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 each of them:
  - Corpus of Russian spoken in Daghestan
  - Ustja River Basin Corpus

Precursors of the project

Nina Dobrushina

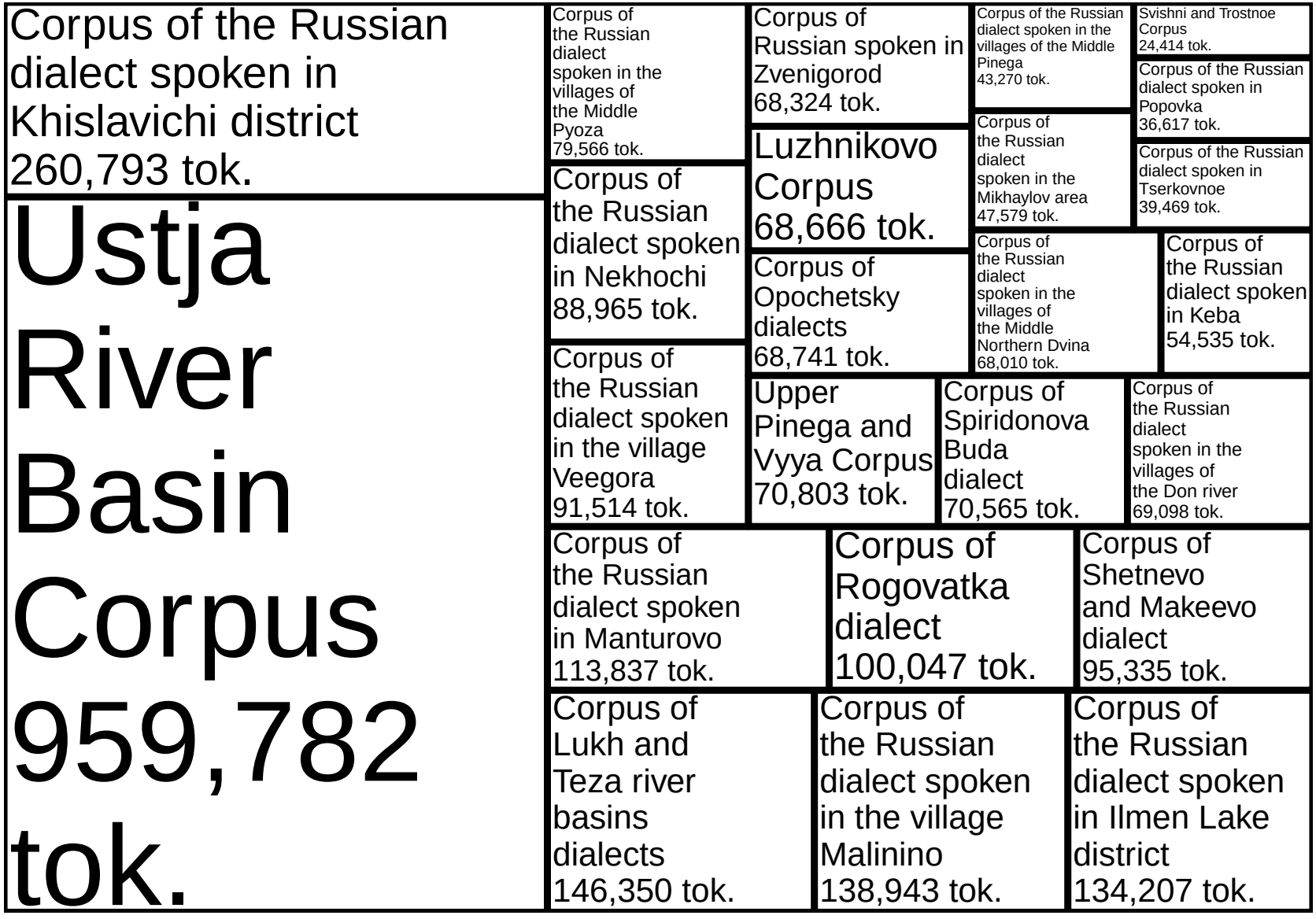
Michael Daniel

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

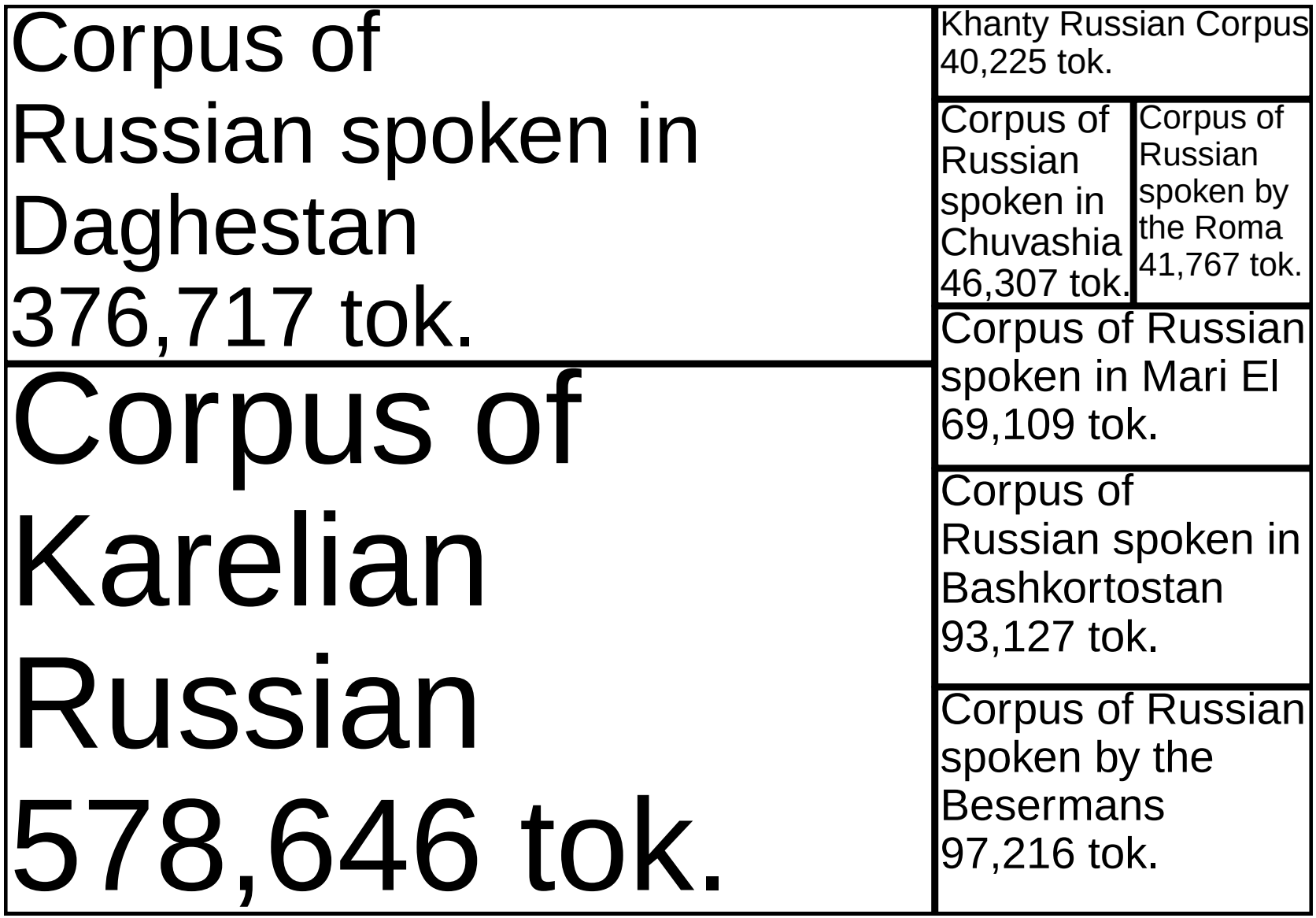
Resources of the Linguistic Convergence Laboratory

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

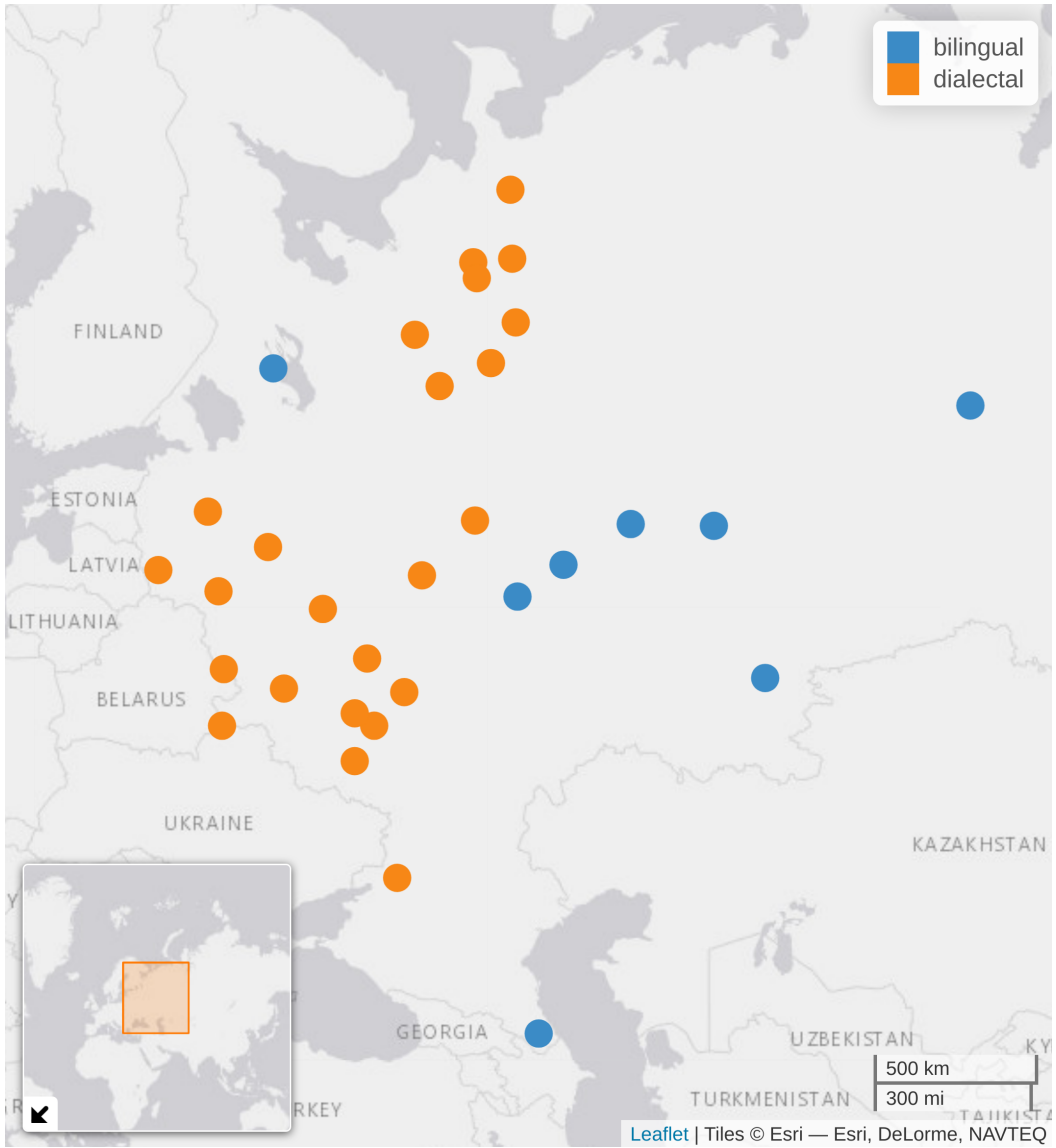
Dialectal Corpora



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

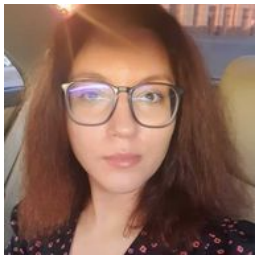
- Daghestanian Russian [Daniel et al., 2010, Naccarato et al., 2021, Panova and Philippova, 2021]
- Russian of Erzya speakers [Shagal, 2016]
- Russian of Kazakh speakers [Rakhilina and Kazkenova, 2018]
- Contact Russian of Northern Siberia and the Russian Far East [Stoynova, 2019, 2021]
- Russian of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

DiaL2

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 syntactic annotation
- filter the result table for a particular feature selected by the researcher
- annotate standardness of the utterances
- remove fully-standard speakers
- model the standardness of the utterances using sociolinguistic and linguistic features as predictors

