Multi-dimensional analysis of Czech

Václav Cvrček

Zuzana Komrsková

David Lukeš

Petra Poukarová

Anna Řehořková

Adrian Jan Zasina

24th September 2018

CZECH NATIONAL CORPUS

Multi-dimensional analysis

Principles of multi-dimensional analysis (MDA)

- systemic & *functional* variability (× random, sociolinguistic…)
- motivated by context & situation
- *registers × genres* (~ intratextual × extratextual perspective)

- text production process involves *interrelated choices*

  → multi-dimensional approach seems a good fit (Biber 1991; Biber & Conrad 2009)

- 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

Methodology

1. corpus compilation
2. features: operationalization & extraction
3. statistical analysis (**factor analysis**, FA)
4. interpretation of results

Data

Koditex corpus

- guiding principles: *diverse*, contemporary, *text length* control
  - "diversified" stratified sampling
  - after 1990, majority from 2007–2014
  - text excerpts = **chunks** (not whole texts)
- annotation: lemmas, tags, multi-word unit & named-entity recognition
- tools: KonText, MorphoDiTa, NameTag
- 10 million words, 3,334 text chunks, 2000–5000 words each
- 3 modes – `wri`, `spo`, `web`
  - 8 divisions, 45 classes, $\approx$ 200,000 words per class
  - letters, administrative texts: shorter chunks (1000 words)
  - posts (`web-mul`): aggregated according to author and time
  - `spo`: one speaker within conversation

Features

Features and their operationalization I

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

- phonetics – narrowing *é > í*, diphthongization *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…

Features and their operationalization II

- pragmatics – contact expressions, fillers, intensifiers, downtoners…
- syntax – types of attributes, clusters of POS, types of dependent clauses…
- text/discourse – questions, phraseology, word repetition…

Type-based features – inventories of pronouns, prepositions, conjunctions (relativized using **zTTR**, Cvrček & Chlumská 2015)

Lexical richness – Yule's K, **thematic concetration** (Popescu et al. 2007), unigrams & bigrams (zTTR)

