CZECH NATIONAL
CORPUS

CZECH NATIONAL CORPUS

30+ years of corpus-based language variation studies. Experiences, challenges and inspirations

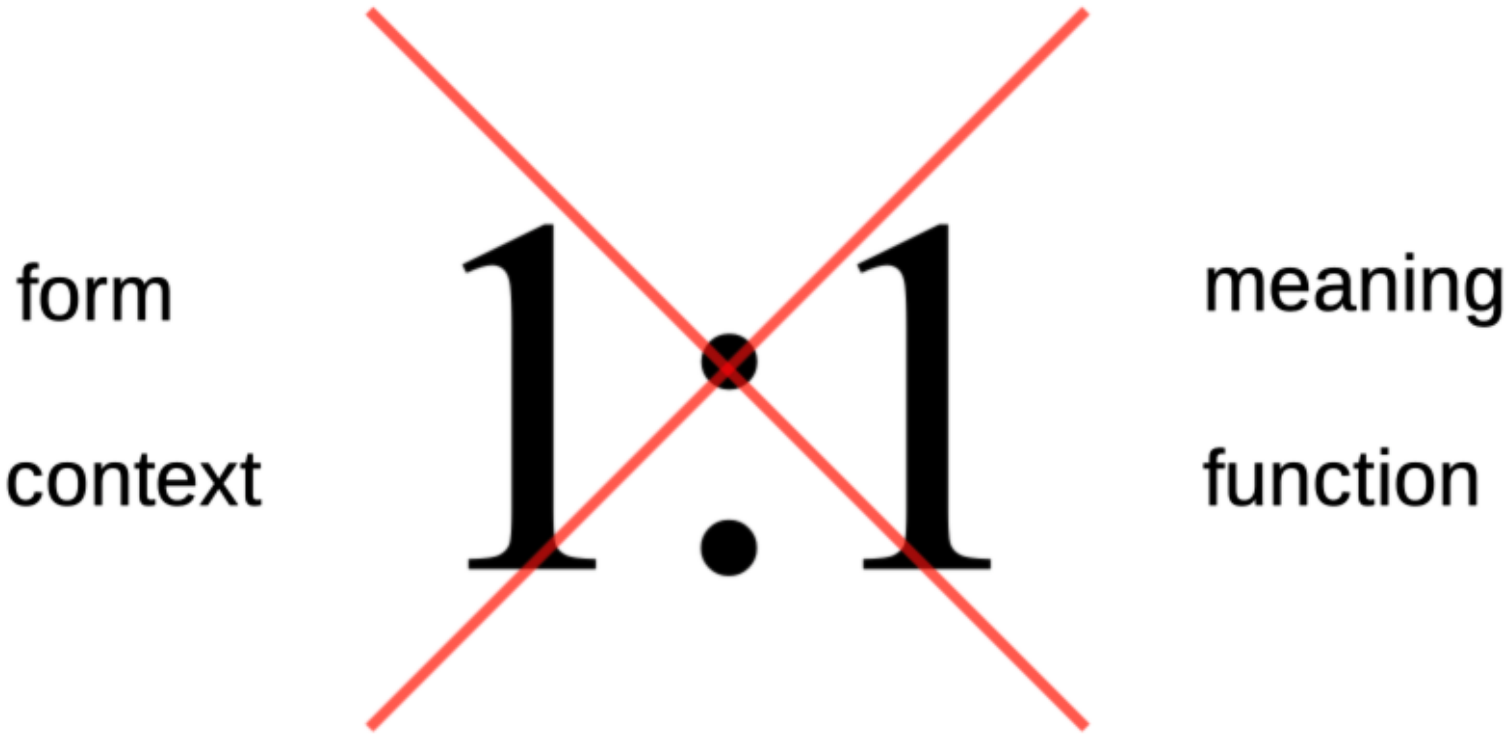
Václav Cvrček

Slovko 2019

Bratislava, October 24

1:1


form
1:1
meaning
context
function


form
context
1:1
meaning
function

Variation in language

Absence of 1:1 correspondence between form–function

- synonymy (more forms for one function)
  - *splendid* – *smashing*, *strong* – *powerful*
  - *robiť* – *drieť* (make, labour)
  - *lidma* – *lidmi* (people$^{Inst.pl.}$)

Variation in language

Absence of 1:1 correspondence between form–function

- synonymy (more forms for one function)
  - *splendid* – *smashing*, *strong* – *powerful*
  - *robiť* – *drieť* (make, labour)
  - *lidma* – *lidmi* (people$^{Inst.pl.}$)
- homonymy/polysemy (more functions of one form)
  - *stavení* (building$^{\{N,G,D,A,V,L\}sg.,\{N,G,A,V\}pl.}$)
  - *left* (leave, not right)

Variants of variation

Language levels

- phonology, morphematics – phonemes, morphemes
- morphology, derivation – indicators of variety
- lexicon, syntax – meaning/function
- text – register/style, sociolect

Perspectives

- synchronic (sociolinguistic, register)
- diachronic (dialectal)

Variation and linguistics

Variation and linguistics

Variation and linguistics

...isn't linguistics all about variability?

How do we cope with variation...

- ...by describing it – range & principles of variation (H. Kučera)

Variation and linguistics

...isn't linguistics all about variability?

How do we cope with variation...

- ...by describing it – range & principles of variation (H. Kučera)
- ...by searching for "invariant" (and ignoring v.) – langue $\times$ parole, corpus annotation (?)

Variation and linguistics

...isn't linguistics all about variability?

How do we cope with variation...

- ...by describing it – range & principles of variation (H. Kučera)
- ...by searching for "invariant" (and ignoring v.) – langue × parole, corpus annotation (?)
- ...by denying/fighting it – prescriptive tendencies
  - but N.B.: variation is natural & *all-pervasive in human language* (Ferguson 1983: 154, cit. Biber-Conrad 2009: 23)

Variation and linguistics

...isn't linguistics all about variability?

How do we cope with variation...