Num constructions

Non-standard numeral constructions in L2 Russian

Chiara Naccarato

George Moroz

Non-standard numeral constructions in L2 Russian

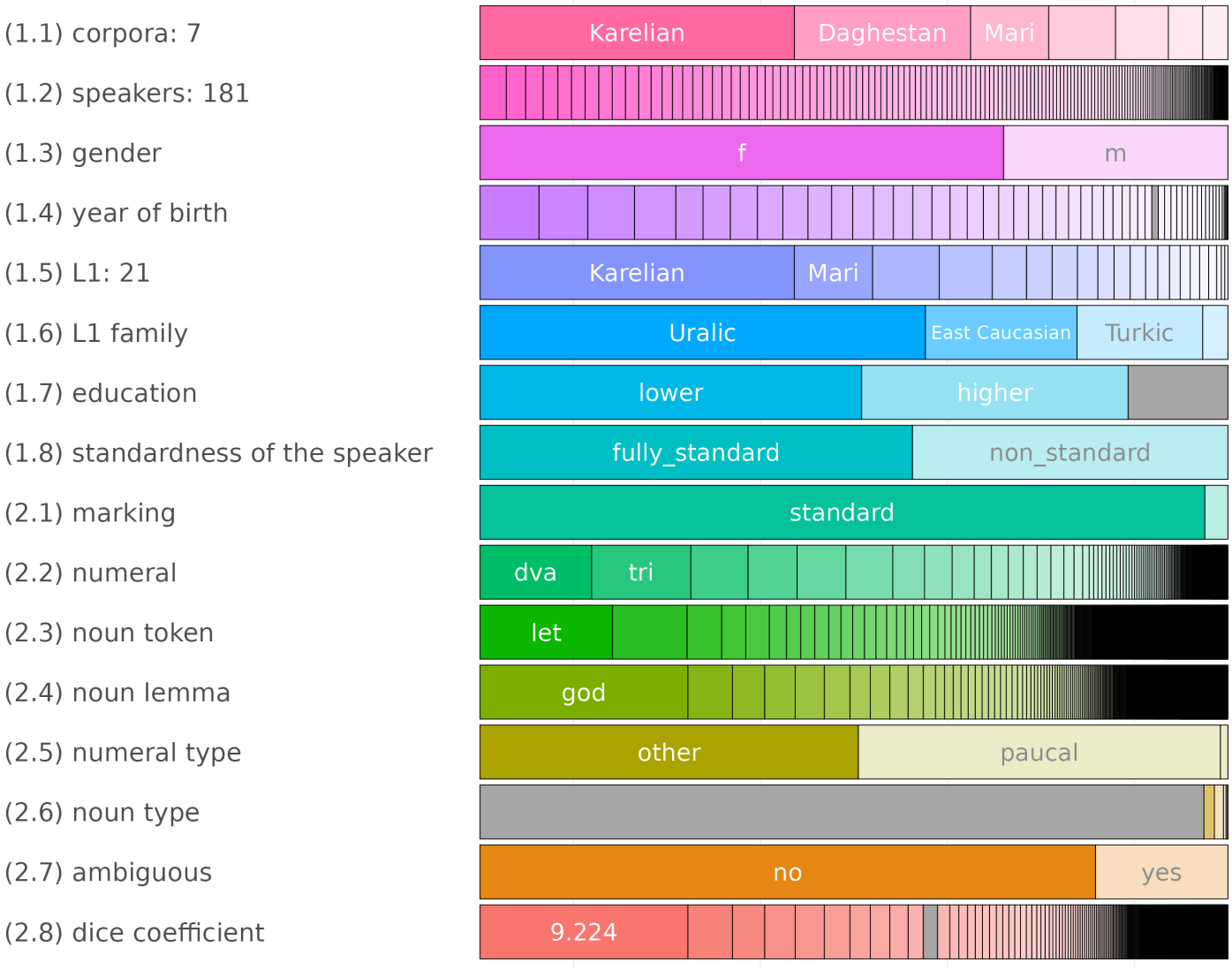
- Variation in numeral constructions (NCs) in bilingual corpora
  - e.g. *dva brat* vs. *dva brata*
- Previous research on other L2 Russian varieties
  - Stoynova (2021) on Nanai and Ulcha Russian: evidence for pattern borrowing
- Also mentioned by
  - Shagal (2016: 369-370) for Erzya Russian
  - Rakhilina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

Research questions

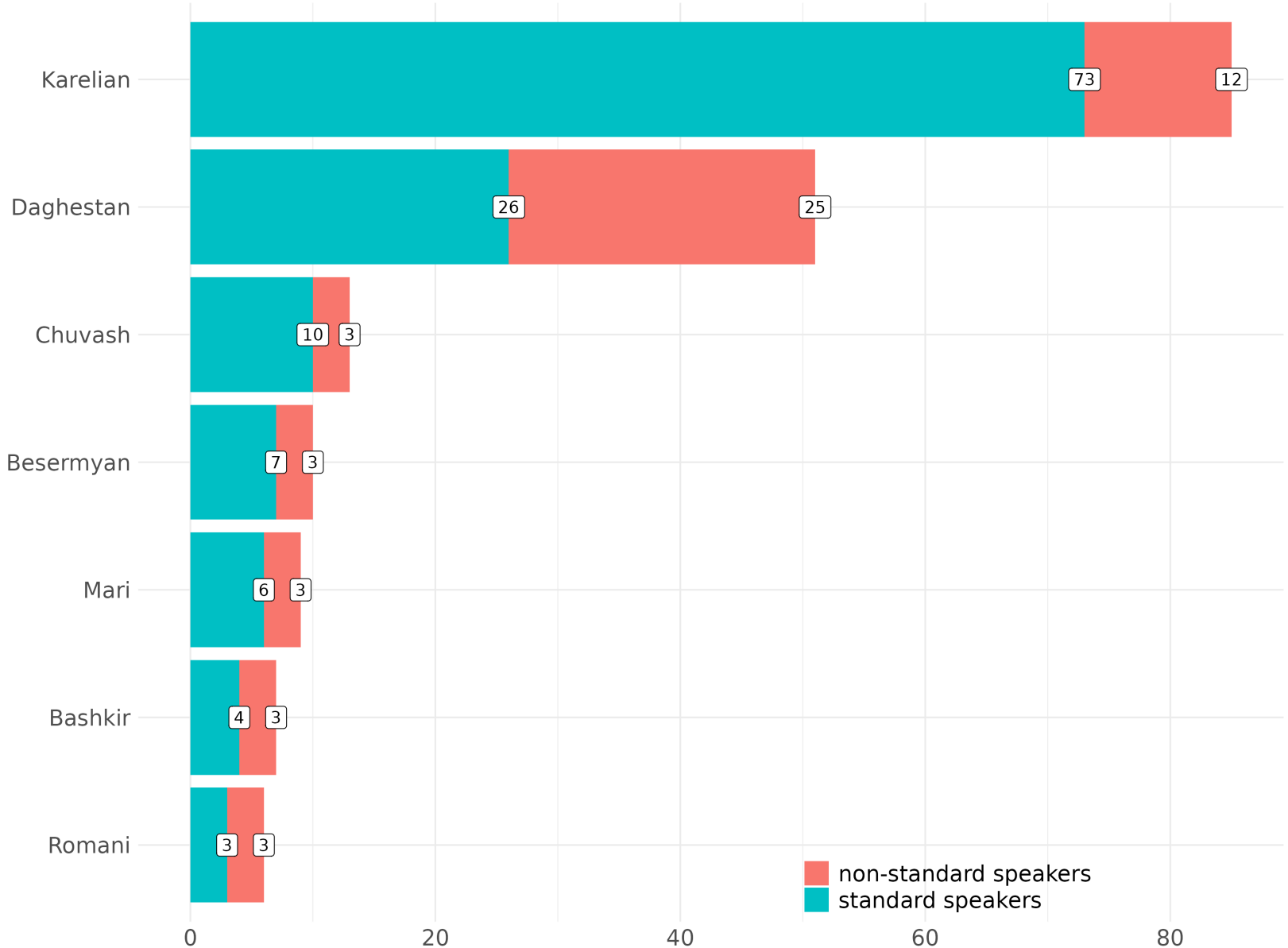
- Does the amount of variation in NCs differ across corpora and/or among speakers of the same variety?
- Can variation in NCs be explained in terms of contact influence?
- Do other factors promote or hinder variation in NCs?

The database and parameters of data annotation

4,144 observations

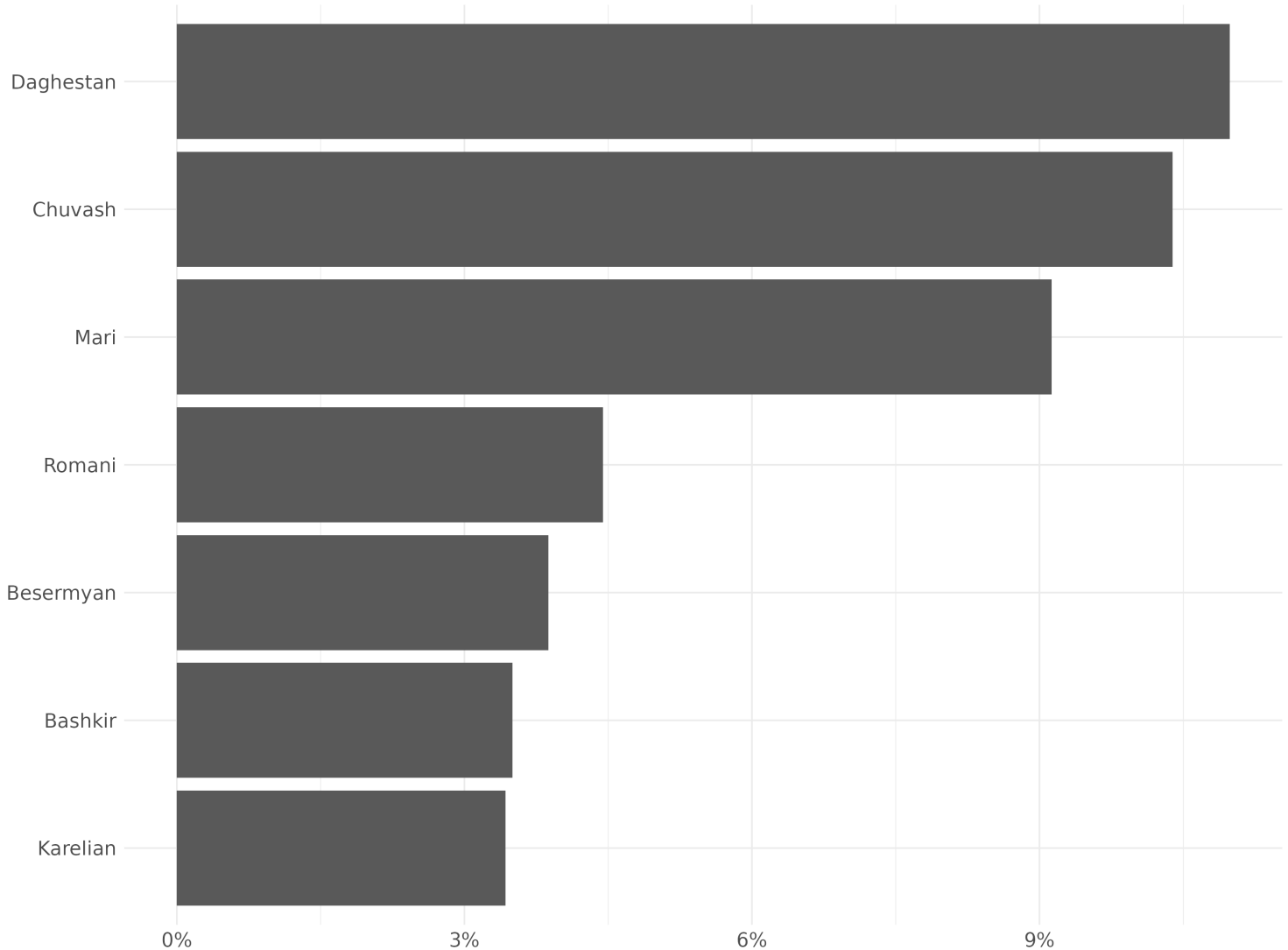


Fully standard (71.3%) vs. non-standard speakers (28.7%)



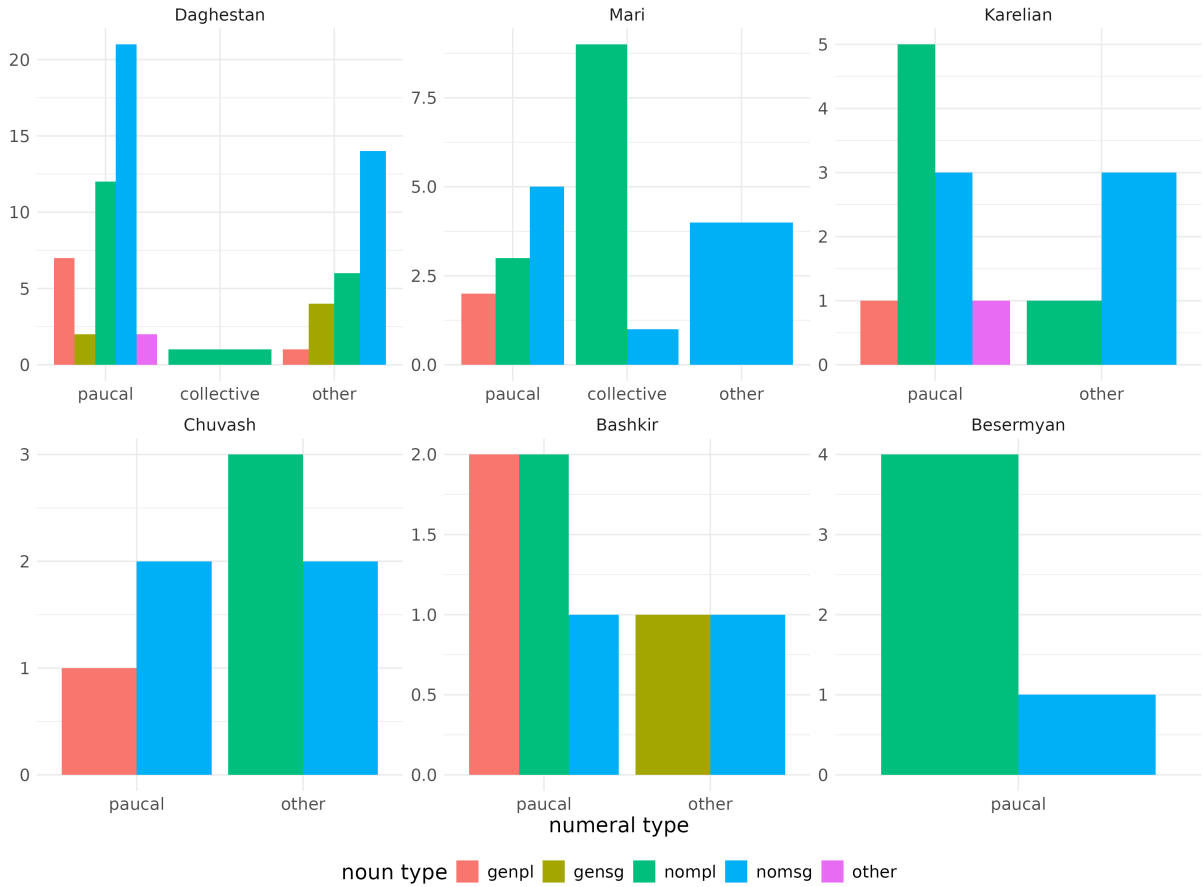
Proportion of non-standard occurrences per corpus

