

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 Ustyia
- 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

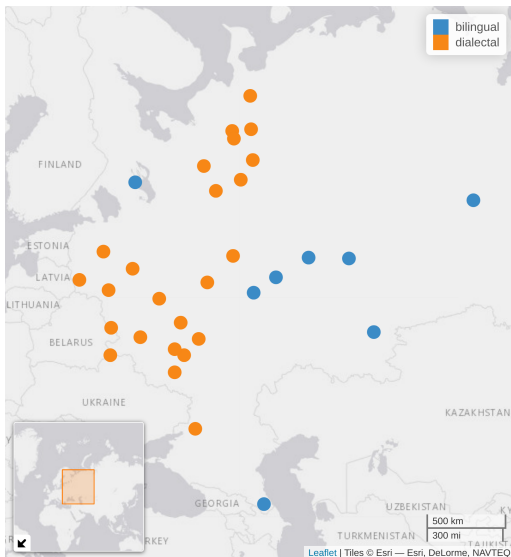
Dialectal Corpora

<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.	
		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 Manturovo 113,837 tok.	Corpus of Rogovodka 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 Russian spoken in Mari El 69,109 tok.	
	Corpus of Russian spoken in Bashkortostan 93,127 tok.	
Corpus of Karelian Russian 578,646 tok.	Corpus of Russian spoken by the Besermans 97,216 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?

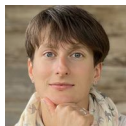
Can we analyze variation of linguistic features
across all corpora?

What are the factors that influence variation?

Can we find different variation patterns?

The Dial2 project

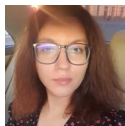
The DiaL2 team



Maria Ermolova



Anna Grishanova



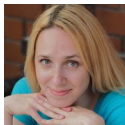
Natalia Koshelyuk



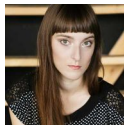
George Moroz



Chiara Naccarato



Anastasia Yakovleva



Svetlana Zemicheva

The Dial2 pipeline

The Dial2 results

The Dial2 sideproject

Future plans

References I

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