- ...by describing it – range & principles of variation (H. Kučera)
- ...by searching for "invariant" (and ignoring v.) – langue × parole, corpus annotation (?)
- ...by denying/fighting it – prescriptive tendencies
  - but N.B.: variation is natural & *all-pervasive in human language* (Ferguson 1983: 154, cit. Biber-Conrad 2009: 23)
- ...by studying it – variability on lower levels is used on higher ones (emphasises hierarchical nature of language)

Variation as a pointer

- "free variation" does not exist (in the long run)
  - alternative forms → functional (or semantic) differentiation
  - alternative meanings → formal (or contextual) differentiation
- if there is a variability ⇒ language will employ it
- variation is a *pointer* to a (hidden) *function* (usually on a higher level)

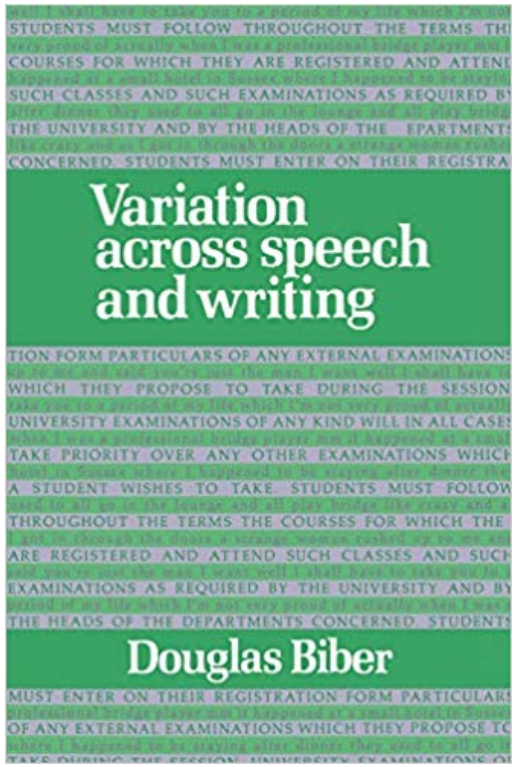
Variation and corpora

Corpus-based approaches to variation

- (annotation – lemmatization, tagging – as a way of coping with variability)
- variation is an empirical phenomenon *par excellence* – most of the variation cannot be captured by intuition
- finding invariant is parallel with searching for *pattern* ($\leftarrow$ very CL concept)
- $\Rightarrow$ frequency is crucial in describing variation (*SyD*, *Word at a Glance*)
- corpora are necessary for *identification* areas of variation as well as for describing their *principles*, *range* and *inventory*

30+ years of corpus-based...

Douglas Biber (1988): *Variation across speech and writing*. Cambridge: Cambridge University Press.



Variation in texts

Variability of texts

Invariant: information/message

Variability of texts

Invariant: information/message

Traditionally described by stylistics

- qualitative (what is general and what is specific?)
- absence of scaling (what is dominant and what is marginal?)

Two perspectives

Emphasised in CL approaches to text variation

- intratextual – *dough* – **register** (linguistic properties)
- extratextual – *cake* – **genre** (conventional categorization)

Multi-dimensional analysis (MDA)

Principles of MDA

Multi-dimensional analysis (Biber 1988; Biber & Conrad 2009)

- systemic & **functional** variability
- motivated by context & situation
- registers ($\sim$ intratextual) perspective
- assumption: text production involves **interrelated choices** $\rightarrow$ groups of features $\rightarrow$ dimensions of variation
- what is used, how often and together with what (bottom-up empirical approach)

Methodology of MDA

1. corpus compilation

Methodology of MDA

1. corpus compilation
2. list of features

Methodology of MDA

1. corpus compilation
2. list of features
3. operationalization

Methodology of MDA

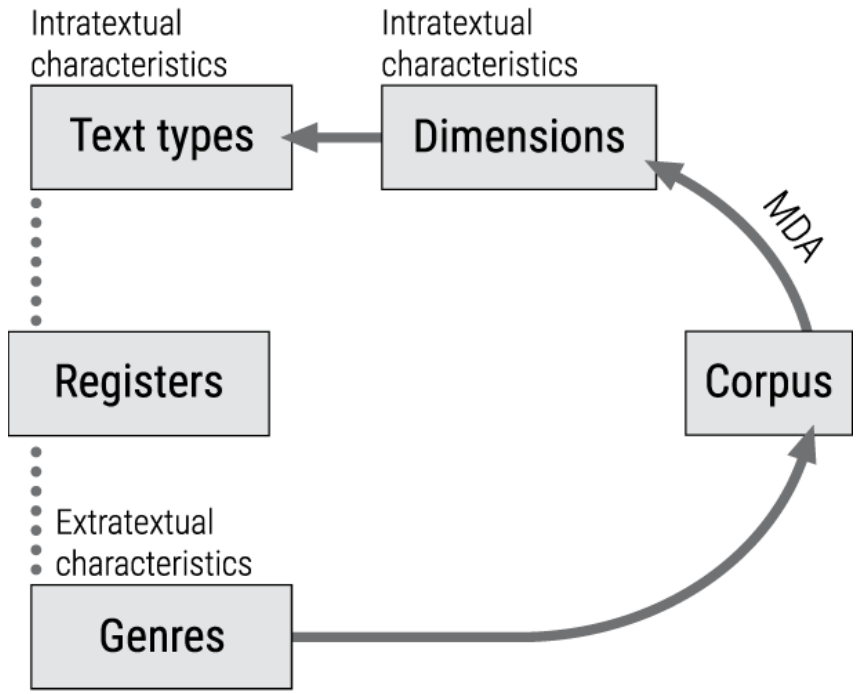
1. corpus compilation
2. list of features
3. operationalization
4. statistical evaluation (factor analysis)

Methodology of MDA

1. corpus compilation
2. list of features
3. operationalization
4. statistical evaluation (factor analysis)
5. interpretation $\rightarrow$ dimensions of variation, registers...

Methodology of MDA

1. corpus compilation
2. list of features
3. operationalization
4. statistical evaluation (factor analysis)
5. interpretation → dimensions of variation, registers...



MDA of Czech

MDA of Czech

Václav Cvrček

Zuzana Komrsková

David Lukeš

Petra Poukarová

Anna Řehořková

Adrian Jan Zasina

MDA of Czech

Expected challenges / highlights of MDA...

- ...in Czech – situation bordering on **diglossia** (Bermel 2014): Literary × Common Czech
- ...in Slavic languages – specific morphology, **inflection**, free word order
- ...in 21st century – how to include the web data (Biber & Egbert 2016; Sharoff 2018)

Results published in:

- Cvrček, V. et al. (2018a): From Extra- to Intratextual Characteristics: Charting the Space of Variation in Czech through MDA. *Corpus Linguistics and Linguistic Theory*.
- Cvrček, V. et al. (2018b): Variabilita češtiny: multidimenzionální analýza. *Slovo a slovesnost* 79, 293–321.