1,748 observations

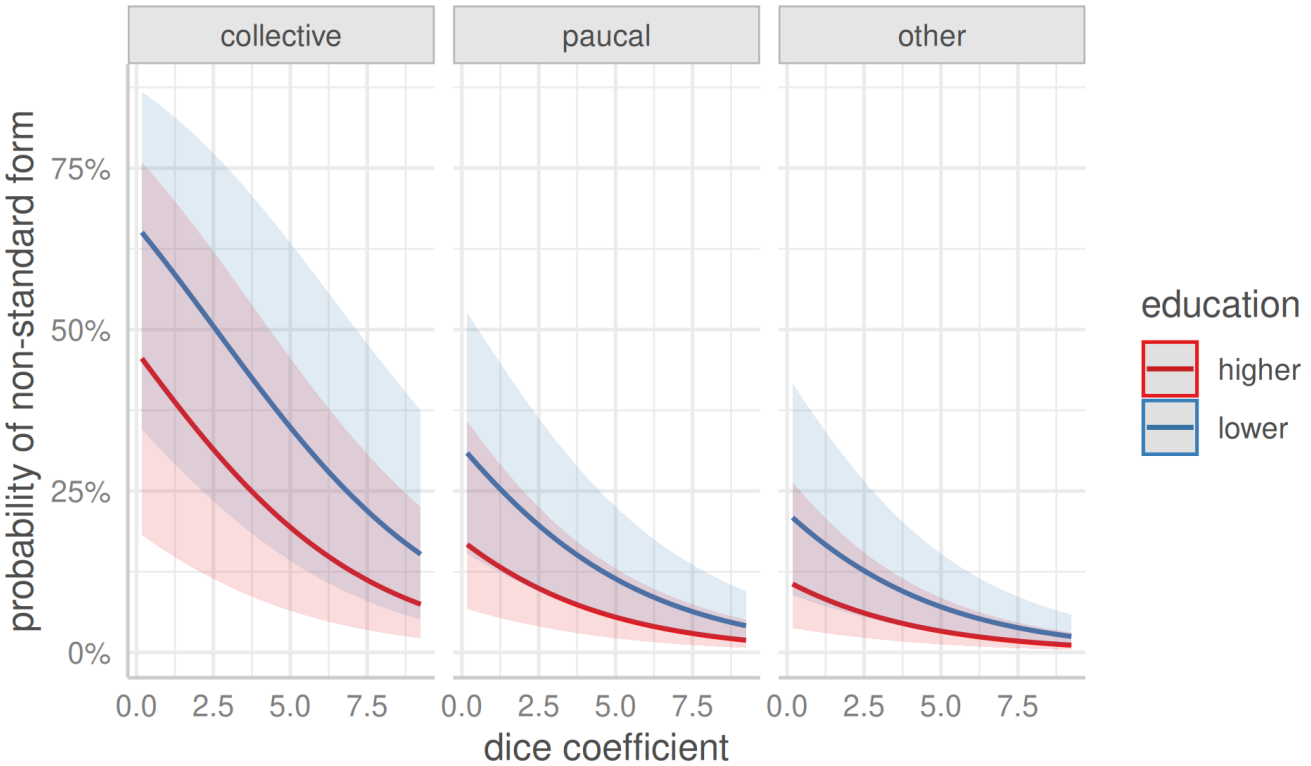


Distribution of n-std forms with different types of numerals

- 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

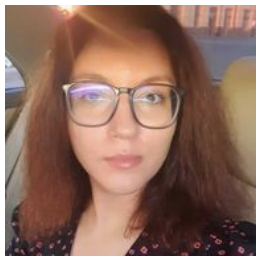
Conclusions

- 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

Preposition drop

Preposition drop in Russian spoken by Mari and Beserman bilinguals

Anastasia Yakovleva



Natalia Koshelyuk

George Moroz

Aims and Research Questions

- a corpus-based study of preposition drop (p-drop) in the speech of Mari-Russian and Beserman-Russian bilinguals compared to the speech of Russian monolinguals;
- demonstrate that the prepositions *v* 'in', *k* 'to', *s* 'with' are omitted in the speech of bilinguals more often than in monolinguals' speech;
- propose some possible explanations for the variation attested across different bilingual speakers.

Methods and Data

- spontaneous oral speech of 20 Beserman-Russian and Mari-Russian bilinguals;
- total number of tokens ~166,000;
- in comparison with the speech of nine Russian monolinguals.

