

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 of the project

Precursors of the project



Nina Dobrushina



Michael Daniel

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

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 Ustyia
- 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

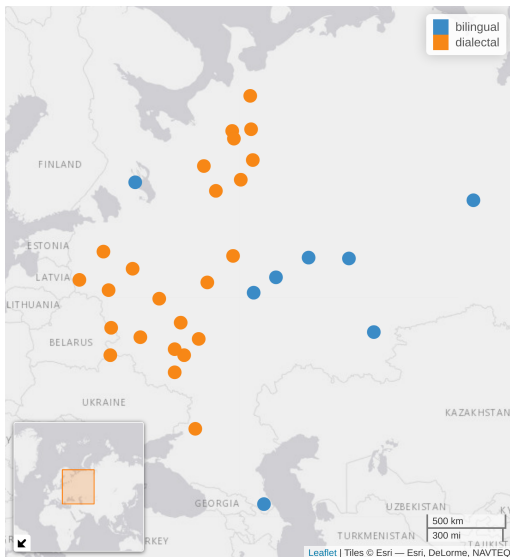
Dialectal Corpora

<p>Corpus of the Russian dialect spoken in Khislavichi district 260,793 tok.</p> <p>Ustja River Basin Corpus 959,782 tok.</p>	Corpus of the Russian dialect spoken in the villages of the Middle Pyoza 79,566 tok.	Corpus of Russian spoken in Zvenigorod 68,324 tok.	Corpus of the Russian dialect spoken in the villages of the Middle Pinega 43,270 tok.	Sivshini and Irostoie Corpus 24,414 tok.
	Corpus of the Russian dialect spoken in Nekhoichi 88,965 tok.	Luzhnikovo Corpus 68,666 tok.	Corpus of the Russian dialect spoken in the Mikhaylov area 47,579 tok.	Corpus of the Russian dialect spoken in Popovka 36,617 tok.
			Corpus of the Russian dialect spoken in the villages of the Middle Northern Dvina 68,010 tok.	Corpus of the Russian dialect spoken in Tserkovnoe 39,469 tok.
	Corpus of the Russian dialect spoken in the village Veegora 91,514 tok.	Corpus of Opochetsky dialects 68,741 tok.	Corpus of the Russian dialect spoken in the villages of the Middle Northern Dvina 68,010 tok.	
	Corpus of the Russian dialect spoken in Manturovo 113,837 tok.	Upper Pinega and Vyva Corpus 70,803 tok.	Corpus of Spiridonova Buda dialect 70,565 tok.	Corpus of the Russian dialect spoken in the villages of the Don river 69,098 tok.
		Corpus of the Russian dialect spoken in the village Malinino 138,943 tok.	Corpus of Rogovodka dialect 100,047 tok.	Corpus of Shetnevo and Makeevo dialect 95,335 tok.
		Corpus of the Russian dialect spoken in the village Ilmen Lake district 134,207 tok.	Corpus of the Russian dialect spoken in the village Ilmen Lake district 134,207 tok.	Corpus of the Russian dialect spoken in the village 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 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.	
Corpus of Karelian Russian 578,646 tok.		

Bilingual and Dialectal Corpora



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?

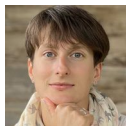
Can we find different variation patterns?

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](#)]
- Russian of Nganasan speakers [[Khomchenkova, 2020](#)]
- Russian of Nanai and Ulcha speakers [[Stoynova, 2021](#)]

The Dial2 project

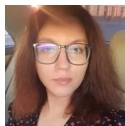
The DialL2 team



Maria Ermolova



Anna Grishanova



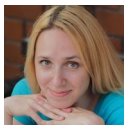
Natalia Koshelyuk



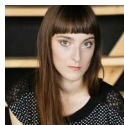
George Moroz



Chiara Naccarato



Anastasia Yakovleva



Svetlana Zemicheva

The Dial2 pipeline

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
 - Stoyanova (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

