Text as data

Petr Ocelík and Lukáš Lehotský

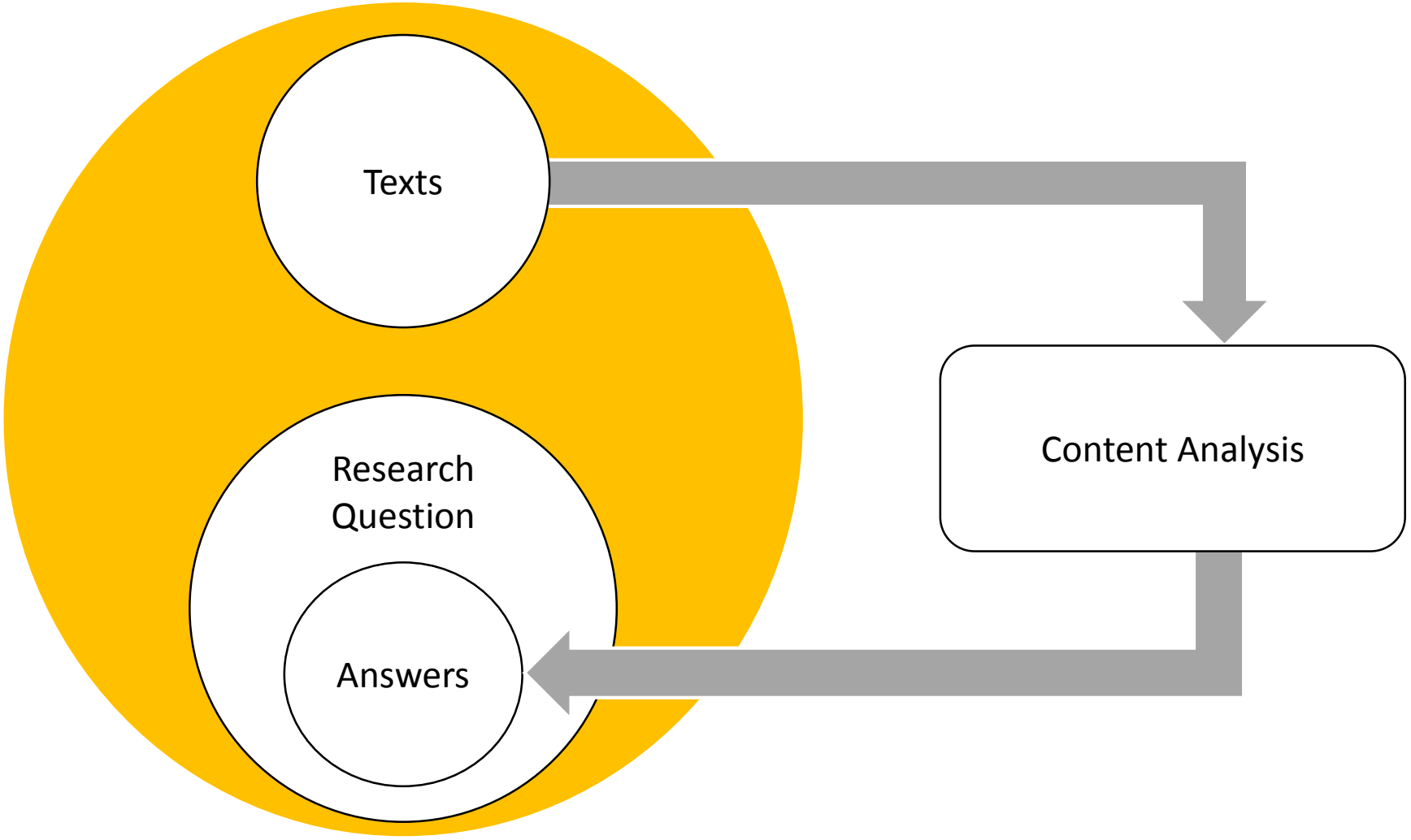
Text analysis as approach

- Content analysis
  - Supervised methods
  - Unsupervised methods
- Discourse network analysis
- Grounded theory
- Qualitative discourse analysis
- …

Content analysis

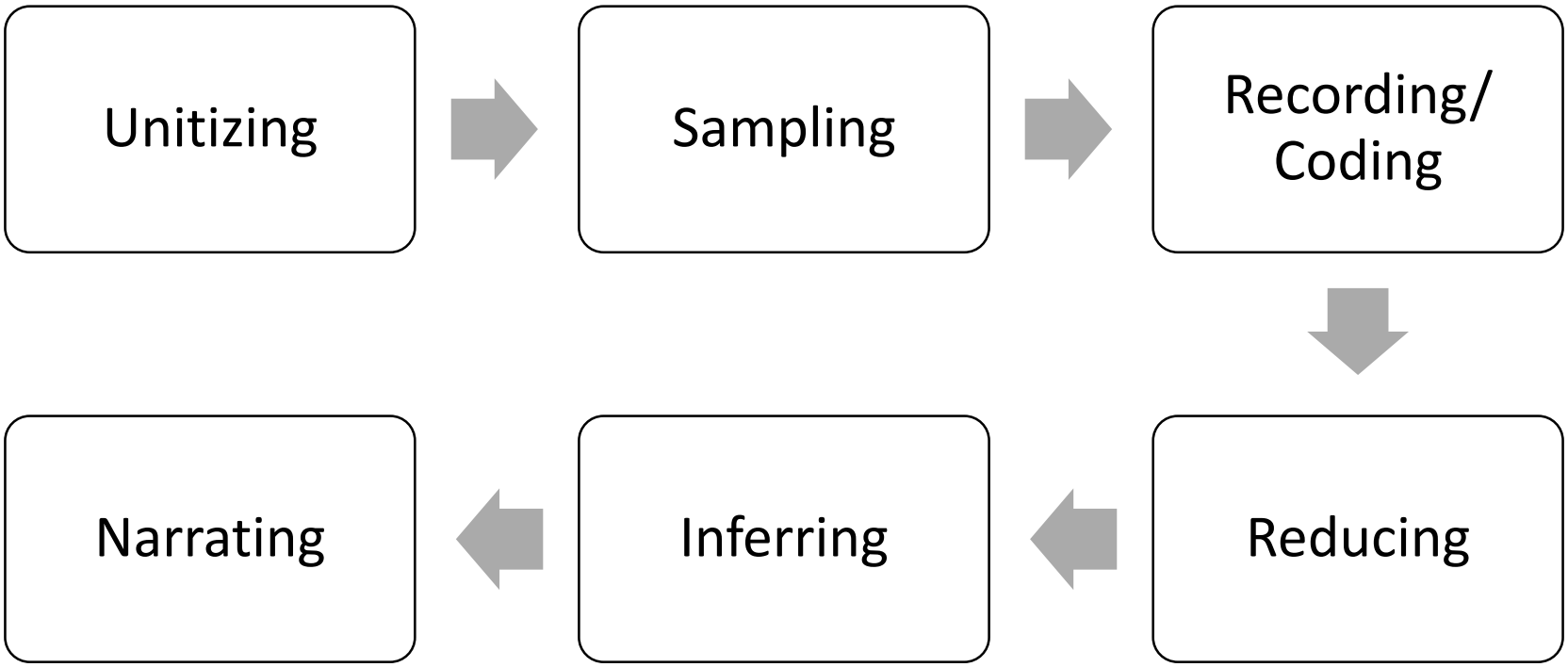
- Question of manifest vs. latent content
  - Berelson (1952) – objective, systematic, and quantitative description of manifested content of communication
  - Many other commit themselves to notion of content being present in the message (Riffe, Lacy & Fico 1998; Gerbner 1985; Shapiro & Markoff 1997)
  - Other criticize this assumption (Krippendorff 2013; Holsti 1969)