Methods and Data

(1.1) corpora: 3

(1.2) speakers: 29

(1.3) gender

(1.4) year of birth

(1.5) years of education

(2.1) preposition lemma

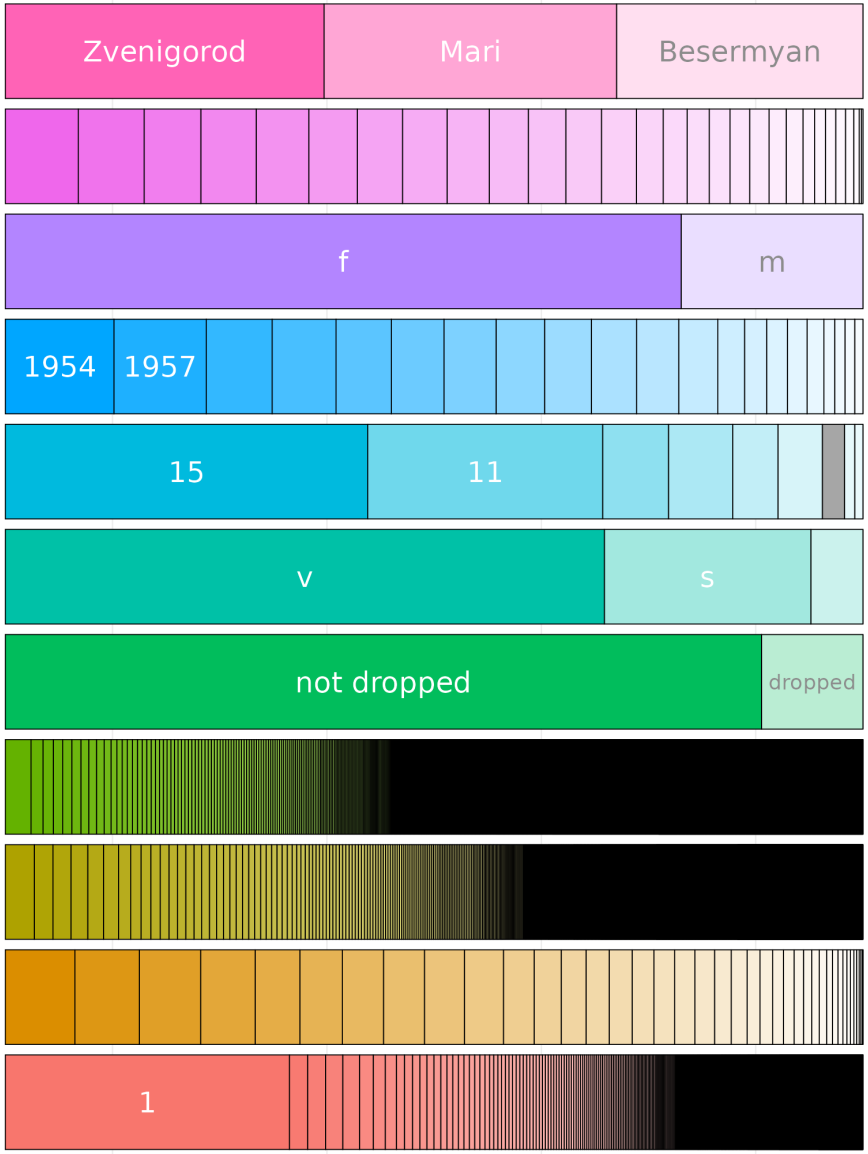
(2.2) preposition drop

(2.3) following wordform's token

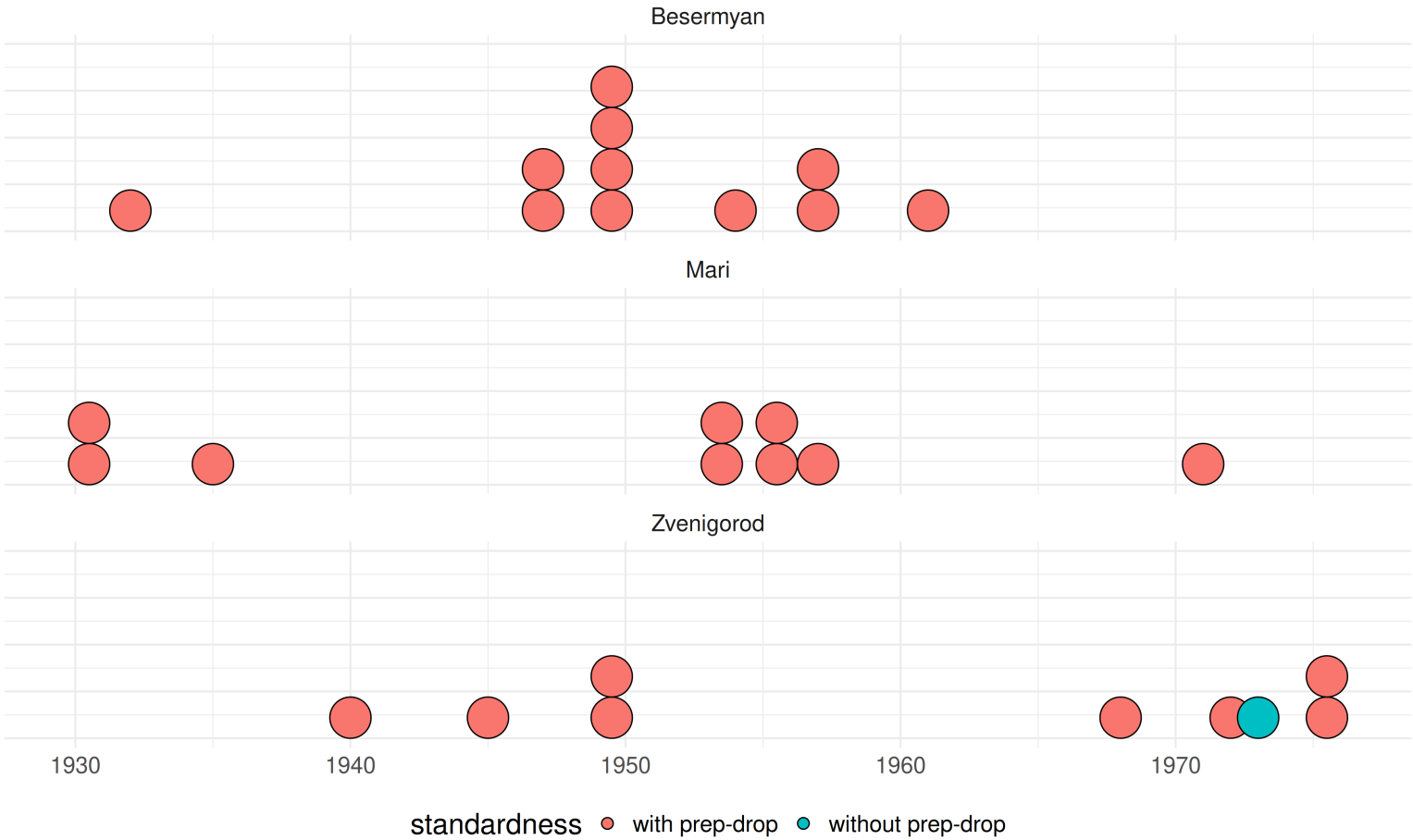
(2.4) following wordform's lemma

(2.5) following wordform's first sound

(2.6) dice coefficient

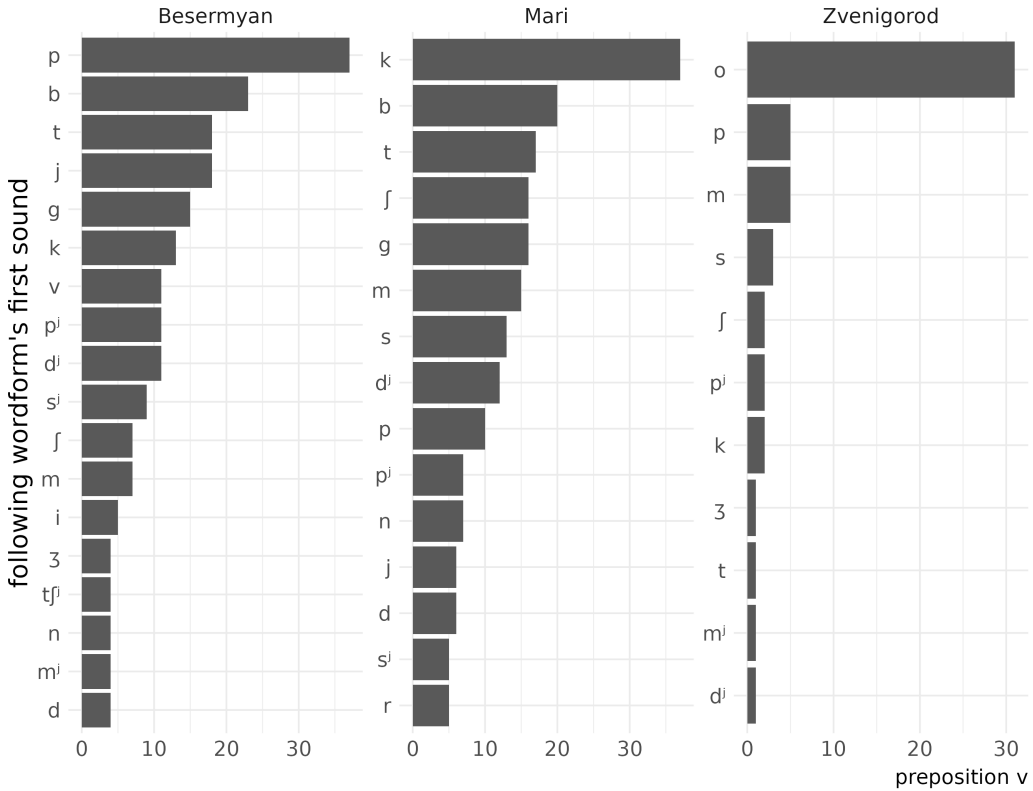


Findings



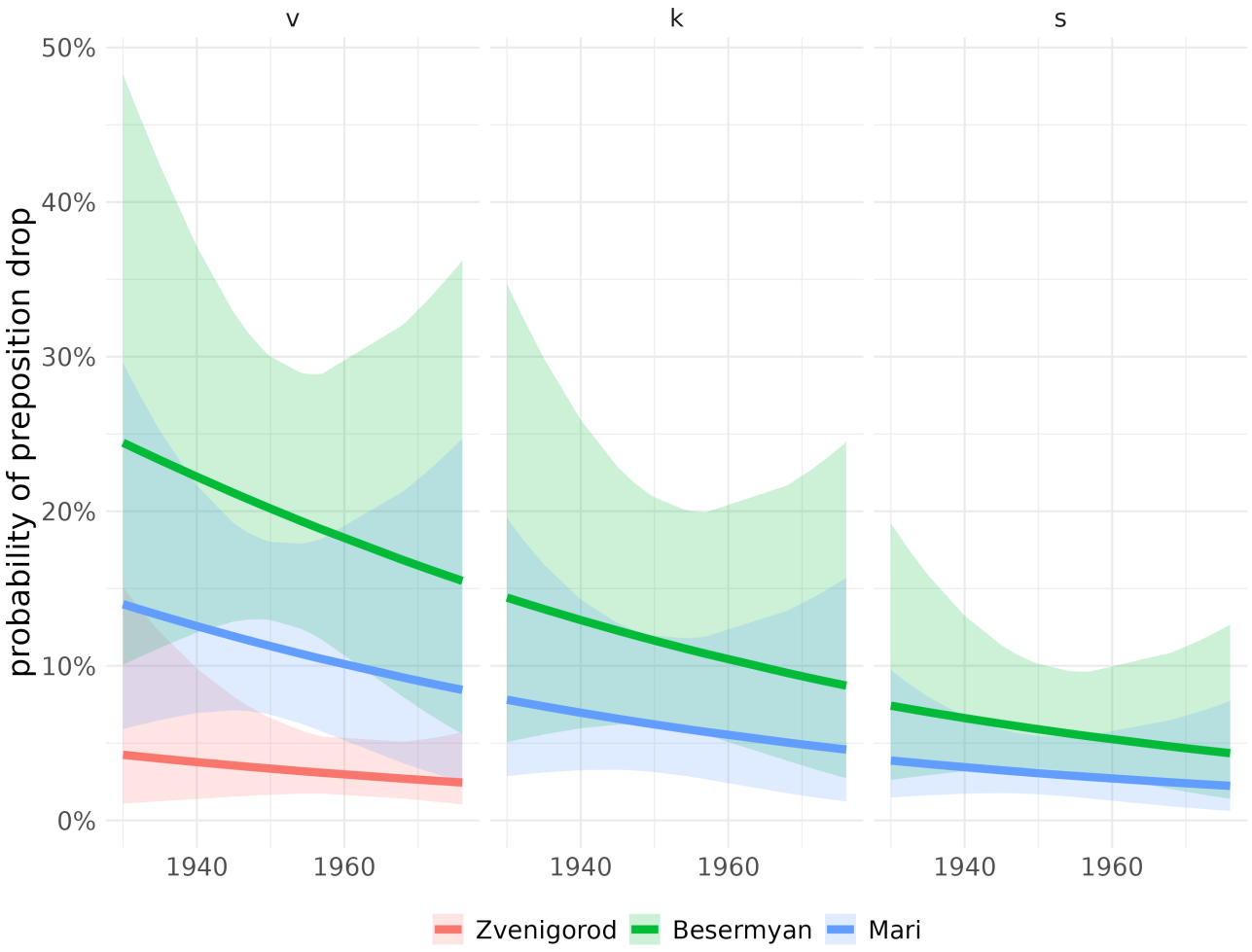
- both bilinguals and monolinguals omit prepositions;
- however, they do so in different contexts and for different reasons

Findings



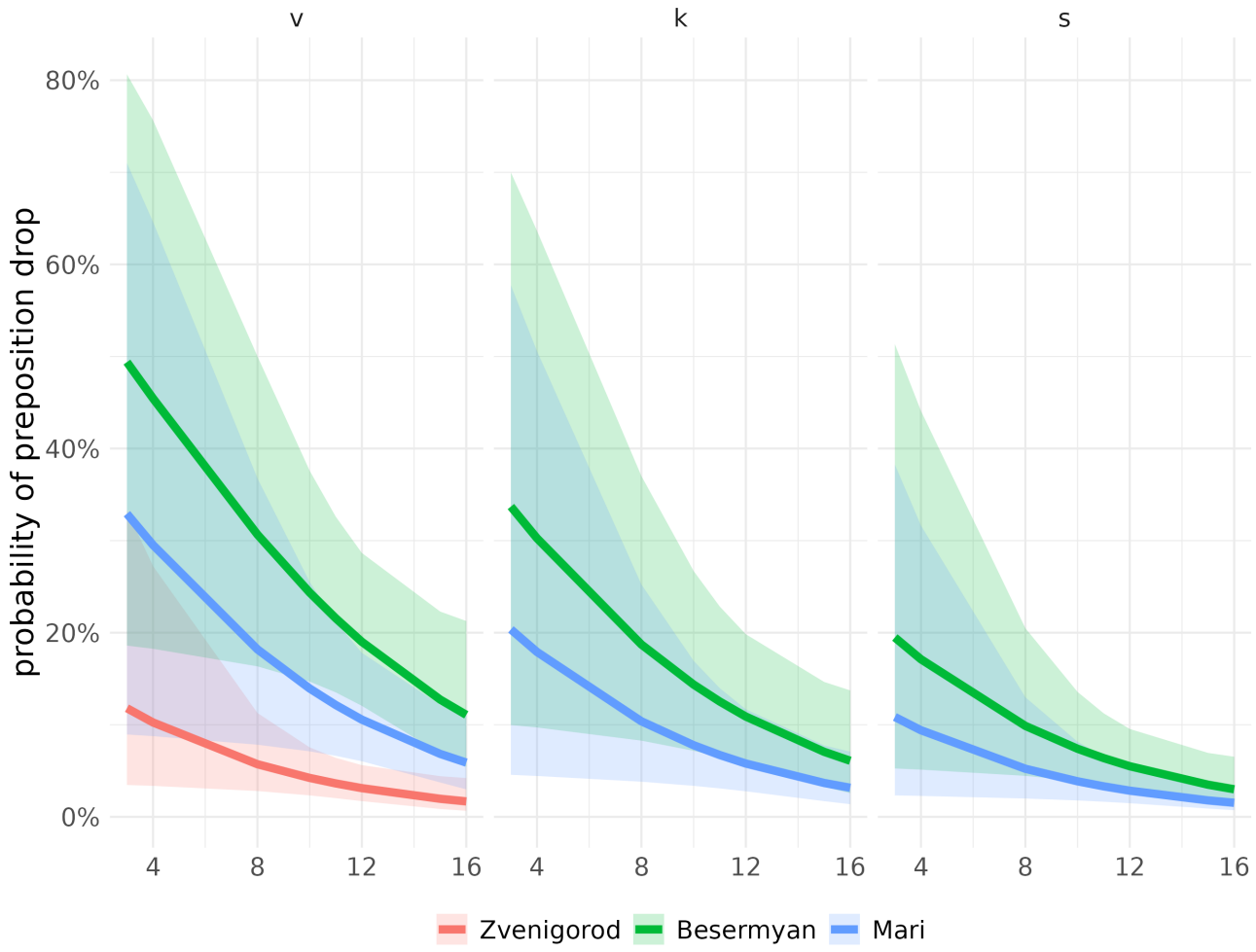
- In bilinguals' speech, p-drop can possibly be explained by phonetic interference from their native languages (avoiding consonant cluster);
- monolinguals mainly omit prepositions in lexicalized expressions (*v obščem* 'in general').

Findings: sociolinguistic factors



The older and less educated a person is, the higher the probability of p-drop in their speech.

Findings: sociolinguistic factors



The older and less educated a person is, the higher the probability of p-drop in their speech.

Preposition Drop in Chuvash

Anna Grishanova

Preposition drop in Chuvash Russian

- The following prepositions were dropped at least once: *v* 'in/to', *u* 'at', *na* 'on/to', *s* 'with/from/off', *iz* 'from/of', *do* 'up to/until', *čerez* 'via/through/later'.
- This includes syllabic prepositions (such as *na* 'on') that do not tend to drop in the speech of Mari and Beserman bilinguals:
- [*na*] *každom povorote stolba stoit* 'A pillar stands on every turn'

Preposition drop in Chuvash Russian

- The omission of syllabic prepositions and the sporadic nature of the drop of certain prepositions compelled a more semantic approach to this data. Thus, every context in the dataset was annotated according to four big semantic groups: location, direction, source and time.
- Another interesting feature found in Chuvash Russian speech was non-standard case usage, in particular the expansion of the Nominative case. Thus, we also annotated instances of non-standard case usage and used the following word's case as a factor in statistical modeling.

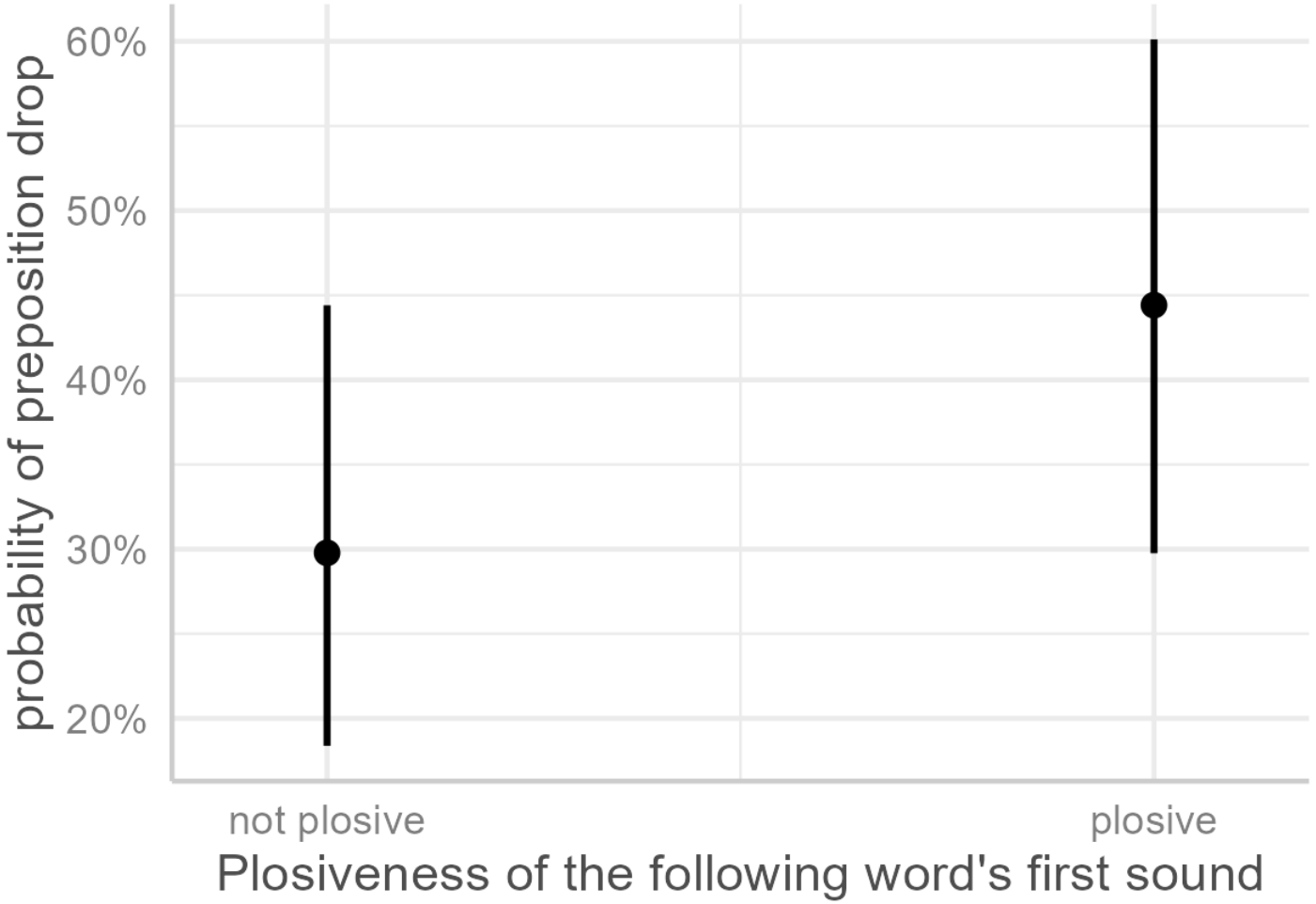
Preposition drop in Chuvash Russian

We implemented a nested mixed effect logistic regression model.