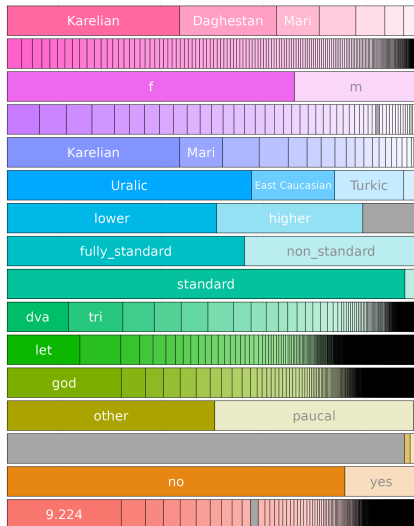
(2.4) noun lemma

(2.5) numeral type

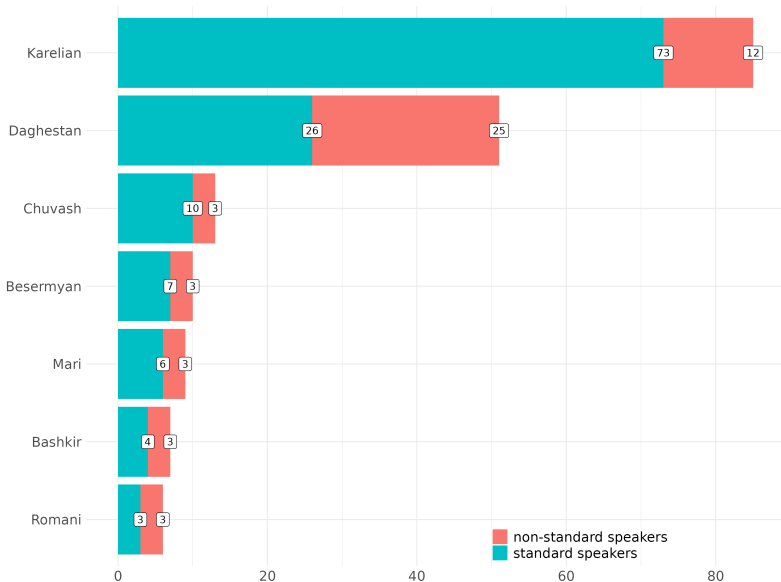
(2.6) noun type

(2.7) ambiguous

(2.8) dice coefficient

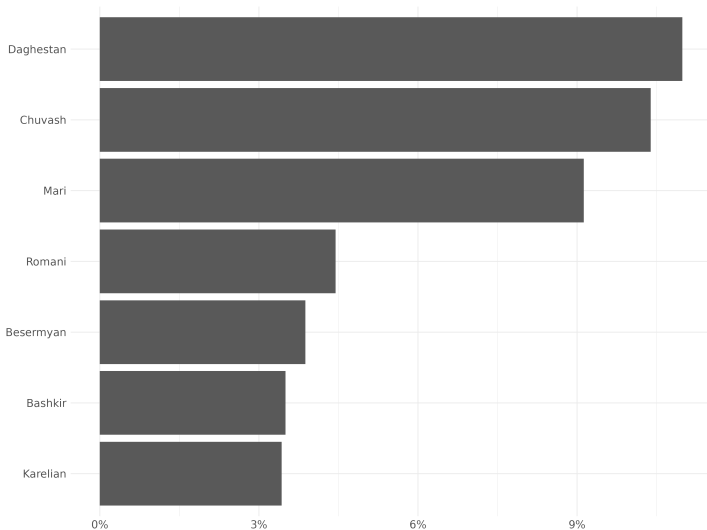


Fully standard (71.3%) vs. non-standard speakers (28.7%)



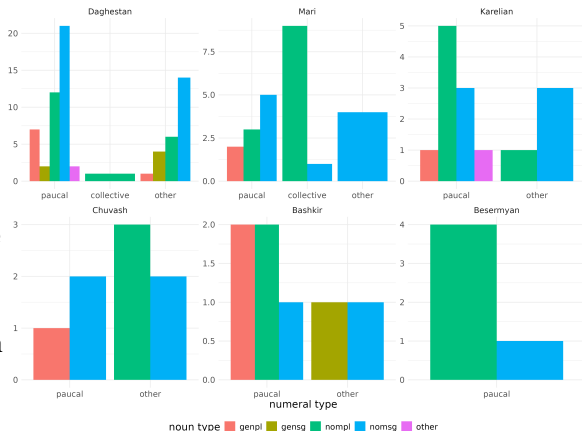
Proportion of non-standard occurrences per corpus

1,748 observations

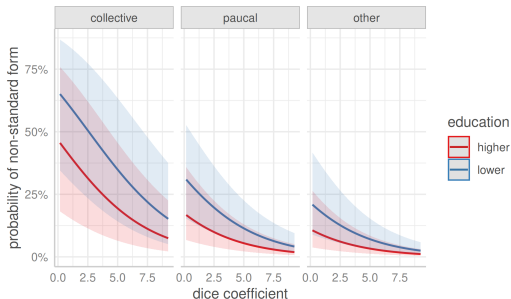


Distribution of n-std forms with different types of numerals

- 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

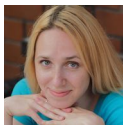


- **Logistic regression:** $\text{standardness} \sim \text{Dice coefficient} + \text{year of birth} + \text{education} + \text{numeral type} + \text{gender} + (1|\text{L1 family/speaker id})$
- **Conditional importance of the variables in our model (generalized R squared):** collocationality (Dice coefficient)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

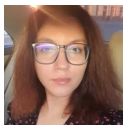
Conclusions

- 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



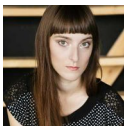
George Moroz

Propositional Drop in Chuvash



Anna Grishanova

Dialect Genitive Plural Forms in Numeral Constructions



Svetlana Zemicheva



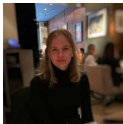
Chiara Naccarato



George Moroz

The Dial2 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.
- 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.
- 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.

References II

- 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. Stoyanova. 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.