Data: Corpus Koditex

- 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
- 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

Features and their operationalization

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

Features and their operationalization

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...
- 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 concentration (Popescu et al. 2007), unigrams & bigrams (zTTR)

Evaluation & statistics

Text-linguistic approach to variation

- frequency of all features in each text
- co-occurrence of features
- factor analysis: *latent factors* influencing use of features
- latent factors = dimensions of variation (major forces in shaping a text)
- dimensions are not equally important (hierarchy)

Factor analysis outputs

- loadings – ”correlations”of features and dimensions
  - participation of a feature on a dimension
- factor scores – positions of texts within dimensions
  - linguistic characteristics of a text
- 8 dimensions identified
- variance explained: 56 %

Factor analysis outputs

- loadings – "correlations"of features and dimensions
  - participation of a feature on a dimension
- factor scores – positions of texts within dimensions
  - linguistic characteristics of a text
- 8 dimensions identified
- variance explained: 56 %

Interpretation follows these questions:

- what are the loadings of individual features (prominent vs. inert)?
- what is the position of individual text (based on factor scores)?
- what is the position of genres (groups of texts)?

Feature loadings – 1st dimension

Description	Loading
verbs: past tense	0.977
verbs	0.960
verbs: indicative forms	0.952
finite verbs	0.946
verbal aspect (perfective)	0.934
3rd person pronouns (personal + possessive)	0.778
semantically bleached verbs	0.721
function words	0.712
adverbs of time	0.687
pronouns	0.684
verbs: 1st person	0.682
reporting verbs (verba dicendi)	0.665

Feature loadings – 1st dimension

Description	Loading
verbs: past tense	0.977
verbs	0.960
verbs: indicative forms	0.952
finite verbs	0.946
verbal aspect (perfective)	0.934
3rd person pronouns (personal + possessive)	0.778
semantically bleached verbs	0.721
function words	0.712
adverbs of time	0.687
pronouns	0.684
verbs: 1st person	0.682
reporting verbs (verba dicendi)	0.665

Description	Loading
nominal post-modifiers without agreement	-0.792
adjectives	-0.781
noun pre-modifiers with agreement	-0.723
abstract nouns	-0.723
nouns: genitive	-0.723
adjective clusters	-0.705
noun clusters	-0.694
clusters of same-case adjectives	-0.675
average word length (number of syllables)	-0.674
nouns	-0.672
verbal nouns	-0.625

Qualitative double-check

„Opravdu si myslíš, že ti dovolím odplout?“ zeptal se vévoda, objal ji a přitáhl si ji k sobě. Na okamžik Valeria vůbec nedokázala uvěřit, že se něco takového děje. Pak však jeho rty zajaly její a on ji políbil a celý svět se náhle zatočil. Líbal ji něžně, ale majetnicky, stejně jako posledně. Když pak cítila, že v ní začíná narůstat extáze, zvedl hlavu a velmi tiše se zeptal: „Kdy si mě vezmeš, má lásko?“ Valeria na něj jen beze slova hleděla. Obličej se jí rozzářil, jako by v ní někdo zapálil tisíc svící.

(Cartland, Barbara: Ve víru lásky, `wri-fic-nov-lov`)

Qualitative double-check

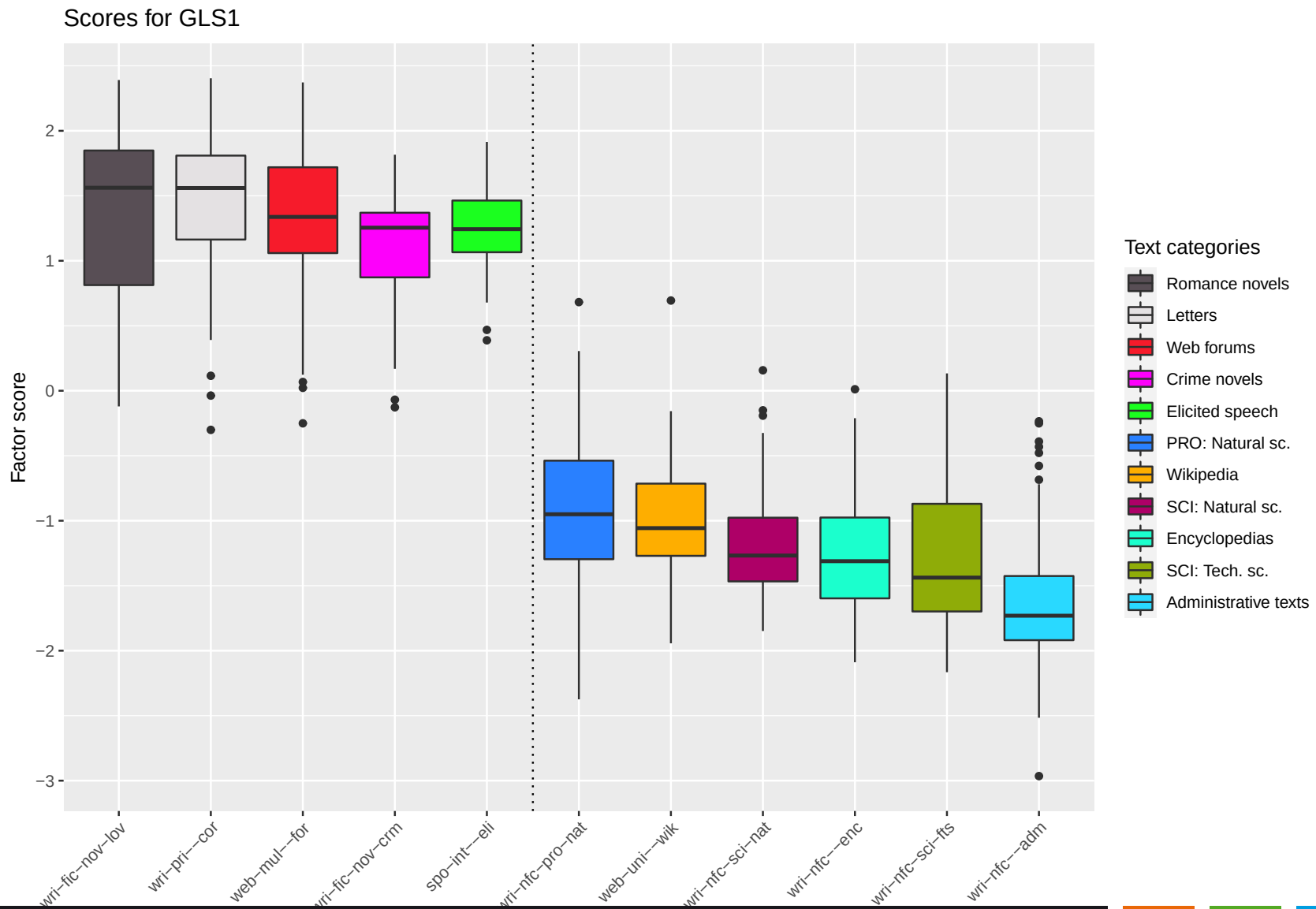
„Opravdu si myslíš, že ti dovolím odplout?“ zeptal se vévoda, objal ji a přitáhl si ji k sobě. Na okamžik Valeria vůbec nedokázala uvěřit, že se něco takového děje. Pak však jeho rty zajaly její a on ji políbil a celý svět se náhle zatočil. Líbal ji něžně, ale majetnicky, stejně jako posledně. Když pak cítila, že v ní začíná narůstat extáze, zvedl hlavu a velmi tiše se zeptal: „Kdy si mě vezmeš, má lásko?“ Valeria na něj jen beze slova hleděla. Obličej se jí rozzářil, jako by v ní někdo zapálil tisíc svící.

