The DiaL2 project: pipeline, results, news and future work

George Moroz Olga Gich Anna Grishanova Natalia Koshelyuk
Chiara Naccarato Anna Panova Anastasia Yakovleva
Svetlana Zemicheva

17.09.2024

Precursors

Precursors of the project

Nina Dobrushina

Michael Daniel

- Multiple sociolinguistic expeditions to Daghestan
- Several dialect expeditions to Ustya

Precursors of the project

Nina Dobrushina

Michael Daniel

- Multiple sociolinguistic expeditions to Daghestan
- Several dialect expeditions to Ustya
- Online corpora available for everyone:
  - Corpus of Russian spoken in Daghestan
  - Ustja River Basin Corpus

Precursors of the project

Nina Dobrushina

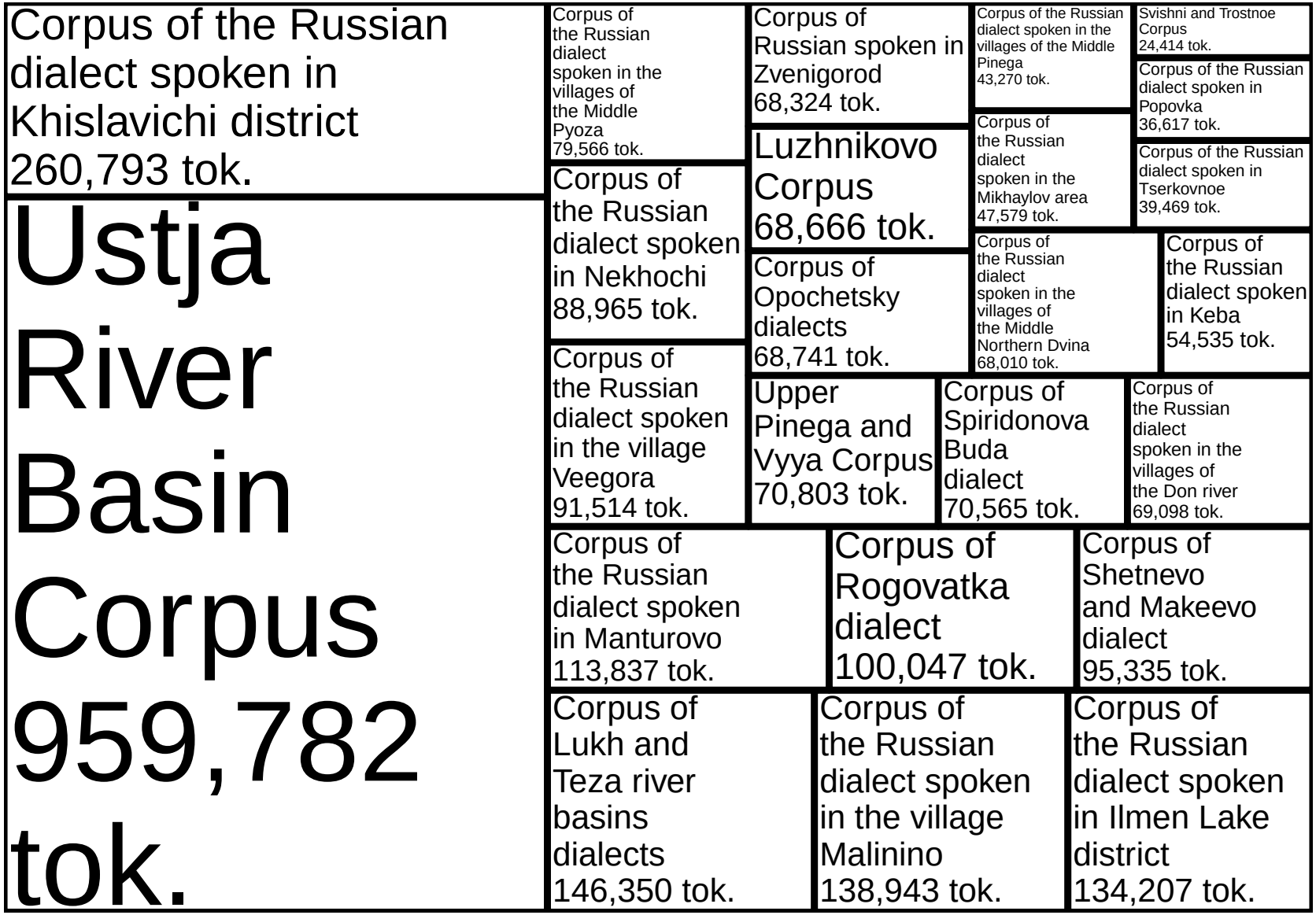
Michael Daniel

- Multiple sociolinguistic expeditions to Daghestan
- Several dialect expeditions to Ustya
- Online corpora available for everyone:
  - Corpus of Russian spoken in Daghestan
  - Ustja River Basin Corpus
  - ... and other bilingual and dialect corpora

