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

Corpus of the Russian dialect spoken in Khislavichi district
260,793 tok.

Ustja River Basin Corpus 959,782 tok.

Corpus of the Russian dialect spoken in the villages of the Middle Pyoza
79,566 tok.

Corpus of the Russian dialect spoken in Nekhochi
88,965 tok.

Corpus of the Russian dialect spoken in the village Veegora
91,514 tok.

Corpus of Russian spoken in Zvenigorod
68,324 tok.

Luzhnikovo Corpus
68,666 tok.

Corpus of Opochetsky dialects
68,741 tok.

Upper Pinega and Vyya Corpus
70,803 tok.

Corpus of the Russian dialect spoken in the villages of the Middle Pinega
43,270 tok.

Corpus of the Russian dialect spoken in the Mikhaylov area
47,579 tok.

Corpus of the Russian dialect spoken in the villages of the Middle Northern Dvina
68,010 tok.

Corpus of Spiridonova Buda dialect
70,565 tok.

Svishni and Trostnoe Corpus
24,414 tok.

Corpus of the Russian dialect spoken in Popovka
36,617 tok.

Corpus of the Russian dialect spoken in Tserkovnoe
39,469 tok.

Corpus of the Russian dialect spoken in Keba
54,535 tok.

Corpus of the Russian dialect spoken in the villages of the Don river
69,098 tok.

Corpus of the Russian dialect spoken in Manturovo
113,837 tok.

Corpus of Rogovatka dialect
100,047 tok.

Corpus of Shetnevo and Makeevo dialect
95,335 tok.

Corpus of Lukh and Teza river basins dialects
146,350 tok.

Corpus of the Russian dialect spoken in the village Malinino
138,943 tok.

Corpus of the Russian dialect spoken in Ilmen Lake district
134,207 tok.

Bilingual Corpora

Corpus of Russian spoken in Daghestan 376,717 tok.	Khanty Russian Corpus 40,225 tok.
	Corpus of Russian spoken in Chuvashia 46,307 tok. / Corpus of Russian spoken by the Roma 41,767 tok.
Corpus of Karelian Russian 578,646 tok.	Corpus of Russian spoken in Mari El 69,109 tok.
	Corpus of Russian spoken in Bashkortostan 93,127 tok.
	Corpus of Russian spoken by the Besermans 97,216 tok.

Bilingual and Dialectal Corpora

Created with `lingtypology` [Moroz, 2017]
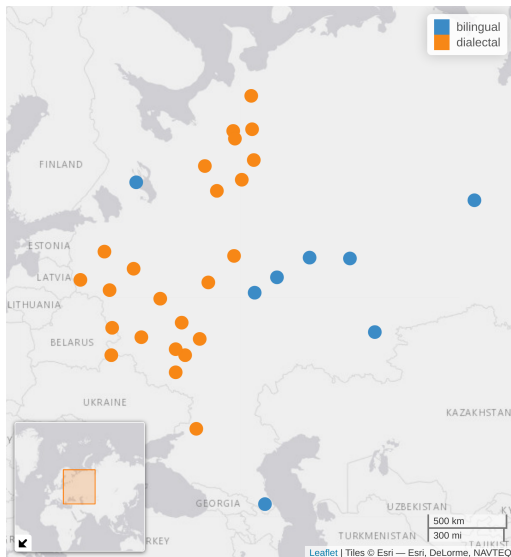
Can we analyze variation of linguistic features across all corpora?

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]
- Dialect of Ustja River Basin [Daniel et al., 2019]
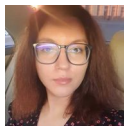
DiaL2

The DiaL2 team
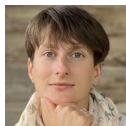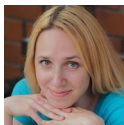

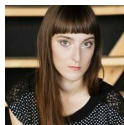Maria Ermolova   Anna Grishanova   Natalia Koshelyuk   George Moroz

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 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

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

(1.1) corpora: 7

(1.2) speakers: 181

(1.3) gender

(1.4) year of birth

(1.5) L1: 21

(1.6) L1 family

(1.7) education

(1.8) standardness of the speaker
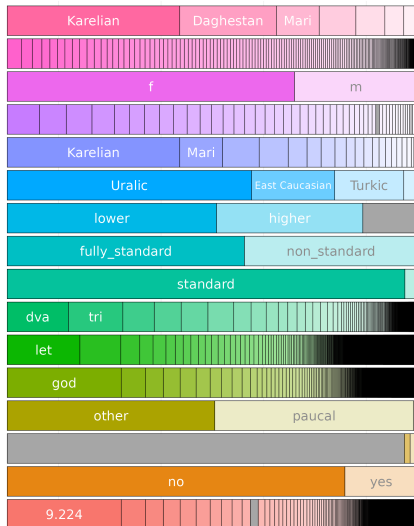
(2.1) marking

