Using Register Analysis to Establish Style Markers in Dictionaries

Václav Cvrček

27th June 2023

Introduction

Register evaluation in lexicon?

Language corpora as a source of information for

- meaning
- typical context / collocations
- frequency
- …
- why not style/usage markers?

Style/usage markers/labels

- variety – standard vs. non-standard (× Homoláč & Mrázková 2014)
- communicative situations – corpus metadata (text types, genres)
  - text types contain mixture of styles (cf. newspaper)
  - what about corpora without metadata? (web-crawled, Baroni et al. 2009; Benko 2014; Jakubíček et al. 2013; Davies 2018)
- *register affiliation*

Outline of the talk

1. multi-dimensional analysis and establishing registers
2. how to measure affinity of a lexeme and register
3. empirical verification
    - extraction of register-typical words
    - verification of established labels (case of ASSČ)

Multi-dimensional analysis

Principles of multi-dimensional analysis (MDA)

Biber 1995; Biber & Conrad 2009

- model of systemic & *functional* variability (× random, sociolinguistic…)
  - motivated by context & situation
- text production process involves *interrelated choices*
- dimensions of variation (“intratextual” perspective)
- establishing registers (clusters of text with similar linguistic features)
- model of register variation can be used for evaluation of texts (“additive” MDA)

Methodology of MDA

Establishing MD model:

1. corpus compilation
2. features: operationalization & extraction
3. statistical evalaution: factor analysis → *dimensions*
4. interpretation of results (labelling dimensions)
5. clusters of texts → **registers**

MDA of Czech

CNC: MDA team

MDA of Czech

Mini-portal **https://www.korpus.cz/mda**

Cvrček, V. – Komrsková, Z. – Lukeš, D. – Poukarová, P. – Řehořková, A. – Zasina, A.J. (2021): *From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA*. Corpus Linguistics and Linguistic Theory 17(2), p. 351-382.

Cvrček, V. – Komrsková, Z. – Lukeš, D. – Poukarová, P. – Řehořková, A. – Zasina, A. J. (2018): *Variabilita češtiny: multidimenzionální analýza*. Slovo a slovesnost 79, (p. 293–321).

Cvrček, V. - Laubeová, Z. - Lukeš, D. - Poukarová, P. - Řehořková, A. - Zasina, A. J. - Benko, V. (2020): *Comparing web-crawled and traditional corpora*. Language Resources & Evaluation 54, p. 713–745.

Cvrček, V. – Laubeová, Z. – Lukeš, D. – Poukarová, P. – Řehořková, A. – Zasina, A. J. (2020): *Registry v češtině*. Praha: Nakladatelství Lidové noviny, (233 p.).

MDA of Czech

- inspiration from English and other languages
- expected challenges / highlights of MDA…
  - … in Slavic languages – specific morphology, *inflection*, free word order
  - … in Czech – situation bordering on **diglossia** (Bermel 2014): Literary × Common Czech

Data: **Koditex** corpus

- “traditional” carefully designed corpus covering all available text types
- guiding principles: *diverse*, contemporary, *text length* control
  - text excerpts = **chunks** (not whole texts)
  - 3 modes – `wri`, `spo`, `web`
    - 8 divisions, 45 classes, ≈ 200,000 words per class

Category	#
Tokens	10,8 M
Words (excl. punct.)	9 M
Lemmata (types)	204 K
Text chunks	3 334

Koditex: composition

Features and their operationalization

Originally 140+ features, final list *122*, e.g.:

- **phonetics** – narrowing *é* > *í*, vowel breaking *ý* > *ej*, average word length…
- **morphology** – freq. of cases, numbers, moods, tenses…
- **derivation** – adjectives denoting similarity, verbal nouns, diminutives…
- **lexicon** – indefinite pronouns, reporting verbs, verbs of thinking, semantically bleached nouns…
- **pragmatics** – contact expressions, fillers, intensifiers, downtoners…
- **syntax** – types of attributes, clusters of POS, types of dependent clauses…
- **text/discourse** – questions, phraseology, word repetition…

Statistical evaluation: Factor analysis

- 122 features × 3292 text chunks
- *factor analysis*:
  - identifying latent factors
  - **R** environment, using `fa` function from **psych** package
  - parameters:
    - rotation: *promax* (oblique)
    - factoring method: *generalized weighted least squares* (GLS)
    - number of factors/dimensions: *8*
    - variance explained: *56 %*