(Cartland, Barbara: Ve víru lásky, `wri-fic-nov-lov`)

Speciální pedagog získává odbornou kvalifikaci vysokoškolským vzděláním získaným studiem v akreditovaném magisterském studijním programu v oblasti pedagogických věd zaměřené na speciální pedagogiku. (…) Psycholog získává odbornou kvalifikaci vysokoškolským vzděláním získaným studiem v akreditovaném magisterském studijním programu psychologie…

(Michalík, Jan: Katalog posuzování míry speciálních vzdělávacích potřeb; `wri-nfc-pro-ssc`)

Aggregated factor scores – 1st dimension



Interpretation – 1st dimension

Dimension 1: **dynamic** (+) vs. **static** (-)

- verbal (+related) vs. nominal (+related) constructions
- opposing strategies: elaboration of clause members (-) or adding new clauses (+) → *clausal* vs. *phrasal* (Biber 2014)
- inert feats: dim 1 is indifferent to preparedness of speakers/writers
- (+) factor scores: two shades of "verbality" – *narrative* (e.g. various kinds of novels) + *reflective* (verbs of thinking in `pri-cor` or `web` forums)
- (-) factor scores: information-dense texts – official documents, hard science papers, encyclopaedias
- most variance explained

Feature loadings – 2nd dimension

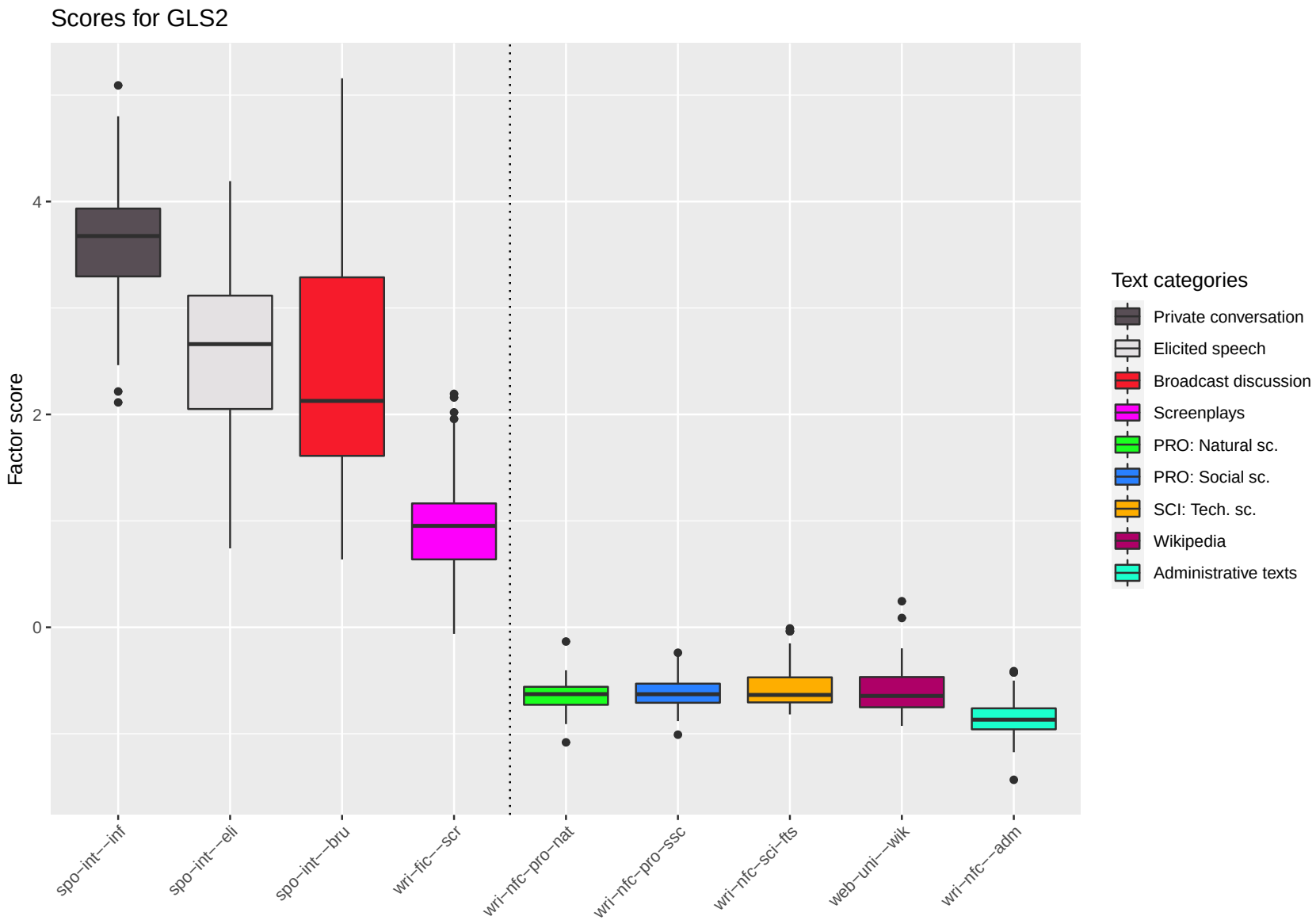
Description	Loading
contact expressions	0.974
fillers	0.854
interjections	0.824
demonstrative pronouns (excl. 'to')	0.821
expressive particles	0.795
pronoun non-dropping	0.793
vowel breaking *ý* > *ej* in endings	0.778
demonstrative adverbs	0.776
word repetition	0.767
locative adverbs	0.763
narrowing *é* > *í*/*ý* in endings	0.747