Sociolinguistic factors like **age** and **years of education** appeared to be not significant. The significant factors turned out to be:

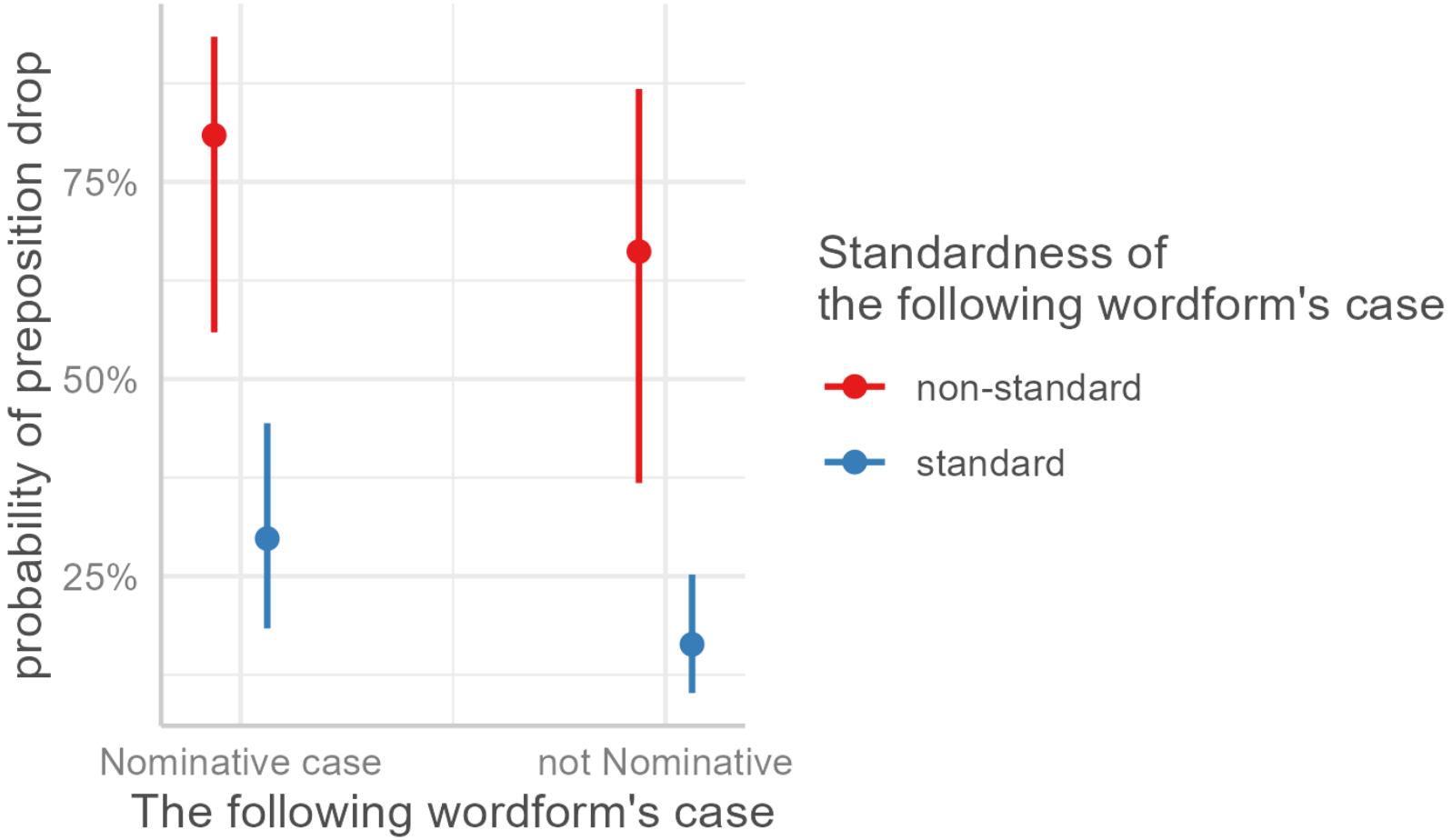
- semantics of the prepositional phrase
- the plosiveness of the following word's first sound
- the Nominative case of the following word
- dice coefficient
- the standardness of the following word's case

Preposition drop in Chuvash Russian



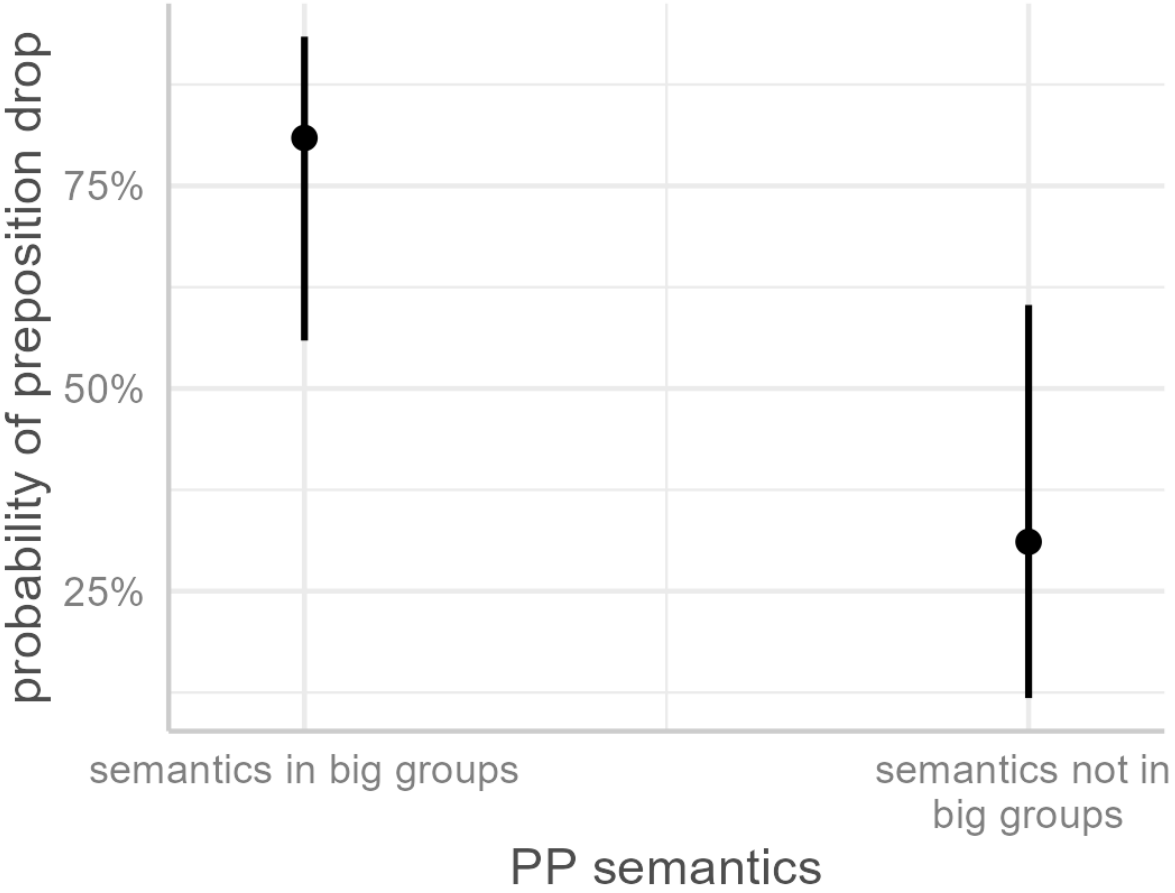
The plosiveness of the following word's first sound increases the probability of preposition drop

Preposition drop in Chuvash Russian



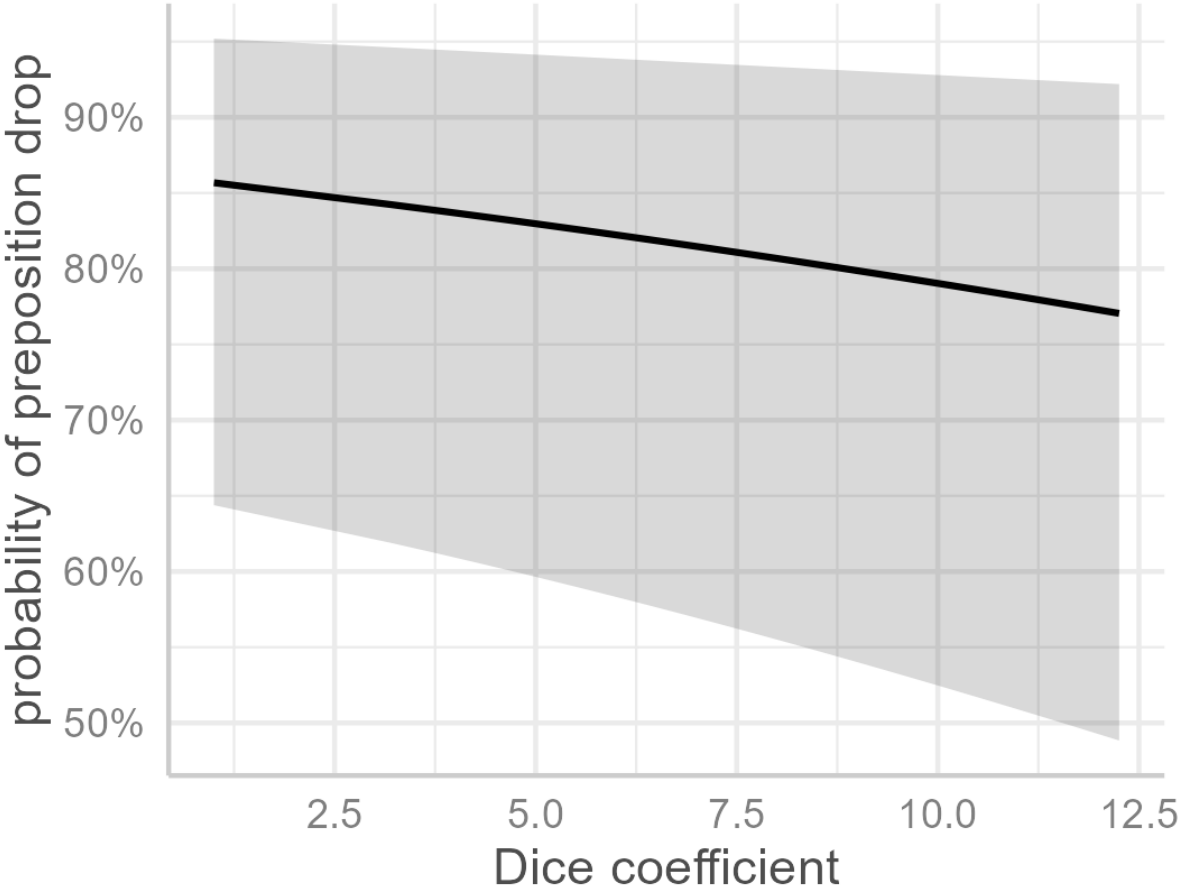
The Nominative or a non-standard case of the following wordform increases the probability of preposition drop

Preposition drop in Chuvash Russian



On the other hand, the semantics of the prepositional phrase not being that of location, direction, source and time decreases the probability of a preposition drop.

Preposition drop in Chuvash Russian



Finally, just like in the study of preposition drop in the speech of Mari and Beserman bilinguals, higher degree of collocationality (dice coefficient) decreases the probability of omission of a preposition.

Gen Pl Forms

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 show different tendencies as compared to other spoken varieties of Russian

Examples

- Num(not-paucal)$_{\text{nom,acc,gen}}$ + N$_{\text{m,n-Gen.Pl}}$
  - pjat' xozjaev-ov (dial. five owner-Gen.Pl)
  - pjat' xozja-ev-Ø (stnd. five owner-Pl-Gen)
- Num(paucal)$_{\text{acc,gen}}$ + 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 using dialect Gen.Pl forms in numeral constructions?
  - Overall frequency of dialect Gen.Pl in different contexts
  - Noun stem
  - Numeral-Noun collocationality
  - Numeral form (Nom/Gen)
  - Year of birth
  - Education level
  - Gender
- Does dialect "overuse" [Kasatkin, 2005] of the *-ov* ending affect cases like *kilogramm* vs *kilogramm-ov*?