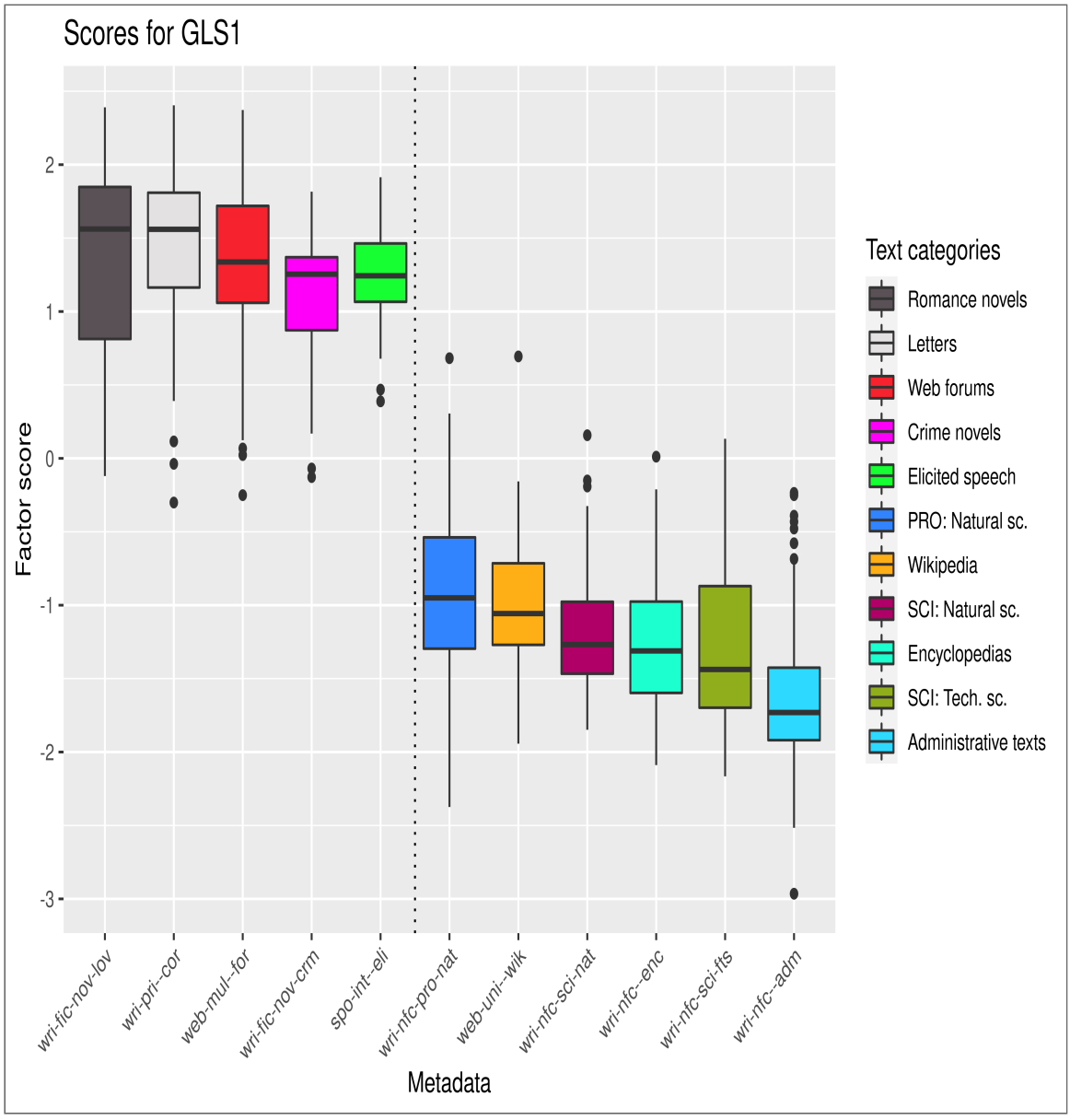
Factor analysis – output

- **loadings** – “correlations” of features and dimensions
  - participation of a feature on a dimension
- **factor scores** – positions of texts (chunks) within dimensions
  - linguistic characteristics of a text
- variance explained: *56 %*

Interpretation: Dimensions of variability

1. *dynamic* (+) × *static* (-): verbal/clausal × nominal/phrasal constructions
2. *spontaneous* (+) × *prepared* (-): hit-and-miss redundant coding × carefully worded formulations
3. *higher* (+) × *lower* (-) *level of cohesion*: propensity to use connecting devices and means of intratextual reference
4. *polythematic* (+) × *monothematic* (-): lexically rich × repetitive texts
5. *higher* (+) × *lower* (-) *amount of addressee coding*: explicit references to communication partners
6. *general* (+) × *particular* (-): description of general qualities × discussion of particular referents
7. *prospective* (+) × *retrospective* (-): present and future tense, non-narrative × past tense, narrative
8. *attitudinal* (+) × *factual* (-): degree of explicit epistemic certainty, higher × lower amount of hedging

Dim 1: dynamic (+) × static (-)



Features dim 1: dynamic (+) × static (-)

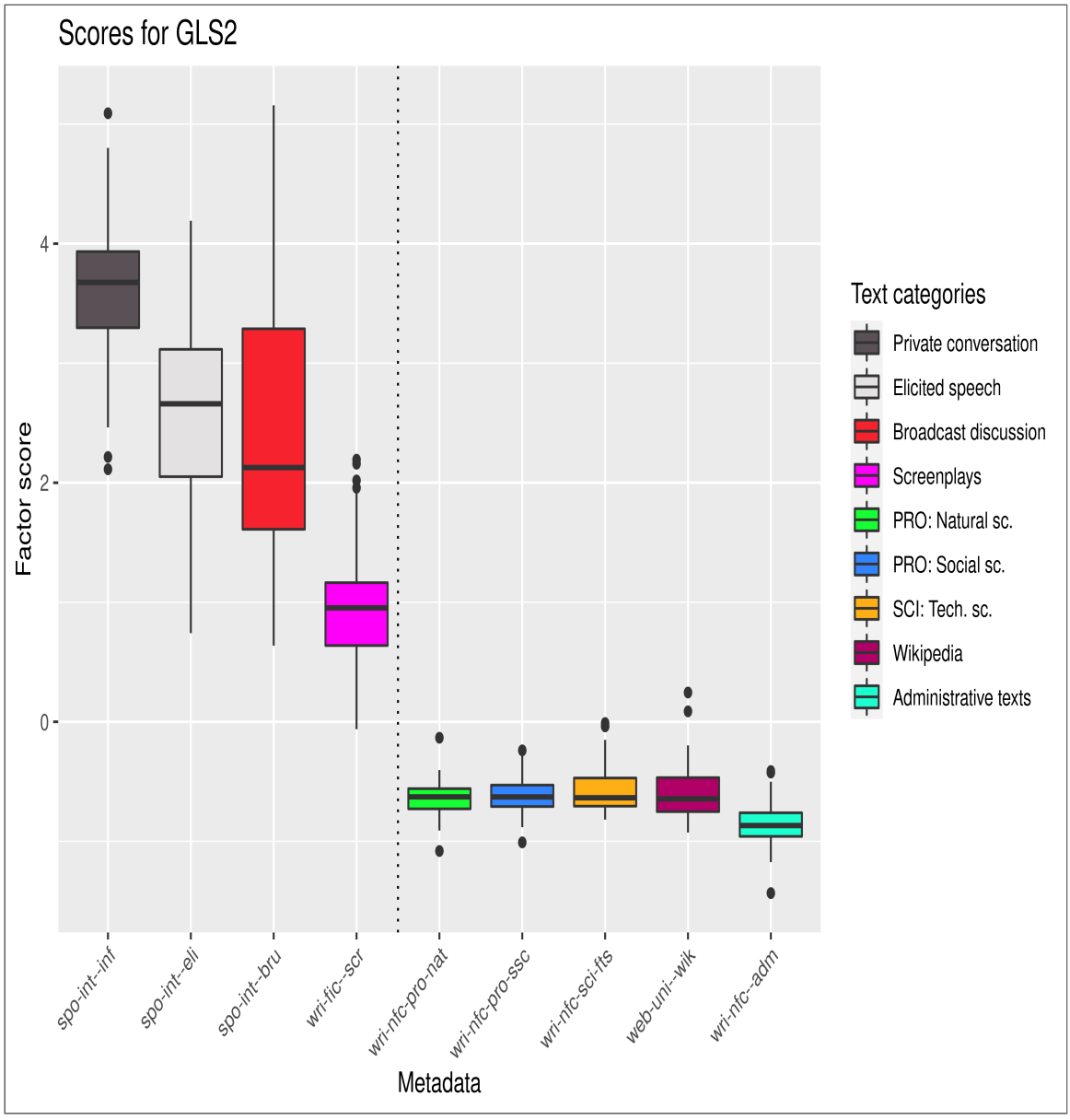
Positive loading features:

- past tense verbs, finite verbs, verbs: indicative forms, perfective verbs, 3rd person pronouns (personal + possessive), semantically bleached verbs, function words (pron., num., prep., conj. & part.), adverbs of time, pronouns, verbs: 1st person, verba dicendi, sentence negation…

Negative loading features:

- nominal post-modifiers without agreement, adjectives, abstract nouns, noun pre-modifiers with agreement, nouns: genitive, adjective clusters, noun clusters, clusters of same-case adjectives, nouns, average word length (number of syllables), verbal nouns complex prepositions…

Dim 2: spontaneous (+) × prepared (-)



Features dim 2: spontaneous (+) × prepared (-)

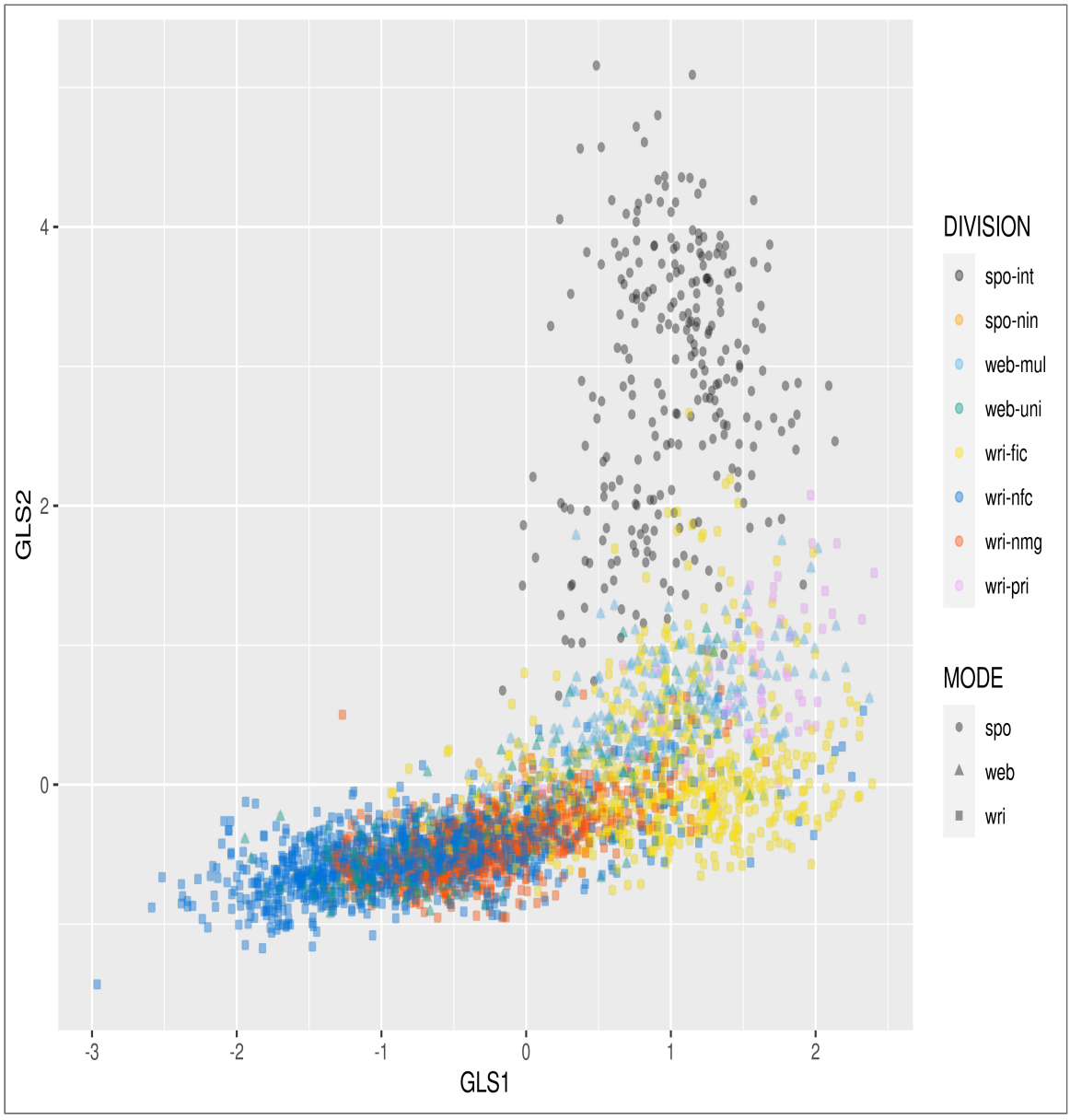
Positive loading features:

- features marking a) interactivity and online production (contact expressions, fillers, demonstratives, pronouns, word repetition) and b) informality (expressive particles, interjections) attract c) conventionalized non-standard Common Czech morphonological variants, symptomatic of diglossia

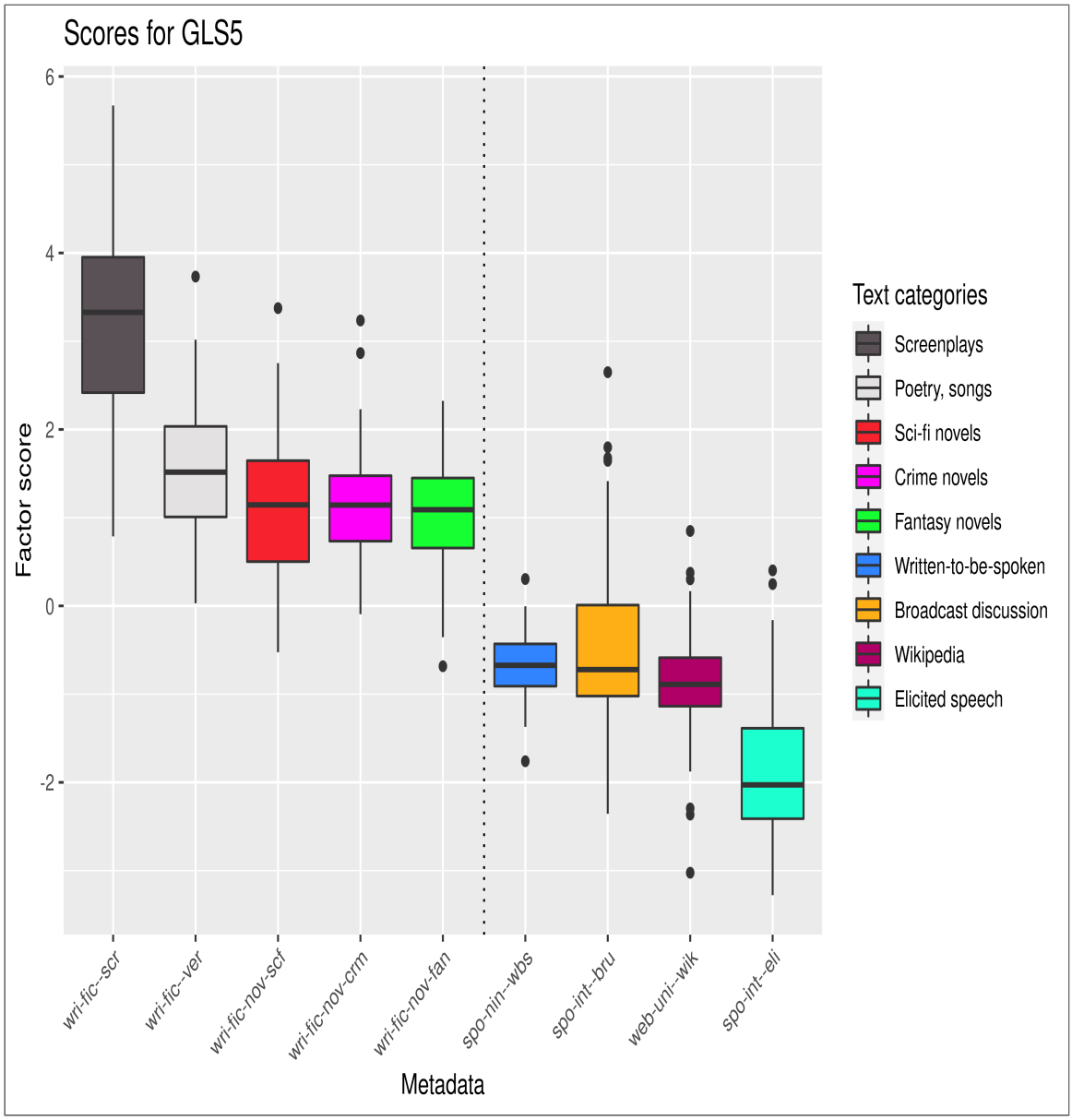
Negative loading features:

- both frequency and type inventory of prepositions, clauses with interrogative or relative adverbs, lexical richness (zTTR), nouns, longer words

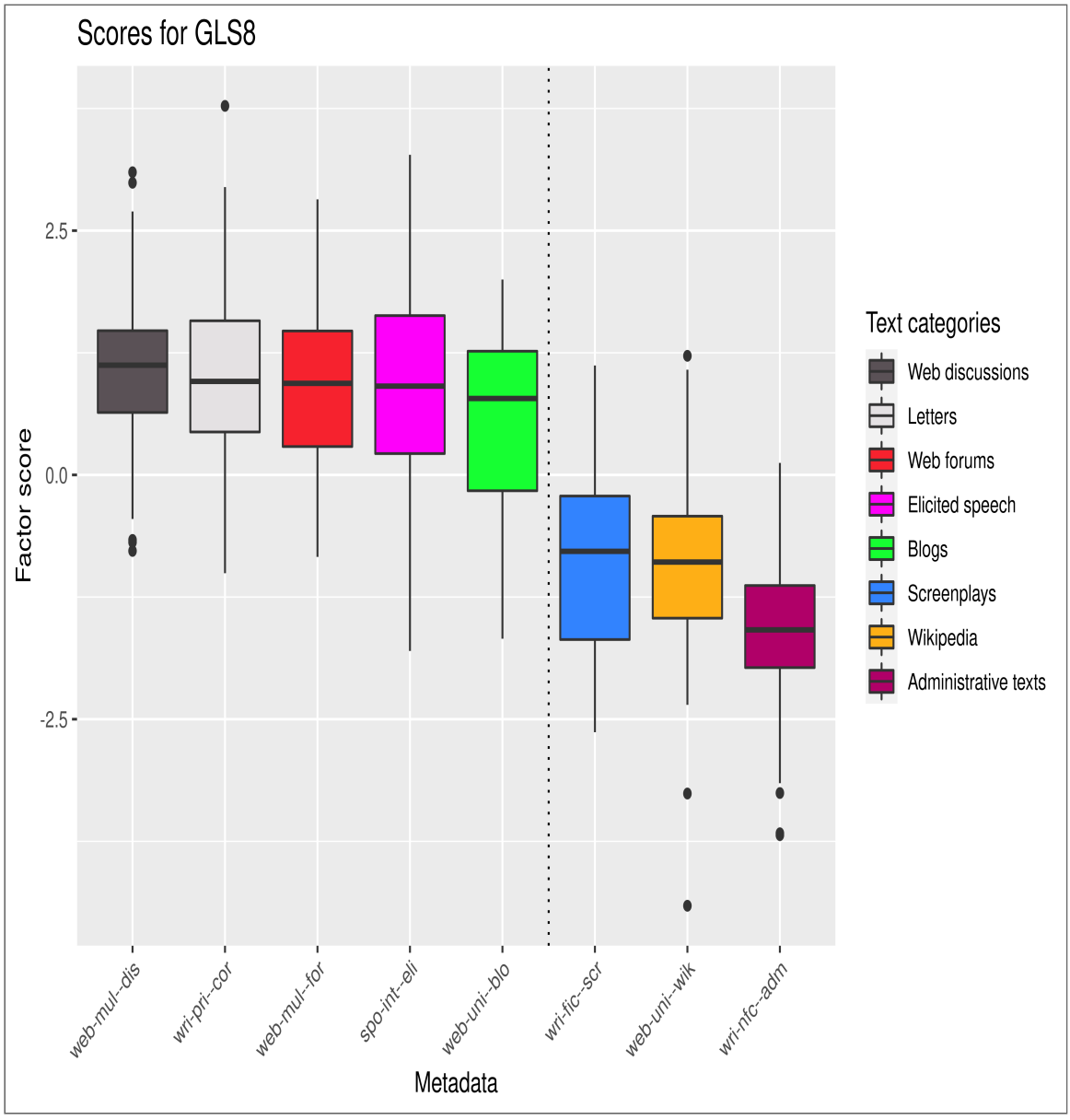
2D-plot: dim 1 and dim 2



Dim 5: higher (+) × lower (-) amount of addressee coding



Dim 8: attitudinal (+) × factual (-)



Applications of MDA

1. Projection on MD model

Additive MDA (cf. Berber Sardinha et al. 2019, 165):

- using previously established general model for new data
- texts from other corpora (e.g. web-crawled) can be analysed with the same set of features that were used for the original MDA
- relative frequencies of features can be converted to positions on dimensions

$\Rightarrow$ register information for any text of a language

2. Establishing registers

Register

- a variety defined by (a) set of distinct pervasive and functional linguistic features and (b) communicative situation
- operationalization: set (cluster) of texts sharing the same position on dimensions