Statistical evaluation

Factor analysis

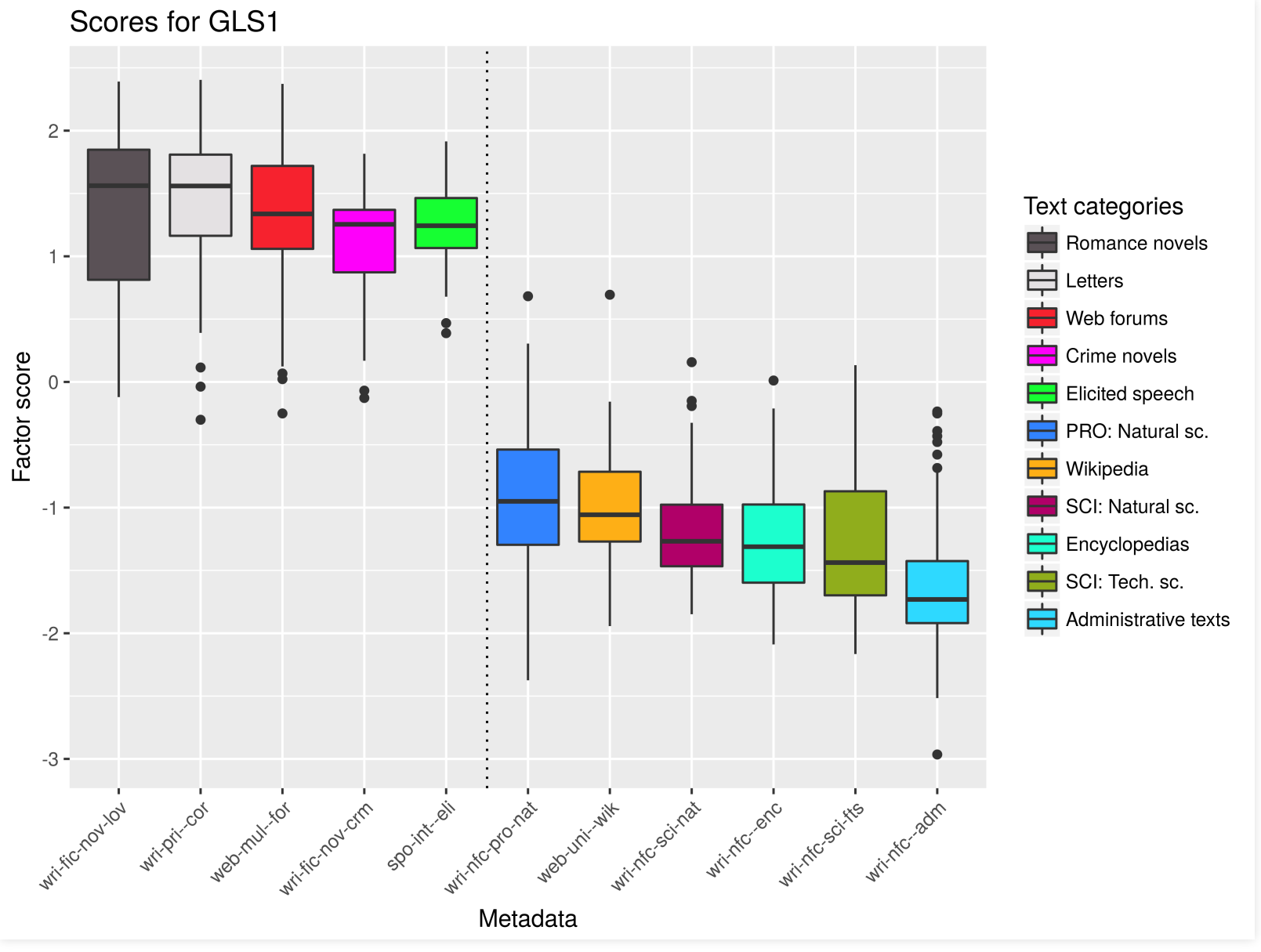
- *input*: 122 features × 3292 text chunks
- *factor analysis*:
  - R environment, using `fa` function from `psych` package
  - parameters:
    - rotation: promax (oblique)
    - factoring method: generalized weighted least squares (GLS)
    - number of factors/dimensions (!): 8
- *output*:
  - **loadings** – "correlations" of features and dimensions
  - **factor scores** – positions of texts (chunks) within dimensions
- variance explained: *56%*