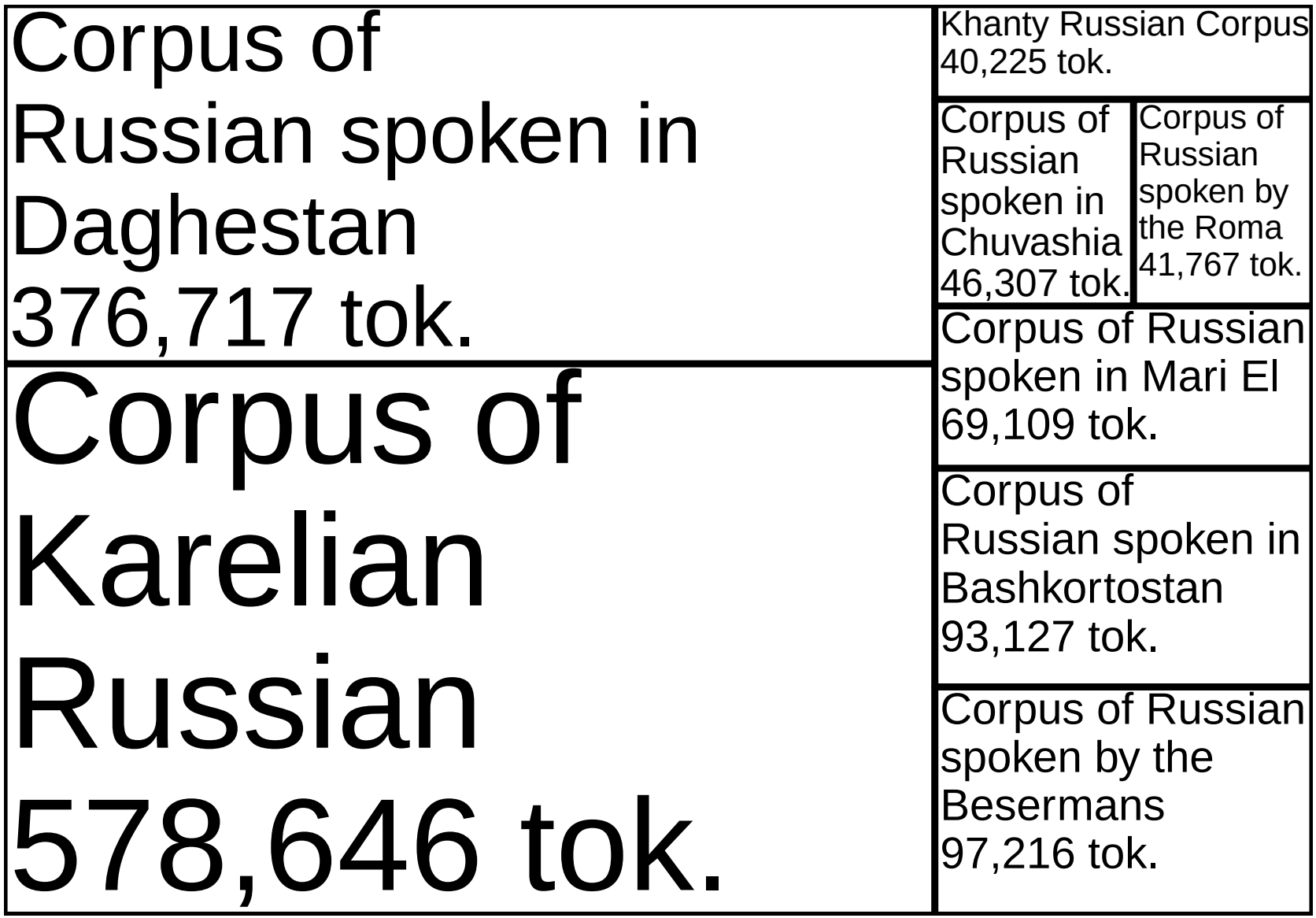
Resources of the Linguistic Convergence Laboratory

- https://lingconlab.ru/
- 24 dialectal corpora
- 8 bilingual corpora

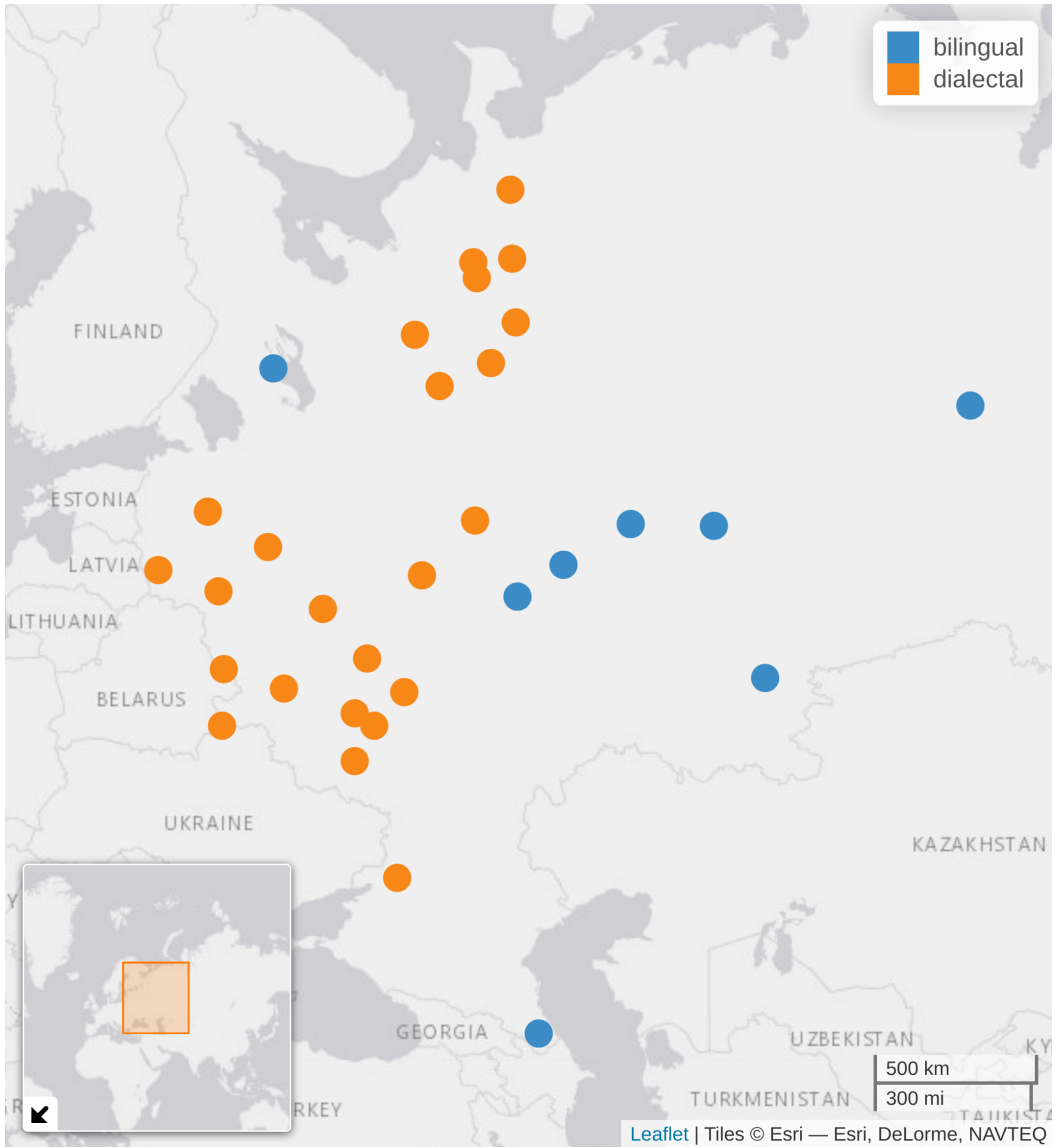
Dialectal Corpora



Bilingual Corpora



Bilingual and Dialectal Corpora



Created with `lingtypology` [Moroz, 2017]

Can we analyze variation of linguistic features across all corpora?

Can we analyze variation of linguistic features across all corpora?

What are the factors that influence variation?

Can we analyze variation of linguistic features across all corpora?

What are the factors that influence variation?

Can we find different variation patterns?