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

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

No

lexeme	*-ov*	Ø	total	percentage of Ø forms
hectare	2	25	27	93%
kilogram	2	76	78	97%
gram	0	58	58	100%

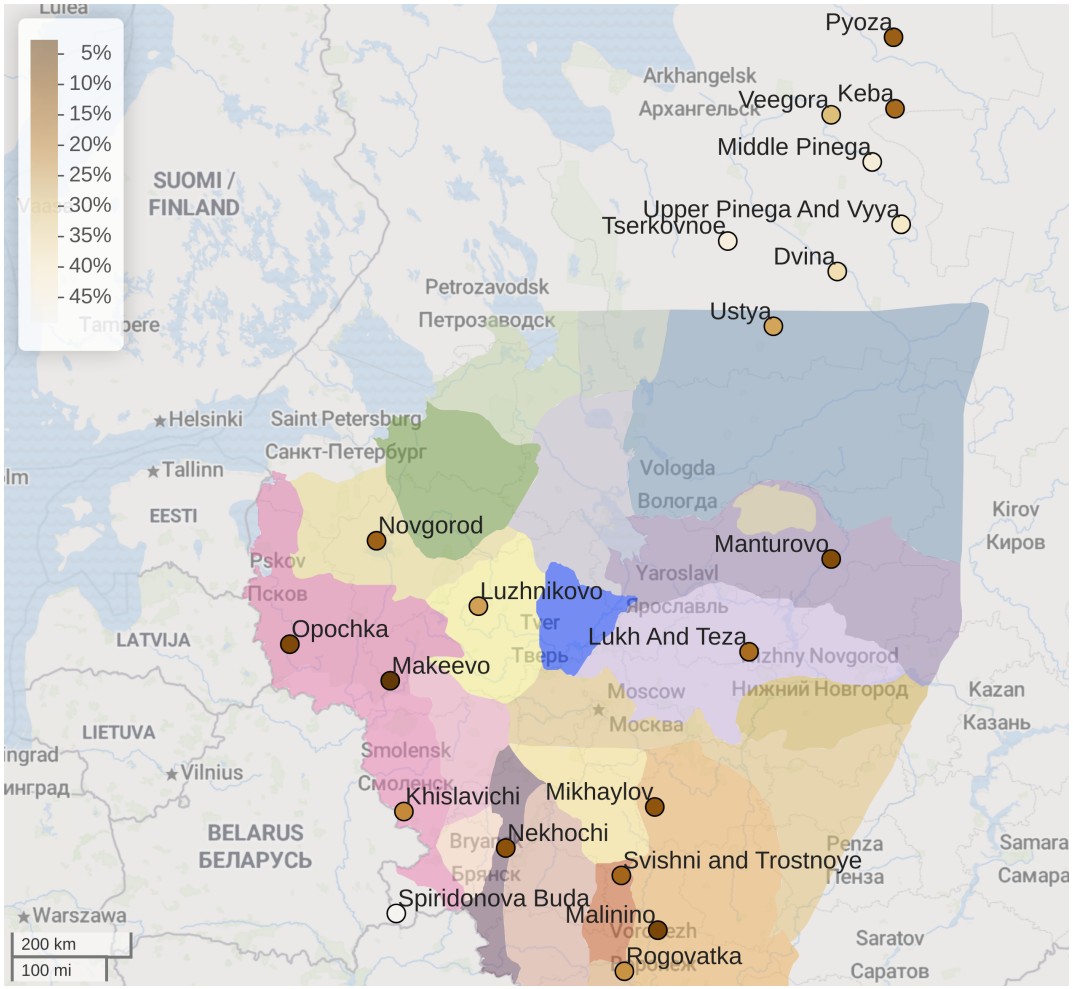
Does noun stem play a role?

Yes

stem type	dialect	standard	total	dialect share
1 – t	503	1297	1800	28%
2 – t’	20	356	376	5%
3 – g	3	278	281	1%
3* – g	3	20	23	13%
4 – ž	1	6	7	14%
5 – č	21	56	77	27%
6 – a	0	2	2	0%
7 – I	0	1	1	0%

Fisher exact test p-value = 0.0004998

Geographical distribution



Dialect polygones by I. A. Marchenko, created with `lingytpology` [Moroz, 2017]

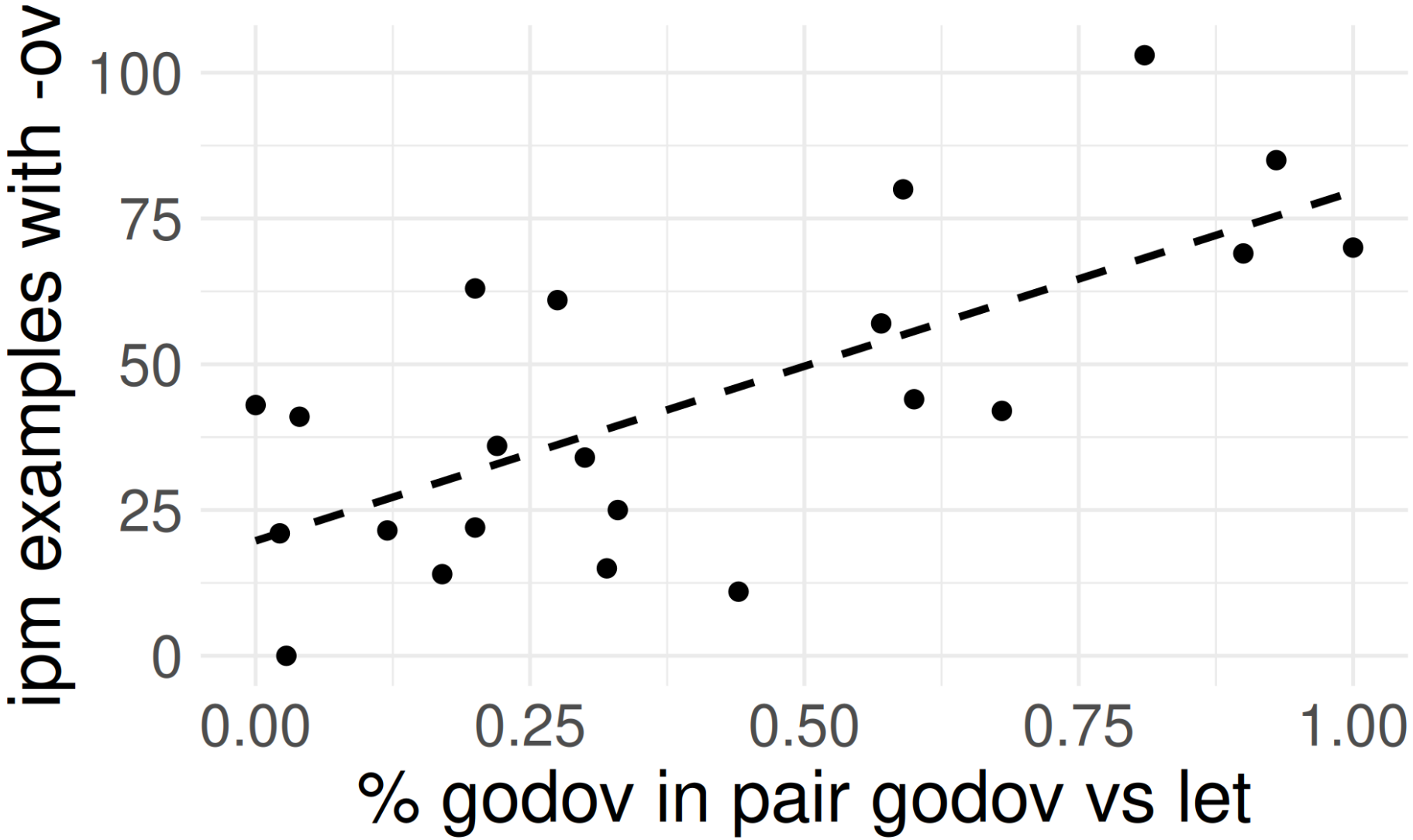
No geographical pattern observed (but, see Spiridonova Buda).

Is overall frequency of dialect Gen.Pl significant?

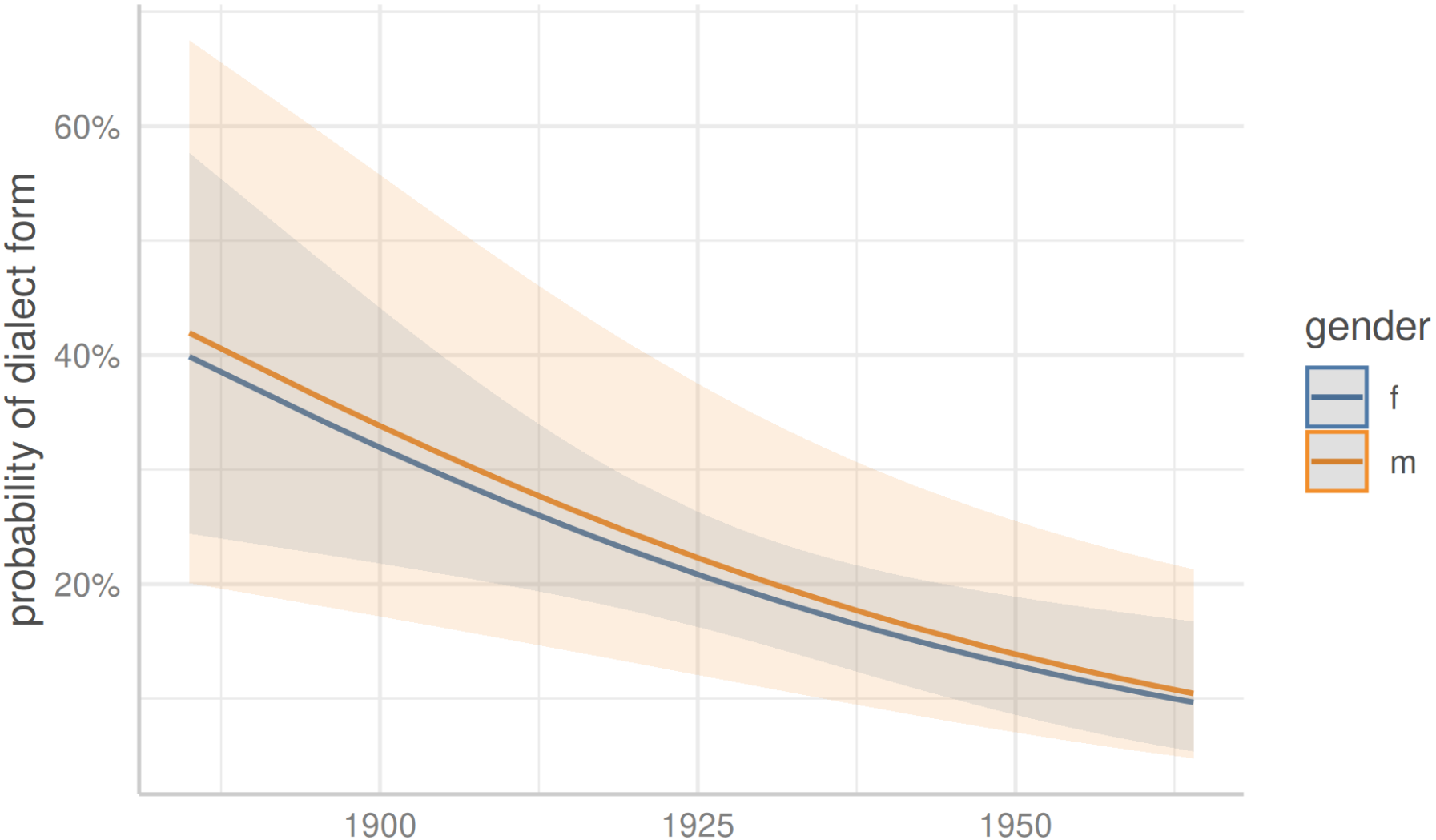
- corpus: Dvina
- *godov*: 25
- *let*: 8
- % *godov*: 75
- *-ov* in other than numeral contexts: 9
- corpus size: 68,010
- ipm *-ov*: 132

Example of 'other' contexts: *ne bylo cerkvov* (stand *cerkvej*) 'there were no churches'