Interpretation

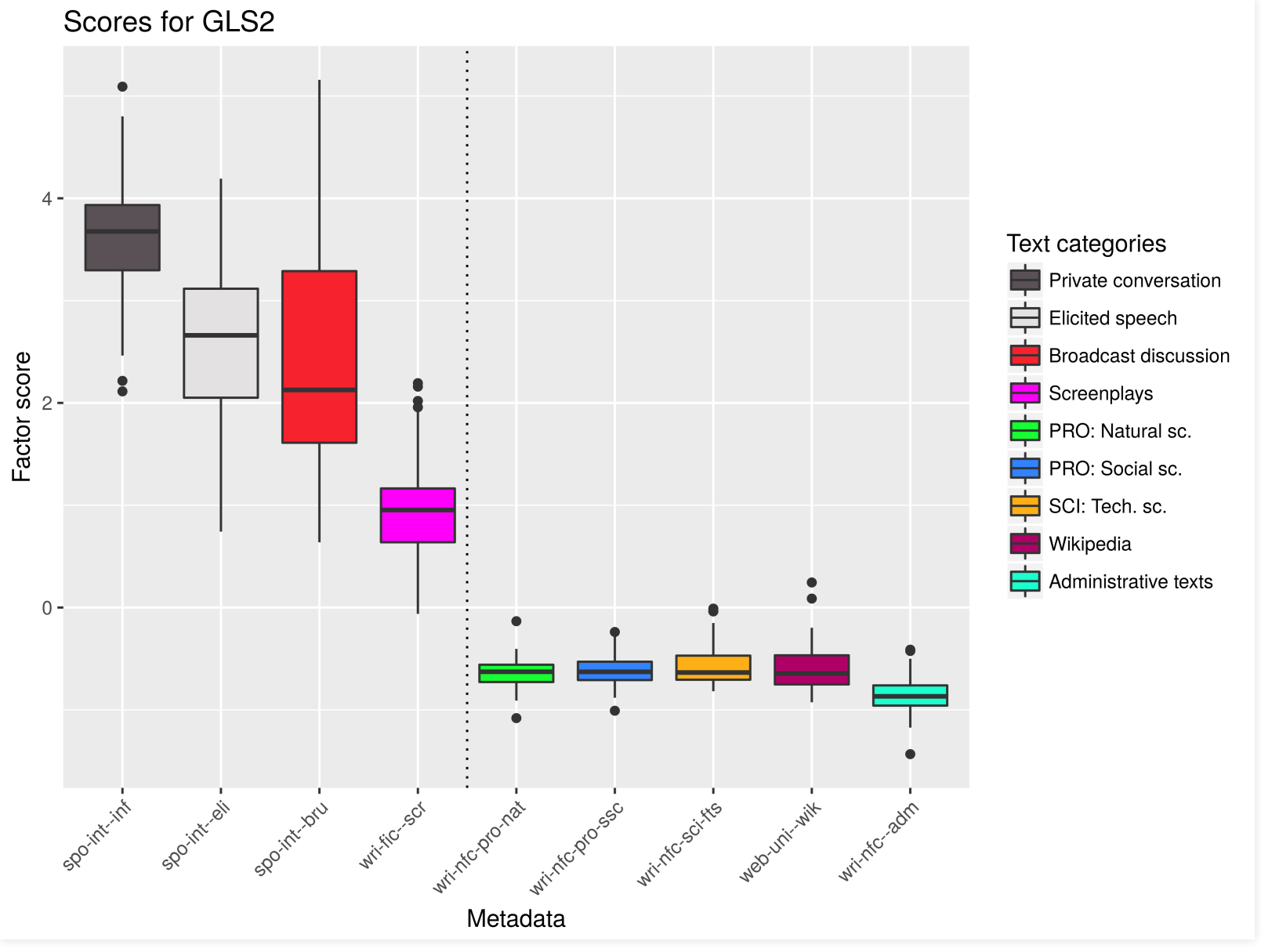
Dimensions of variability

1. *dynamic* (+) vs. *static* (-)
2. *spontaneous* (+) vs. *prepared* (-)
3. higher (+) vs. lower (-) level of cohesion
4. polythematic (+) vs. monothematic (-)
5. *higher* (+) vs. *lower* (-) *amount of addressee coding*
6. general/intension (+) vs. particular/extension (-)
7. prospective (+) vs. retrospective (-)
8. *attitudinal* (+) vs. *factual* (-)