Design of CA research



Krippendorff 2013, p. 36

Design of CA research



Krippendorff 2013, p. 86

Unitizing

- Natural units
  - Sentences
  - Quasi-sentences
  - Paragraphs
  - Texts
- Constructed units
  - N words
  - N sentences
  - …

Oživení české ekonomiky v posledních dvou letech akcelerovalo i požadavky našich odborů. Ty na dnešek svolaly manifestační mítink v rámci své aktivity nazvané *Konec levné práce*. Odboráři se snaží přesvědčit veřejnost, že čeští zaměstnanci patří neoprávněně mezi nejhůře placené v rámci celé EU a že za tíženým růstem mezd v posledních dvou letech stojí právě jejich práce. Ani jedno z toho není pravdou.

Odbory v prvním bodě své argumentace často začínají porovnáním nákladů práce. Náklady práce sestávající se z mezd a platů a nemzdových nákladů (sociální příspěvky zaměstnavatelů, které jsou u nás významné, plus daně mínus dotace) jsou v ČR osmé nejnižší mezi všemi zeměmi EU-28. Podle Eurostatu dosahují za rok 2015 výše 9,9 euro na hodinu, zatímco průměr v celé EU činí 25 euro a např. v Německu to je 32 euro. Neboli, české mzdy, resp. mzdové náklady jsou jen na 40 % evropského průměru a na 30 % Německa. To dle odborů dává dostatečný prostoru pro zvyšování mezd a je to prý jen sobeckost a hamounství podnikatelů, které musí překonat až odborářský mzdový nátlak. Srovnání mezd, byť v přepočtu na společnou měnu, je však bez vztahu k ekonomickým parametrům země irelevantní. Korekcí o tzv. paritu kupní síly se náklady práce dostávají již přibližně na 60 % unijního průměru a zhruba na 50 % německých mezd. Rozdíl zůstává. Je způsoben odlišnou produktivitou práce. Upravíme-li data rovněž o produktivitu práce, která je u nás oproti unijnímu průměru menší o třetinu a oproti Německu poloviční, jsou rozdíly ve mzdách téměř smazány. Mzdy v České republice s ohledem na vyspělost naší země a produktivitu práce nikterak nepřirozeně nevybočují.

O růst mezd v posledních dvou letech se nezasloužily ani tak vyjednávací schopnosti odborů, jako opět růst produktivity práce. Základní ekonomickou zákonitostí je, že se dlouhodobě reálné mzdy (tj. mzdy očištěné o vliv inflace, berouc v úvahu jejich skutečnou kupní sílu) vyvíjejí v souladu s produktivitou

Sampling

- Sampling comparable to choosing a sample in survey
- Random/non-random selection
- Selection
  - Analysis of whole population will not yield additional insights
  - Practical considerations