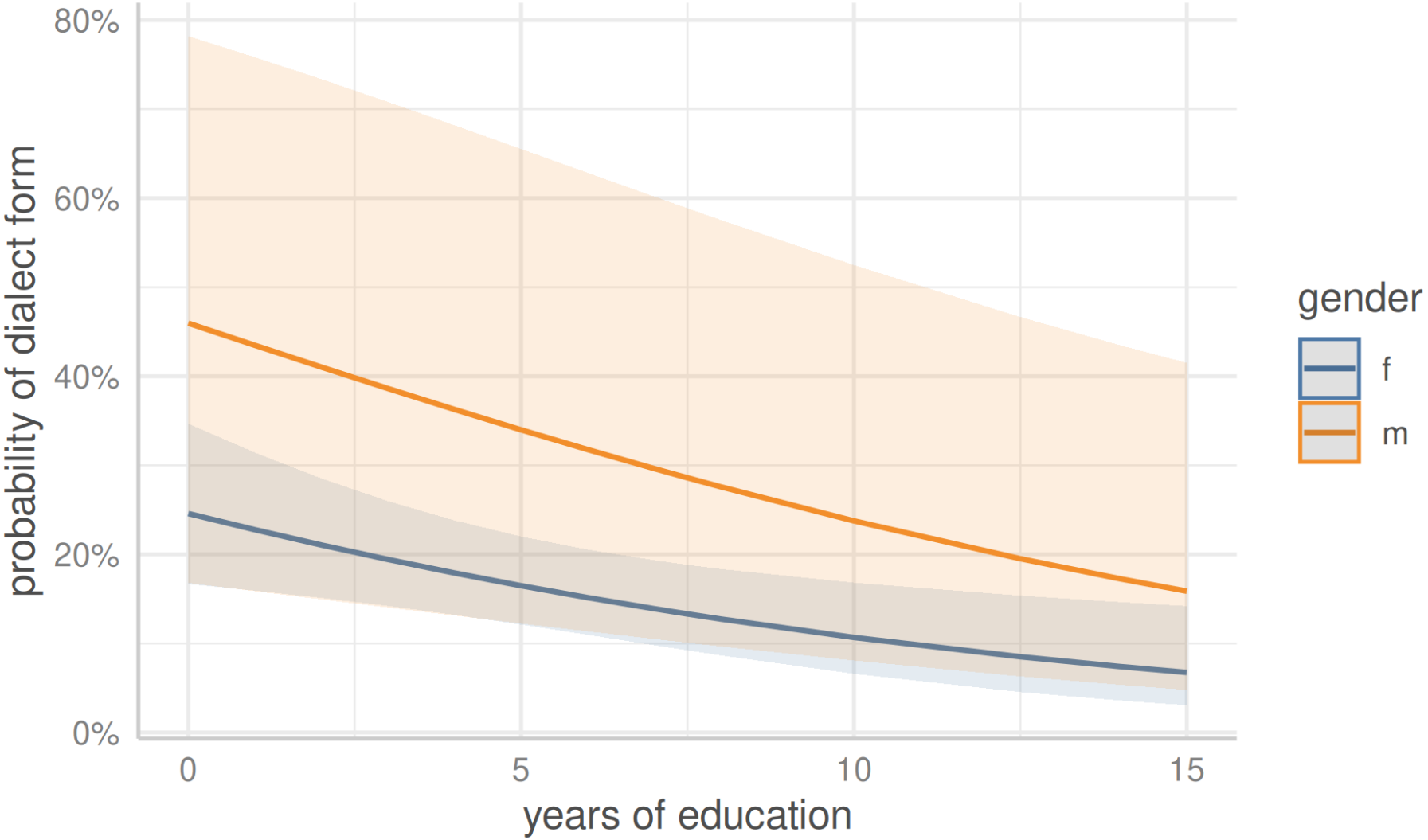
Pearson correlation coefficient: 0.699



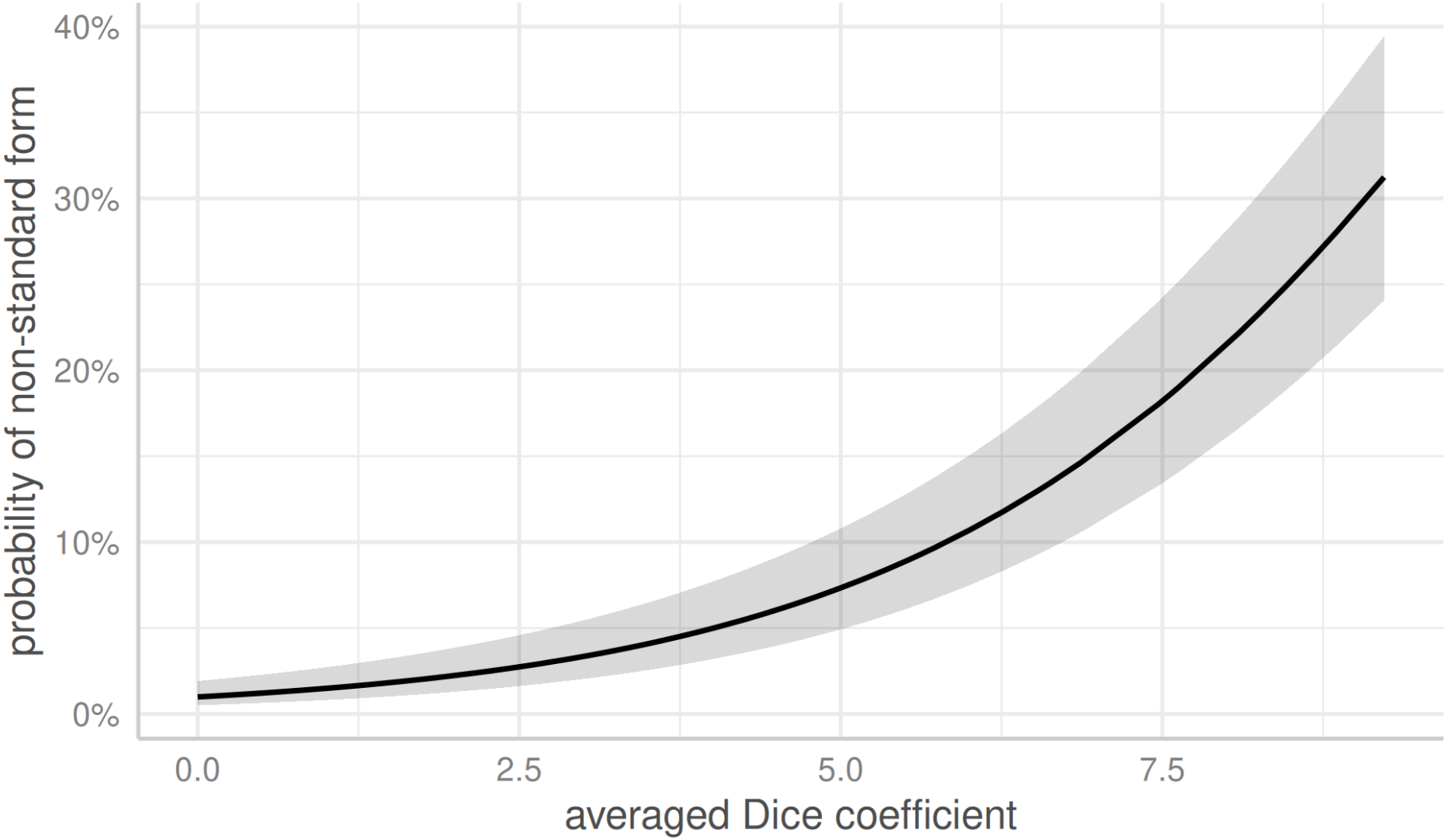
Is year of birth significant?



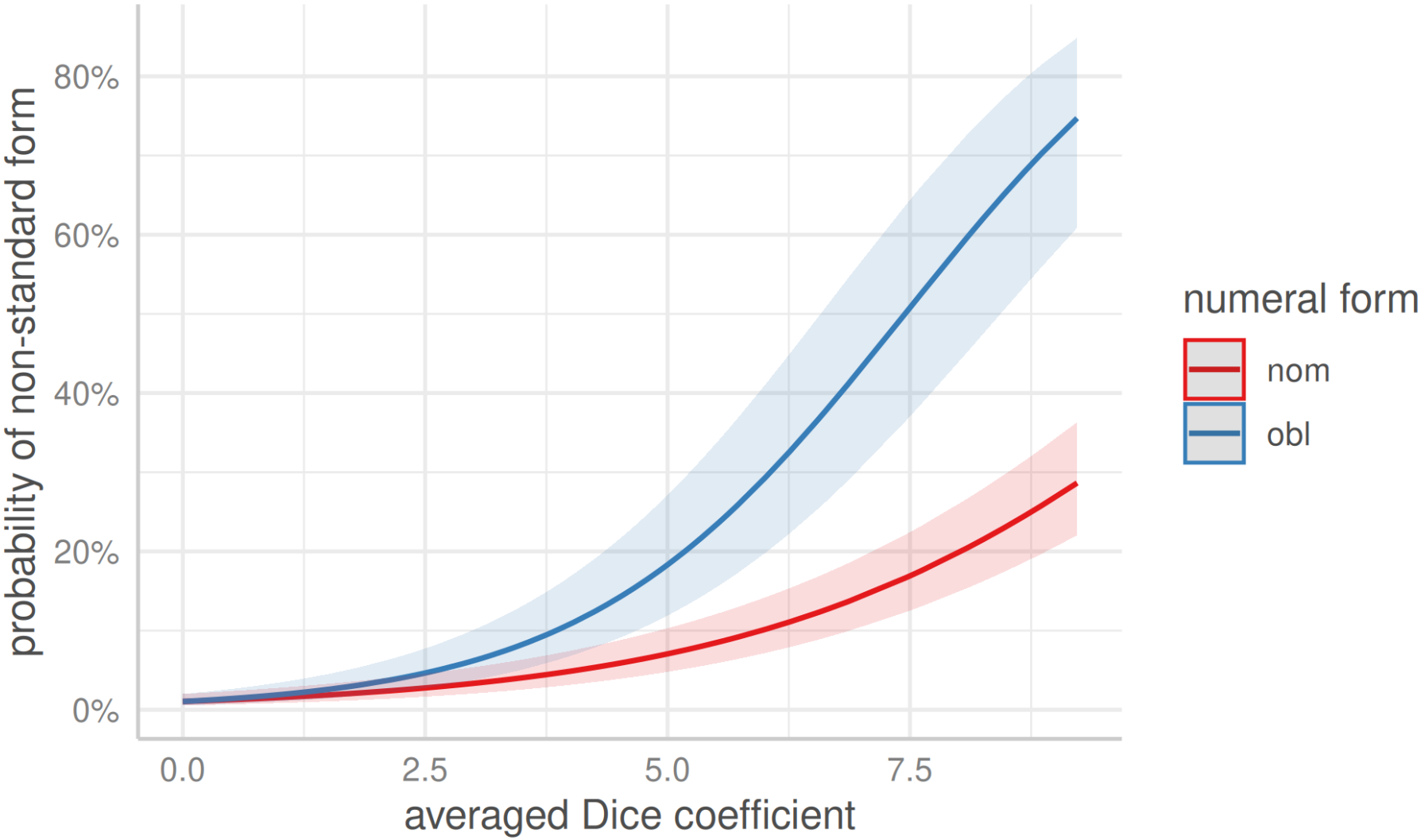
Is education level significant?



Is collocationality significant?



Is the numeral form significant?



Conclusions

- Dialect "overuse" of the *-ov* ending does not affect measure words: zero inflection forms (*kilogramm*) are found in more than 90% of contexts
- Year of birth and education level affect the probability of using dialect Gen.Pl forms (older people and people with a low education level use dialect forms more frequently)
- 'The most frequent pairs survive' (Chiara): the probability of using dialect forms is higher for nouns which often co-occur 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

Neg Exist constructions

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*

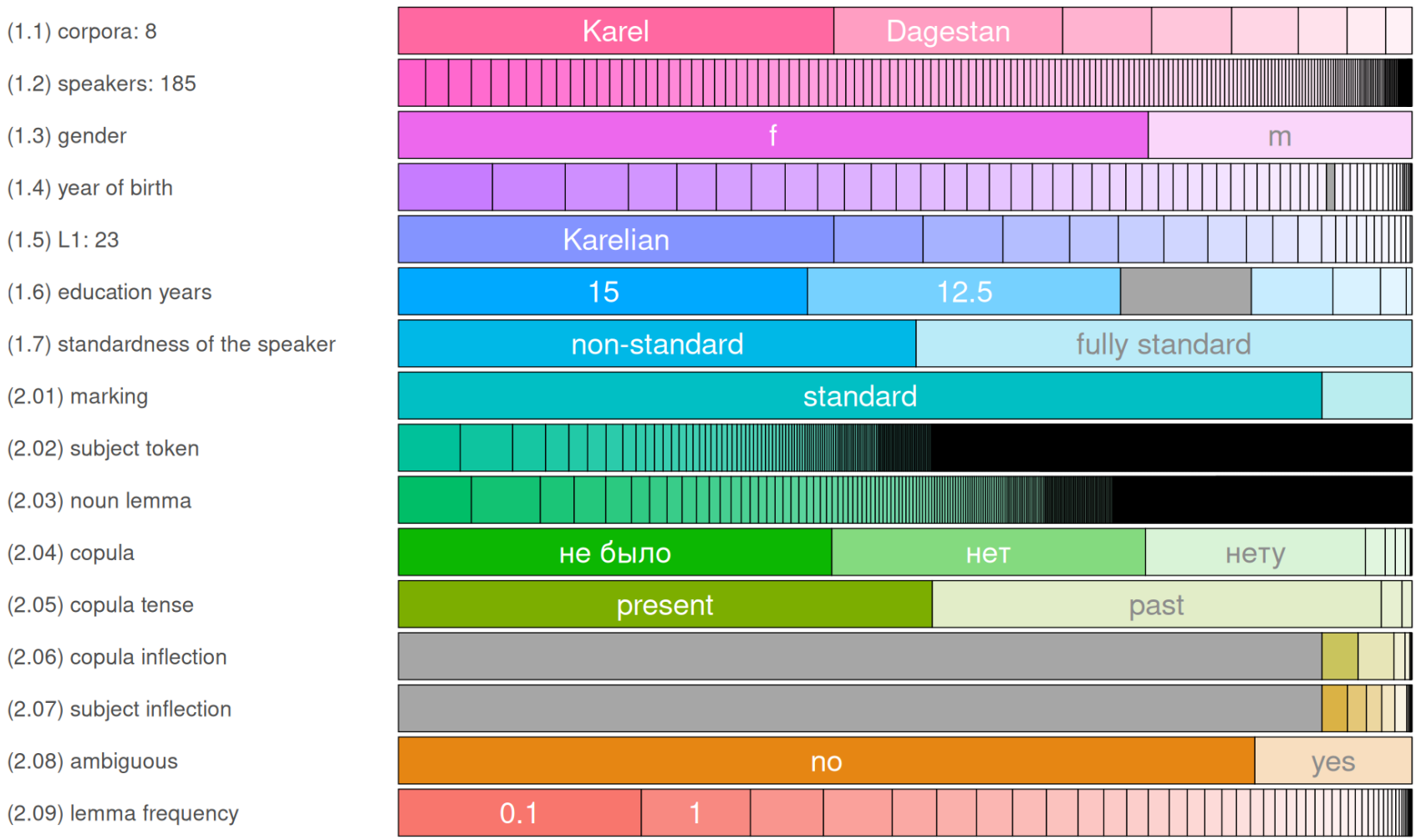
- Previous research on other L2 Russian varieties
  - Nanai and Ulcha Russian [Stoynova, 2019, 27]
  - Moksha Russian [Kashkin, 2020, 116]
  - Hill Mari Russian [Kashkin, 2022, 39]
- Usually treated as a contact phenomenon because in the L1s of Russian bilinguals who display this trait there is no genitive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