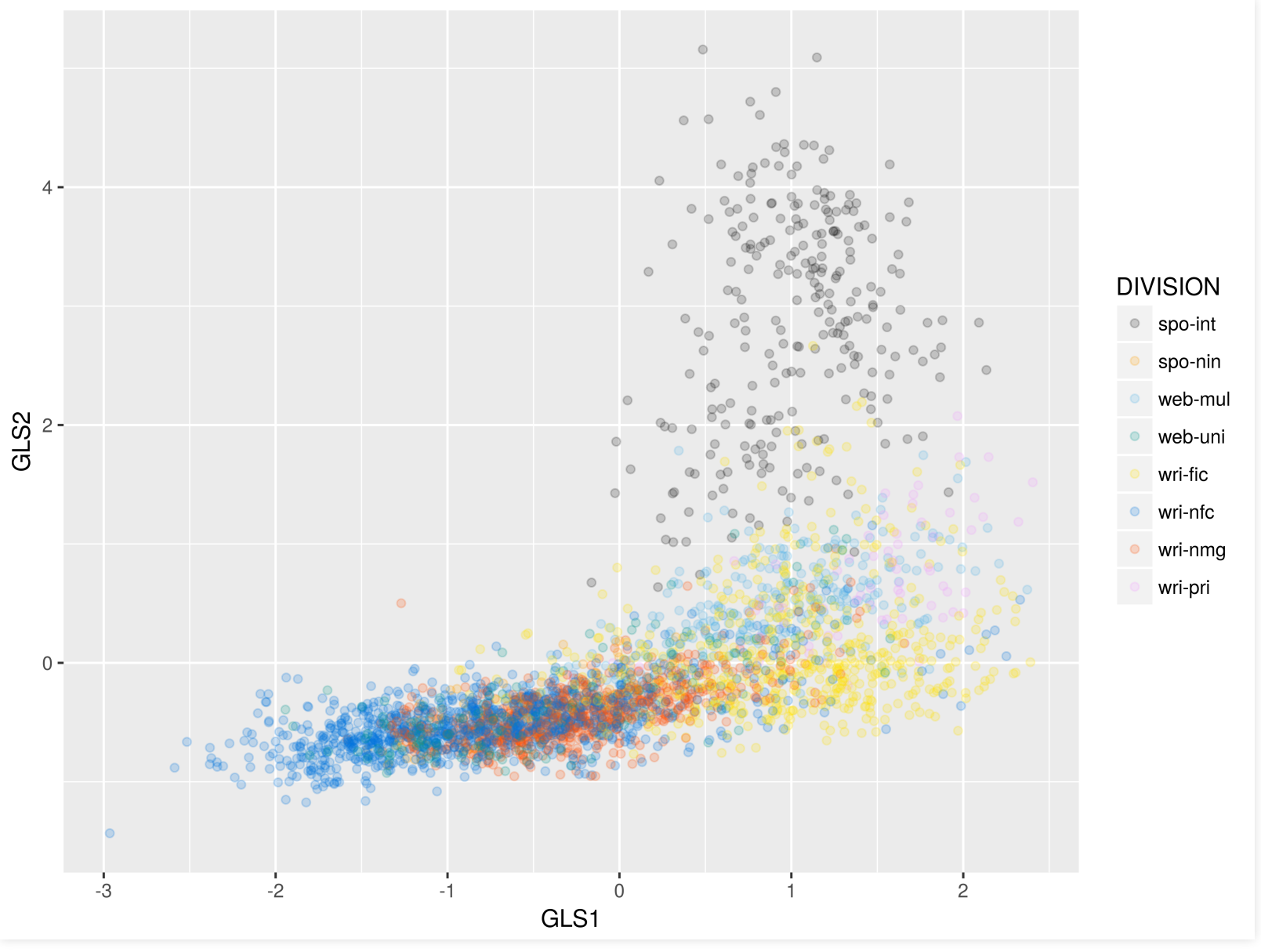
Dim 1: dynamic (+) × static (-)



