Register versus idiolect variation

Václav Cvrček

Zuzana Komrsková

David Lukeš

Petra Poukarová

Anna Řehořková

Adrian Jan Zasina

24th September 2018

CZECH NATIONAL CORPUS

Introduction

MDA framework

- language variability can be caused by several different phenomena, e.g.
  - register
  - author/speaker
- MD model based on the *Koditex* corpus
  - 8 dimensions
  - explains 56% of variability
  - position of a text in 8D space

Research questions

How can the MD model help:

- in analyzing elicited data?
- in quantifying differences between elicitation scenarios?
- in analyzing register versus idiolect variation?

CPACT data

Speakers

- data collected within CPACT project (GA ČR 16-19087S, D. Kučera)
- 200 native speakers of Czech
  - proportionate stratified sampling
  - selected according to the criteria of age (15–24, 25–34, 35–55 and 55+), gender, highest achieved level of education
  - rich psychological metadata – results from several psychological tests, e.g. *Big Five personality traits* (extraversion, conscientiousness, openness to experience, agreeableness, neuroticism), *DASS 21* (Depression, Anxiety, Stress Scale) etc.

Texts

- each participant wrote 4 texts within one day
- requirements:
  - length: 180–200 words
  - form: letter
  - follow task scenario
- scenarios describe situational setting + motivation
- defined by 2 criteria, **formality** and *interpersonal stance*:

	Informal	**Formal**
Dominant	Letter from vacation	Letter of complaint
Submissive	Letter of apology	Cover letter