Feature loadings – 2nd dimension

Description	Loading
contact expressions	0.974
fillers	0.854
interjections	0.824
demonstrative pronouns (excl. 'to')	0.821
expressive particles	0.795
pronoun non-dropping	0.793
vowel breaking *ý* > *ej* in endings	0.778
demonstrative adverbs	0.776
word repetition	0.767
locative adverbs	0.763
narrowing *é* > *í*/*ý* in endings	0.747

Description	Loading
nominal cases with prepositions	-0.624
clauses with wh-adverbs	-0.567
prepositions	-0.559
verbal aspect (perfective)	-0.493
unigrams	-0.463
nouns: nominative-accusative	-0.460
nouns	-0.367
repertoire of prepositions	-0.360
average word length (number of syllables)	-0.357
nouns: instrumental	-0.349
nouns: locative	-0.307

Factor scores – 2nd dimension



Interpretation – 2nd dimension

Dimension 2: **spontaneous** (+) vs. **prepared** (-)

- reflects differences in conditions of production: `wri` (editing and refining possible) vs. `spo` (online production)
- positive features mark:
  1. interactivity (contact exp., fillers, demonstratives, pronouns, word repetition)
  2. informality (expressive particles, interjections)
  3. conventionalised non-standard Common Czech morphonological variants
- (+) texts: `spo-int-inf`, `pri-cor`, `web-mul` (`fcb` / `for`)
- (-) texts: administrative texts, Wikipedia, `sci-fts`, `pro-nat`

2D graph


GLS2
4
2
0
-3
-2
-1
0
1
2
GLS1
DIVISION
spo-int
spo-nln
web-mul
web-unl
wrl-flc
wrl-nfc
wrl-nmg
wrl-prl
MODE
spo
web
wrl

All dimensions

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

Note: not all dims are equal – most important: 1, 2, 5, 8

MDA summary

MDA of Czech – outcomes

- hierarchical description of variation
  - projection of low-level features (e.g. morphology) on higher levels (register)
  - relative importance of dimensions and features

MDA summary

MDA of Czech – outcomes

- hierarchical description of variation
  - projection of low-level features (e.g. morphology) on higher levels (register)
  - relative importance of dimensions and features
- better description of features (systemic functional variation)

MDA summary

MDA of Czech – outcomes

- hierarchical description of variation
  - projection of low-level features (e.g. morphology) on higher levels (register)
  - relative importance of dimensions and features
- better description of features (systemic functional variation)
- applications of MD model
  - landscape description (registers)
  - sources of variation (idiolect vs. register)
  - practical implications (corpus design etc.)

Establishing registers

Intratextual classification

Registers

- classification based on features used (rather than convention or tradition)
- clusters of texts in 8-D space (distance $\sim$ similarity)

Intratextual classification

Registers

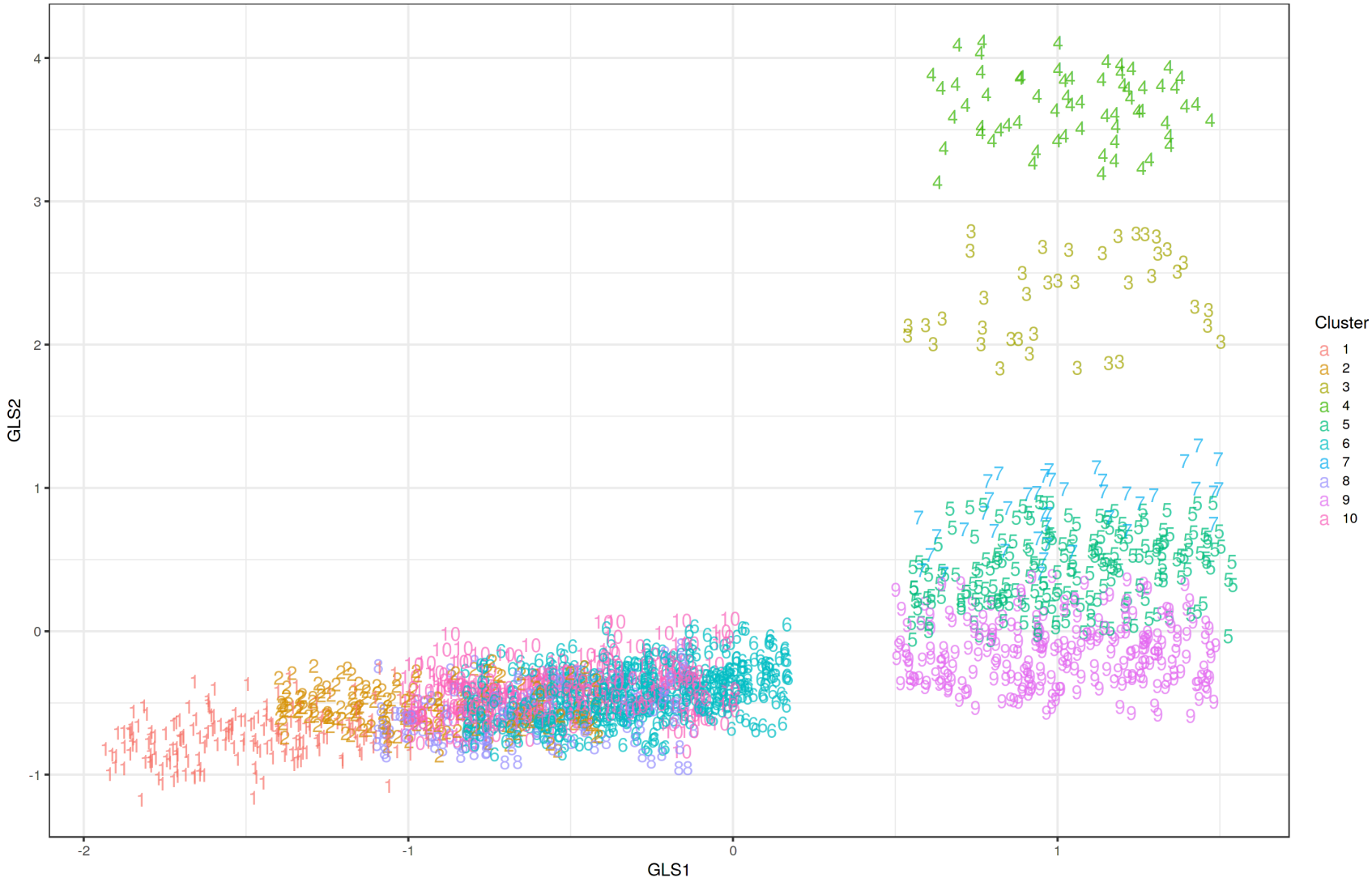
- classification based on features used (rather than convention or tradition)
- clusters of texts in 8-D space (distance $\sim$ similarity)