(2.2) numeral
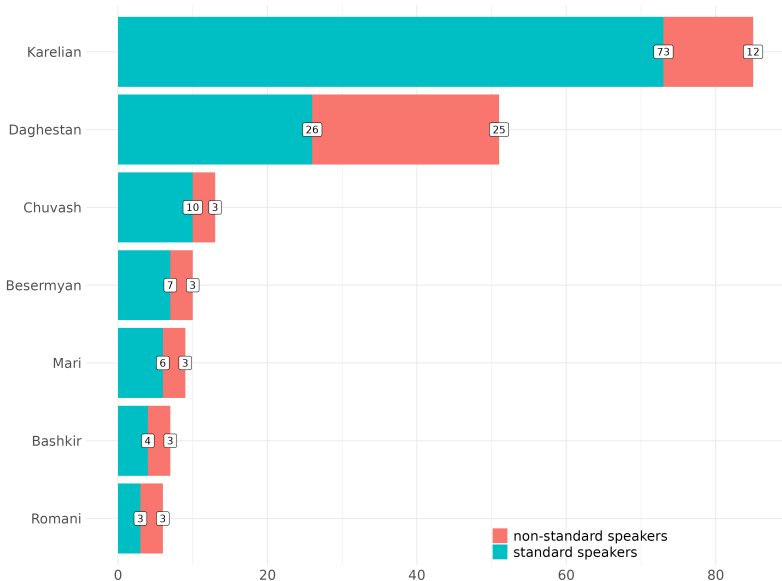
(2.3) noun token

(2.4) noun lemma

(2.5) numeral type

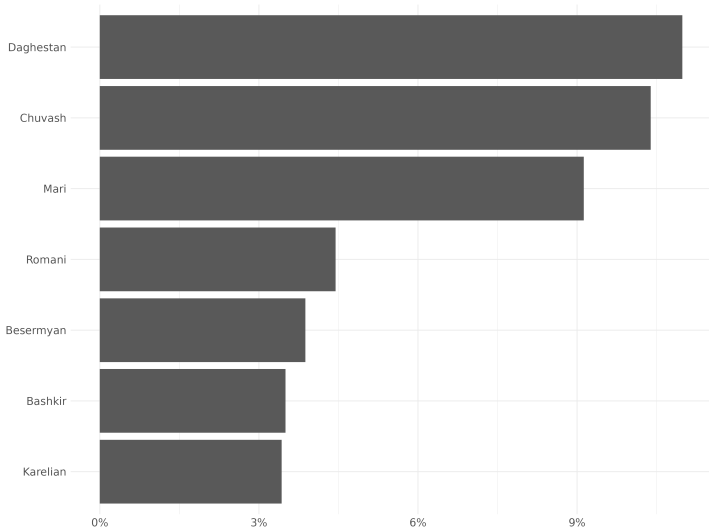(2.6) noun type

(2.7) ambiguous

(2.8) dice coefficient

Fully standard (71.3%) vs. non-standard speakers (28.7%)
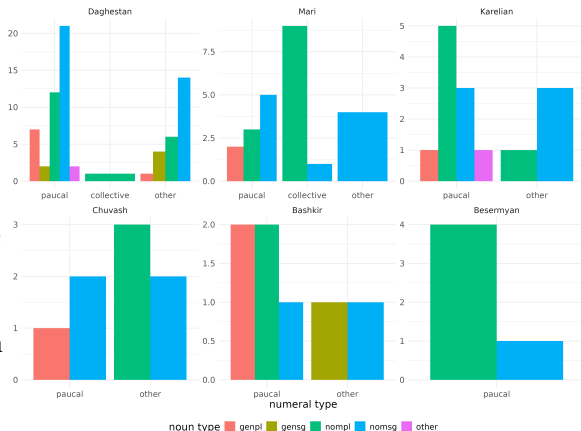
Proportion of non-standard occurrences per corpus
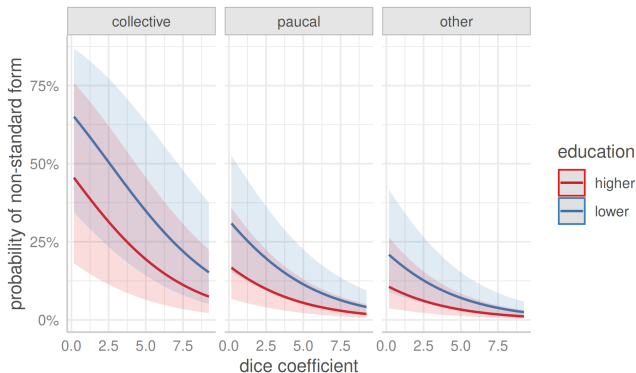
1,748 observations