Register classification

Static registers

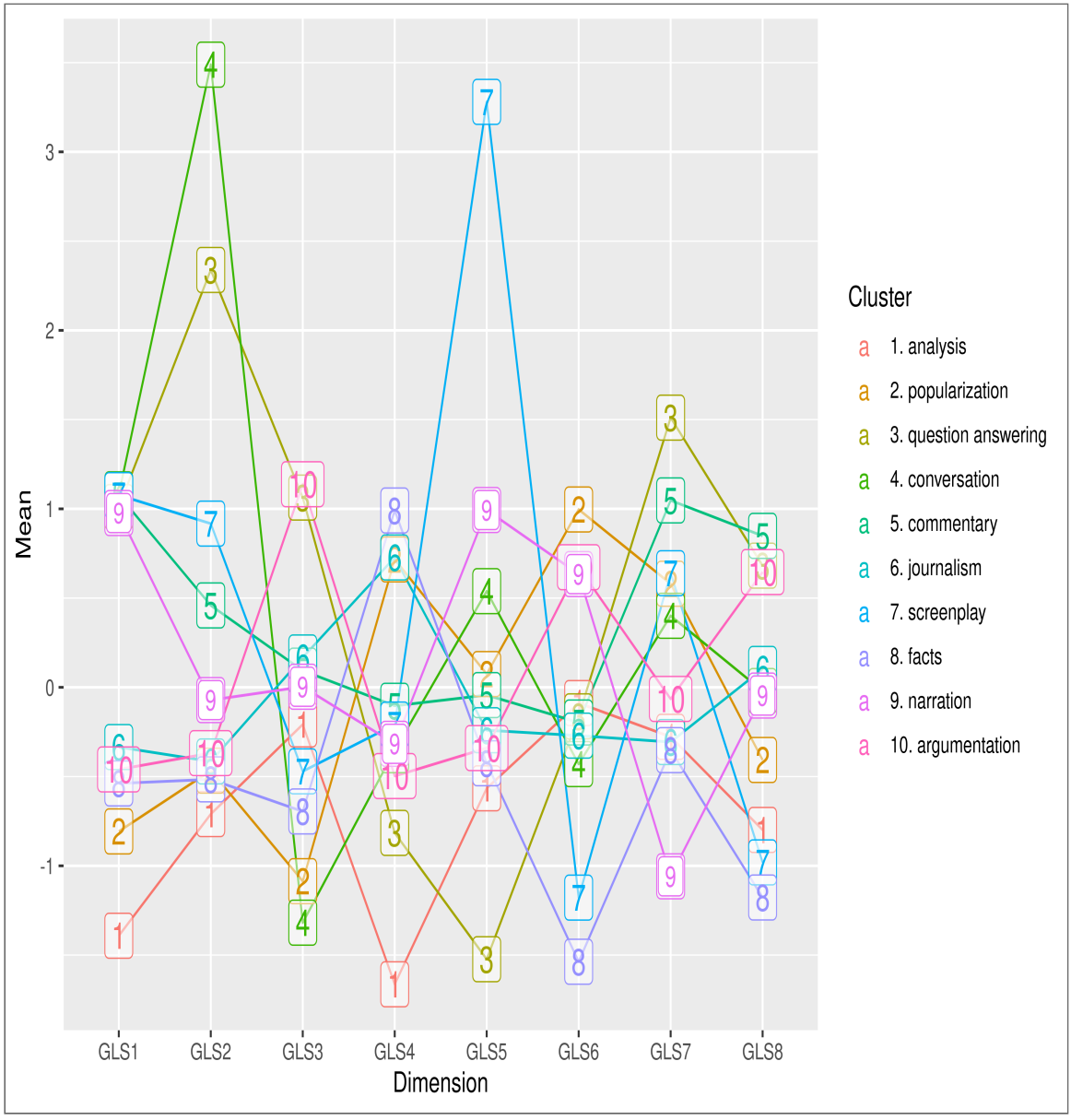
- analysis: static monothematic
- popularization: static polythematic general
- journalism: static mixed/indeterminate
- facts: static polythematic particular
- argumentation: static cohesive

Dynamic registers

- question answering: dynamic without addressee coding
- conversation: dynamic spontaneous
- commentary: dynamic attitudinal
- screenplay: dynamic with addressee coding
- narration: dynamic retrospective

Narration: dynamic retrospective register

Average position of texts within the narrative cluster (9)



Words in registers

Words in registers

Usage based approach to style labels:

- preference/affiliation/affinity of words to a register $\sim$ *association*

$\Rightarrow$ using frequency and association measures

Association with registers

Instead of co-occurrence of words *x* and *y* we work with word *x* and its presence in texts of register *R*:

$$\text{MI-score} = \log_2 \frac{f(xy) \times N}{f(x) \times f(y)} \Rightarrow \log_2 \frac{f(xR) \times N}{f(x) \times f(R)}$$

$$\text{logDice} = 14 + \frac{2f(xy)}{f(x) + f(y)} \Rightarrow 14 + \frac{2f(xR)}{f(x) + f(R)}$$

where:

- $f(xR)$ …occurrences of *x* in texts of register *R*
- $f(R)$ …number of tokens in texts of register *R*

Top 3 associations

lemma	Register	lemma	logDice	Regist
a ‘and’	analysis	kontrolér ‘inspector’	9.35	analy
v ‘in’	analysis	honitba ‘hunt’	9.24	analy
být ‘be’	analysis	skartační ‘shredding’	8.70	analy
ten ‘this’	question answering	se ‘filler’	10.32	quest
být ‘be’	question answering	svědek ‘witness’	9.50	quest
že ‘that’	question answering	svědčit ‘testify’	9.38	quest
a ‘and’	argumentativní	svěřenský ‘trust’	9.57	argun
být ‘be’	argumentativní	zatáčení ‘turning’	9.33	argun
se ‘refl. pron.’	argumentativní	přitažení ‘dragging’	9.23	argun

logDice: too common/indefinite

MI-score: too specialized

source: Koditex

Simple modification

Number of texts (chunks) instead of occurrences/frequencies

$$\text{MI-score} = \log_2 \frac{texts(xR) \times texts(corpus)}{texts(x) \times texts(R)}$$

$$\text{logDice} = 14 + \frac{2 \times texts(xR)}{texts(x) + texts(R)}$$

where:

- $texts(xR)$ …number of texts of register *R* containing *x*
- $texts(R)$ …number of texts of register *R*

Modified results

lemma	Register analysis	logDice
příslušný ‘relevant’	rezistor ‘resistor’	12.93
uvedený ‘mentioned’	vysílačka ‘walkie-talkie’	12.85
stanovený ‘stated’	tiskopis ‘print’	12.80
ee ‘filler’	ehm ‘filler’	13.28
tenhleten ‘that’	tentamto ‘this’	13.03
ňák ‘somehow’	nějak ‘somehow’	12.99
důsledek ‘consequence’	přitažení ‘dragging’	12.77
proces ‘process’	decentralizace ‘decentralizati	12.62
význam ‘meaning’	harmonizovaný ‘harmonized’	12.60

logDice: usable (?)