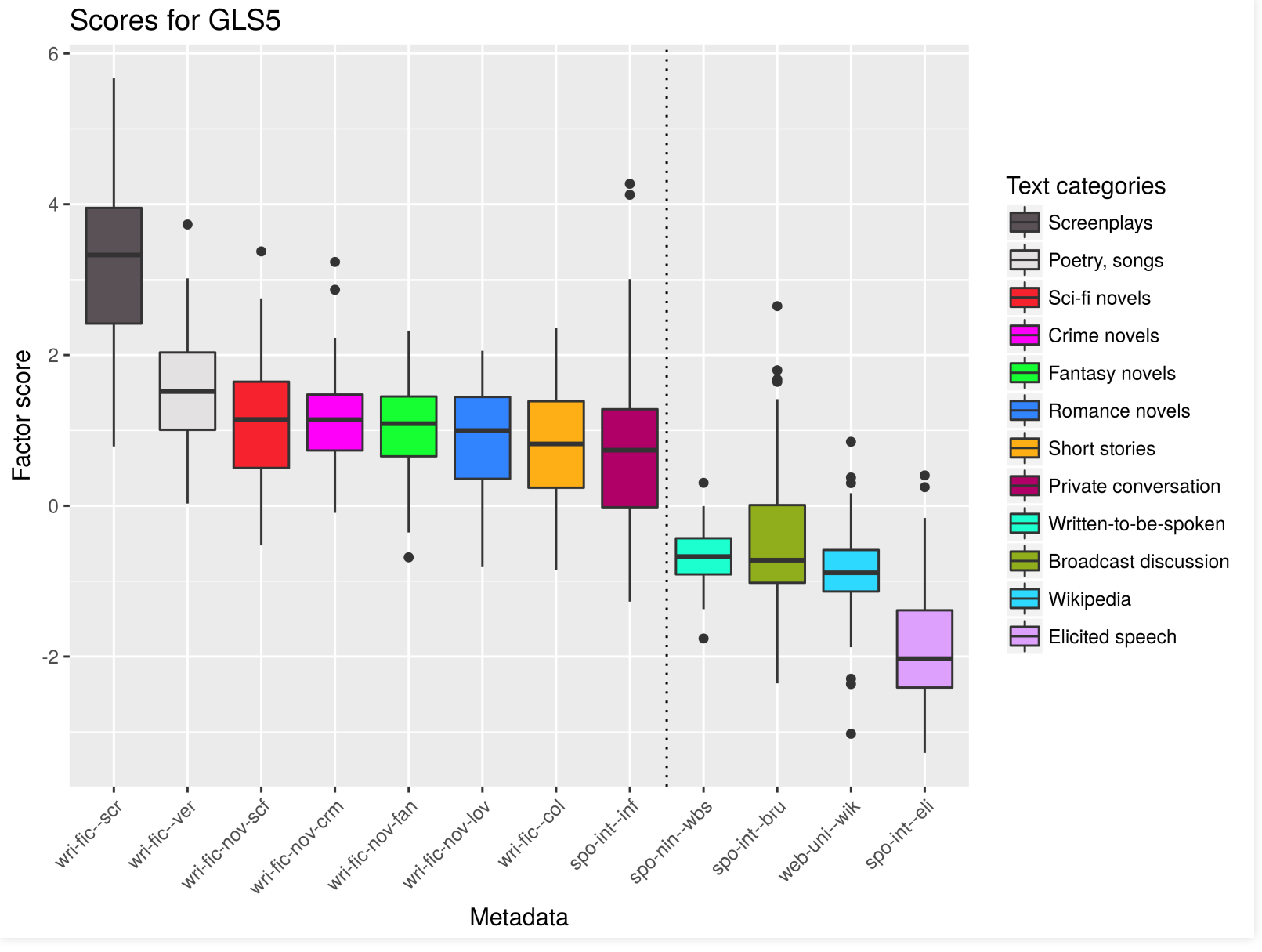
Dim 2: spontaneous (+) × prepared (-)