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

Prepositional drop

Preposition drop in Russian spoken by Mari and Beserman bilinguals
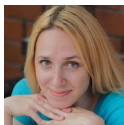
Anastasia Yakovleva
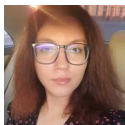
Natalia Koshelyuk

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
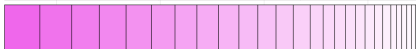
(1.3) gender
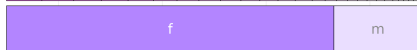
(1.4) year of birth

(1.5) years of education

(2.1) preposition lemma

(2.2) preposition drop
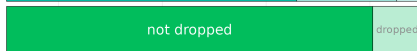
(2.3) following wordform's token

(2.4) following wordform's lemma

(2.5) following wordform's first sound

(2.6) dice coefficient

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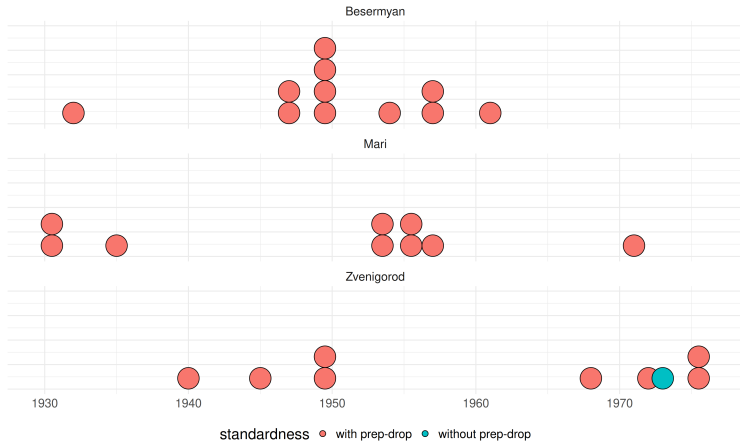
- both bilinguals and monolinguals omit prepositions;
- however, they do it in different contexts and for different reasons

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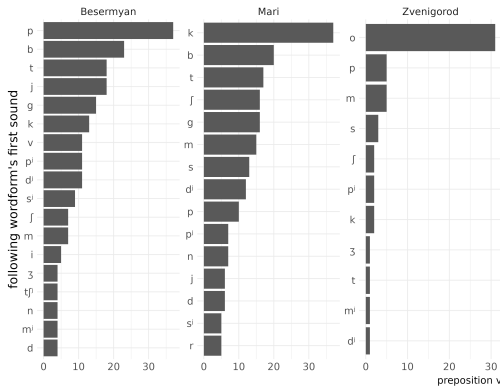
- In bilinguals, p-drop can possibly be explained by phonetic interference from their native languages (avoiding consonant cluster);
- monolinguals mainly omit prepositions in lexicalized expressions (v obschem 'in general').

Findings: sociolinguistic factors
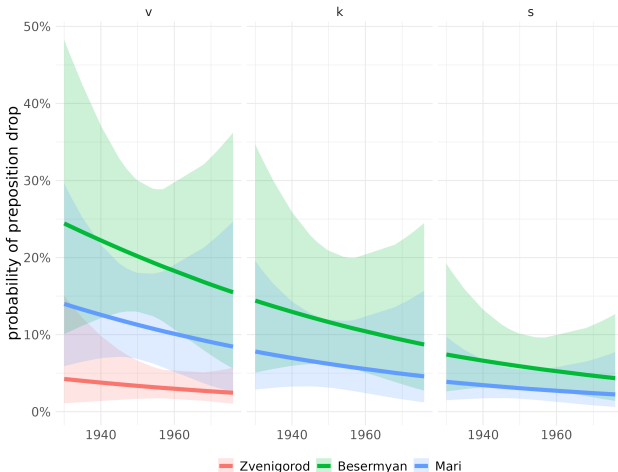
The older and less educated a person is, the higher the probability of prep-drop in their speech.