Research questions

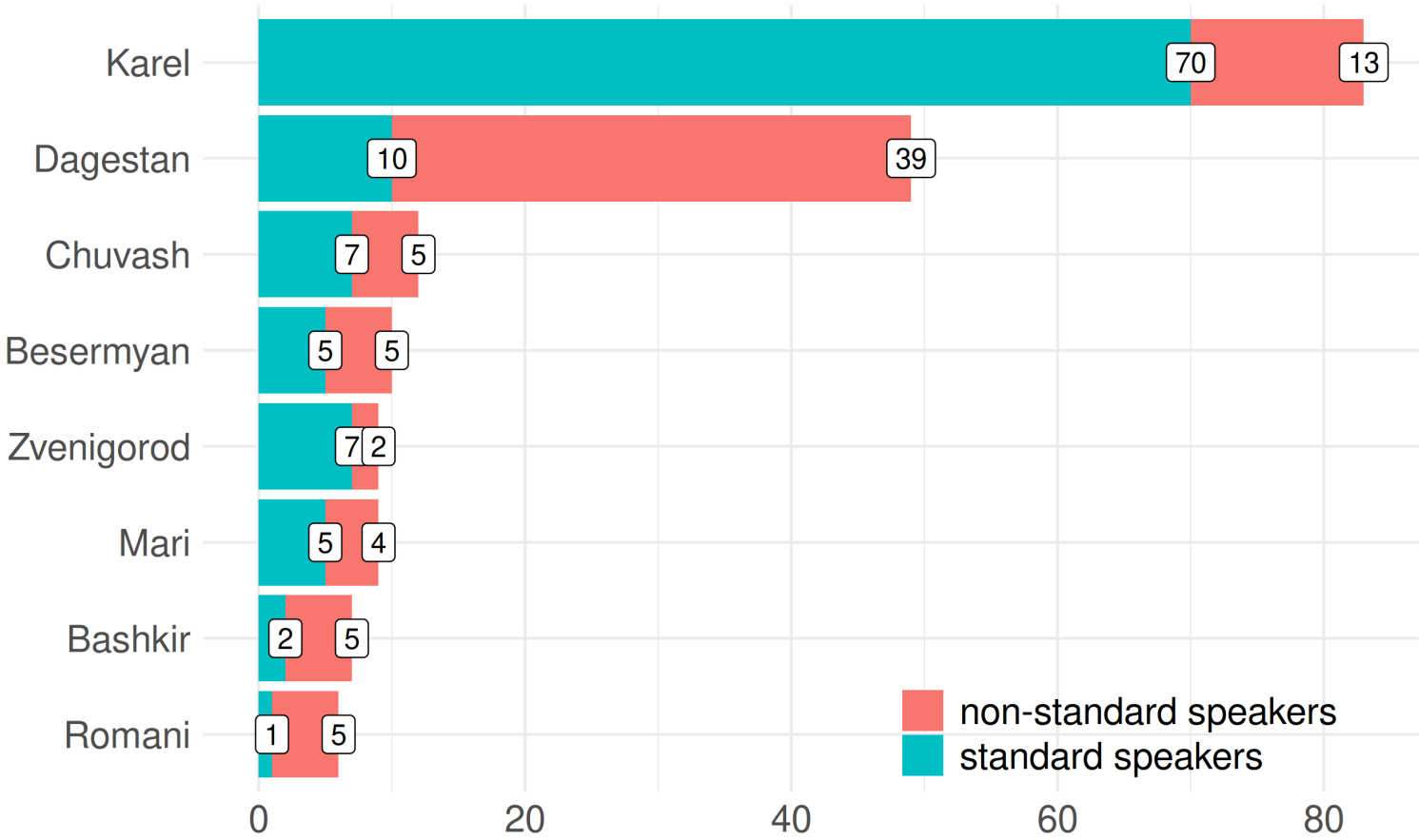
- Does the amount of variation in NECs differ across corpora and/or among speakers of the same variety?
- Can variation in NECs be explained in terms of contact influence?
- Do other factors promote or hinder variation in NECs?

The database and parameters of data annotation

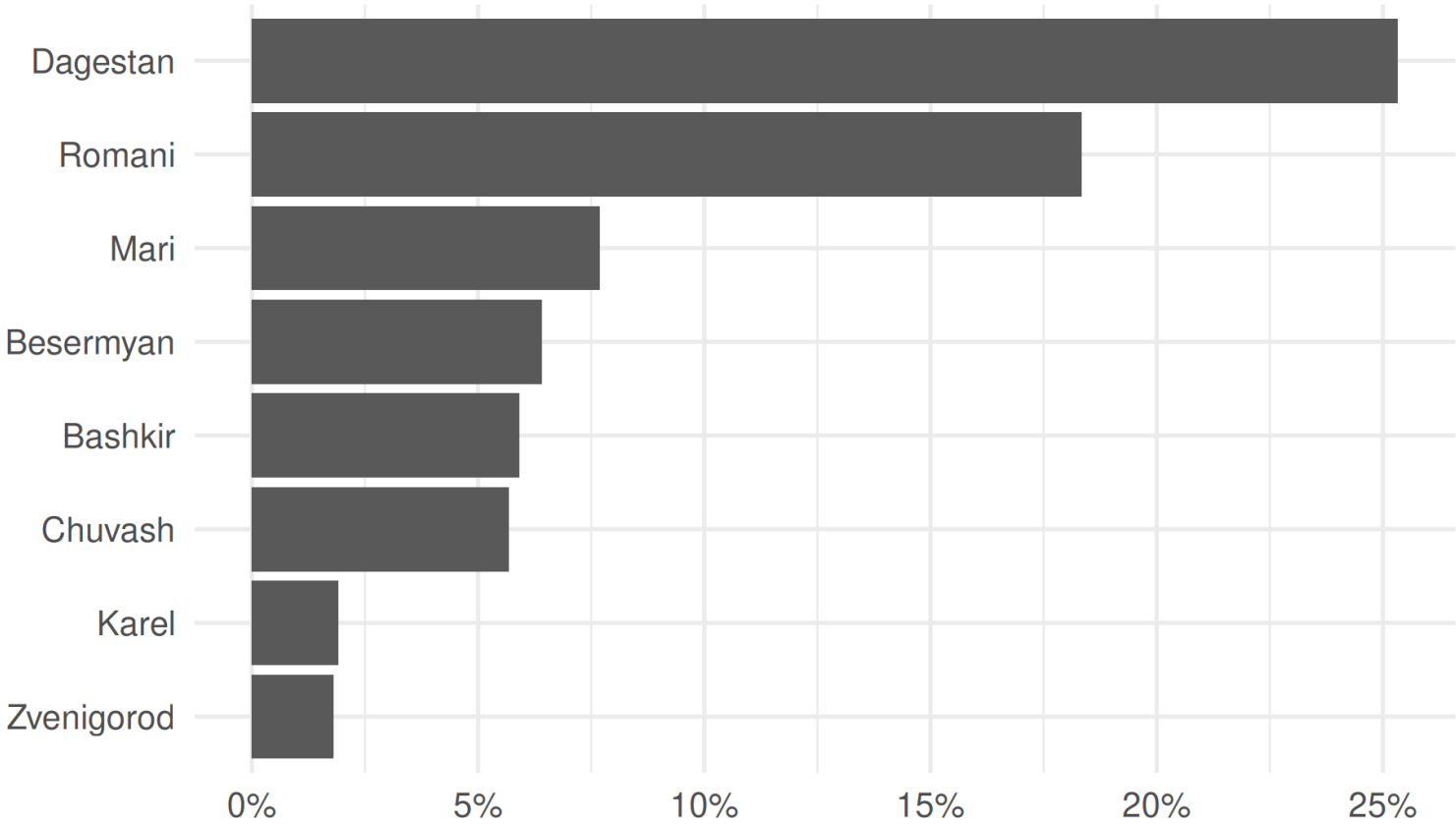
2,309 observations



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



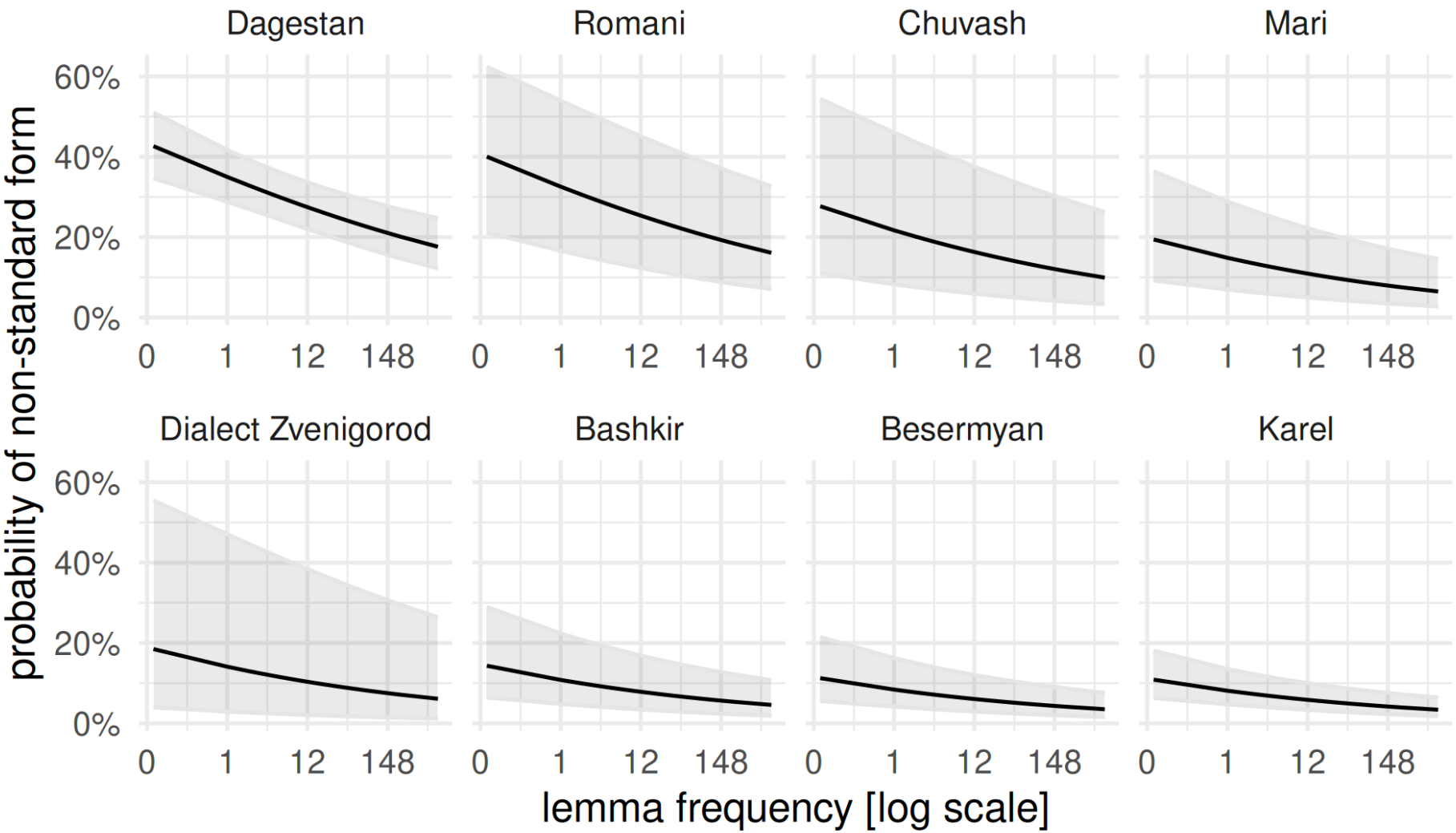
Proportion of non-standard occurrences per corpus



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

Aim of the project

- Typos and spelling errors in corpora complicate the search
- Looking for mistakes in the corpora manually is time-consuming
- Standard spell-checking algorithms are ineffective for cleaning data due to the presence of non-standard wordforms
- Spell-checkers for Standard Russian treat dialectal wordforms as spelling errors
  - *Они все разъехадчи* —> *Они все разъехались*
- Objective: create a mechanism for automatic search for errors and typos
- It is crucial for the future mechanism to distinguish between a typo and a dialectal wordform

Hapax Legomena

- First Approach: Analyze hapax legomena (tokens that appear only once in the corpora) to identify typos
- Pipeline:
  - Select all hapax legomena
  - Put them in the normal form
  - Check with a spellchecker and dictionaries
- Results: Low effectiveness, most part of the hapax legomena are not typos
  - Out of 2,000 hapax legomena, only 10 were typos

Utilizing LLMs and Machine Learning

- First Idea: Utilize prompts for Chat GPT-4o Mini and Gemini 1.5 Pro models
- Testing: Verification on test sentences
- The results were satisfactory, especially when combining models
- Further Steps:
  - Use free and local models
  - Develop the method further

Machine Learning and BERT

- Objective: Train the model to predict if a word is spelled correctly or if it is dialectal
- Pipeline:
  - Annotation:
    - D — dialectal wordform
    - E — typo
    - O — absence of dialectal wordforms or typos
  - Create a synthetic dataset of typos to balance the classes
  - Train a model to assign labels to words in a sentence
  - Example of a row in the dataset:

```
[('вот', 'O'), ('сейчас', 'O'), ('вот', 'O'), ('я', 'O'), ('тоже', 'O'),
('ездиала', 'E'), ('на', 'O'), ('автобусе', 'O'), ('ну', 'O'),
('но', 'O'), ('уже', 'O'), ('наверно', 'D'), ('с', 'O'), ('месяц', 'O'),
('не', 'O'), ('ездию', 'D'), ('ну', 'O')]
```

Machine Learning and BERT

- Results: promising but imperfect, the training dataset needs to be augmented
- Further Steps:
  - Train the BERT model on the annotated corpus
  - Evaluate the accuracy of the model on test data
  - Apply the model to clean data from typos

Future Plans

Future Plans

- create more corpora
  - we can do this within the Lab too (thanks to `whisper` speech to text model)

Future Plans

- create more corpora
  - we can do this within the Lab too (thanks to `whisper` speech to text model)
- annotate more features

Future Plans

- create more corpora
  - we can do this within the Lab too (thanks to `whisper` speech to text model)
- annotate more features
- make an Atlas of analyzed features

Future Plans

- create more corpora
  - we can do this within the Lab too (thanks to `whisper` speech to text model)
- annotate more features
- make an Atlas of analyzed features
- analyze how features interact within the speaker

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.

M. Daniel, R. von Waldenfels, A. Ter-Avanesova, P. Kazakova, I. Schurov, E. Gerasimenko, D. Ignatenko, E. Makhlina, M. Tsfasman, S. Verhees, A. Vinyar, V. Zhigulskaya, M. Ovsyannikova, S. Say, and N. Dobrushina. Dialect loss in the Russian North: Modeling change across variables. *Language Variation and Change*, 31(3):353–376, 2019.

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.

References II

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.

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.

References III

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

C. Naccarato, A. Panova, and N. Stoynova. Word-order variation in a contact setting: A corpus-based investigation of Russian spoken in Daghestan. *Language Variation and Change*, 33(3):387–411, 2021.

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.

References IV

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.

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.