Previous publications

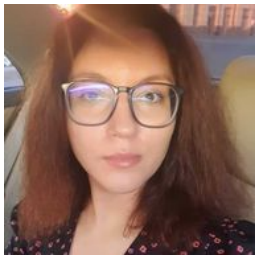
- Daghestanian Russian [Daniel et al., 2010, Panova and Philippova, 2021]
- Russian of Erzya speakers [Shagal, 2016]
- Russian of Kazakh speakers [Rakhilina and Kazkenova, 2018]
- Contact Russian of Northern Siberia and the Russian Far East [Stoynova, 2019, 2021]
- Russian of Moksha speakers [Kashkin, 2020]
- Russian of Hill Mari [Kashkin, 2022]
- Russian of Nganasan speakers [Khomchenkova, 2020]

The DiaL2 project

The DiaL2 team

Maria Ermolova

Anna Grishanova



Natalia Koshelyuk

George Moroz

Chiara Naccarato

Anastasia Yakovleva

Svetlana Zemicheva

The DiaL2 pipeline

- collect all `.eaf` files
- extract transcriptions using the `phonfieldwork` R package [Moroz, 2023]
- use the `udpipe` package in order to gather morphological and syntactical annotation
- filter the result table for particular feature selected by a researcher
- annotate standardness of the utterances
- remove fully-standard speakers
- model the standardness of the utterances using sociolinguistic and linguistic features as predictors

The DiaL2 results

Some results

- Non-standard numeral constructions in L2 Russian
- Propositional Drop
- Propositional Drop in Chuvash
- Dialect Genitive Plural Forms of Masculine and Neuter Nouns in Numeral Constructions
- Negative Existential Constructions

Non-standard numeral constructions in L2 Russian

Chiara Naccarato

George Moroz

Non-standard numeral constructions in L2 Russian

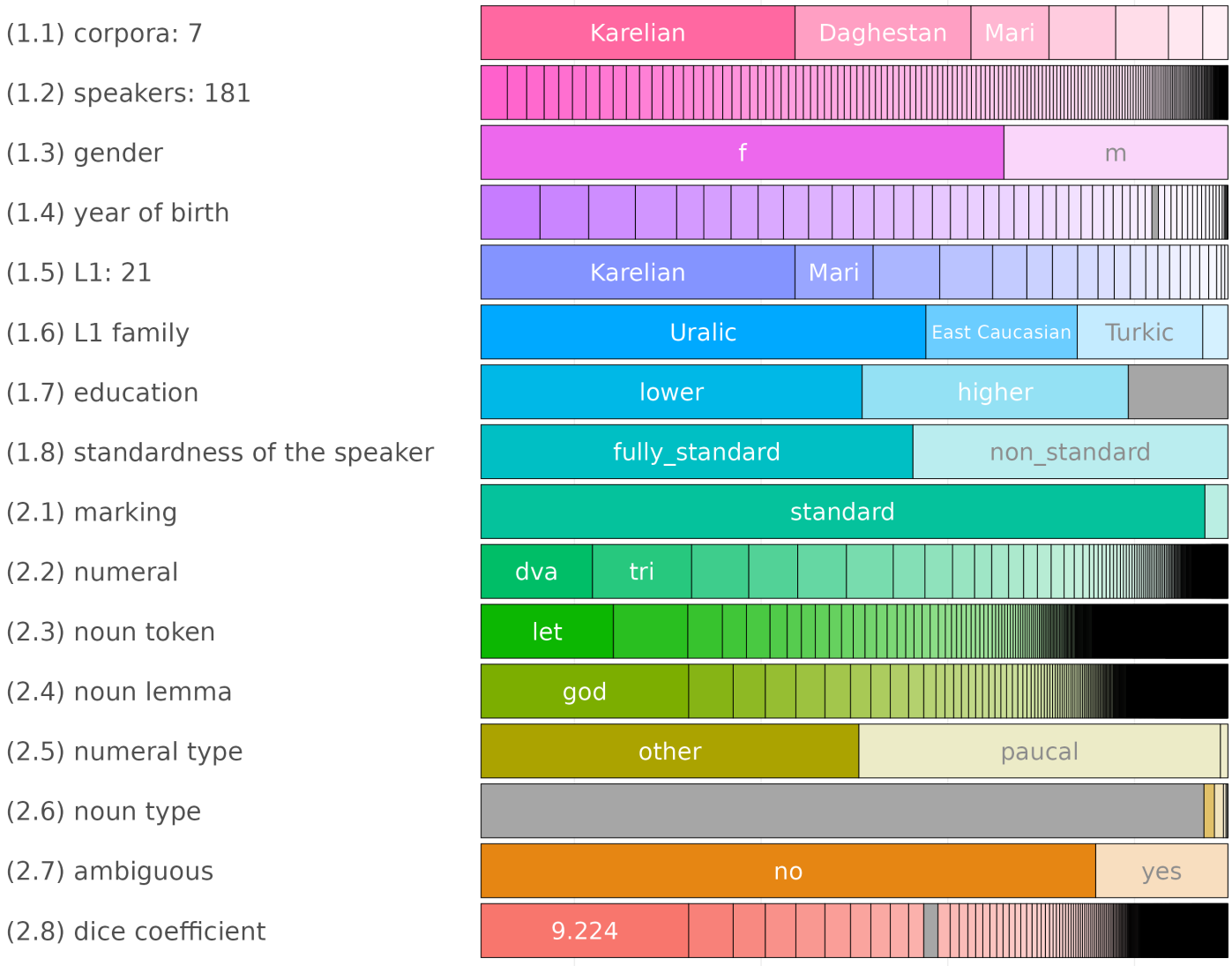
- Variation in numeral constructions (NCs) in bilingual corpora
  - e.g. dva brat vs. dva brata
- Previous research on other L2 Russian varieties
  - Stoynova (2021) on Nanai and Ulcha Russian: evidence for pattern borrowing
- Also mentioned by
  - Shagal (2016: 369-370) for Erzya Russian
  - Rakhilina & Kazkenova (2018: 610) for Kazakh Russian