MI: still problematic

source: Koditex

Examples from a larger corpus

Applied to SYN2015 (100m corpus of written Czech)

Argumentation (static cohesive):

pojem ‘*concept*’, důsledek ‘*consequence*’, teorie ‘*theory*’, proces ‘*process*’, obecný ‘*general*’, jev ‘*phenomenon*’, daný ‘*given*’, určitý ‘*certain*’, princip ‘*principle*’, -li ‘*if*’, příklad ‘*example*’, hledisko ‘*viewpoint*’, aspekt ‘*aspect*’, předpoklad ‘*assumption*’, obecně ‘*general (adv.)*’

Screenplay (dynamic with addressee coding):

teda ‘*then*’, prominout ‘*sorry*’, prdel ‘*shit/ass*’, hele ‘*hey*’, jo ‘*yeah*’, sakra ‘*damn*’, kurva ‘*fuck/whore*’, kouknout ‘*look*’, viď ‘*see/right*’, dneska ‘*today*’ tvůj ‘*your*’, hm ‘*huh*’, aha ‘*oh*’, koukat ‘*stare*’, ahoj ‘*hi*’

Narration (dynamic retrospective):

zeptat ‘*ask*’, ty ‘*you*’, tvář ‘*face*’, dveře ‘*door*’, oko ‘*eye*’, rameno ‘*shoulder*’, slyšet ‘*hear*’, vlas ‘*hair*’, tvůj ‘*your*’, odpovědět ‘*answer*’, hlas ‘*voice*’, sedět ‘*sit*’, zvednout ‘*raise*’, tenhle ‘*this*’, usmát ‘*smile*’

Post-hoc verification of labels

Verification of established labels

Labels in *Akademický slovník současné češtiny* (Academic dictionary of contemporary Czech) – currently compiled at the CLI (words: A–G)

Usage labels in ASSC

Style characteristics:

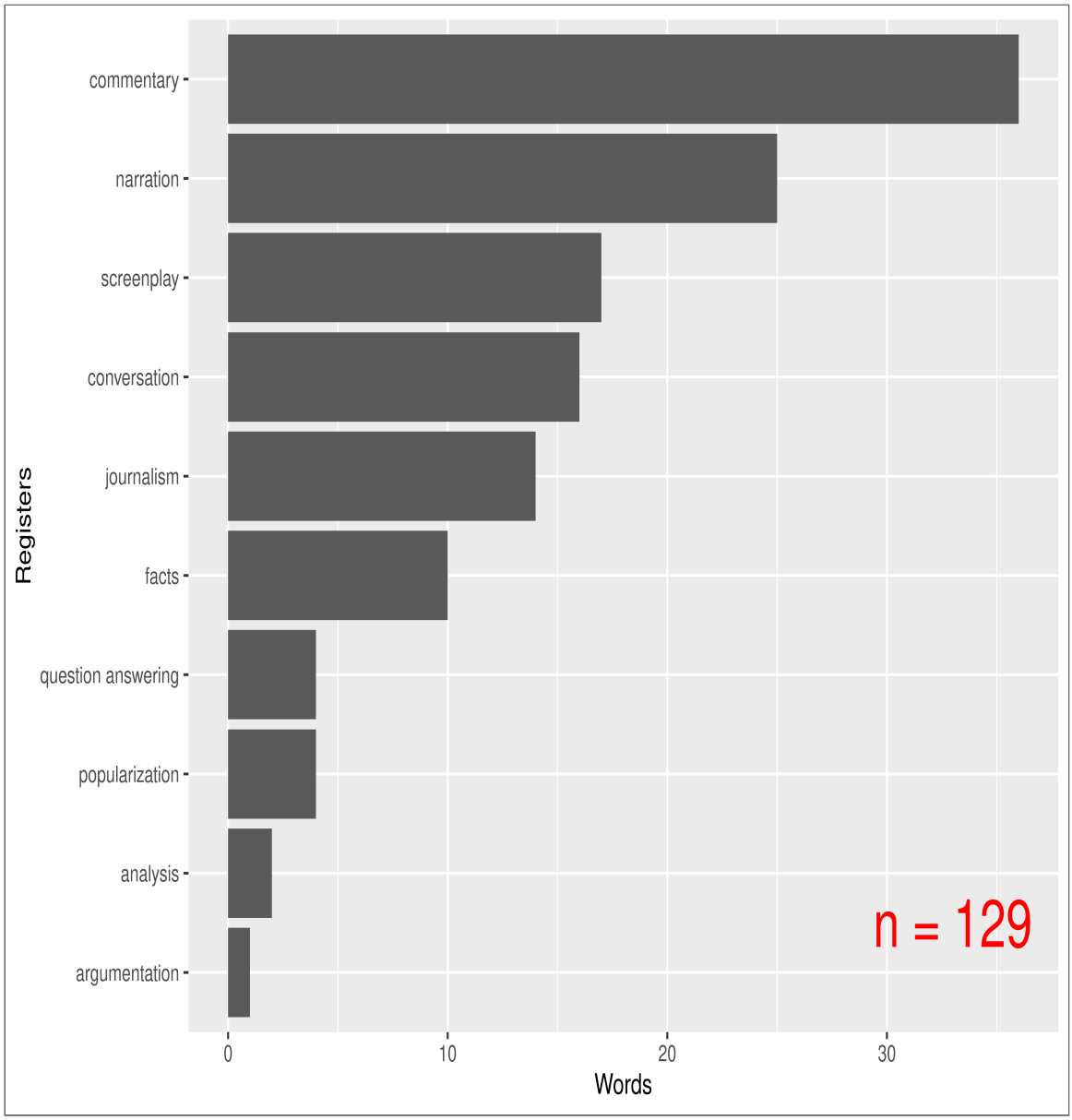
- colloquial
- colloquial with a tendency to become neutral
- slang and professional terms
- vulgar expressions (too few occ. in corpus)
- pejorative

RQ: Where can we typically find them?

Data: Koditex (SYN2015)