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

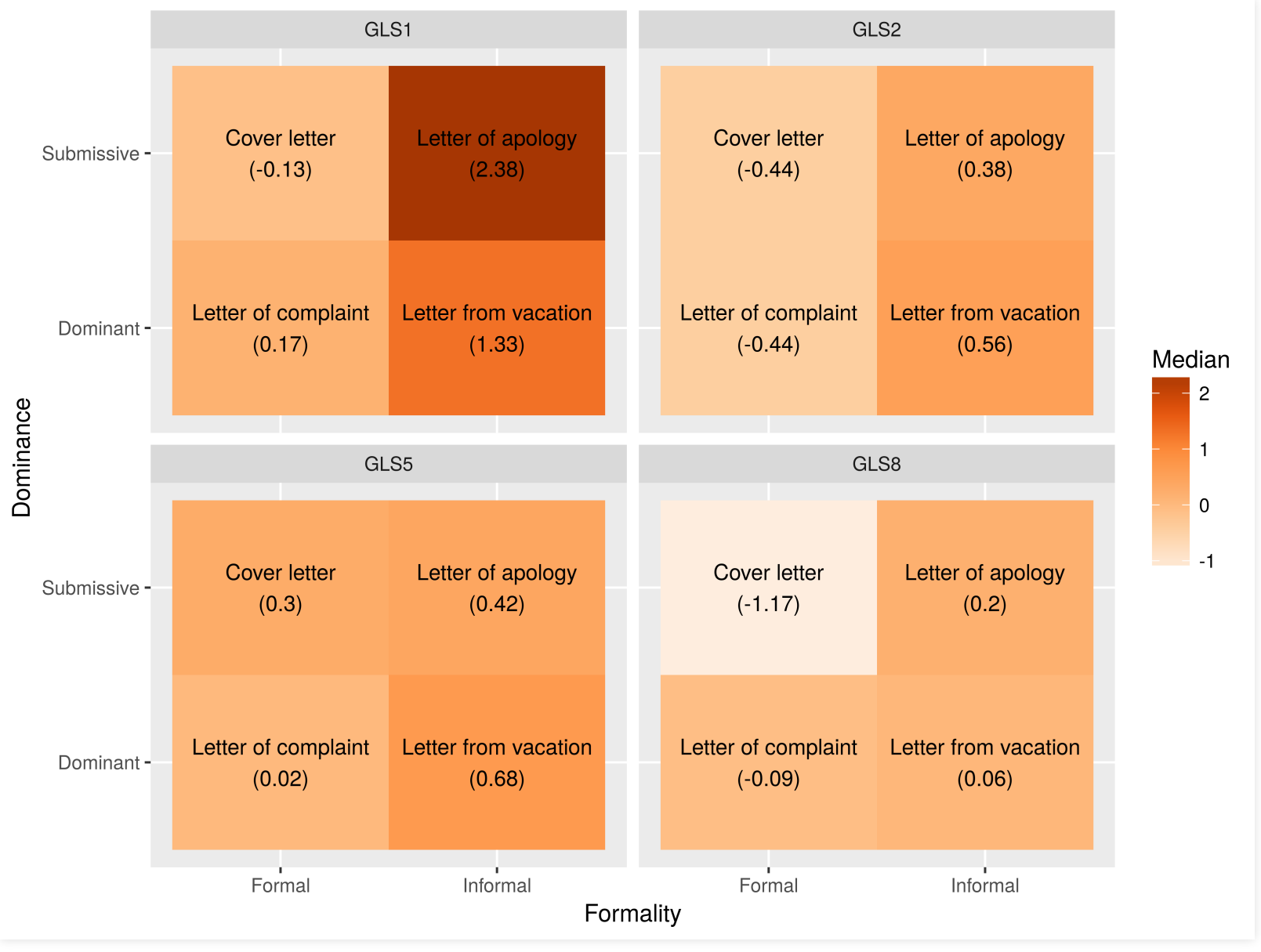
Register variability of elicited texts

Validity of elicited texts

- psychologically oriented analyses of linguistic behavior often based on texts elicited through carefully devised scenarios (e.g. Knapp, et al., 1974; Fisher, 1993; Newman, et al., 2003)
  - advantages: controlled environment, diversity of scenarios $\Rightarrow$ diversity of elicited data

- Are there systematic differences between texts elicited by different scenarios?
- How to design scenarios to be meaningfully distinct in terms of the language they elicit?
- Which scenarios contribute to the diversity of the resulting material?

Elicited texts in 4 dimensions



Differences between scenarios

- position of elicited texts (CPACT) corresponds to similar non-elicited texts (Koditex)
- each task = specific situation, setting
  - formal tasks (cover letter, letter of complaint): large overlap in the MD model
  - *letter from vacation*: greatest within-task variability
- scenarios differ in “degrees of freedom” given to author
  - the more rigid and highly conventionalized a type of text is, the less room for individual variability

Register versus idiolect

Motivation and overview

CPACT data:

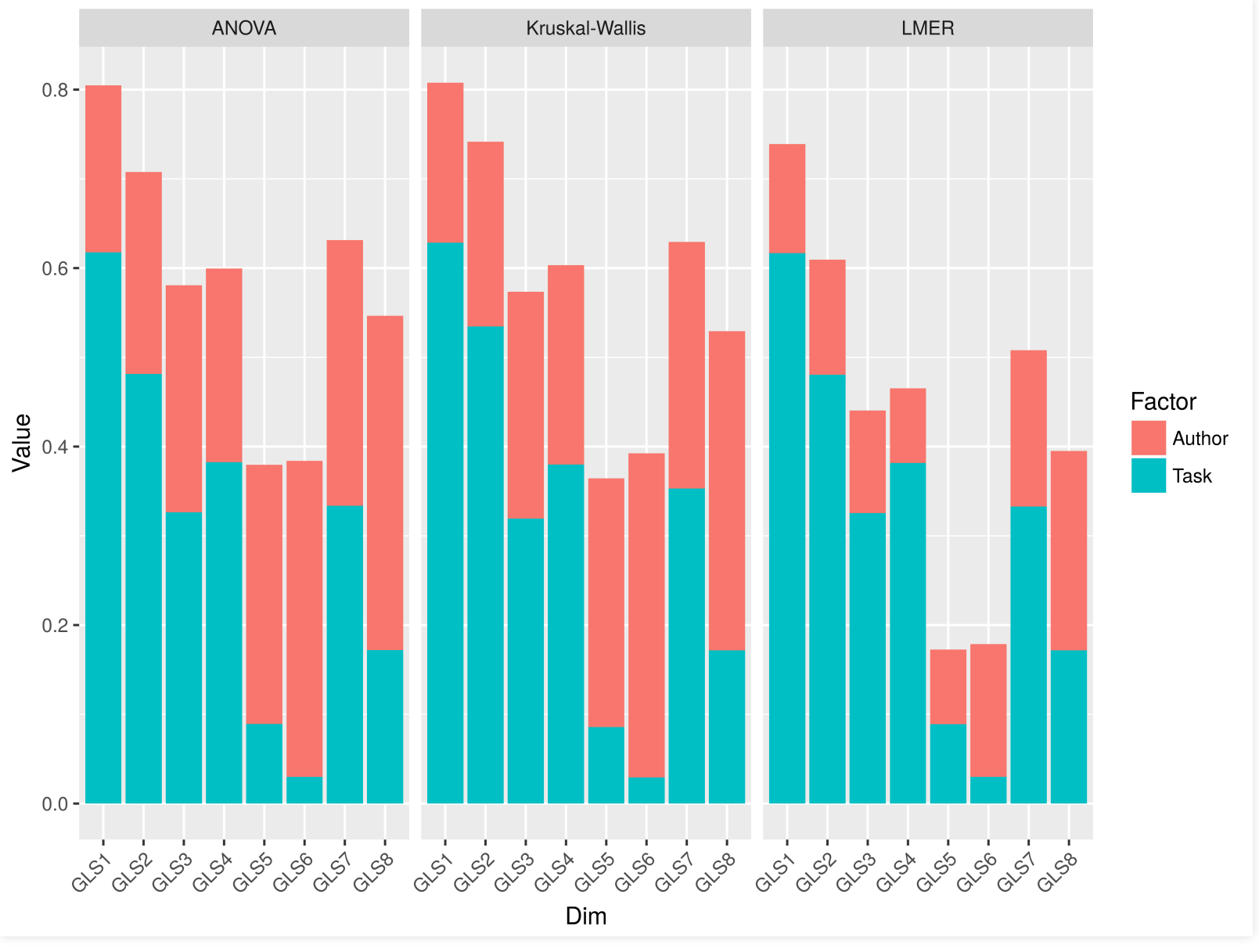
- a unique opportunity to study the interference of idiolect
- approximations:
  - *idiolect* = 200 speakers – balanced stratified sampling
  - *register* = 4 scenarios (tasks) – situation (formality), context (stance)
  - all texts are letters – the difference between scenarios should be subtle (!)

Methods

- Statistical modeling:
  - ANOVA – effect size (η)
  - Kruskal-Wallis test – effect size (E_R^2)
  - Linear Mixed-effects models (LMER) – coefficient of determination (R^2)
- Intuitive approach:
  - determining average distance in MD space

Results – statistical models

Surprisingly (?) similar results:



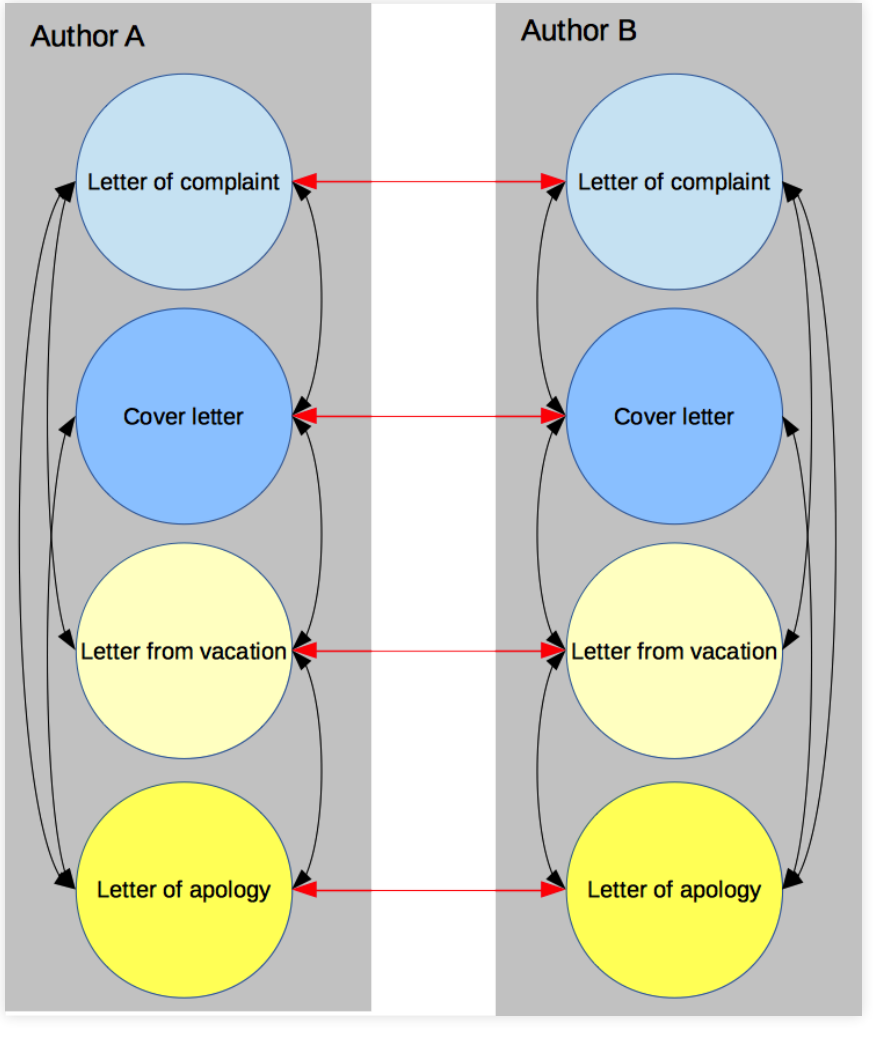
Results – statistical models

Proportion of variation accounted for by `task` and `author` in all dimensions (weighted with regards to the "importance" of individual dimensions):

Method	wAverage.Task	wAverage.Author
ANOVA	0.6119	0.3881
Kruskal-Wallis	0.6279	0.3721
LMER	0.7273	0.2727

Intuitive approach: distances between texts

Same author (black) vs. identical register (red)



Results – distances

Central tendency and spread of distances:

group	mean	median	sd	mad
same_auth	1.460	1.418	0.539	0.546
same_task	1.096	1.048	0.375	0.359

This is in agreement with the statistical models:

- texts of one author differ more than texts based on one scenario
- → scenario (task) is a better predictor of position within MD space

Conclusions

Conclusions

- an MD model can help in assessing the *ecological validity* of elicited texts
  - are they what/where we expected them to be?
- projecting elicited data onto the MD model can reveal key differences (or overlaps) between scenarios
- controlled settings of CPACT data allow for the combined study of *idiolect and register variation*

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, B. & Conrad, S. (2009): *Register, Genre, and Style*. New York, NY: Cambridge University Press.
- Cvrček, V. & Chlumská, L. (2015): Simplification in translated Czech: a new approach to type-token ratio. *Russian linguistics* 39/3, (p. 309–325).
- Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research* 20(2), (p. 303–315).
- Knapp, M. L., Hart, R. P., & Dennis, H. S. (1974). An Exploration of Deception as a Communication Construct. *Human Communication Research* 1(1), (p. 15–29).
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: predicting deception from linguistic styles. *Personality & Social Psychology Bulletin* 29(5), (p. 665–675).
- 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,
- and the **CPACT** research project funded by the *Czech Science Foundation*, grant nr. GA ČR 16-19087S.

It builds upon work made possible by the **Czech National Corpus** project (LM2015044) funded by MEYS within the framework of *Large RDI Infrastructures*.

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

MINISTRY OF EDUCATION,
YOUTH AND SPORTS