Research questions

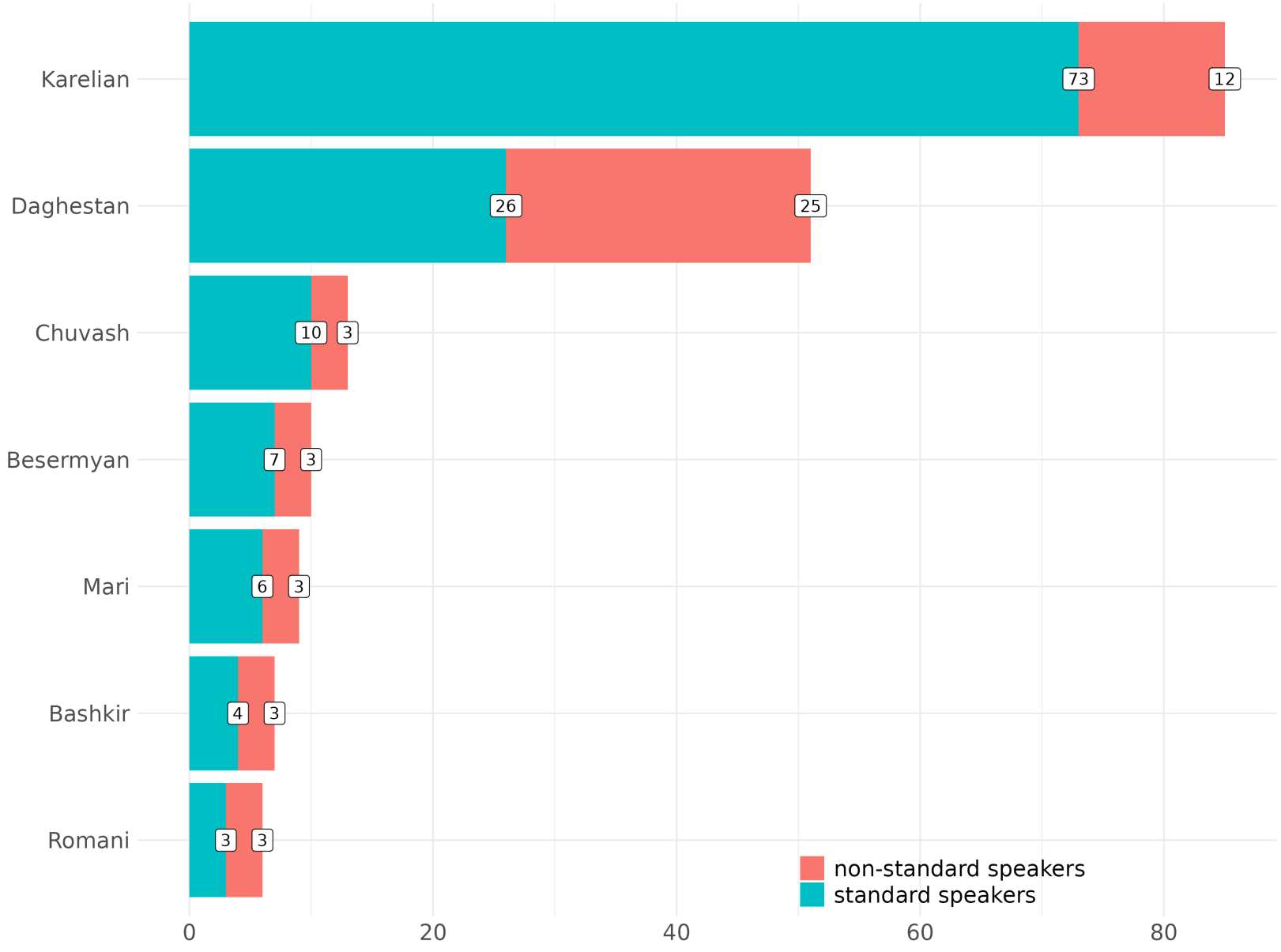
- Does the amount of variation in NCs differ across corpora and/or among speakers of the same variety?
- Can variation in NCs be explained in terms of contact influence?
- Do other factors promote or hinder variation in NCs?

The database and parameters of data annotation

4,144 observations

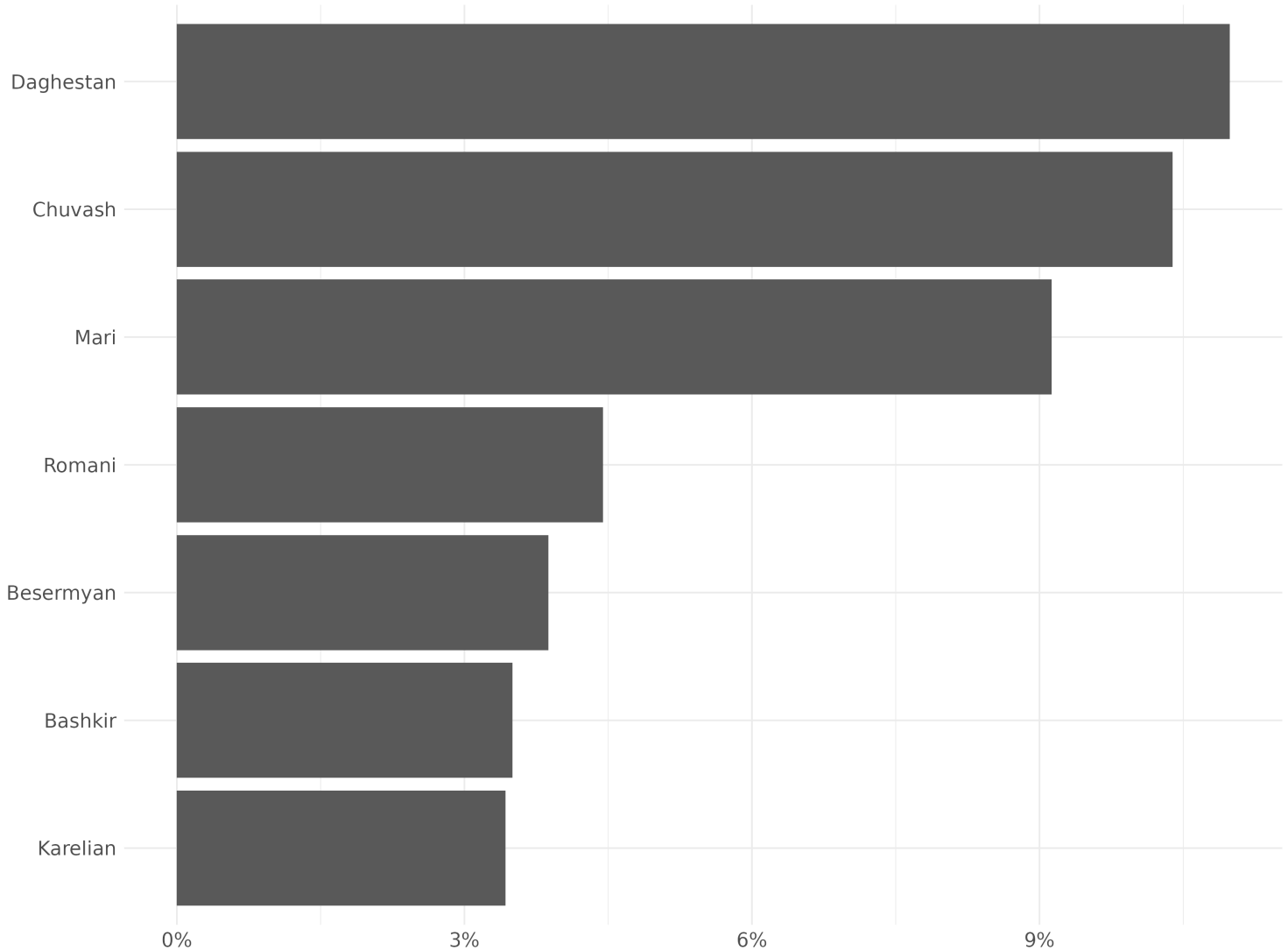


Fully standard (71.3%) vs. non-standard speakers (28.7%)