Findings: sociolinguistic factors
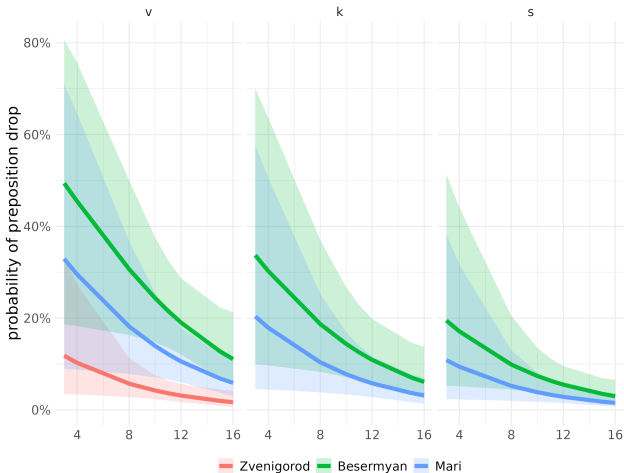
The older and less educated a person is, the higher the probability of prep-drop in their speech.

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] kazhdom povorote stolba stoit 'A pillar stands on every turn'

Preposition drop in Chuvash Russian

- The omission of syllabic prepositions and 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
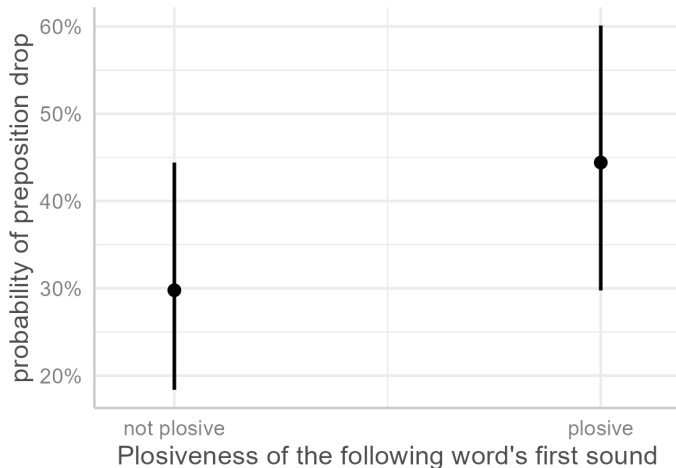
The plosiveness of the following word's first sound increases the probability of a preposition drop

Preposition drop in Chuvash Russian
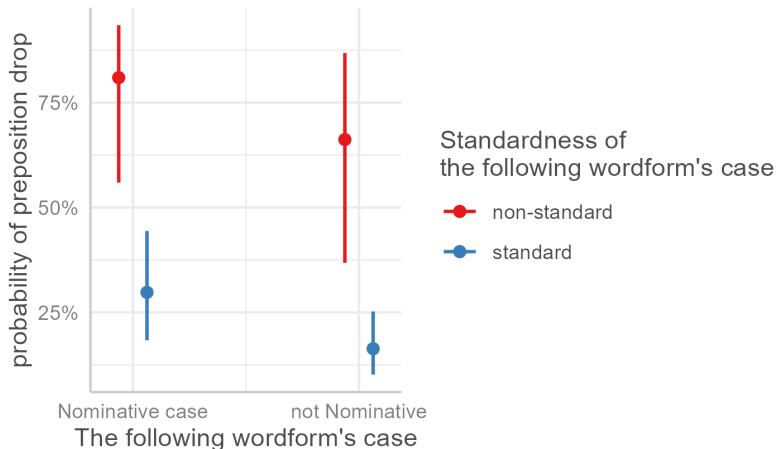
The Nominative or a non-standard case of the following wordform
increase the probability of a preposition drop

Preposition drop in Chuvash Russian
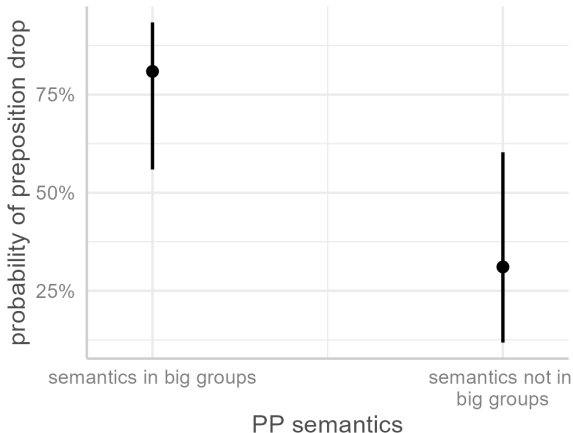
On the other hand, the semantics of the prepositional phrase not being that of location, direction, source and time decrease the probability of a preposition drop.

