

Associations in discourse analysis

Workshop on DH and corpus-linguistic and AI methods for
analyzing discourse
April 4, 2025

Václav Cvrček & Masako Fidler
Charles University, Brown University

Main goals of this talk

1. Corpus methods in discourse analysis
2. Data: ONLINE corpus
3. Associations and how to find them in the data
4. Using associations in discourse analysis
5. Collocations vs. associations



Corpus-assisted discourse studies (CADS)

- ▶ Exploratory, less hypothesis-driven
- ▶ Quantitative (× qualitative interpretation, annotation etc.)
- ▶ “data mining” ⇒ hypothesis (pattern) ⇒ experiment ⇒ hypothesis (pattern) ⇒ experiment ⇒ ... ⇒ conclusion

Our questions and our suspicions

- ▶ What do we mean by “discourse of X” (discourse of racism, discourse of migration...)?
- ▶ Frequently used methodology to study “discourse of X”: keywords + collocation analysis (cf. Partington & Duguid 2013:307, 315; Heritage & Baker 2022: 443; Philip 2011: 26, 64)
 - ▶ “Collocations create connotations” (Stubbs, 2005: 14)
- ▶ Hypothesis: collocations might not provide a complete picture, esp. in contrast to methods for harvesting associations
 - ▶ Market Basket Analysis (Cvrček & Fidler 2022)
 - ▶ Companions (Cvrček & Fidler 2024; Fidler & Cvrček 2024)



Data: ONLINE corpus



ONLINE corpus

- ▶ monitor corpus for mapping the dynamic content of the Czech internet
- ▶ Generation 1 (Cvrček & Procházka 2020)
 - ▶ data for 2/2017 – 3/2021
 - ▶ online media
 - ▶ discussions
 - ▶ internet forums
 - ▶ social networks (facebook, twitter)
- ▶ Generation 2 (Cvrček, Jeziorský & Henyš 2022)
 - ▶ 4/2021 onwards
 - ▶ online media only
 - ▶ several anti-system web portals were blocked after Feb 25, 2022
 - ▶ updated on a daily basis (ca 1–4.5 mil. words)
 - ▶ annotation – lemmatization and tagging

Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)

Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals



Media typology

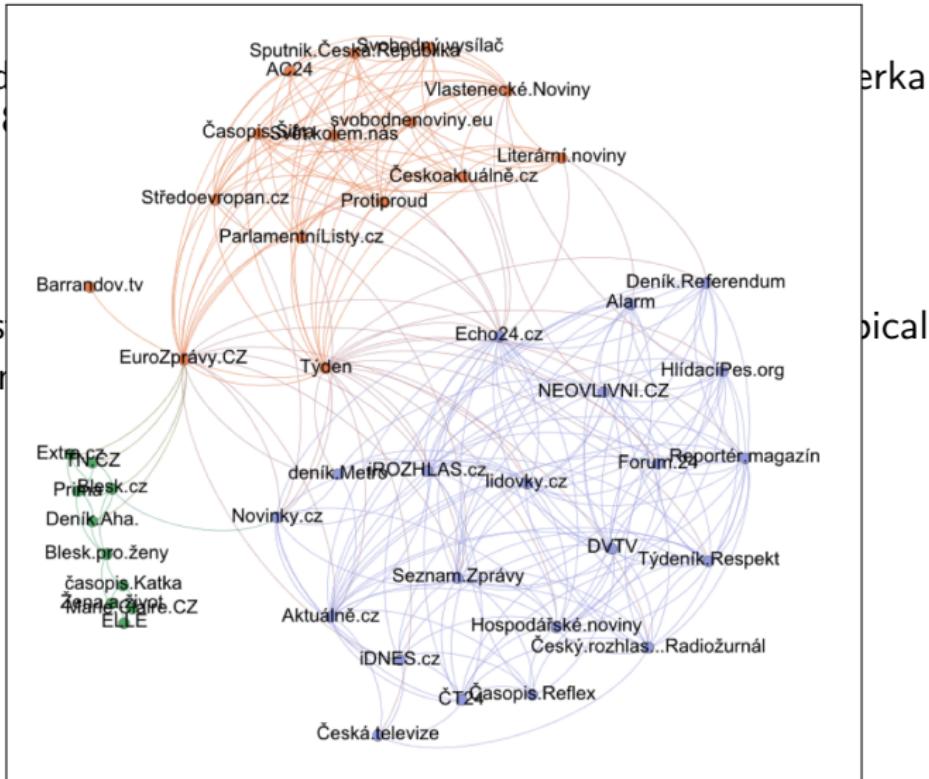
- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals
 - ▶ alexaRank: visitors of websites

Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals
 - ▶ alexaRank: visitors of websites
 - ▶ crowdTangle: shares and likes on social networks

Media typology

- media 2013
-
-
-
- clusters examined



Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals
 - ▶ alexaRank: visitors of websites
 - ▶ crowdTangle: shares and likes on social networks
- ▶ clusters of media portals – labelling based on prototypical examples
- ▶ Mainstream media (MS)

Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals
 - ▶ alexaRank: visitors of websites
 - ▶ crowdTangle: shares and likes on social networks
- ▶ clusters of media portals – labelling based on prototypical examples
- ▶ Mainstream media (MS)
- ▶ Antisystem media (ANTS)



Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals
 - ▶ alexaRank: visitors of websites
 - ▶ crowdTangle: shares and likes on social networks
- ▶ clusters of media portals – labelling based on prototypical examples
- ▶ Mainstream media (MS)
- ▶ Antisystem media (ANTS)
 - ▶ pro-Kremlin, anti-EU, anti-NATO, anti-establishment narratives

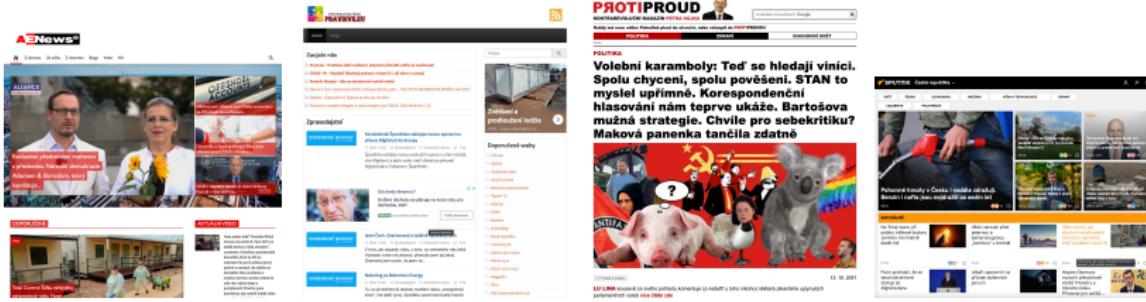
Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals
 - ▶ alexaRank: visitors of websites
 - ▶ crowdTangle: shares and likes on social networks
- ▶ clusters of media portals – labelling based on prototypical examples
- ▶ Mainstream media (MS)
- ▶ Antisystem media (ANTS)
 - ▶ pro-Kremlin, anti-EU, anti-NATO, anti-establishment narratives
 - ▶ fast cloning of texts

Media typology

- ▶ media classes based on the similarity of audiences (Šlerka 2018)
 - ▶ similarweb: links between portals
 - ▶ alexaRank: visitors of websites
 - ▶ crowdTangle: shares and likes on social networks
- ▶ clusters of media portals – labelling based on prototypical examples
- ▶ Mainstream media (MS)
- ▶ Antisystem media (ANTS)
 - ▶ pro-Kremlin, anti-EU, anti-NATO, anti-establishment narratives
 - ▶ fast cloning of texts
 - ▶ coordinated sharing

Anti-system web portals in the Czech Republic



Characteristic features:

- ▶ spread of disinformation and pro-Russian propaganda
- ▶ unclear ownership
- ▶ anonymous authors and/or editorial team
- ▶ sources are often missing or are irrelevant (non-existent "experts")

(source: SIS, MI CR, EFIJ – Online media portals rating,
<https://www.nfnz.cz/rating-medii/zpravodajske/>)



Established methods

Methods of CADS

- ▶ frequency and dispersion of units (words)
- ▶ concordance lines close-reading
- ▶ collocations and co-occurrences
- ▶ keyword analysis (KWA)
 - ▶ KWA compares target text/corpus and reference corpus
 - ▶ Identifies prominent units – keywords (KWs) (Scott & Tribble 2006)
 - ▶ Based on differences in frequencies: statistical significance (log-likelihood or chi² tests) + effect size (DIN)
 - ▶ KWs: what the text is about (topics), genre/register (× cultural keywords, search terms)
- ▶ very few methodological innovations (cf. effect size – Gabrielatos & Marchi, 2012; dispersion – Egbert & Biber, 2019; KMA – Fidler & Cvrček 2019; DKL – Gries, 2024)

Top 10 keywords

Period: 10/2017–10/2018

Ref. corpus: SYN2015 (fiction + non-fiction + journalism)

Mainstream

Lemma	texty
podle 'according to'	2730
prezident 'president'	2312
vláda 'government'	2262
uvést 'say, state'	1797
Babiš [CzR PM]	1587
volba 'election'	1540
soud 'court'	1382
strana 'party'	1324
Zeman [CzR president]	1200
předseda 'chairman'	1155

Top 10 keywords

Period: 10/2017–10/2018

Ref. corpus: SYN2015 (fiction + non-fiction + journalism)

Mainstream

Lemma	texty
podle 'according to'	2730
prezident 'president'	2312
vláda 'government'	2262
uvést 'say, state'	1797
Babiš [CzR PM]	1587
volba 'election'	1540
soud 'court'	1382
strana 'party'	1324
Zeman [CzR president]	1200
předseda 'chairman'	1155