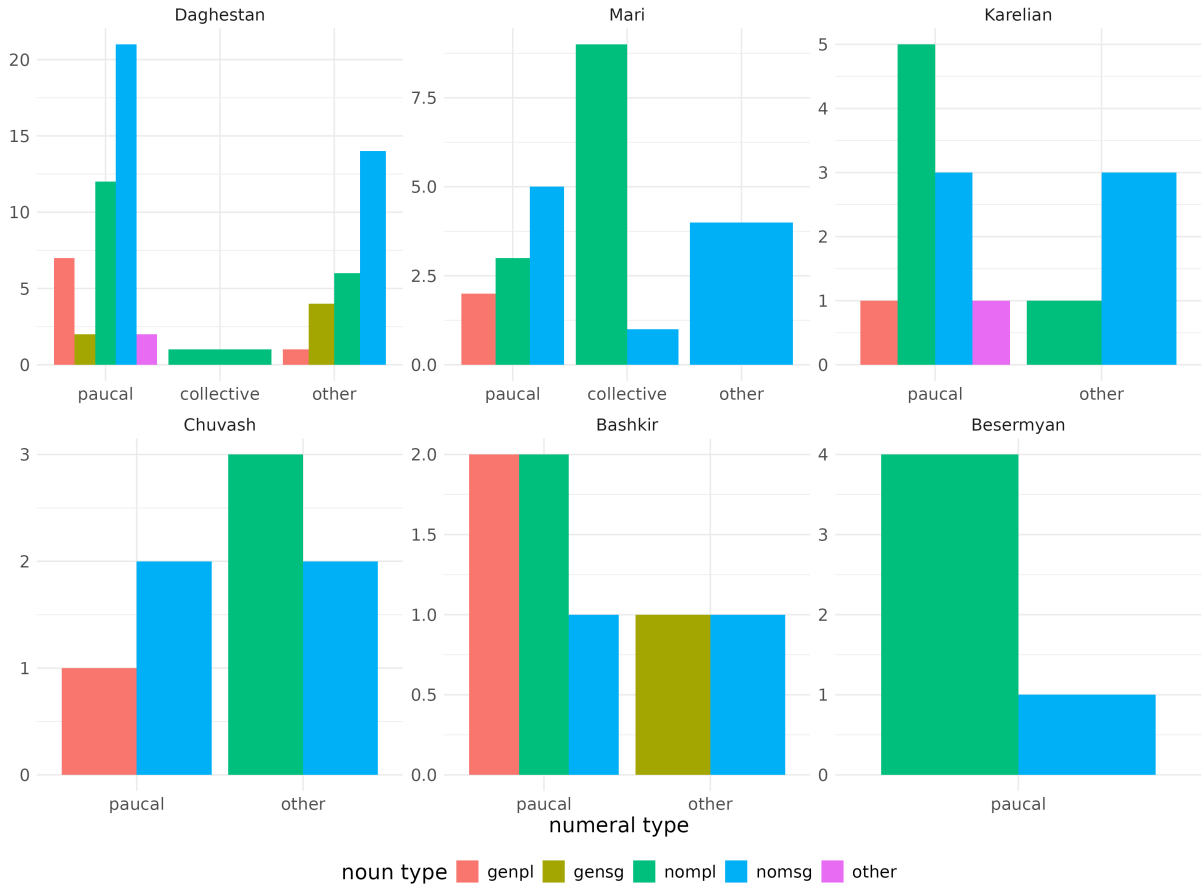
Proportion of non-standard occurrences per corpus

1,748 observations

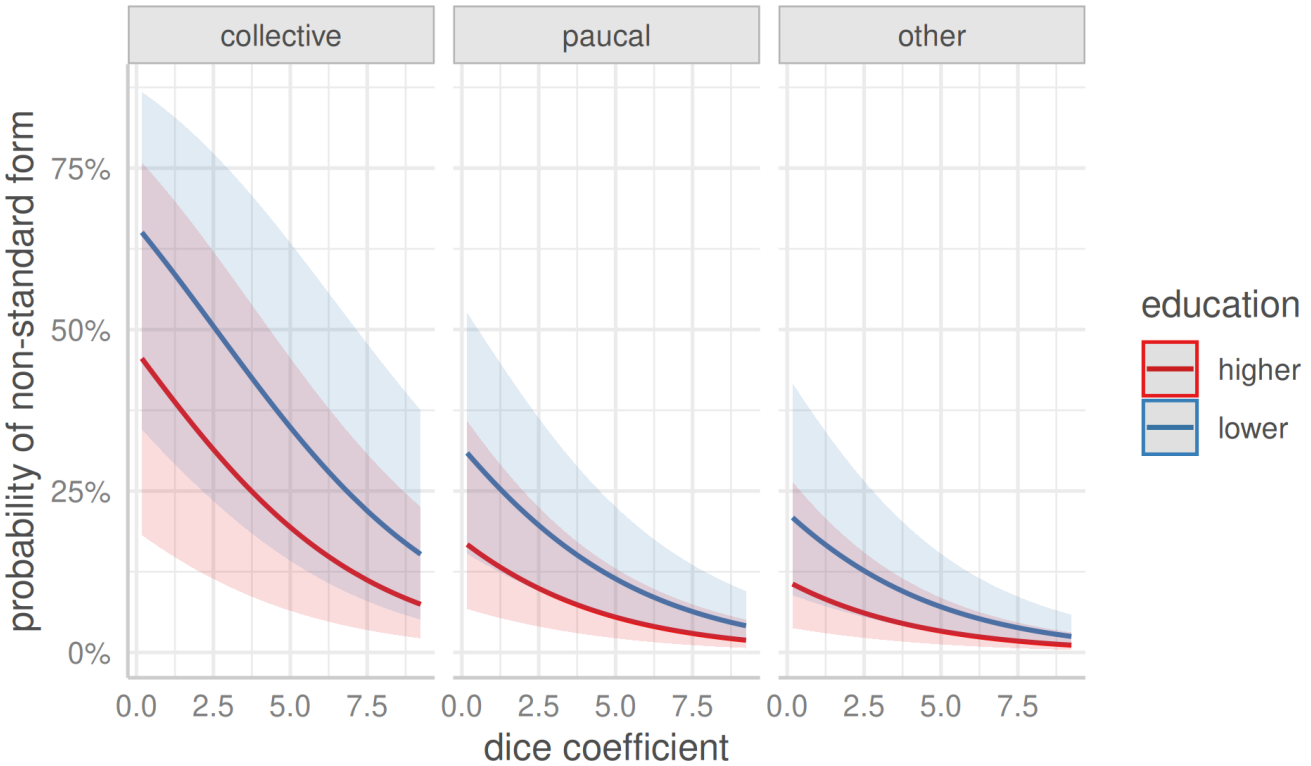


Distribution of n-std forms with different types of numerals

- NOM instead of GEN is frequent both with paucals and other numerals
- n-std GEN is attested sporadically
- other case forms are even less frequent
- only ~45% of n-std expressions could in principle be explained by L1 pattern borrowing