Particular methods

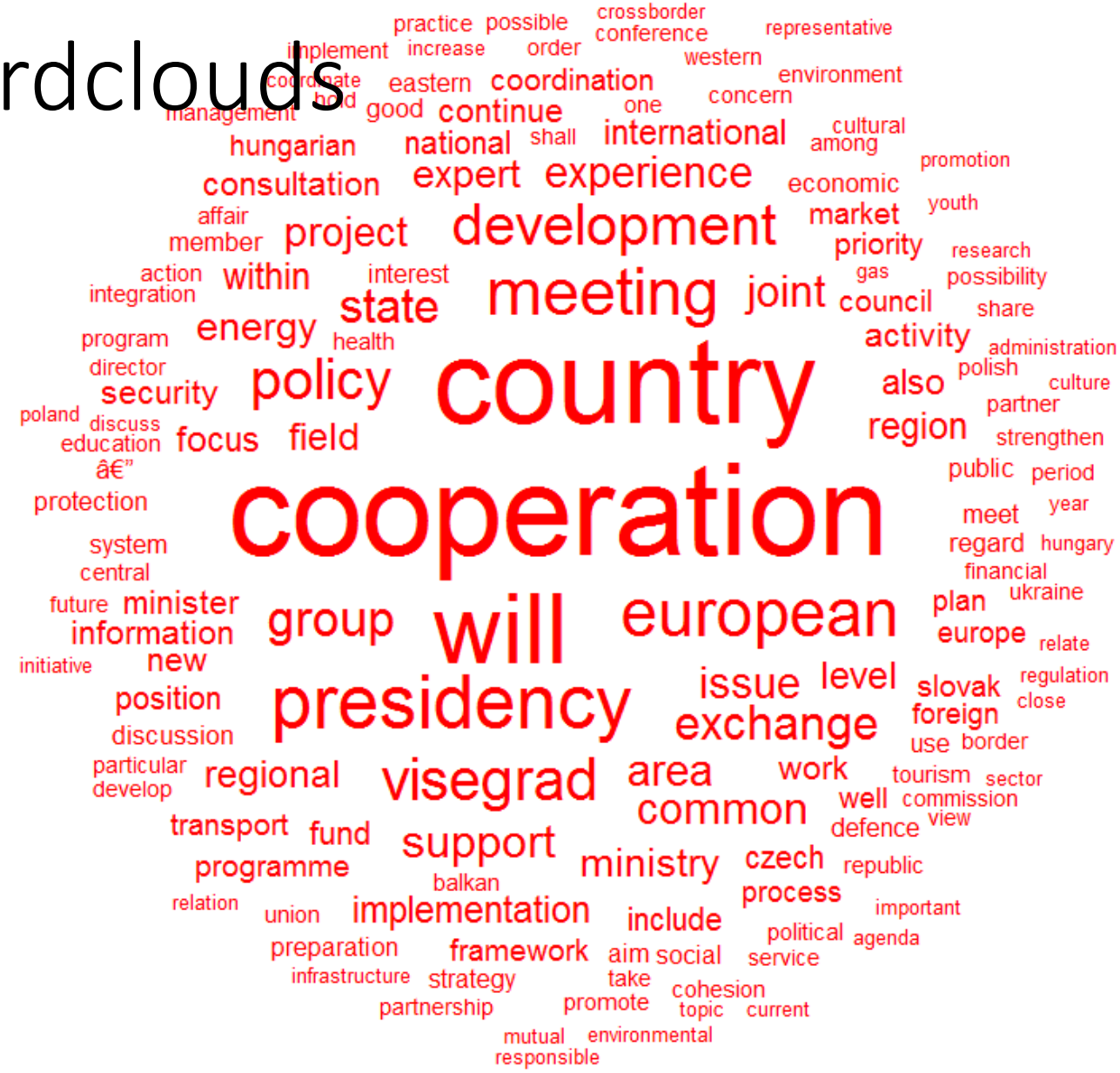
Word frequencies

- Basic exploration of text corpus
- Bag-of-words assumption
- Assumption that words convey meaning
- Assumption that certain words describe particular concept

Word frequencies
Term-document matrix

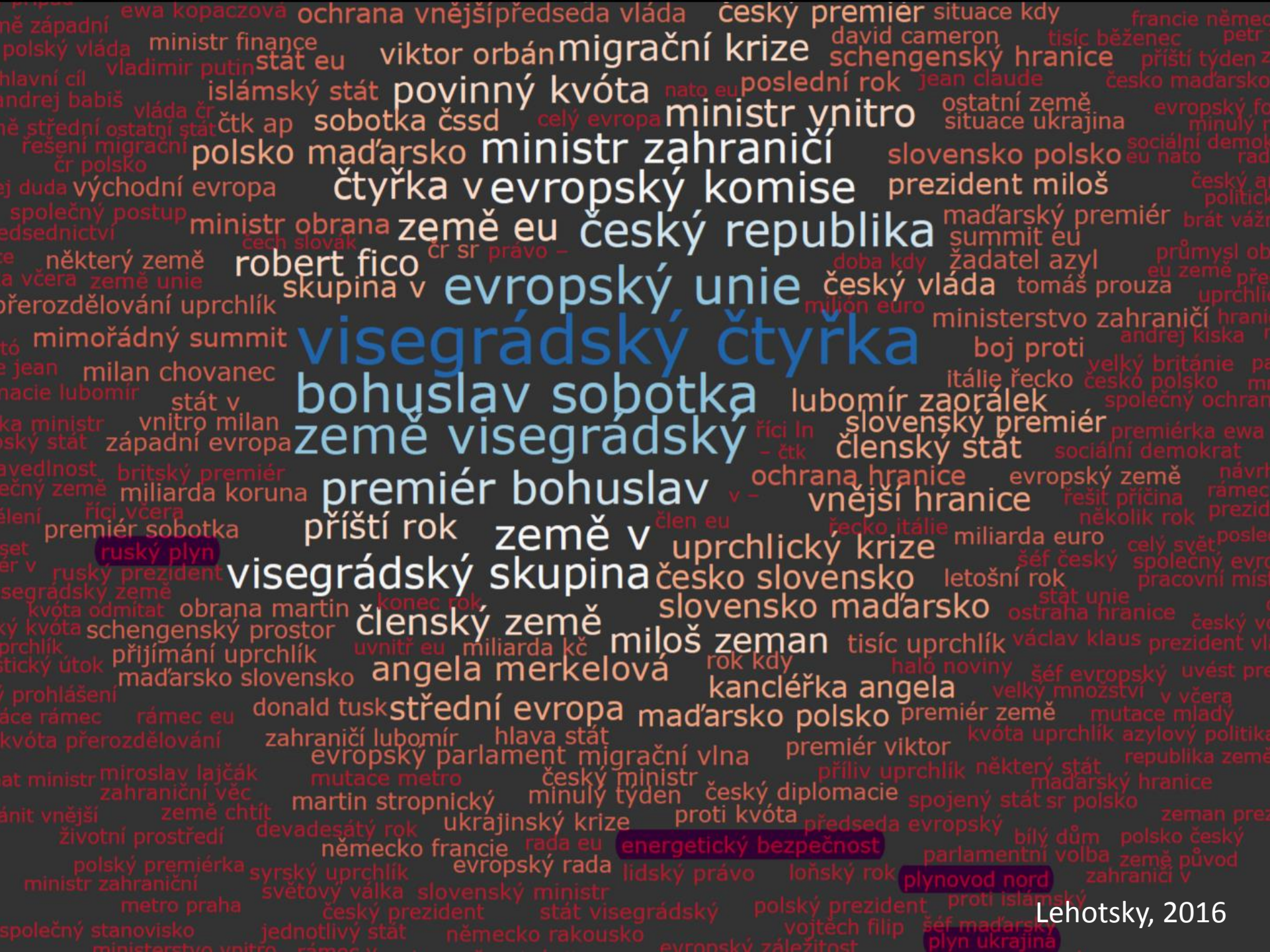
	2003-2004-cz	2004-2005-pl	2005-2006-hu	2006-2007-sk	2007-2008-cz	Sum
agriculture	3	6	2	5	3	**19**
aim	4	2	7	12	6	**31**
area	11	8	8	28	26	**81**
base	1	2	2	2	5	**12**
border	5	9	9	3	3	**29**
central	2	3	6	3	5	**19**
cohesion	3	1	7	4	4	**19**
commission	2	7	3	2	4	**18**
common	10	9	17	8	17	**61**
community	2	2	3	3	6	**16**
concern	9	13	12	18	6	**58**

Wordclouds



Lehotsky, 2016

Lehotsky, 2016



Lehotsky, 2016

Term correlations

- Co-occurrence of terms within unit
- Bag-of-words assumption
- Assumption of semantic proximity of co-occurring words
- Correlation coefficients such as Pearson's correlation coefficient

Coding

- Discovering concepts in text data
- Inductive (Glaser & Strauss, 1967)
  - Generation of codebook from available textual data
    - Open
    - Axial
- Deductive (Lacewell, Volkens & Werner, 2014)
  - Based on existing codebook
  - CMP

Coding

- Best practices
  - Two independent coders
  - Method of agreement
  - Measures of validity
    - Krippendorff's Alpha
- Misclassification (Mikhaylov, Laver & Benoit, 2011)
- Experimental approaches (Tosti-Kharas & Conley, 2016)
  - Crowdsourcing academic tasks

- Countries of the Visegrad Group are facing similar challenges in the energy sector, and are aware of the utmost importance of the issue of energy security.
- Energy policy concerns could be better dealt with on the basis of regional cooperation, especially in diversifying the natural gas supply of the countries of Central-, East- and South-East-Europe.
- Due to the lack of adequate interconnections and limited possibilities of reverse flow among the countries of the region, the development of infrastructure of transmission and storage of natural gas, crude oil and electricity is necessary in order to eliminate barriers in the transmission of energy among the counties in this region and to enable solidarity reaction to crises.
- Participants decided to further discuss the ways of improving the energy security situation of our countries and to adopt the necessary measures which may help weaken the impacts of any possible disruption of supply in the future.
- They expressed their support to strengthen cooperation in further integrating their gas networks, including the Southern Corridor and the Nabucco Project as well as the North-South transportation corridor through the region, the planned Croatian and Polish Liquefied Natural Gas terminals and the NETS project.
- Participants declared their willingness to provide support for the missing interconnectors, including joint efforts for a higher allocation of EU financial resources to all projects with the potential of increasing the energy security of the region.
- Cooperation of energy companies of the region will also be further encouraged.