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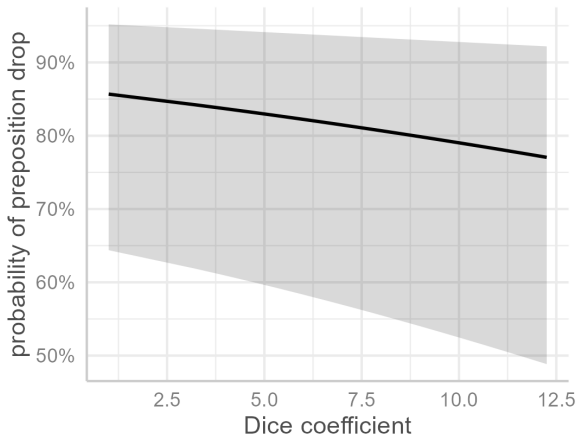
Finally, just like in the study of preposition drop in the speech of Mari and Beserman bilinguals, higher degree of collocationality (dice coefficient) decreases the probability of the omission of a preposition.

Gen Pl Forms

Dialect Genitive Plural Forms in Numeral Constructions
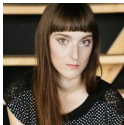
Svetlana Zemicheva

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 be controversial to the tendencies observed in
  other colloquial varieties

Examples

- Num(not-paucal)$_{nom,acc,gen}$ + N$_{m,n\text{-Gen.Pl}}$
    - pjat' xozjaev-ov (dial. five owner-Gen.Pl)
    - pjat' xozja-ev-Ø (stnd. five owner-Pl-Gen)
- Num(paucal)$_{acc,gen}$ + N$_{m,n\text{-Gen.Pl}}$
    - 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

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

(1.1) corpora: 8	Karel	Dagestan					
(1.2) speakers: 185							
(1.3) gender	f				m		
(1.4) year of birth							
(1.5) L1: 23	Karelian						
(1.6) education years	15		12.5				
(1.7) standardness of the speaker	non-standard		fully standard				
(2.01) marking	standard						
(2.02) subject token							
(2.03) noun lemma							
(2.04) copula	не было		нет		нету		
(2.05) copula tense	present		past				
(2.06) copula inflection							
(2.07) subject inflection							
(2.08) ambiguous	no		yes				
(2.09) lemma frequency	0.1	1					

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

Proportion of non-standard occurrences per corpus
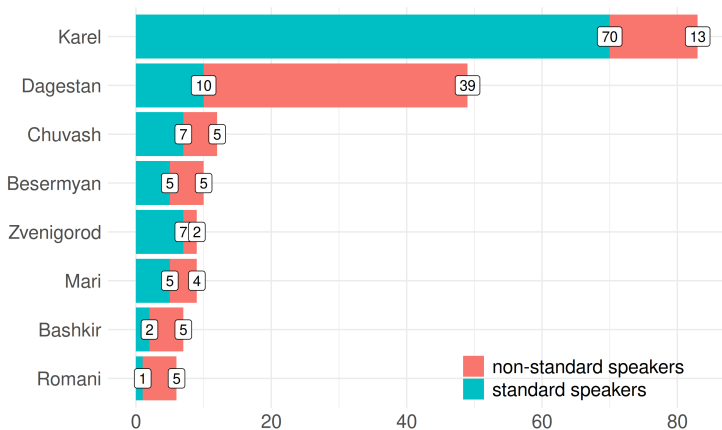
Types of non-standard marking

- 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
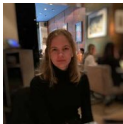
Anna Panova

Olga Gich

George Moroz

Plans

Plans

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

Plans

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

Plans

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

Plans

- create more corpora
  - we can do 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.

Michael Daniel, Ruprecht von Waldenfels, Aleksandra Ter-Avanesova, Polina Kazakova, Ilya Schurov, Ekaterina Gerasimenko, Daria Ignatenko, Ekaterina Makhlina, Maria Tsfasman, Samira Verhees, Aleksei Vinyar, Vasilisa Zhigulskaya, Maria Ovsyannikova, Sergey Say, and Nina 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.

References II

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.

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.

References III

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.

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.

References IV

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.

N. Stoynova. Nestandartnye količestvennye konstrukcii v russkoj reči nositelej nanajskogo i uľčskogo 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.