Statistical modelling



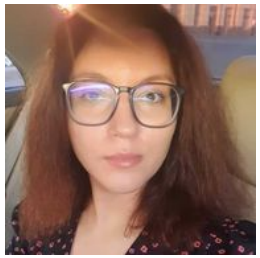
- Logistic regression: `standardness ~ Dice coefficient + year of birth + education + numeral type + gender + (1|L1 family/speaker id)`
- Conditional importance of the variables in our model (generalized R squared): collocationality (Dice coefficient) > education > year of birth > numeral type > gender

Conclusions

- Variation in NCs is attested in all L2 corpora, but not to the same extent in each of them
- Daghestanian Russian as a more uniform variety, probably due to a lower pervasiveness of Russian in every-day life, especially in the more isolated communities of the highlands
- The variables that turned out to be statistically significant are all logically related to L2 proficiency and exposure to the input, but there is no robust evidence for a contact explanation

Propositional Drop

Anastasia Yakovleva



Natalia Koshelyuk

George Moroz

Propositional Drop in Chuvash

Anna Grishanova

Dialect Genitive Plural Forms in Numeral Constructions

Svetlana Zemicheva

Chiara Naccarato

George Moroz

Motivation

- widespread feature
- was studied in standard Russian & bilingual varieties
- in standard Russian some special numerative forms tend to appear in numeral constructions (Kholodilova, forthcoming)
- dialect speech may be controversial to the tendencies observed in other colloquial varieties

Examples

- $\text{Num(not-paucal)}_{\text{nom,acc,gen}} + \text{N}_{\text{m,n-Gen.Pl}}$
  - pjat' xozjaev-ov (dial. five owner-Gen.Pl)
  - pjat' xozja-ev-Ø (stnd. five owner-Pl-Gen)
- $\text{Num(paucal)}_{\text{acc,gen}} + \text{N}_{\text{m,n-Gen.Pl}}$
  - tridcat-i dv-ux god-ov (dial. thirty-Gen two-Gen year-Gen.Pl)
  - tridcat-i dv-ux let-Ø (stnd. thirty-Gen two-Gen year-Gen.Pl)

Research questions

- What factors may affect the probability of use of dialect Gen.Pl forms in numeral constructions?
  - Overall frequency of dialect Gen.Pl in different contexts
  - Noun stem
  - Numeral-Noun collocationality level
  - Numeral form (Nom/Gen)
  - Year of birth
  - Education level
  - Gender
- Does dialect "overuse" [Kasatkin, 2005] of the *-ov* ending affect cases like *kilogram* vs *kilogram-ov*?

Does dialect “overuse” of the *-ov* ending affect measure words?

No

Does noun stem play a role?

Geographical distribution

Does overall frequency of dialect Gen.Pl significant?

Is year of birth significant?

Is education level significant?

Is the level of collocationality significant?

Is the numeral form significant?

Conclusions

- Dialect "overuse" of the *-ov* ending doesn't affect mesuare words: zero inflexion forms (*kilogram*) are found in more than 90% of contexts
- Early year of birth & low education level make the probability of using dialect Gen.Pl form higher
- 'The most frequent pairs survive' (Chiara): the probability of dialect form higher for nouns which often cooccur with numerals
- The type of noun stem seems to be significant
- Potential contact influence in the case of Spiridonova Buda? (Belorussian)
- The correlation between the frequency of dialect Gen.Pl forms in numeral constructions and other types of contexts is questionable

Negative Existential Constructions

Chiara Naccarato

George Moroz

Negative Existential Constructions

- Existential negation = negation strategies used in existential sentences of the type *there is/are no* X (*somewhere*), in which the subject is typically non-referential
- We use the terms "existential negation" and "negative existential constructions" (NECs) in a wider sense to include constructions that are sometimes referred to as "locative negation" (*X is/are not in some place*, in which X is a definite subject) and "possessive negation" (*Y does/do not have X*); cf. [Veselinova, 2013, 110–111]
- All of them predicate absolute absence rather than relative absence, and Russian employs one and the same strategy in all three cases, which is different from the strategy employed in standard negation, i.e. negation of overt verb predicates

Non-standard marking in NECs

- Variation in NECs in bilingual corpora (+ comparison with the monolinguals' variety of Russian spoken in Zvenigorod)