102 Foreign Special Relationships: Negative	Negative mentions of particular countries with which the manifesto country has a special relationship.
103 Anti-Imperialism - comprised of:	
103.1 State Centred Anti-Imperialism	Negative references to imperial behaviour and/or negative references to one state exerting strong influence (political, military or commercial) over other states. May also include: • Negative references to controlling other countries as if they were part of an empire; • Favourable references to greater self-government and independence for colonies; • Favourable mentions of de-colonisation.
103.2 Foreign Financial Influence	Negative references and statements against international financial organisations or states using monetary means to assert strong influence over the manifesto or other states. May include: • Statements against the World Bank, IMF etc.; • Statements against the Washington Consensus; • Statements against foreign debt circumscribing state actions.
104 Military: Positive	The importance of external security and defence. May include statements concerning: • The need to maintain or increase military expenditure; • The need to secure adequate manpower in the military; • The need to modernise armed forces and improve military strength; • The need for rearmament and self-defence; • The need to keep military treaty obligations.

Lehmann et. al., 2017

Coding

2. MODERNIZACE ČESKÉ SPOLEČNOSTI	NA
2.1. ZDRAVÁ EKONOMIKA – základ dlouhodobé prosperity	NA
Globální prostředí	NA
Globalizace zasahuje do všech oblastí našeho života a nelze jí uniknout.	107
Projevuje se například zvyšující se rychlostí kapitálových toků	408
nebo čím dál tím větším podílem levného spotřebního zboží z dovozu na trzích vyspělejších zemí.	401
Globalizace je nejen příležitostí, ale také hrozbou.	107
Konkurenceschopnost asijských ekonomik se opírá o ekologický a sociální dumping.	503
Na mezinárodním poli budeme proto usilovat o postupné prosazování minimálních sociálních a ekologických standardů	503
kombinovaných s promyšlenou (tzv. kvalifikovanou) ochranou evropského trhu.	406

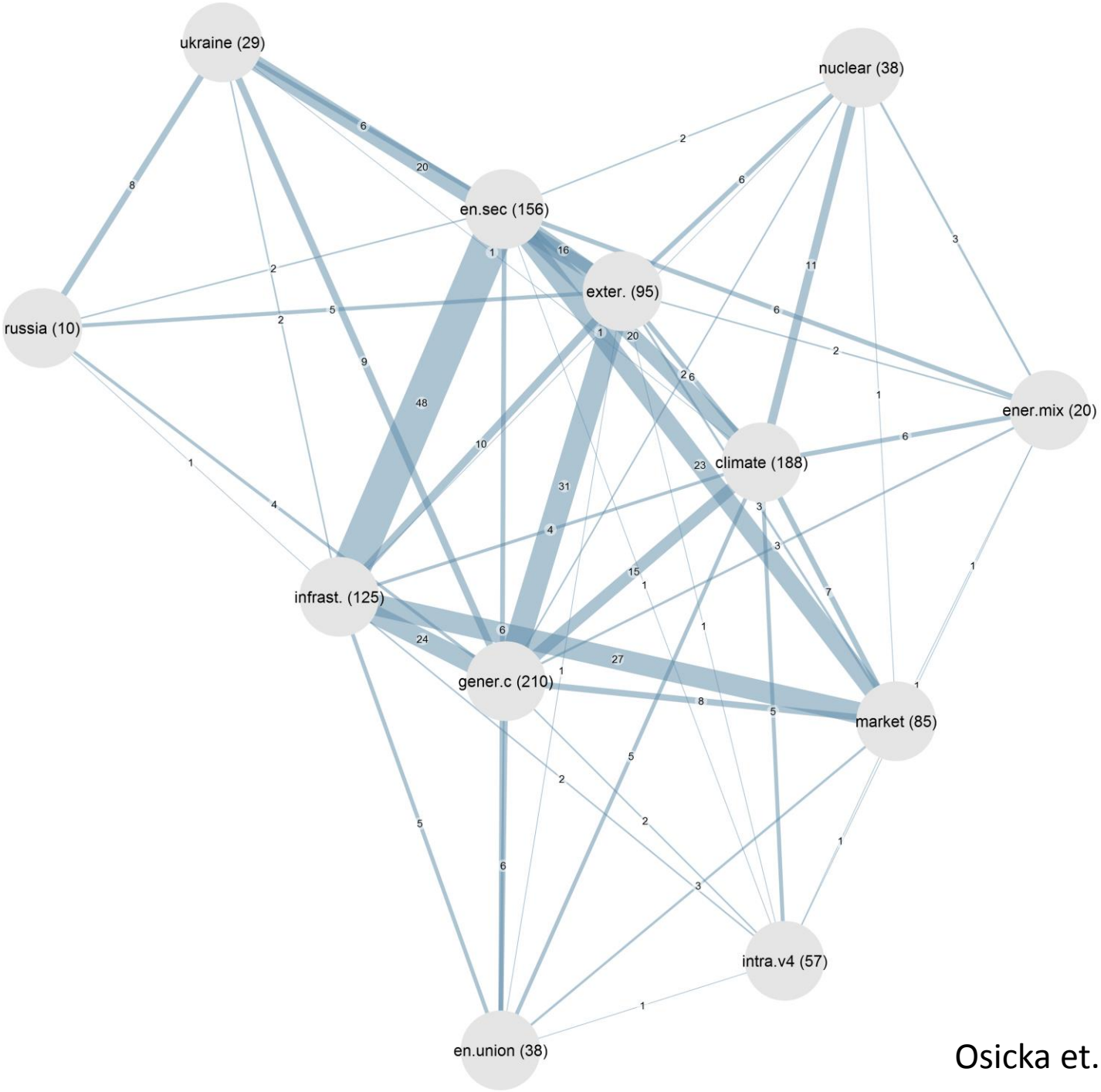
Lehmann et. al., 2017

Keywords in Context (KWIC)