Anti-system

Lemma	texty
Rusko 'Russia'	1931
USA	1696
prezident 'president'	1680
ruský 'Russian'	1563
americký 'American'	1472
vláda 'government'	1300
politický 'political'	1212
válka 'war'	1105
země 'country'	1085
EU	1025

(Fidler & Cvrček 2020)

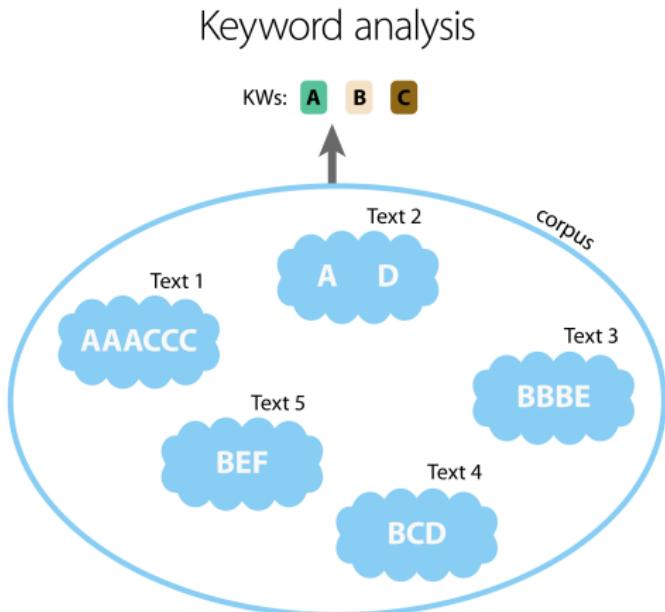
Anti-system unique keywords (samples)

- ▶ neocon
- ▶ anglosionistická 'anglozionist'
- ▶ apostata 'apostate'
- ▶ Russiagate
- ▶ amík 'Ami'
- ▶ havloidní 'Havel-ish (pejor)'
- ▶ dolarizace 'dollarization'
- ▶ židozednářský 'Judeo-Masonic'
- ▶ Armageddon
- ▶ antirusismus 'anti-russian'
- ▶ vazalství 'vassalship'

(Fidler & Cvrček 2020)

KWA – a “bag-of-words” approach

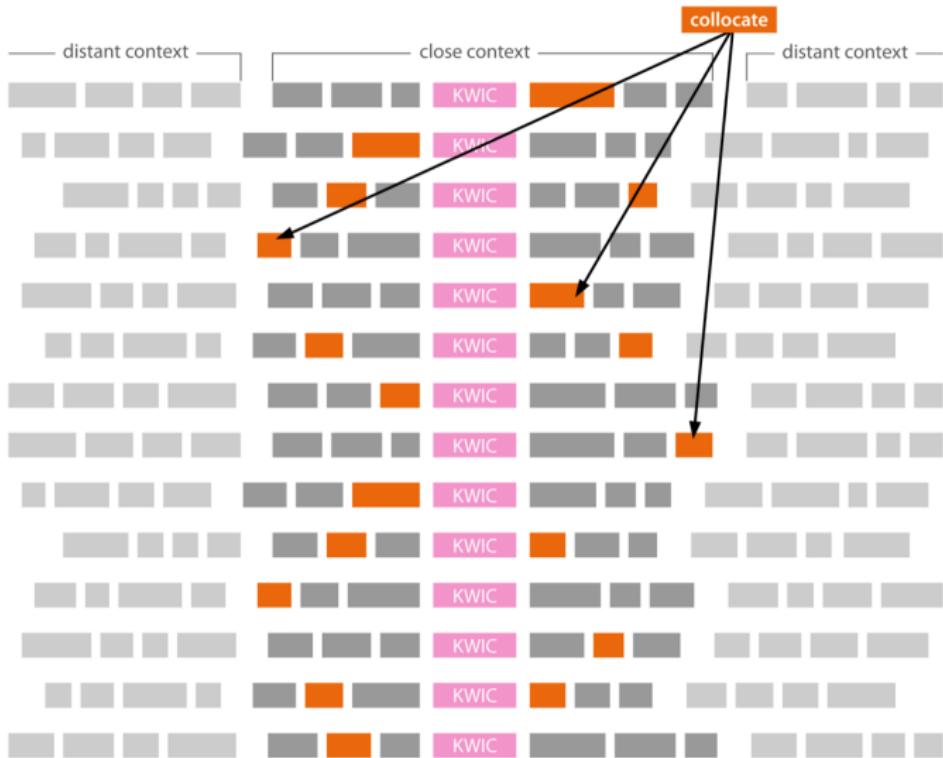
- ▶ Identification of KWs solely on frequency in the corpus
- ▶ Dispersion of units and internal structure is neglected (× Egbert & Biber 2019)
- ▶ List of KWs: devoid of **contextual** information and the strength of **relationship** among KWs
 - ▶ collocations
 - ▶ KW links





Mining associations among topics

Collocations: adding a context to KWs



Market basket analysis (MBA)

Frequently Bought Together

Color: Black

Customers buy this item with Bodum 1548-01US Brazil 8-Cup (34-Ounce) Coffee Pre:



Price For Both: \$39.47



Add both to Cart



Add both to Wish List

These items are shipped from and sold by different sellers. [Show details](#)

Customers Who Bought This Item Also Bought

Color: Black



Bodum Chambord

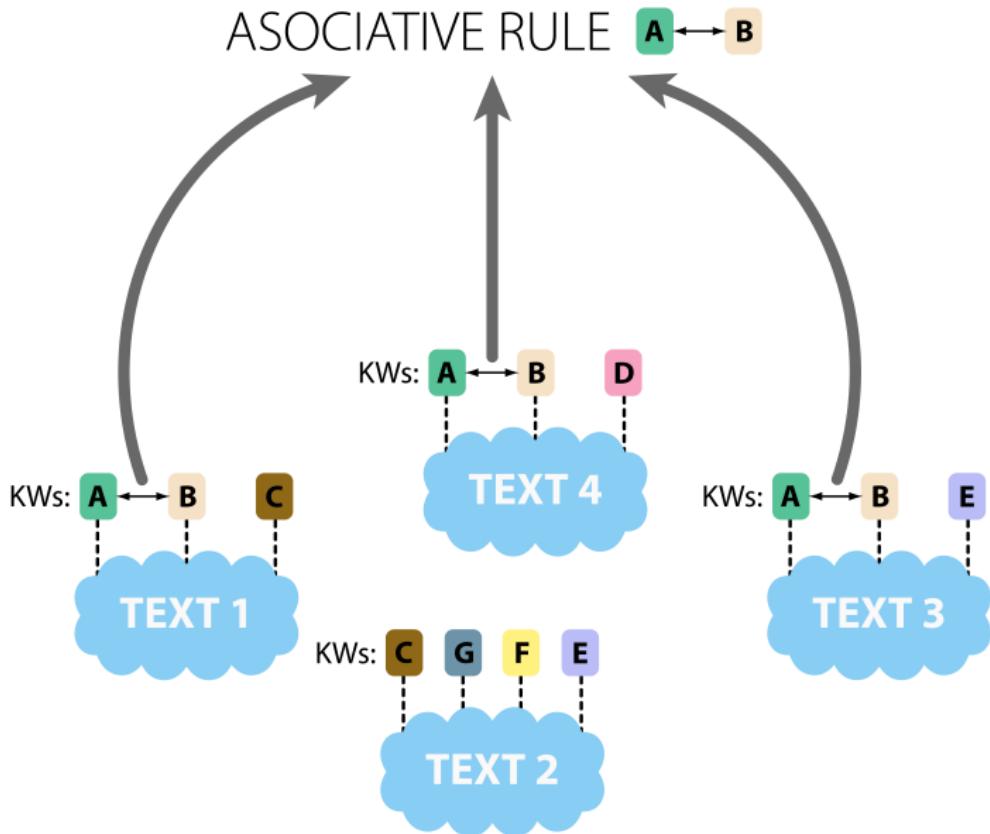


Bodum 1548-01US



Wooden Coffee Grinder

(Market basket analysis)



Associative links extracted from MBA

- ▶ which topics are interrelated in discourse
- ▶ each associative link: Antecedent (LHS) → Consequent (RHS)
- ▶ Associative Links are based on:
 - ▶ KWs of individual texts
 - ▶ Co-occurrence of KWs in texts within media classes
- ▶ Evaluation of ALs:
 - ▶ Support = range of texts with both associated words
 - ▶ Confidence = systematic co-occurrence of words
 - ▶ Lift = strength of association between words

(Han et al. 2011: 244; Miner 2012: 917; Cvrček & Fidler 2022)

Associative links and associative arrays

Associative links with KW *migrant* in anti-system (6–9/2020)

antecedent	consequent
(ne)legální 'illegal'	→ migrant
azyl 'assylum', země 'country'	→ migrant
evropský 'european', migrace 'migration'	→ migrant
EU, evropský 'european', hranice 'border'	→ migrant
	...

Associative links and associative arrays

Associative links with KW *migrant* in anti-system (6–9/2020)

antecedent		consequent
(ne)legální 'illegal'	→	migrant
azyl 'assylum', země 'country'	→	migrant
evropský 'european', migrace 'migration'	→	migrant
EU, evropský 'european', hranice 'border'	→	migrant

...

Associative array (AA): azyl, členský, EU, evropský, hranice,
(ne)legální, migrace, právo, unie, uprchlík, země...