Motivation

- "register matters" (cf. Biber et al. Longman Grammar 1999, Cvrček et al. 2010)
- "know your data" – popularization (non-fiction or journalism?), memoirs (non-fiction, fiction or journalism?)

Clusters – registers

K-means clustering: 10 registers



Registers

- static registers
  - analysis: static monothematic
  - popularization: static polythematic general
  - journalism: static indefinite
  - facts: static polythematic particular
  - reasoning: static cohesive
- dynamic registers
  - survey: dynamic non-addressing
  - conversation: dynamic spontaneous
  - commentary: dynamic attitudinal
  - screenplay: dynamic addressing
  - narration: dynamic retrospective

$\Rightarrow$ further elaboration to subregisters is possible (J. Henyš – 20 web registers)

Proportion of registers within text classes

Web multidirectional (`dis`, `fcb`, `for`)

- commentary (73 %)
- journalism (10 %)
- reasoning (9 %)

Proportion of registers within text classes

Web multidirectional (`dis`, `fcb`, `for`)

- commentary (73 %)
- journalism (10 %)
- reasoning (9 %)

Written fiction (`crm`, `lov`, `scf`, `scr`, `ver`…)

- narration (75 %)
- screenplay (13 %)
- commentary (4 %)

Register versus idiolect

Projecting CPACT data on MD model

CPACT data

- data collected within *CPACT* project (D. Kučera)
- 200 native speakers of Czech – proportionate stratified sampling (age, gender, education)
- rich psychological metadata – *Big Five personality traits*, *DASS 21* (Depression, Anxiety, Stress Scale) etc.
- each participant wrote 4 texts within one day following a scenario (Letter from vacation, Letter of complaint, Letter of apology, Cover letter)
  - form/genre: letter
  - length: 180–200 words

Analysis of CPACT data

- same set of features as used in original MDA
- results projected onto original MD model

Analysis of CPACT data

- same set of features as used in original MDA
- results projected onto original MD model

Statistical modeling:

- ANOVA – effect size (η)
- Kruskal-Wallis test – effect size (E_R^2)
- Linear Mixed-effects models (LMER) – coefficient of determination (R^2)

Response: Text factor score $\sim$ Explanatory: Scenario $+$ Author

Idiolect vs. register (1:2)

Range of variation and corpus design

Representativeness

Corpus representativeness & variation

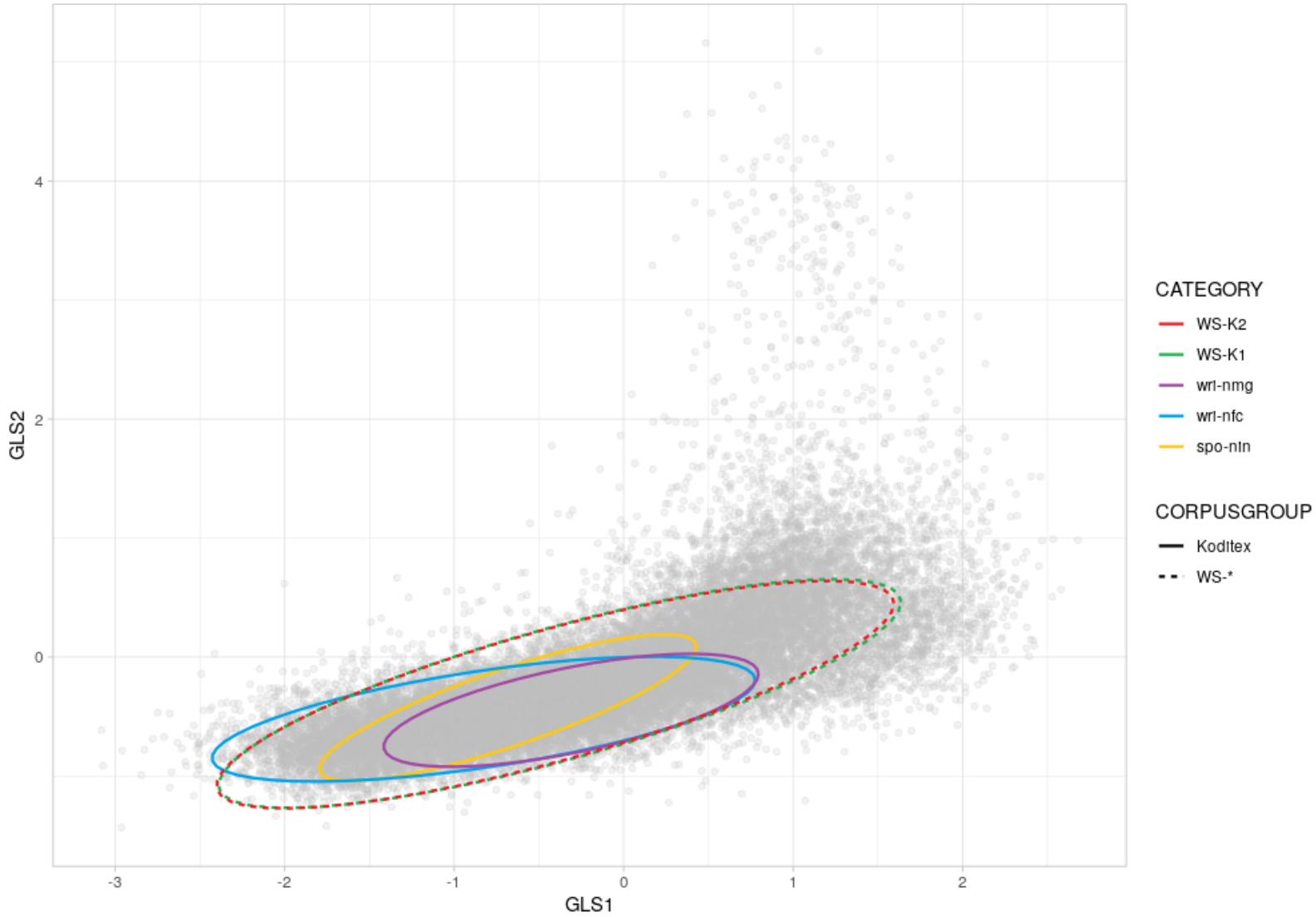
- known issue of CL
- "Representativeness refers to the extent to which a sample includes the full range of variability in a population." (Biber 1993: 243).
- "Thus a corpus design can be evaluated for the extent to which it includes: (1) the range of text types in a language, and (2) the range of linguistic distributions in a language." (Biber 1993: 243).
- $\Rightarrow$ comparing corpora w.r.t. the **variability they cover**