- Allows analysis of concept's original proximate surrounding (linguistic environment)
  - extracted with the concept itself
- Corpus linguistics
- Useful for understanding concepts
- Initial coding might benefit

Keywords in Context (KWIC)

exchange of information about	**energy**	policy and coordination of
and coordination of the	**energy**	policy of V 4
sphere of new EU	**energy**	legislation, especially rule
of trans- European	**Energy**	Networks, concentrate on
with the operation of	**energy**	facility, impact of
the field of the	**energy**	sector, industry and
	Energy	continuation of meeting of
establishment of a common	**energy**	and gas market.
- operation in the	**energy**	sector in the usual



Osicka et. al., 2016

Dictionary-driven methods

- Words categorized in pre-existing dictionary
  - Custom-made categories
  - WordScores (Benoit & Laver, 2002; Laver, Benoit & Garry, 2003)
  - Sentiment analysis (SentiWords)
- Alternative to manual coding
- Strong assumptions
- Open for further analysis

Custom dictionary

- Natural gas – security context
  - Security
  - Supply
  - Geopolitics
  - Interrupt
  - Cut
- Natural gas – market context
  - Liquidity
  - Market
  - Trade
  - Price
  - Exchange

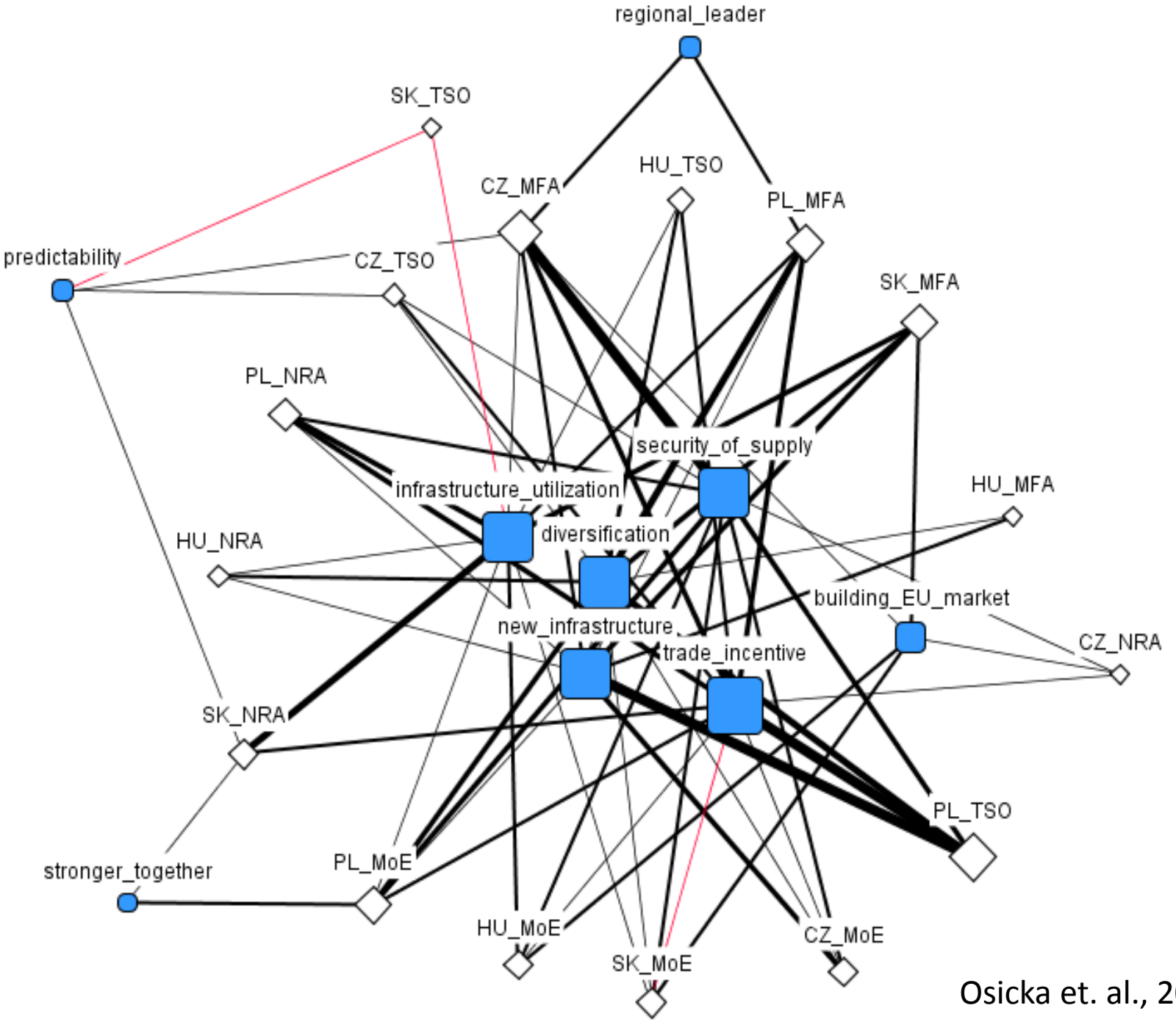
Sentiment analysis

- SentiWords (Guerini, Gatti & Turchi, 2013)

aristotelian_logic#n	0.15793
aristotelian#a	0
aristotelian#n	0
aristotle#n	-0.01819
arithmetic_mean#n	0
happy#a	0.86753
sadist#n	-0.72256
sadness#n	-0.65005
thermonuclear_reactor#n	0
thermonuclear_warhead#n	0