Colloquialisms

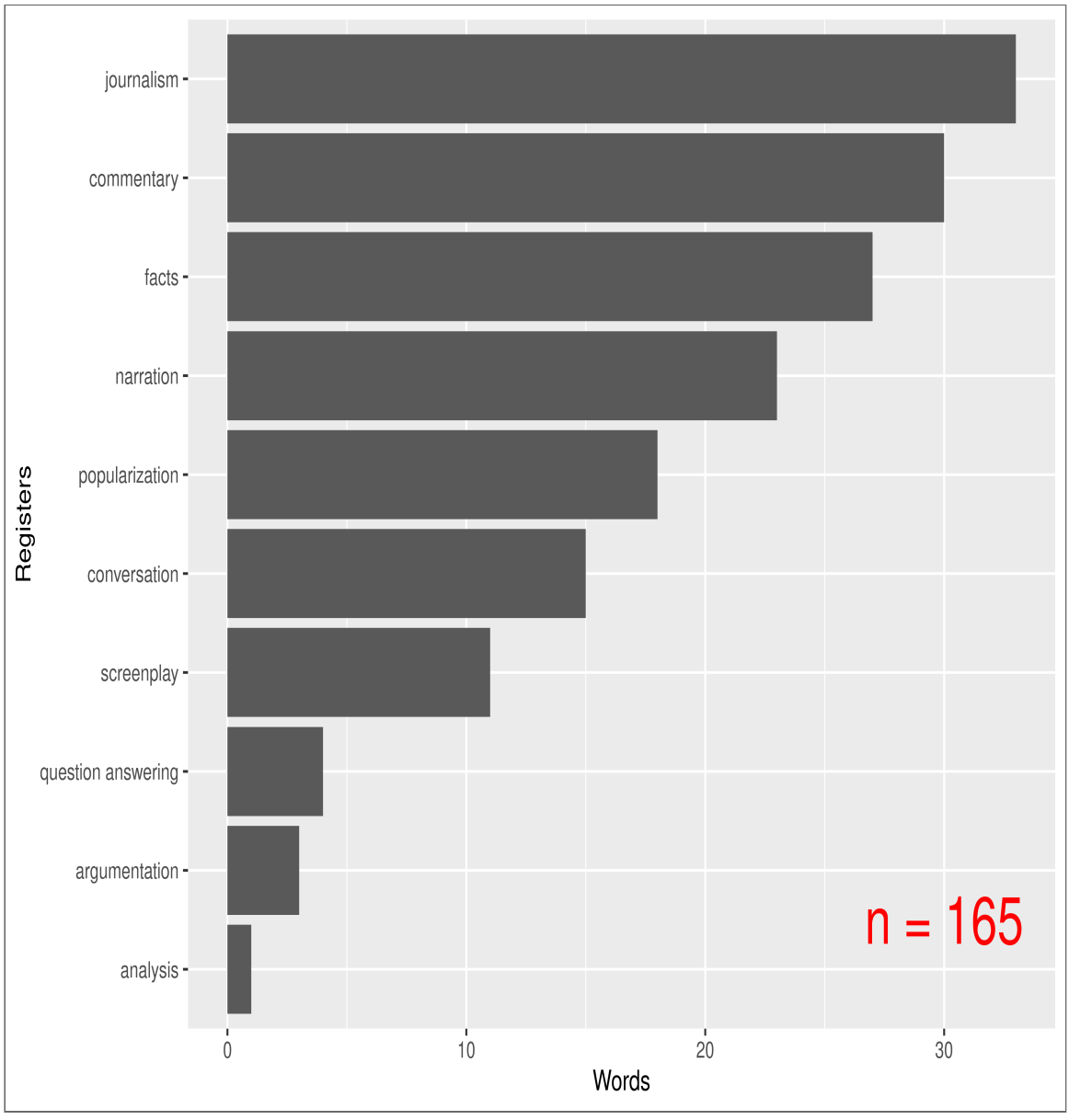
Which register has the strongest association with words labeled as `colloquial` (source: Koditex)?



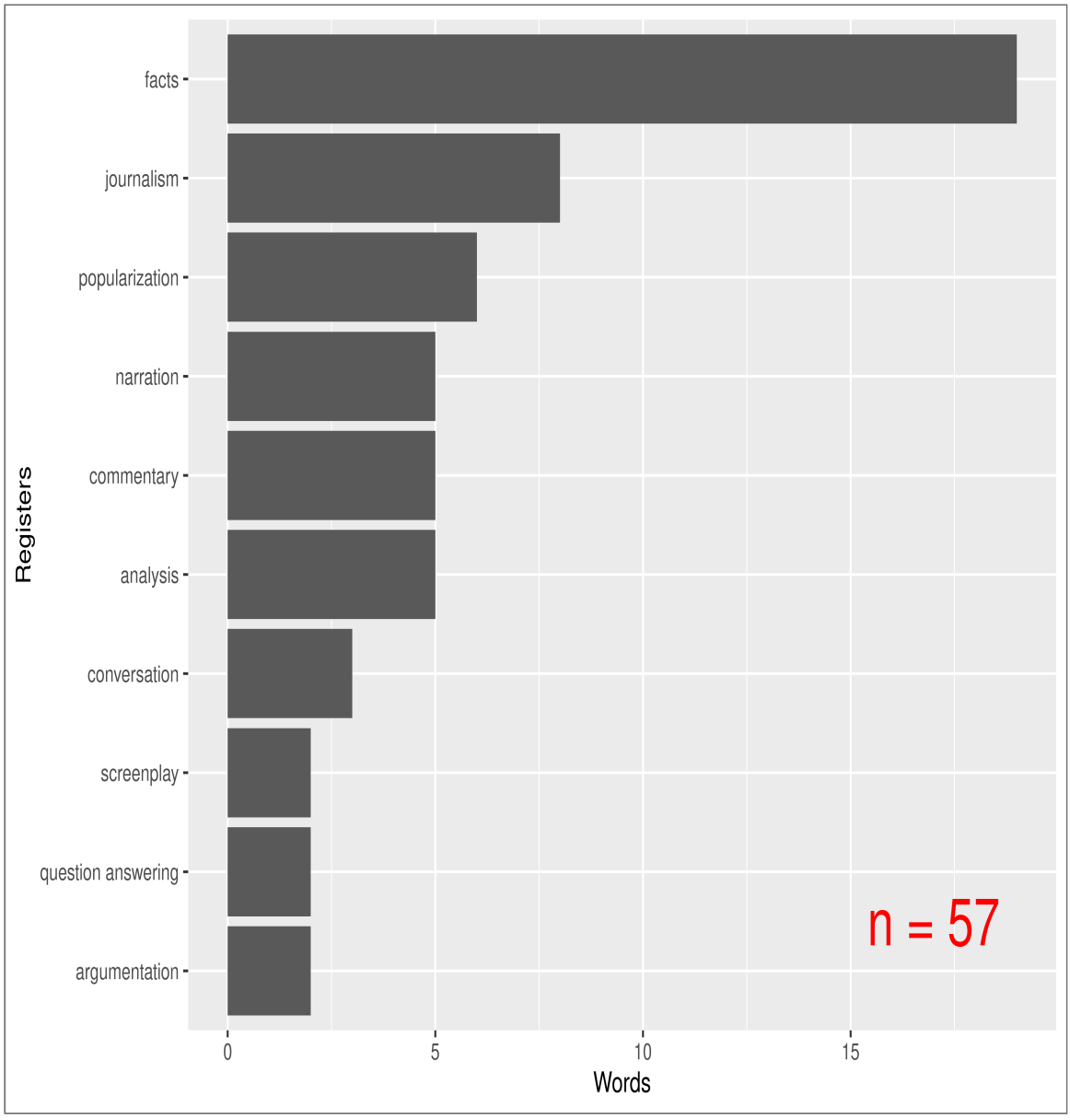
n = 129

Colloquialisms with a tendency to become neutral

Which register has the strongest association with words labeled as
`colloquial with a tendency to become ne`
(source: Koditex)?