Traditional vs. web-crawled corpus

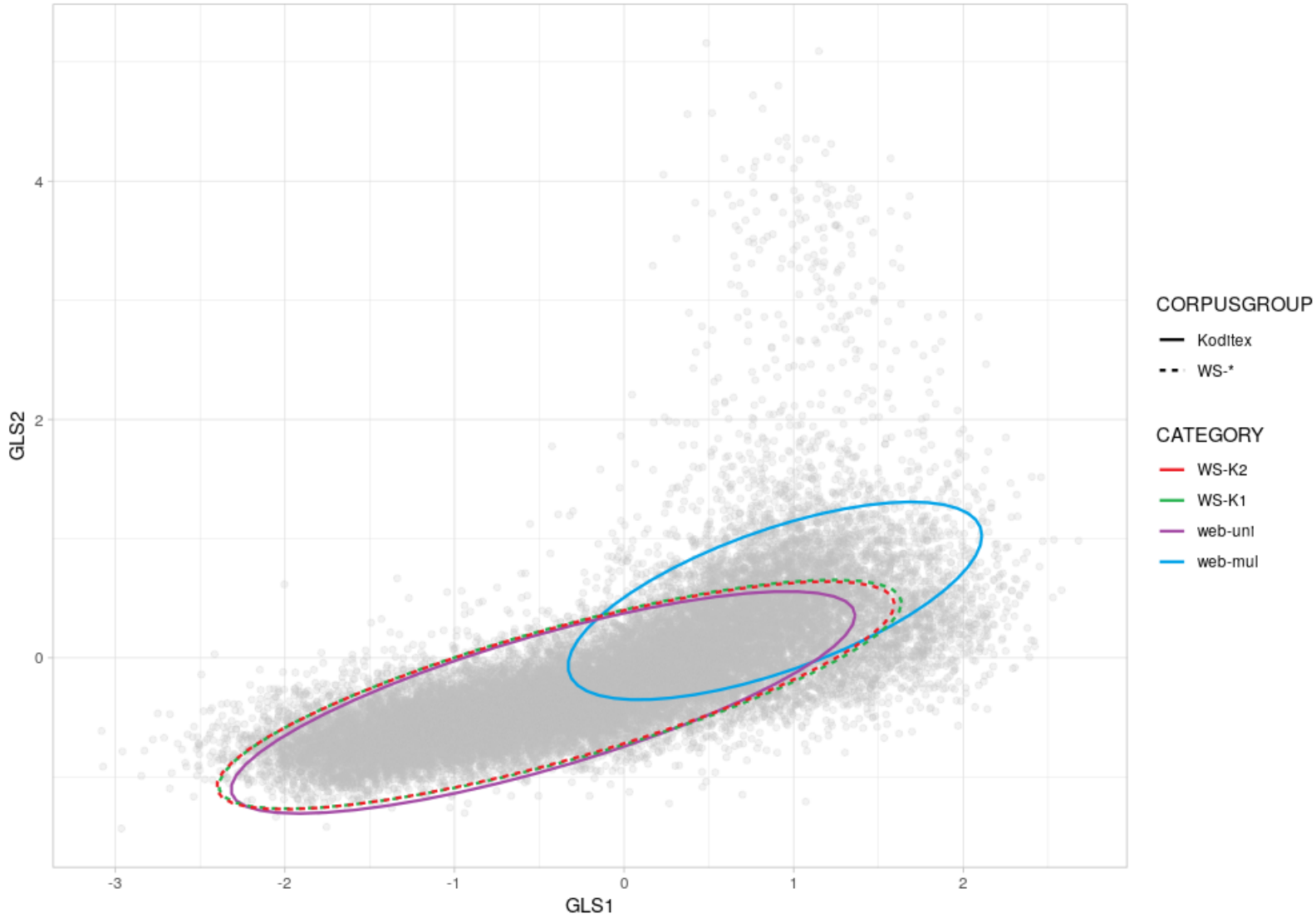
Sampling the Araneum Bohemicum corpus

- Araneum Bohemicum Maximum 15.04 (May and June 2013, 5.4. bln. tokens; Benko 2016)
- opportunistic design
- representation of "searchable"web
- 2 samples (WS-K1, WS-K2 – 5000 texts each)
- text length distributions modelled after Koditex
- subsequent processing analogous to Koditex texts