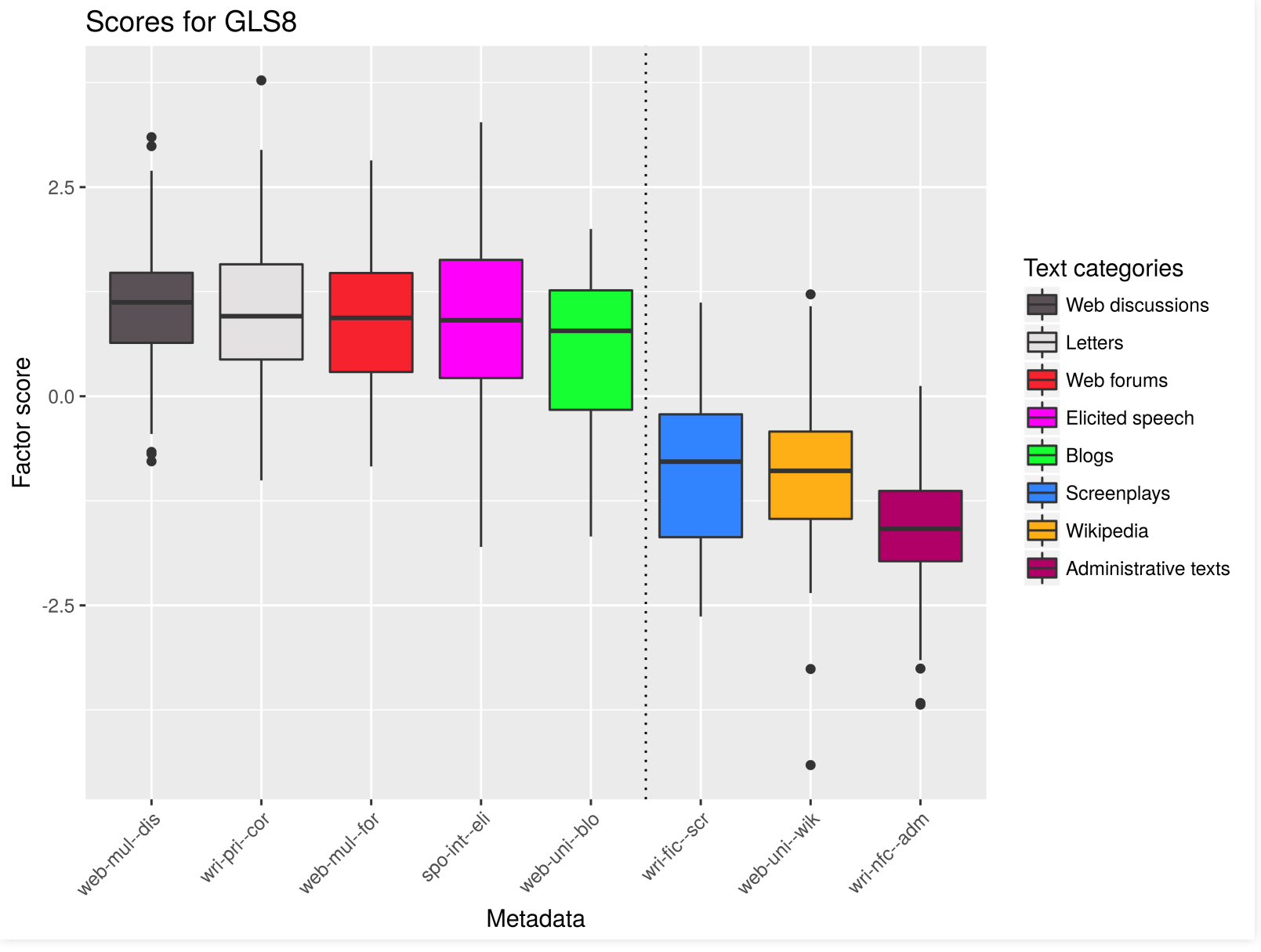
2D-plot: dim 1 and dim 2



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



Dim 8: attitudinal (+) × factual (-)



Final remarks

Conclusion

- Key features of Czech MD model:
  - "universal" dimensions (Biber 2014):
    - *clausal/oral–phrasal/literate* ≈ dynamic × static + spontaneous × prepared (← **diglossia**? spo data?)
    - *narrative–non-narrative* ≈ prospective × retrospective
  - "specific" dimensions:
    - polythematic × monothematic (← **type-based features**)
- Current work:
  - studying *elicited texts* from psychology through MD model
  - comparison of corpora (traditional vs. web-crawled)
- Future plans:
  - how *robust* is the MD model?
  - intratextual classification and annotation in CNC

References

- Bermel, N. (2014): Czech Diglossia: Dismantling or Dissolution? In J. Árokay et al. (eds), Divided Languages? Springer.
- Biber, D. (1991): Variation across speech and writing. Cambridge: Cambridge University Press
- Biber, D. & Conrad, S. (2009): Register, Genre, and Style. New York, NY: Cambridge University Press.
- Biber, D. (2014): Using multi-dimensional analysis to explore cross-linguistic universals of register variation. Languages in Contrast 14(1), 7–34.
- Cvrček, V. & Chlumská, L. (2015): Simplification in translated Czech: a new approach to type-token ratio. Russian linguistics 39(3), 309–325.
- Popescu, I., Best, K. & Altmann, G. (2007): On the dynamics of word classes in texts. Glottometrics 14, (p. 58–71).

Acknowledgments

This research was supported by the *ERDF* project **Language Variation in the CNC** no. CZ.02.1.01/0.0/0.0/16_013/0001758.

It builds upon work made possible by the **Czech National Corpus** project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of *Large Research, Development and Innovation Infrastructures*.

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION,
YOUTH AND SPORTS