Discourse Network Analysis

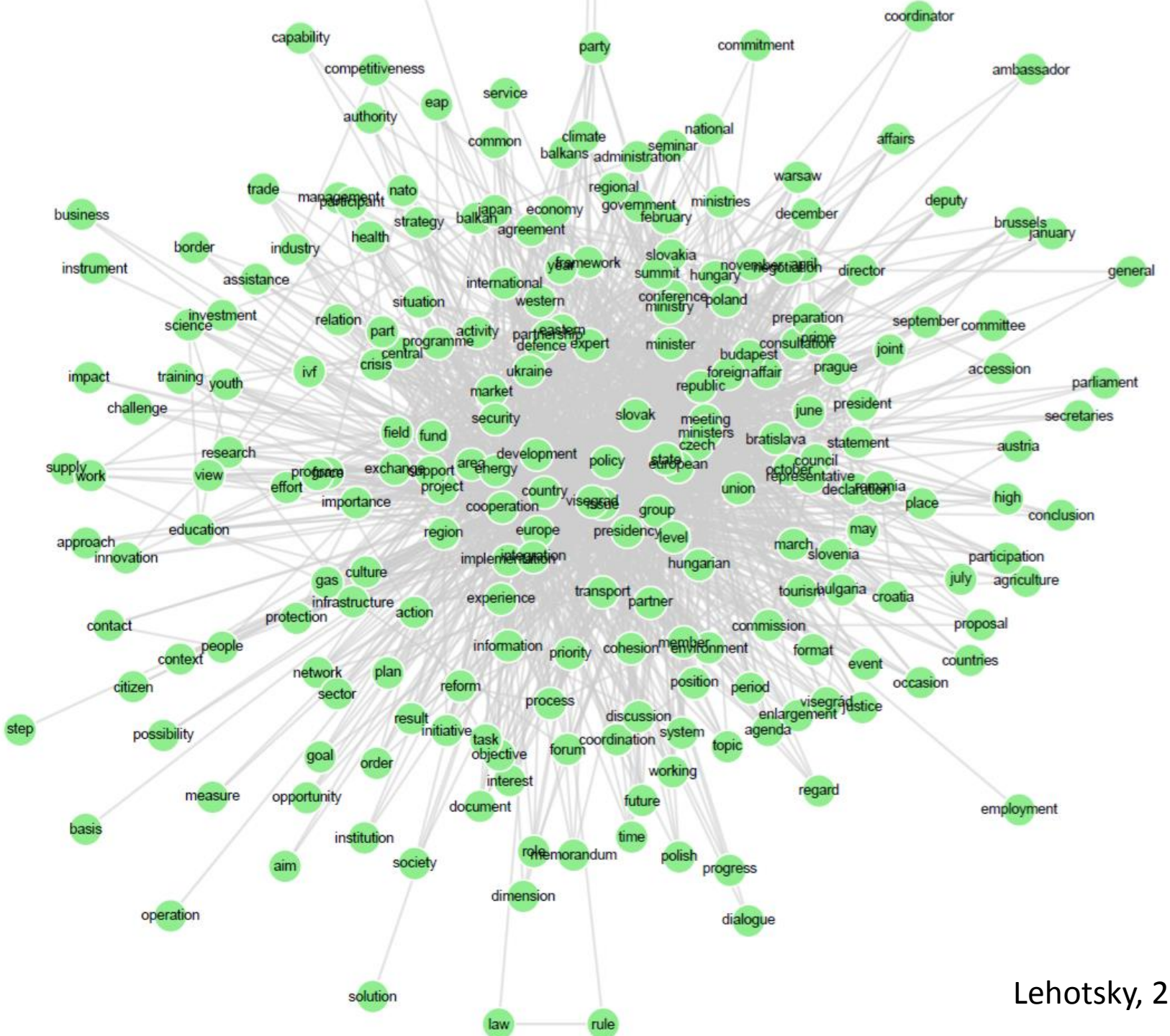
- Actors share beliefs and subscribe to particular concepts (Leifeld & Haunss, 2012)
- Concepts may be captured through textual data
  - Documents (manifestos, press releases…)
  - Interviews
- Text coding
- Output as two-mode network data
  - Actors
  - Concepts (captured in codes)



