


FCS results
0.17
0.16
0.15
0.14
0.13
0.12
0.11
0.10
0.09
0.08
0.07
1990-1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
ODS
OBČAN
POLITIKA
POLITIK
PREMIÉR

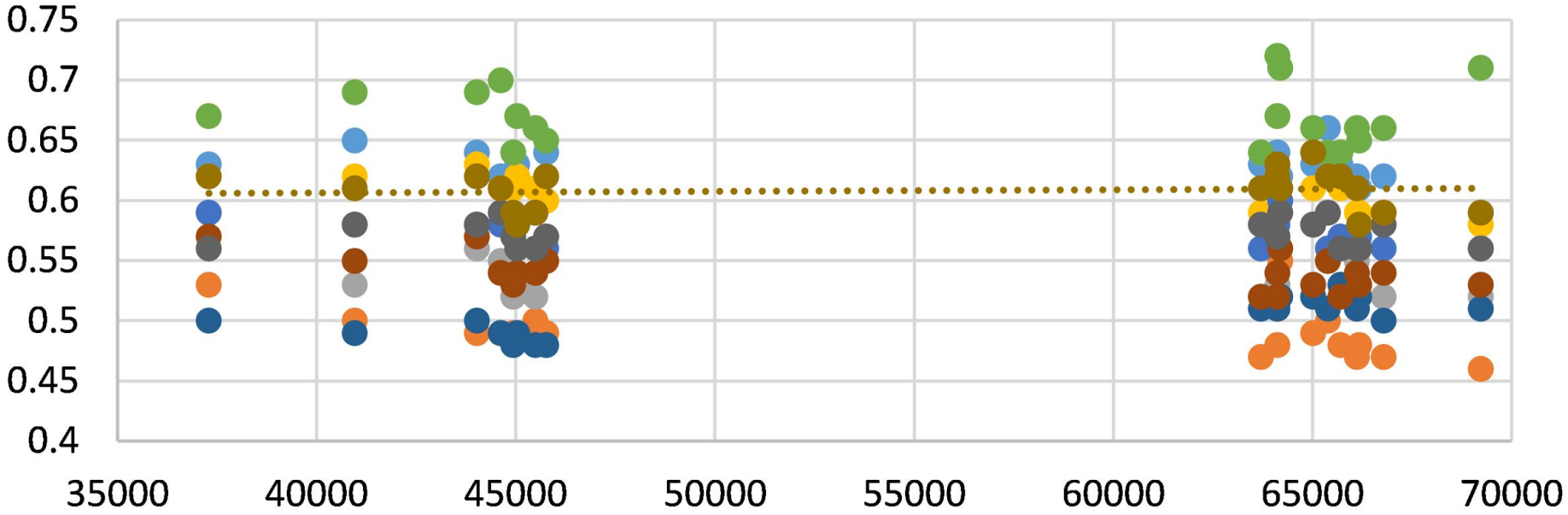

CCS results
0.8
0.7
0.6
0.5
0.4
1990-1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
VLÁDA
PRÁVO
PREZIDENT
MINISTR
MINISTERSTVO


CCS results
0.8
0.7
0.6
0.5
0.4
1990-1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
ODS
OBČAN
POLITIKA
POLITIK
PREMIÉR

NUMBER OF LEMMAS vs FCS



N of lemmas vs. CCS



Conclusion and Discussion

- the concept of lemma specificity can be used for linguistic analysis

Conclusion and Discussion

- the concept of lemma specificity can be used for linguistic analysis
- be careful!!! *FCS*…

Conclusion and Discussion

- the concept of lemma specificity can be used for linguistic analysis
- be careful!!! *FCS*…
- application of the approach to other branches of linguistics
  - critical discourse analysis
  - content analysis
  - stylometry
- the neural network = “black box”
  - parameters are not interpretable

Conclusion and Discussion

- the concept of lemma specificity can be used for linguistic analysis
- be careful!!! *FCS*…
- application of the approach to other branches of linguistics
  - critical discourse analysis
  - content analysis
  - stylometry
- the neural network = “black box”
  - parameters are not interpretable
- however
  - “if a method of this kind is used as a starting point for an analysis which has clear linguistic interpretation (such as CLS and its dynamic development) and it brings valuable results, its application pose a challenge for linguistic research”

thank you for your attention

questions and remarks are welcome