e.g. *gaz ne bylo* vs. *gaza ne bylo*

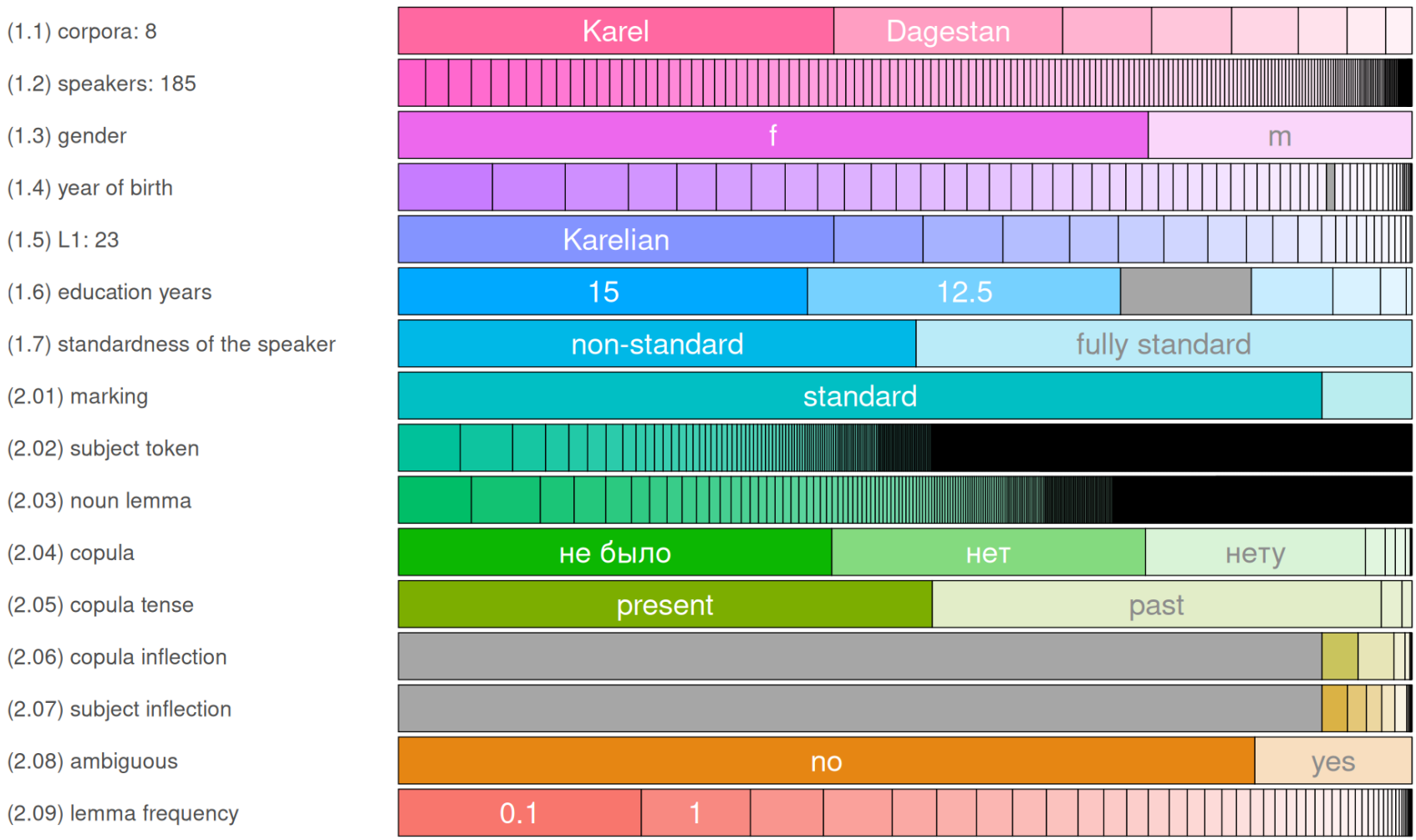
- Previous research on other L2 Russian varieties
  - Nanai and Ulcha Russian [Stoynova, 2019, 27]
  - Moksha Russian [Kashkin, 2020, 116]
  - Hill Mari Russian [Kashkin, 2022, 39]
- Usually treated as a contact phenomenon because in the L1s of Russian bilinguals who display this trait there is no genitive (or any other special) marking of negated subjects

Research questions

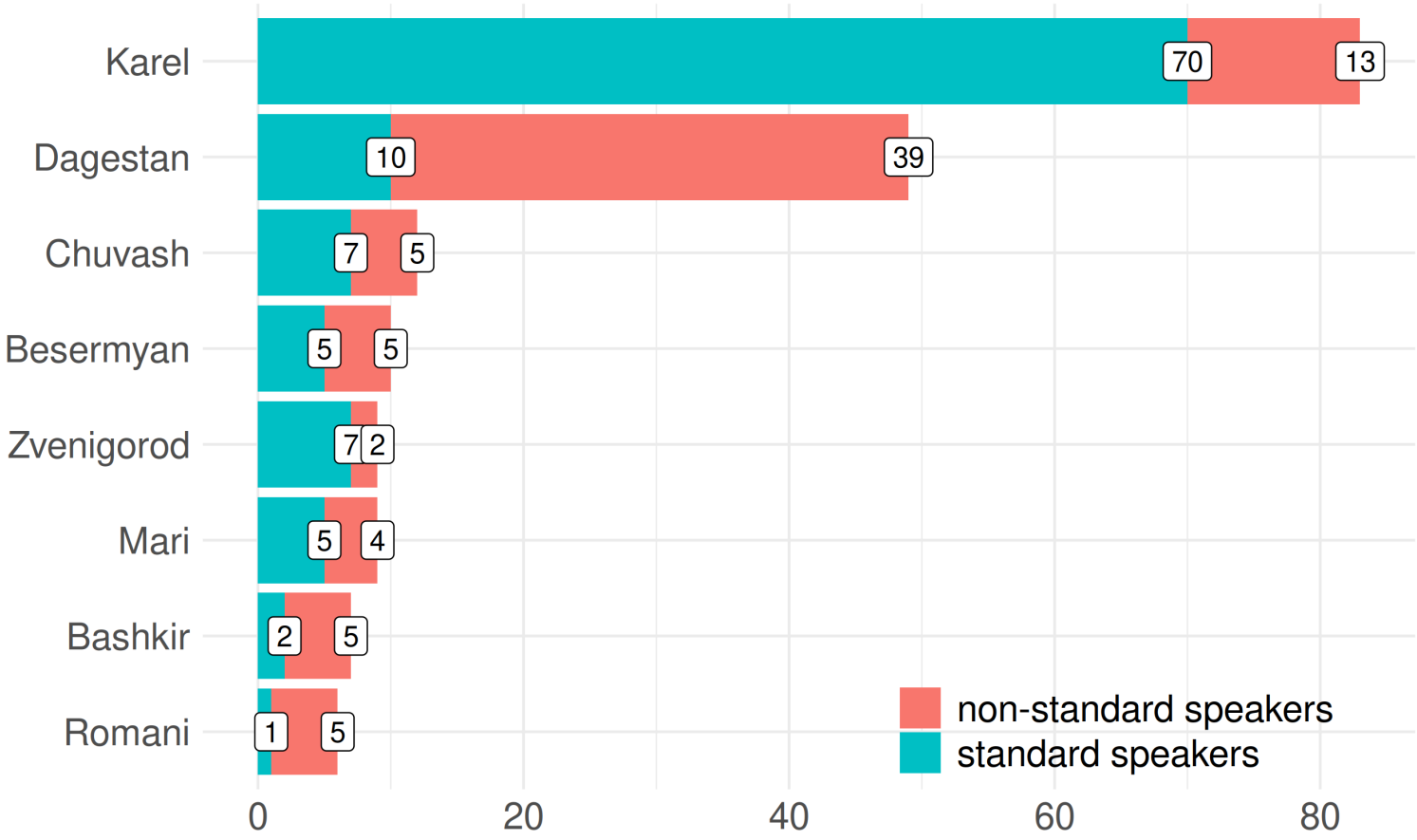
- Does the amount of variation in NECs differ across corpora and/or among speakers of the same variety?
- Can variation in NECs be explained in terms of contact influence?
- Do other factors promote or hinder variation in NECs?