Using MBA

Zone-flooding: migrant

*The Democrats don't matter.
The real opposition is the media.
And the way to deal with them
is to flood the zone with shit.*

Steve Bannon

Zone-flooding: migrant

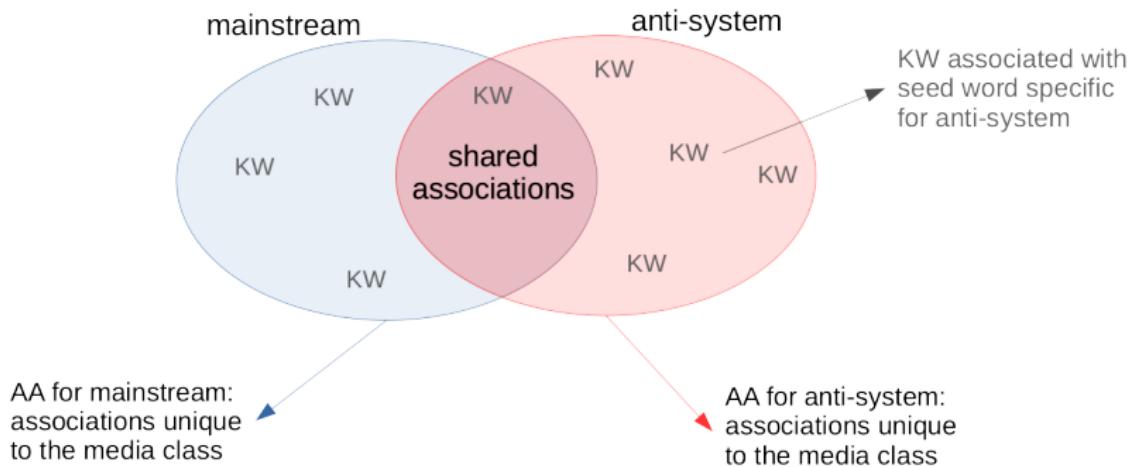
*The Democrats don't matter.
The real opposition is the media.
And the way to deal with them
is to flood the zone with shit.*
Steve Bannon

KW *Migrant* and its associations (10/2017–10/2018)

Segment	Texts	Assoc. links	Median lift
mainstream	256 (2.1 %)	235	14.50
anti-system	472 (6.4 %)	1448	7.01

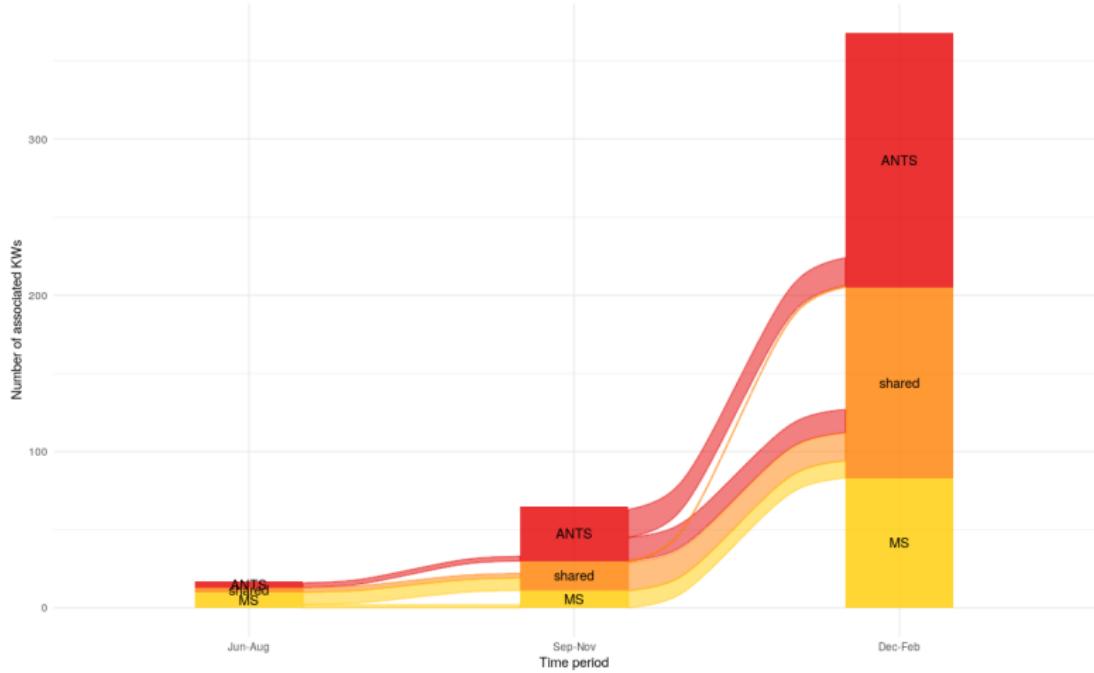
Framing of events: Associative array (AA)

- ▶ Represents the network of associations a word is involved in
- ▶ Contrast in framing between anti-system and mainstream media classes



Sharing & taking over the associations

Associative array of KW *vakcína* 'vaccine'



Period: 6–8/2020, 9–11/2020, 12/2020–2/2021

Parasitic on MS topics

Migration of associations with *vaccine*

- ▶ from MS → ANTS:
 - 1>2 phase, clinical, pandemic, test, healthcare...
 - 2>3 agency, infection, inoculative, Pfizer, number, case, spread...
- ▶ from ANTS → MS:
 - 2>3 Russia, USA, immunity, Prymula, government, Sputnik, V



Parasitic on MS topics

Migration of associations with *vaccine*

- ▶ from MS → ANTS:
 - 1>2 phase, clinical, pandemic, test, healthcare...
 - 2>3 agency, infection, inoculative, Pfizer, number, case, spread...
- ▶ from ANTS → MS:
 - 2>3 Russia, USA, immunity, Prymula, government, Sputnik, V

Associations with *vaccine* specific for one segment

- ▶ within MS:
 - 3 distribution, number, percent, index, degree, Monday...
- ▶ within ANTS:
 - 1,2,3 pharmaceutical
 - 2,3 Gates, army, mandatory, wearing, face mask
 - 3 West, global, Brussels, freedom, right, lockdown, scandal...