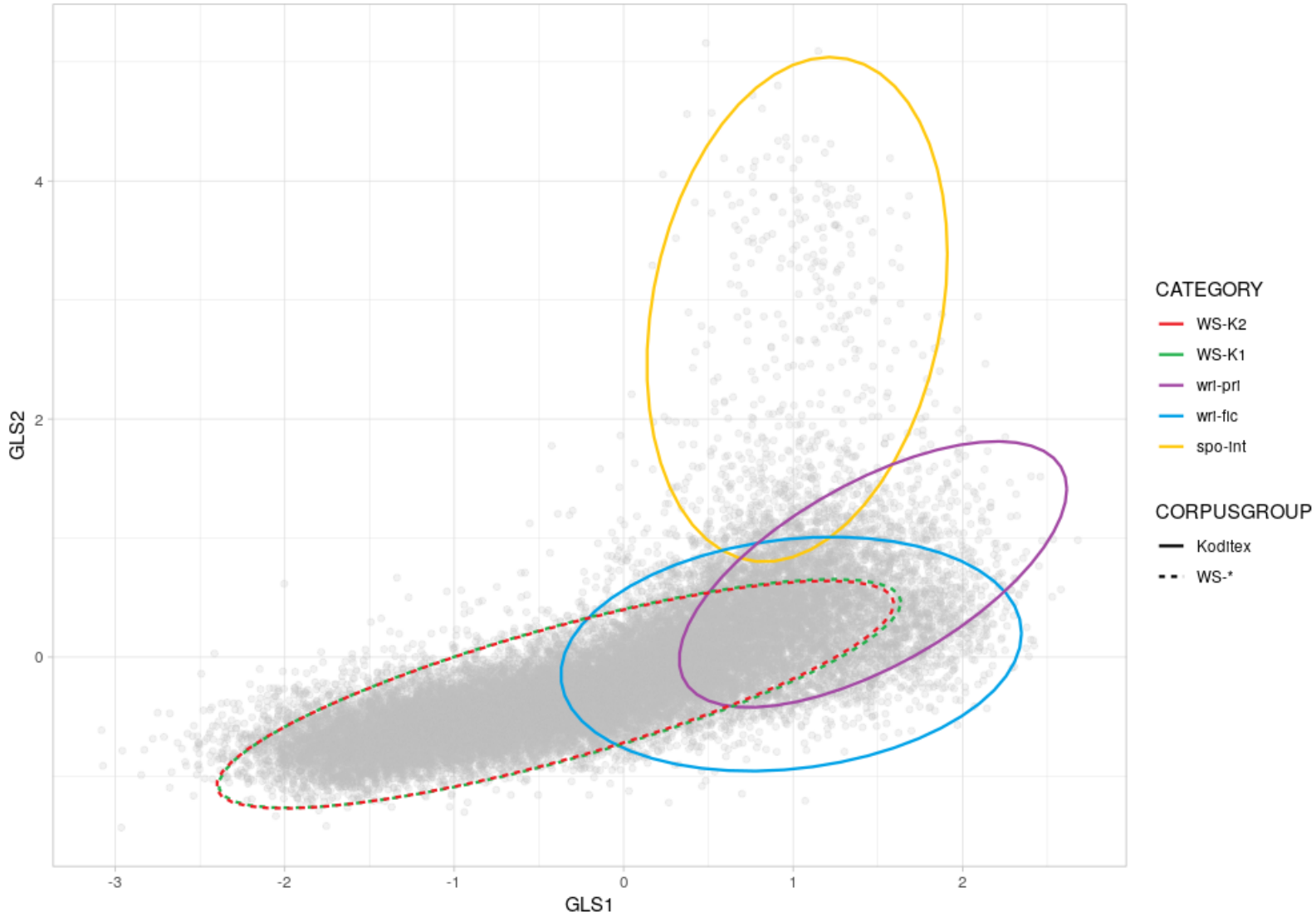
Koditex vs. WebSample in 2D



Koditex vs. WebSample in 2D



Koditex vs. WebSample in 2D



Conclusions

Inspirations

Corpus-based studies of language variation

- reveal the functions of linguistic features, e.g.
  - *vocative* as a typical feature of dialogue (not necessarily spontaneous spoken conversation)
  - *demonstratives* as a correlate of unprepared spoken production

Inspirations

Corpus-based studies of language variation

- reveal the functions of linguistic features, e.g.
  - *vocative* as a typical feature of dialogue (not necessarily spontaneous spoken conversation)
  - *demonstratives* as a correlate of unprepared spoken production
- web = *terra incognita* (J. Henyš – 20 web /sub/registers: review, advise, description, Q&A, how-to, encyclopaedia…)

Inspirations

Corpus-based studies of language variation

- reveal the functions of linguistic features, e.g.
  - *vocative* as a typical feature of dialogue (not necessarily spontaneous spoken conversation)
  - *demonstratives* as a correlate of unprepared spoken production
- web = *terra incognita* (J. Henyš – 20 web /sub/registers: review, advise, description, Q&A, how-to, encyclopaedia…)
- register-sensitive annotation (lemmatization and tagging)

Challenges

- overcoming the stereotypes in variation descriptions
  - based on axiology and prescription
  - non-hierarchical approach of traditional stylistics $\times$ not all factors/dimensions or registers are "born equal"
  - qualitative approach $\times$ distinguishing the marginal and major variants

Challenges

- overcoming the stereotypes in variation descriptions
  - based on axiology and prescription
  - non-hierarchical approach of traditional stylistics $\times$ not all factors/dimensions or registers are "born equal"
  - qualitative approach $\times$ distinguishing the marginal and major variants
- replicability and reliability of MDA
  - the impact of MDA settings (features and texts used) on its results
  - register – topic relationship

Challenges

- overcoming the stereotypes in variation descriptions
  - based on axiology and prescription
  - non-hierarchical approach of traditional stylistics $\times$ not all factors/dimensions or registers are "born equal"
  - qualitative approach $\times$ distinguishing the marginal and major variants
- replicability and reliability of MDA
  - the impact of MDA settings (features and texts used) on its results
  - register – topic relationship
- uncovering the functions of variation

1:1

Thank you for your attention!

Acknowledgement

This study was supported from the ERDF project *Language Variation in the CNC* no. CZ.02.1.01/0.0/0.0/16_013/0001758 and builds upon resources developed during the implementation of 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.

References

- Benko, V. (2016). Two Years of Aranea: Increasing Counts and Tuning the Pipeline. LREC, 4245–4248.
- 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. & Conrad, S. (2009): Register, Genre, and Style. New York, NY: 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.
- Biber, D. & Johansson, S. et al. (1999): Longman Grammar of Spoken and Written English. Longman.
- Biber, D. (1988): Variation Across Speech and Writing. Cambridge University Press.
- 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: 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.
- Cvrček, V. et al. (2018a): From Extra- to Intratextual Characteristics: Charting the Space of Variation in Czech through MDA. Corpus Linguistics and Linguistic Theory [Ahead of print].
- Cvrček, V. et al. (2018b): Variabilita češtiny: multidimenzionální analýza. Slovo a slovesnost 79, 293–321.
- Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J., & Benko, V. (forthcoming a). Comparing web-crawled and traditional corpora.
- Cvrček, V. et al. (forthcoming b): Author and register as sources of variation. A corpus-based study using elicited texts.
- Popescu, I., Best, K. & Altmann, G. (2007): On the dynamics of word classes in texts. Glottometrics 14, (p. 58–71).
- Sharoff, S. (2018): Functional Text Dimensions for the annotation of web corpora. Corpora, 13(1), 65–95.