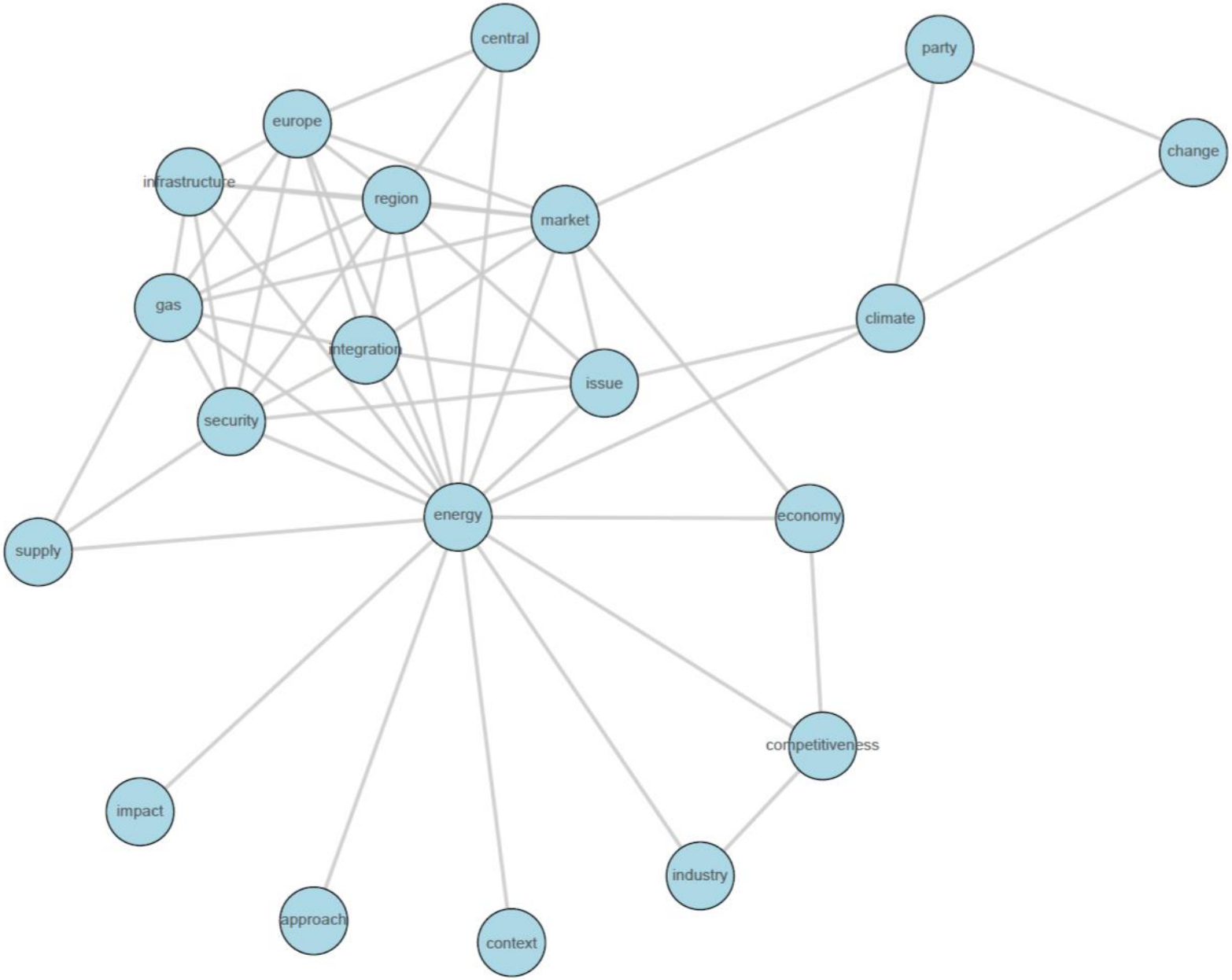
Osicka et. al., 2016

Semantic networks

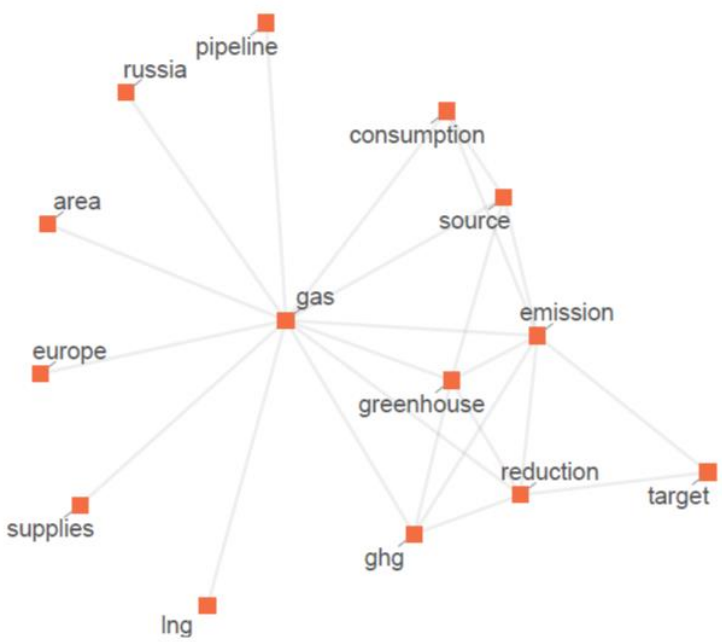
- Unsupervised method (Nerghes & Lee, 2015)
- Co-occurrence of words within a textual unit
- More co-occurrences indicate stronger relation between words
- Clustering algorithms
  - Clusters of words occurring together more frequently than other