Comparing portals via associations (8/2021–1/2022)

- ▶ seed words (2 most discussed topics):
 - ▶ **upcoming crisis** – military training near Ukrainian borders:
Rusko, ruský, Putin, Moskva, Rus
 - ▶ **receding crisis** – vaccination almost complete, number of cases dropping: *covid, očkování, očkovaný, vakcína, pandemie*
- ▶ associations (MBA) → multiple correspondence analysis
(Clarke et al. 2021) → 2D map of KWs and portals

Data

- ▶ 5 Mainstream portals (# of texts): *idnes.cz* (9491), *seznamzpravy.cz* (5890), *ceskatelevize.cz* (4520), *irozhlas.cz* (4055), *novinky.cz* (3397)
- ▶ 5 Anti-system portals (# of texts): *sputniknews.com* (4356), *pravystvor.cz* (1423), *czechfreepress.cz* (553), *svobodny-svet.cz* (632), *novarepublika.cz* (562)

(CzechLabs, Fall 2022)

Multiple correspondence analysis

- ▶ Statistical method for uncovering relations between associative arrays of the same concept (seed words) in different portals
- ▶ Used for keywords (not associations) – discourse of Islam in British press (Clarke et al. 2021)
- ▶ Input data table:
 - ▶ Rows: seed words in portals
 - ▶ Columns: associated KWs
 - ▶ Cell: True/False (association is present/absent)
- ▶ Dimensionality reduction – 2D map of KWs and portals

ANTS portals focusing on Russia-related topics, esp. security threats (pro-Russian narratives, cf. Cvrček & Fidler 2024) not much focus on reporting the state of COVID

(Overt) military & security threats

ANTS portals focusing on Russia-related topics and focusing on the pandemic measures not much on reporting the state of COVID

Negotiations/diplomacy and hidden threats

sputniknews.com

idnes.cz

seznamzpravy.cz

ceskatelevize.cz

novinky.cz

deniknove.cz
svobodny-svet.cz

irochlas.cz

Dim 1

Dim 2

0.4

0.0

-0.4



Questioning anti-covid measures

MS portals interested in reporting on the COVID spread simultaneously focusing on potential security threats

Current state of covid & development

MS portals interested in reporting on the COVID spread, but esp. focusing on the government measures implemented



Collocations versus MBA



East and West – where do we belong?

Data

- ▶ Target corpus: Printed state-wide newspaper articles 1996 vs. 2021 (SYNv11, Křen et al. 2022); approx. 190K or 116K texts/articles
- ▶ Reference corpus: newspaper articles from 1997 + 2020
- ▶ Seed KWs: [lemma=" [Vv] ýchod"] and [lemma=" [Zz] ápad"]
- ▶ Collocations within the span (-3; +3), sorted via LogDice
 - ▶ Prototypical+ CA: top 100 lemmas, min fq. 5 in corpus
- ▶ Associations (MBA support: 0.0001, confidence: 0.2) → associative arrays (AAs)

MBA vs. CA: East + economic concerns

Year	MBA	Both	Collocations
1996	auto, banka, drahý, ekonomika, finance, firma, investice, investiční, kapitál, koncern, koruna, manažer, marka, obchod, podnik, podnikatel, podporovat, prodej, průmyslový, růst, šéf, trh, USD, výrobní, zboží	ekonomický, hospodářský, levný, vývoz	boom, dohánět, expandovat, expanze, export, chudý, konkurence, plynárenský, ropa, rozvoj
2021	ekonomický	–	dovážet, expanze, levný

MBA vs. CA: Russia-related associations

West	MBA	Both	Collocations
1996	Rus, ruský, svaz, ukrajinský, Zjúganov	Gorbačov, Jelcin, Jelcinův, Lebed', Moskva, Rusko, sovětský, SSSR, Ukrajina	Dudajev, Groznyj, Primakov
2021	Putinův, Rus, svaz, ukrajinský	Bělorusko, běloruský, Kreml, Lukašenko, Moskva, Putin, Rusko, ruský, sovětský, Ukrajina	Kyjev, Lavrov, Navalný, protiruský, Stalin