The database and parameters of data annotation

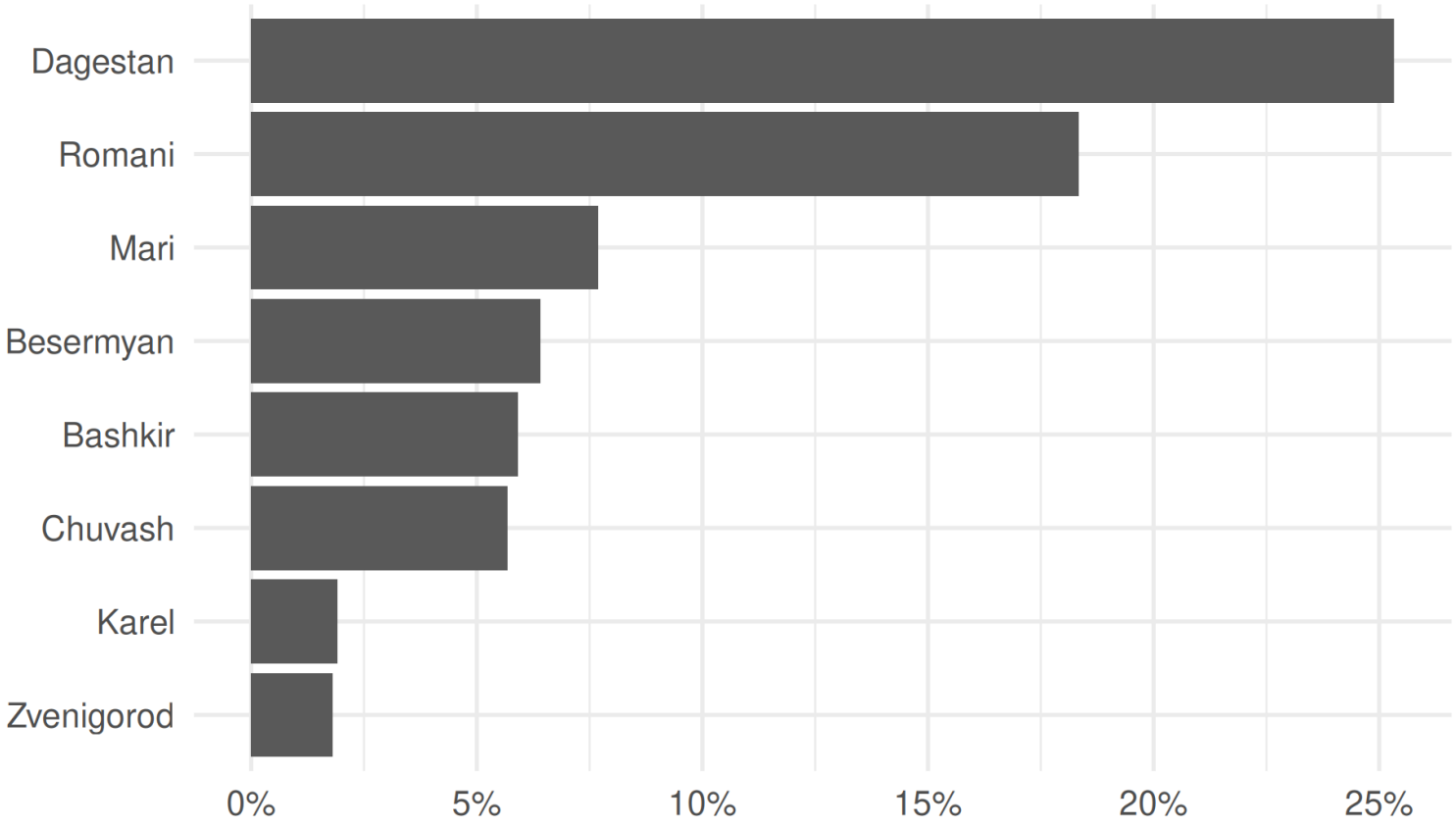
2,309 observations



Fully standard (58%) vs. non-standard speakers (42%)



Proportion of non-standard occurrences per corpus



Types of non-standard marking

- Neuter copula
  - *gaz ne bylo*
- Non-neuter copula
  - *doma ne byl*
- Agreeing subject (could be pattern borrowing for Daghestan)
  - *bogatye ljudi ne byli*

Statistical modelling

Preliminary conclusions

- Findings comparable to those obtained for NCs
- Variation attested in all L2 corpora, but not to the same extent in each of them
- Daghestan as a more uniform variety
- Not all cases of variation can be explained by contact

Sideproject

The DiaL2 sideproject

Anna Panova

Olga Gich

George Moroz

Future plans

References I

M. Daniel, N. Dobrushina, and S. Knyazev. Highlanders' Russian: Case study in bilingualism and language interference in Central Daghestan. *Slavica Helsingiensia*, 40:65–93, 2010.

L. L. Kasatkin. *Russkaya dialectologiya* [*Dialectology of Russian*]. Academia, Moscow, 2005.

E. V. Kashkin. Osobennosti russkoj reči nositelej mokšanskogo jazyka [peculiarities of the russian speech of moksha speakers]. *Trudy instituta russkogo jazyka im V.V. Vinogradova*, 26(4):110–131, 2020.

E. V. Kashkin. O nestandartnom (zametki o russkoj reči gornyx marijcev) [on non-standard features of russian in the grammar and lexicon of hill mari speakers]. *Rodnoj jazyk*, (2):35–51, 2022.

References II

I. Khomchenkova. Contact-induced features in the Russian speech of Nganasans. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 11(2):13–37, 2020.

G. Moroz. Phonetic fieldwork research and experiments with the R package phonfieldwork. In I. Kobozeva, K. Semyonova, A. Kostyuk, L. Zakharov, and N. Svetozarova, editors, *«...Vperyod i vverkh po lestnitse zvuchashey». Sbornik statye k 80-letiyu Olgi Fyodorovny Krivnovoy* [*Festschrift in memoriam to Olga Fyodorovna Krivnova*], pages 376–390. Buki Vedi, Moscow, 2023.

George Moroz. *lingtypology: easy mapping for Linguistic Typology*, 2017. URL https://CRAN.R-project.org/package=lingtypology.

References III

A. Panova and T. Philippova. When a cross-linguistic tendency marries incomplete acquisition: Preposition drop in Russian spoken in Daghestan. *International Journal of Bilingualism*, 25(3):640–667, 2021.

E. V. Rakhilina and A. K. Kazkenova. Zametki o russkom čisle [Notes on Russian number]. *Russian Journal of Linguistics*, 22(3):605–627, 2018.

K. Shagal. Contact-induced grammatical phenomena in the Russian of Erzya Speakers. In *Mordvin languages in the field*, pages 363–377. University of Helsinki, 2016.

N. Stoynova. Russian in contact with Southern Tungusic languages: Evidence from the Contact Russian Corpus of Northern Siberia and the Russian Far East. *Slavica Helsingiensia*, 52:9–36, 2019.

References IV

N. Stoynova. Nestandartnye količestvennye konstrukcii v russkoj reči nositelej nanajskogo i ul'čskogo jazykov [Non-standard syntax of numerals in the Russian of Nanai and Ulcha speakers]. *Russian Linguistics*, 45(3):305–334, 2021.

L. Veselinova. Negative existentials: A cross-linguistic study. *Rivista di linguistica*, 25(1):107–145, 2013.