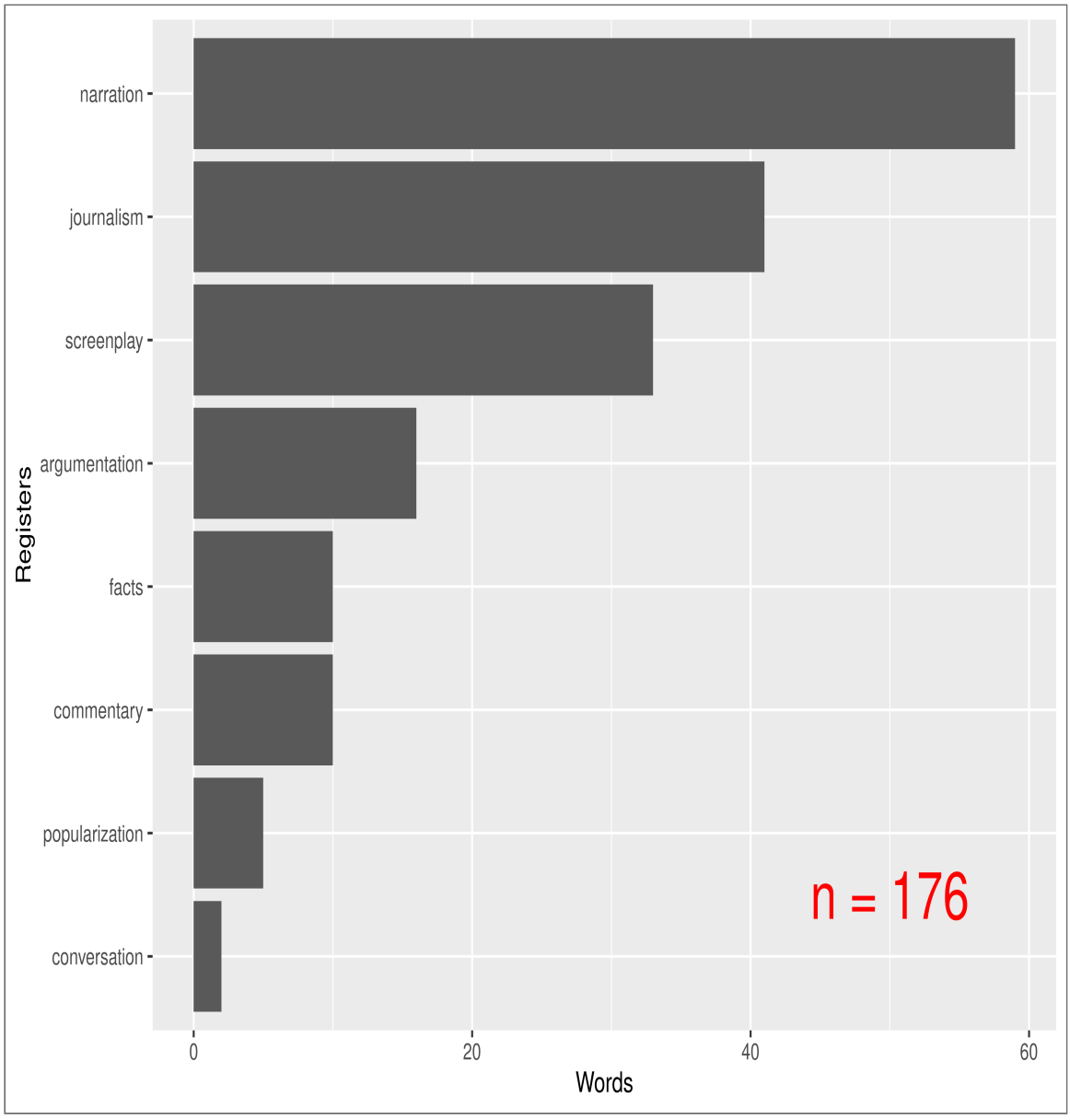
Slang and professional terms



(source: Koditex)

Pejorative expressions

Where are these expressions used in *SYN2015*?



Conclusions

Conclusions I

- What are usage labels?
  - warning signs or
  - descriptions of words' typical habitats
- Parallels with: prescriptive/proscriptive vs. descriptive approach
  - the former is demanded by the users
  - the later is preferred by (empirical) linguists

Conclusions II

Data-driven (descriptive) approach:

- text-linguistic approach (MDA) to variation → registers based on linguistic features → more objective labels
  - applicable to corpora without metadata (web-crawled, cf. Sharoff 2018)
- post-hoc check of the labels – where the words with label X live
- the concept of association can be extended from the lexical level (collocations) to other domains (although modifications are necessary)

References

- Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3), 209–226.
- Benko, V. (2014). Aranea: Yet another family of (comparable) web corpora. International Conference on Text, Speech, and Dialogue, 257–264. Springer.
- Berber Sardinha, T. & Veirano Pinto, M.,& Mayer, C., & Zuppardi, C. & Kauffmann, C. H. (2019). “Adding Registers to a Previous Multi-dimensional Analysis.” In T. B. Sardinha & M. Veirano Pinto (Eds.), Multi-dimensional analysis: Research methods and current issues (pp. 165-186). Bloomsbury Publishing.
- Bermel, N. (2014). Czech Diglossia: Dismantling or Dissolution? In J. Arokay, J. Gvozdanovic, & D. Miyajima (Eds.), Divided Languages? Diglossia, Translation and the Rise of Modernity in Japan, China, and the Slavic World (1st ed., pp. 21–37). Dordrecht: Springer International Publishing.
- Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4), 243–257.
- Biber, D. (1995). Dimensions of Register Variation: A Cross-Linguistic Comparison. Cambridge, England: Cambridge University Press.
- Biber, D., & Conrad, S. (2009). Register, Genre, and Style. Cambridge, England: Cambridge University Press.
- Biber, D., & Egbert, J. (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis.

Journal of English Linguistics, 44(2), 95–137.
- Cvrček, V. et al. (2021). From extra- to intratextual characteristics: Charting the space of variation in Czech through MDA. Corpus Linguistics and Linguistic Theory 17(2), s. 351-382.
- Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J., & Benko, V. (2020). Comparing web-crawled and traditional corpora. Language Resources & Evaluation 54, p. 713–745.
- Davies, M. (2018). The 14 Billion Word iWeb Corpus. Retrieved from **https://www.english-corpora.org/iweb/**
- Homoláč, J. & Mrázková, K. (2014): K stylistickému hodnocení jazykových prostředků, zvláště lexikálních. Slovo a slovesnost 75(1), 3-38.
- Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The tenten corpus family. 7th International Corpus Linguistics Conference CL, 125–127.
- Revelle, W. (2018). **psych: Procedures for Psychological, Psychometric, and Personality Research**.
- Sharoff, S. (2018). Functional Text Dimensions for the annotation of web corpora. Corpora, 13(1), 65–95.