East	MBA	Both	Collocations
1996	Rus, sovětský, SSSR	Moskva, Rusko, ruský, Ukrajina	Primakov
2021	Putin, Rusko, sovětský, ukrajinský	Donbas, Moskva, ruský, Ukrajina	Kreml, Krym, Kyjev, pro-ruský, separatista, separatistický, Vladivostok

MBA vs. CA: weather

West	MBA	Both	Collocations
1996	jasno, mlha, teplota	oblačno, polojasno, přeháňka, studený, vzduch, zataženo	fronta, frontální, chladný, oblačnost, slunce, srážkový, tlak, vlhký, vydatný
2021	jaro	slunce, studený, tlak	bouřka, fronta, oblačnost, povodeň, vedro, vichřice

East	MBA	Both	Collocations
1996	–	–	frontální, oblačno, oblačnost, polojasno, přeháňka, slunce, studený, teplý, tlakový, vát, vzduch
2021	–	–	déšť, slunce, studený



Consistent associations over time

Companions

- ▶ Words/collocations sharing the same frequency development through a period of time
- ▶ Applicable to homogeneous data covering specific events in politics, society... (e.g. monitor corpus)
- ▶ Peaks and valleys mirror the outside world
- ▶ Confirmatory use: extent to which two concepts relate in a time frame



Companions: Time-based associations

Words sharing the same frequency development through a period of time: is *Ukrajina* and *Rusko* (periods: 2/2020–2/2021 and 2/2022–2/2023)?

Before war:



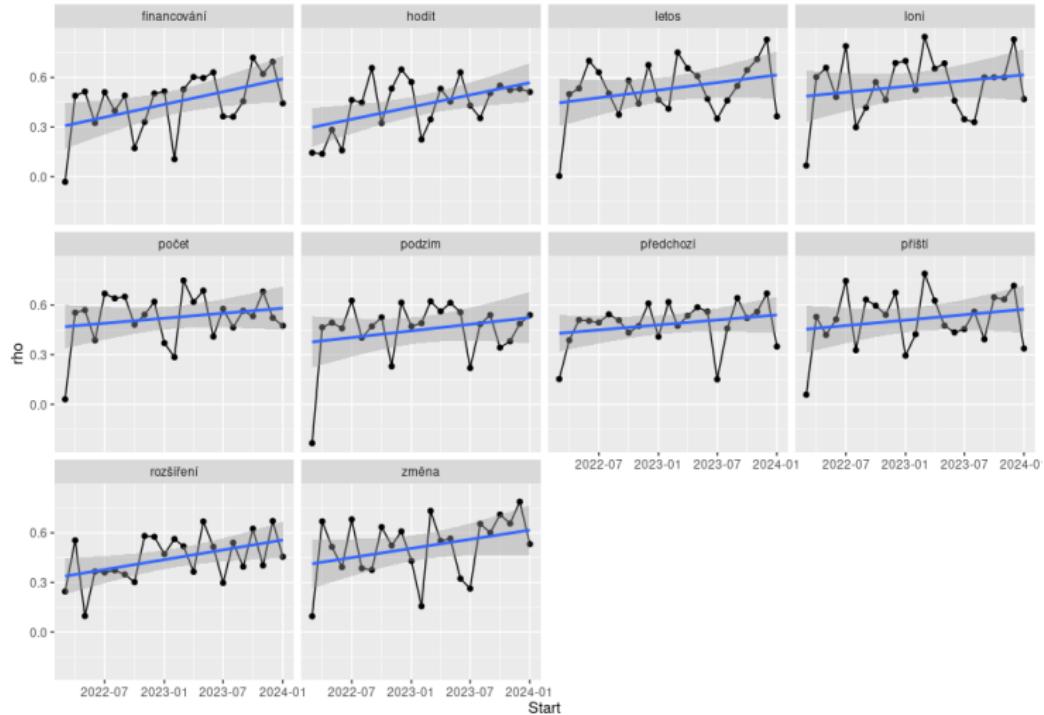
1st year of war:



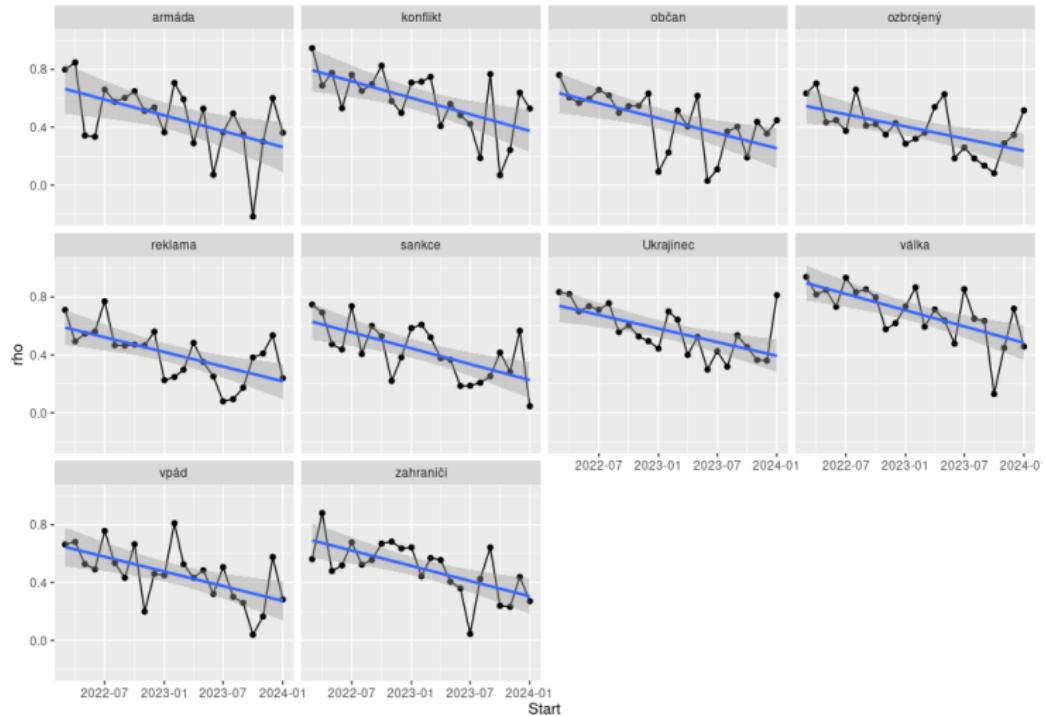
Correlation: $\rho = 0.076$

Correlation: $\rho = 0.846$

Top rising associations in MS with Ukraine (2/2022–2/2024)



Top falling associations in MS with Ukraine (2/2022–2/2024)





Summary

Summary

- ▶ isolated words are of limited use, we need broader context
- ▶ collocations tell only a half of the story → **associations**
- ▶ **MBA**: based on co-occurrence in texts
- ▶ **Companions**: based on co-incidence in time
- ▶ associations help reveal unexpected links between concepts/topics

References

- Baker, P., & McEnery, T. (2005). A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Journal of language and politics*, 4(2), 197–226.
- Clarke, I., McEnery, T., & Brookes, G. (2021). Multiple Correspondence Analysis, newspaper discourse and subregister: A case study of discourses of Islam in the British press. *Register Studies* 3(1). 144–171.
- Cvrček, V., & Fidler, M. (2021). By their associations you will recognize them: Using Market Basket Analysis to probe “alternative” framing of events. *Corpus Linguistics Conference 2021*, Limerick.
- Cvrček, V., & Fidler, M. (2022). No Keyword is an Island: In search of covert associations. *Corpora*, 17(2). 259–290
- Cvrček, V., & Fidler, M. (2023): Identifying political orientation among Czech media classes using Market Basket Analysis. ICLC. Dusseldorf.
- Cvrček, V. & Fidler, M. (2024): From News to Disinformation: Unpacking a Parasitic Discursive Practice of Czech Pro-Kremlin Media. *Scando-Slavica*, 70(1), 32–54.
- Cvrček, V., & Procházka, P. (2020). *ONLINE: monitorovací korpus internetové češtiny*. Ústav Českého národního korpusu. FF UK. www.korpus.cz
- Cvrček, V., Jeziorský, T. & Henyš, J. (2022). *ONLINE2: monitorovací korpus internetové češtiny*. Ústav Českého národního korpusu. FF UK. www.korpus.cz
- Egbert, J. & Biber, D. (2019): Incorporating text dispersion into keyword analyses. *Corpora* 14(1), 77–104.
- Fidler, M. & Cvrček, V. (2019): Keymorph analysis, or how morphosyntax informs discourse. *Corpus Linguistics and Linguistic Theory* 15(1), p. 39–70.
- Fidler, M., & Cvrček, V. (2020). Anti-system web portals and their network of meaning: A corpus-based approach in Czech. *AATSEEL Conference 2020*, San Diego.
- Fidler, M., & Cvrček, V. (2024): Zone-flooding as a discursive strategy of anti-system news portals.
- Gries, S. Th. (2024): *Frequency, Dispersion, Association, and Keyness*. Revising and tupleizing corpus-linguistic measures John Benjamins.
- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Miner, G. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.
- Partington, A. (2004). Corpora and discourse, a most congruous beast. *Corpora and discourse*, 11–20.
- Scott, M. (2010): Problems in investigating keyness, or clearing the undergrowth and marking out trails... In M. Bondi and M. Scott (eds.): *Keyness in Texts*, pp. 43–58. Amsterdam/Philadelphia: John Benjamins.
- Scott, M., & Tribble, C. (2006). *Textual Patterns: Key words and corpus analysis in language education*. John Benjamins.
- Singer, P. W. & Brooking, E. T. (2018). *LikeWar: The weaponization of social media*. Eamon Dolan Books.
- Šlerka, J. (2018). *Typologie domácích zpravodajských webů*. Nadační fond nezávislé žurnalistiky.
<https://www.nfnz.cz/studie-a-analyzy/typologie-domacich-zpravodajskych-webu/>.

Thank you for your attention