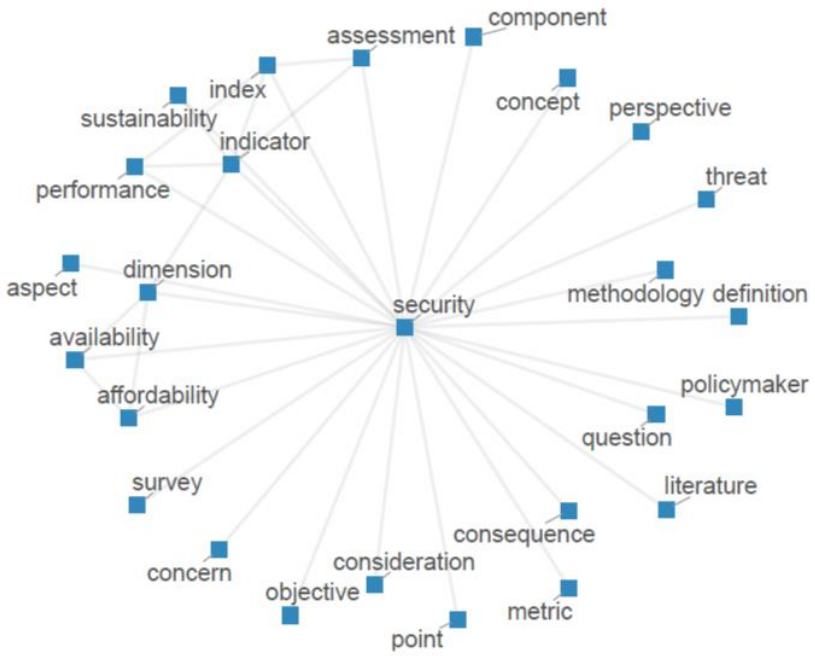


Lehotsky, 2016

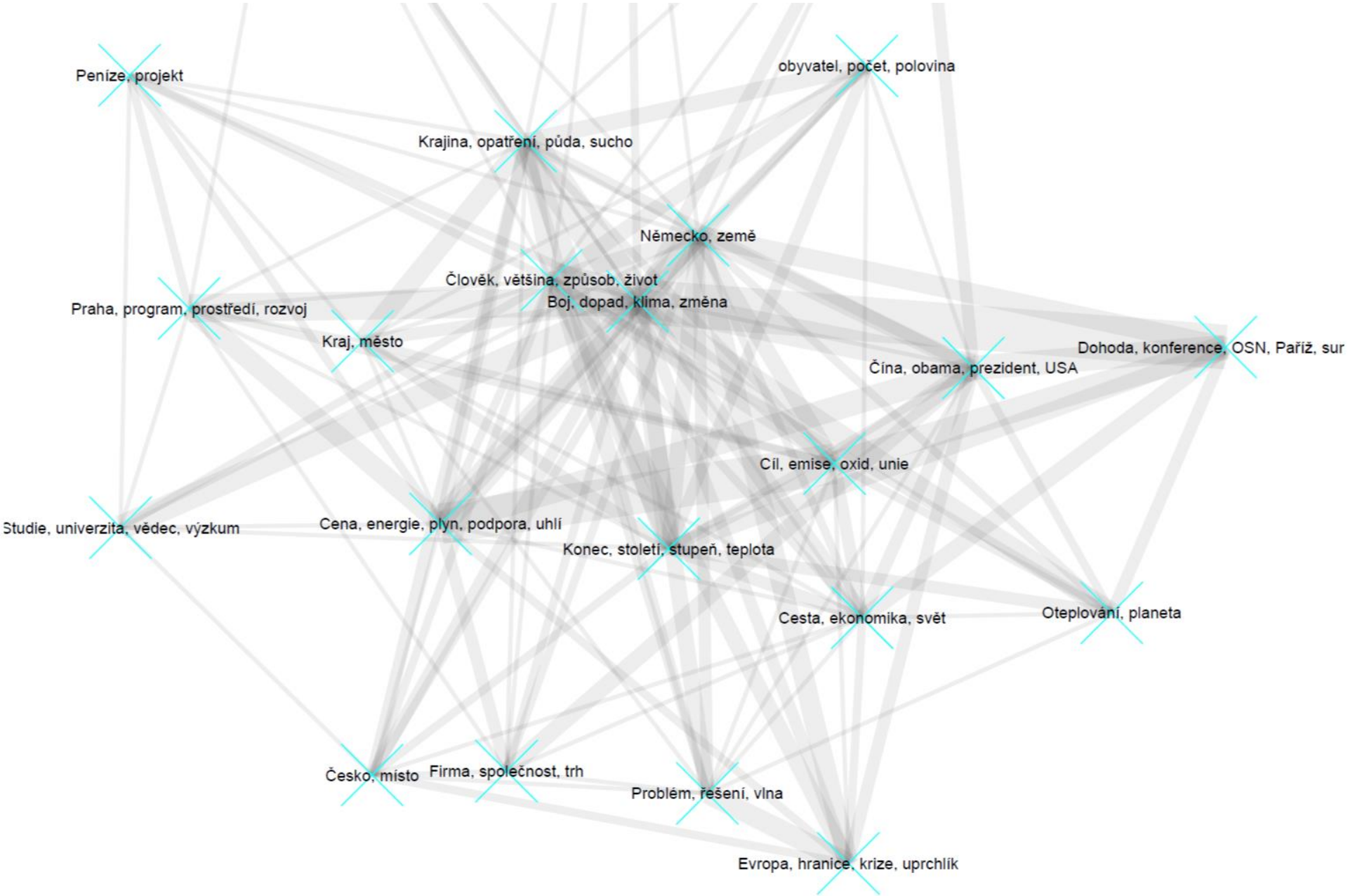


Lehotsky, 2016





Lehotsky, 2017



Lehotsky, 2016

Topic modeling

- Topic model
  - Bag-of-words assumption
  - Latent semantic space
  - Documents are mix of few topics
  - Each word in document belongs to particular topic
  - Approximation of distributions given corpus of data
- Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003)
  - Many other derived models - lifting some LDA assumptions
  - Wide range of issues

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

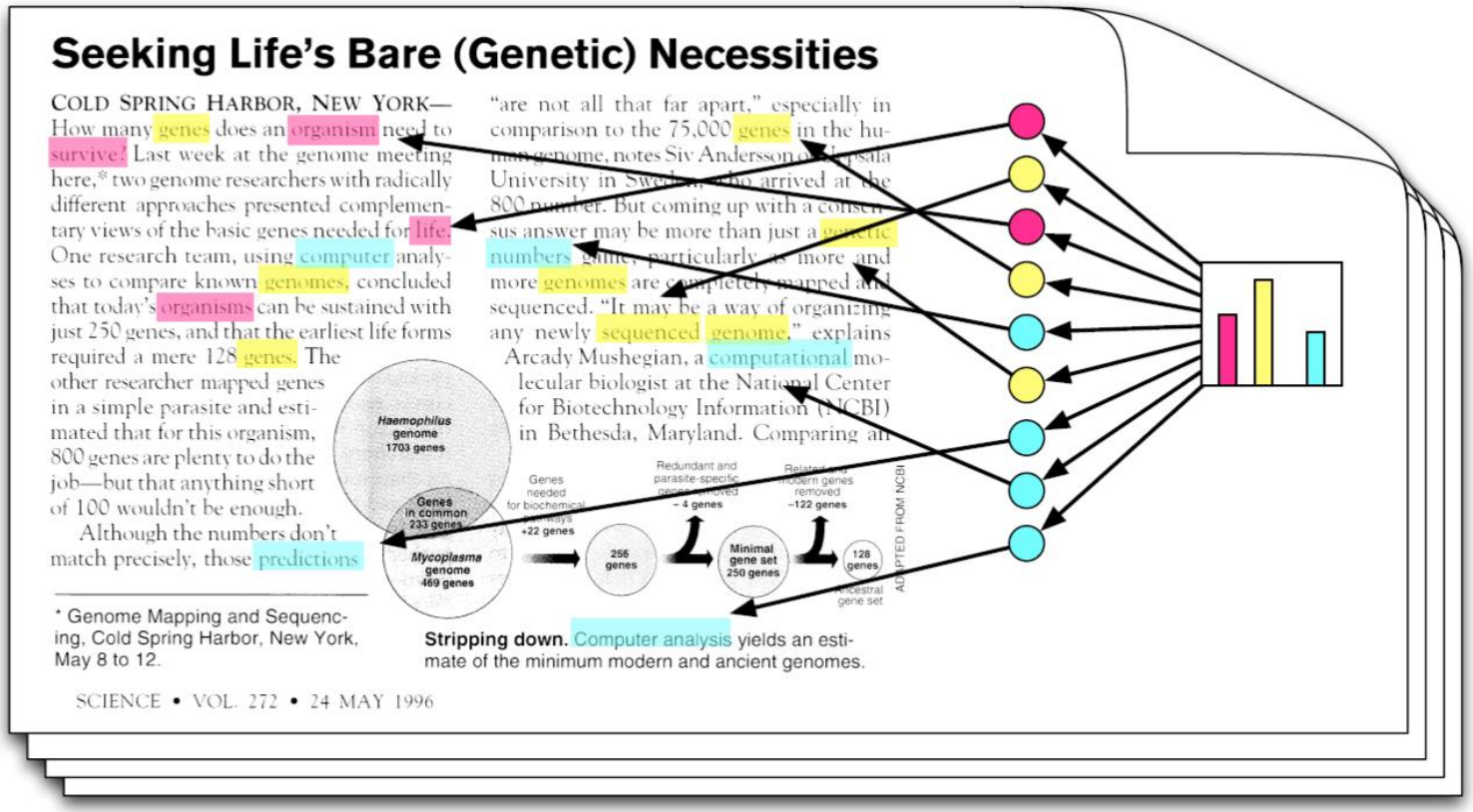
life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Blei, 2011

LDA topic model (9 topics)

elektrárna	vláda	firma	voda	težba	mesto	kraj	dul	clovek
zdroj	ministr	spolecnost	krajina	uhlí	století	práce	uhlí	zeme
uhlí	volba	CEZ	území	limit	Ostrava	region	težba	život
energie	CSSD	skupina	obec	obec	zámek	problém	okd	svet
elektrina	ODS	cena	místo	Jiretín	památka	program	horník	doba
plyn	kraj	uhlí	mesto	prolomení	dum	oblast	spolecnost	díte
cena	návrh	trh	stavba	obyvatel	kostel	rozvoj	clovek	cas
energetika	vec	prodej	težba	clovek	Most	nezamestnanost	práce	válka
zeme	poslanec	zisk	projekt	vláda	hodina	obcan	zamestnanec	praha
výroba	zákon	podíl	most	težar	muzeum	clovek	tuna	evropa
stát	predseda	výroba	oblast	Litvínov	centrum	místo	karviná	rodina
koncepce	politika	akcie	peníze	zásoba	budova	podpora	mluvcí	vetšina
prumysl	KSCM	NWR	metr	dum	areál	doba	šachta	místo
teplo	Stát	podnik	jezero	sdružení	hrad	stát	reditel	žena
Temelín	premiér	investice	zóna	CSA	výstava	pocet	útlum	muž

Ocelik & Lehotsky, 2017

Computer aids

- With graphical user interface
  - Atlas.ti
  - Nvivo
  - MaxQDA
  - WordStat
  - Mallet (LDA)
- Programming languages
  - R
  - Python
  - …