Genealogy vs. contact configuration: argument encoding across Romani dialects in Europe

Kirill Kozhanov, Sergey Say
Universität Potsdam

56th Societas Linguistica Europaea
Athens, 2023 August 29 – September 1

Lithuanian Romani

(1) *uźakir-á* ***tút*** *paše khangir-í*
wait-FUT.1SG 2SG.ACC near chirch-NOM.SG
'I will wait for you by the church'

Polish Romani

(2) *źakir-áva* ***pe túte*** *paś khanger-ý*
wait-FUT.1SG on 2SG.LOC near chirch-NOM.SG
'I will wait for you by the church'

Outline

Introduction

Romani is an Indo-Aryan language that has been spoken in Europe since the Middle Ages.

- As a result of multiple migrations in the 14-15$^{\text{th}}$ centuries, Romani spread across Europe (and beyond).
- Dialectal differences have evolved during the 16–17$^{\text{th}}$ centuries (Matras 2005). Many Romani groups migrated to new territories after the dialectal diffusion took place.
- As of now, Romani is spoken on a vast territory and is influenced by genealogically diverse languages.

- Romani is a perfect candidate to study the effects of language contact.
- Argument encoding is known to be susceptible to contact influence (Grossman et al. 2019; Gaszewski 2020; McAnallen 2021).

Research questions

- How can we assess cross-linguistic (dis)similarities in valency patterns?
- How much variation in valency patterns is observed in contemporary Romani dialects?
- How can genealogical and areal factors explain the observed variation?

Data

2001–2016. *Romani Morpho-Syntax Database*: Dialectological database of Romani (Available online at https://romani.dch.phil-fak.uni-koeln.de/).

- A questionnaire compiled by Yaron Matras and Viktor Elšík includes ca. 300 lexical questions and 700 sentences aimed to elicit morphosyntactic information.
- 119 locations in Europe.

Say, Sergey (ed.). 2020–... *BivalTyp*: Typological database of bivalent verbs and their encoding frames. (Available online at https://www.bivaltyp.info)

- A questionnaire with 130 predicates given in context:

X		Y
Peter	is afraid	of the dog
Peter	likes	this shirt
Peter	needs	money

- 103 languages.
- Valency pattern = combination of argument-encoding devices associated with X and Y in the target idiom: NOM_ACC, DAT_NOM, NOM_pa, etc.

Dataset

- 43 bivalent predicates which appear both in the RMS and the BivalTyp questionnaires.
- 119 Romani idioms.
- 18 contact languages (Indo-European and Uralic).
- Each Romani idiom has at least one contact language, and all of them are represented in BivalTyp.

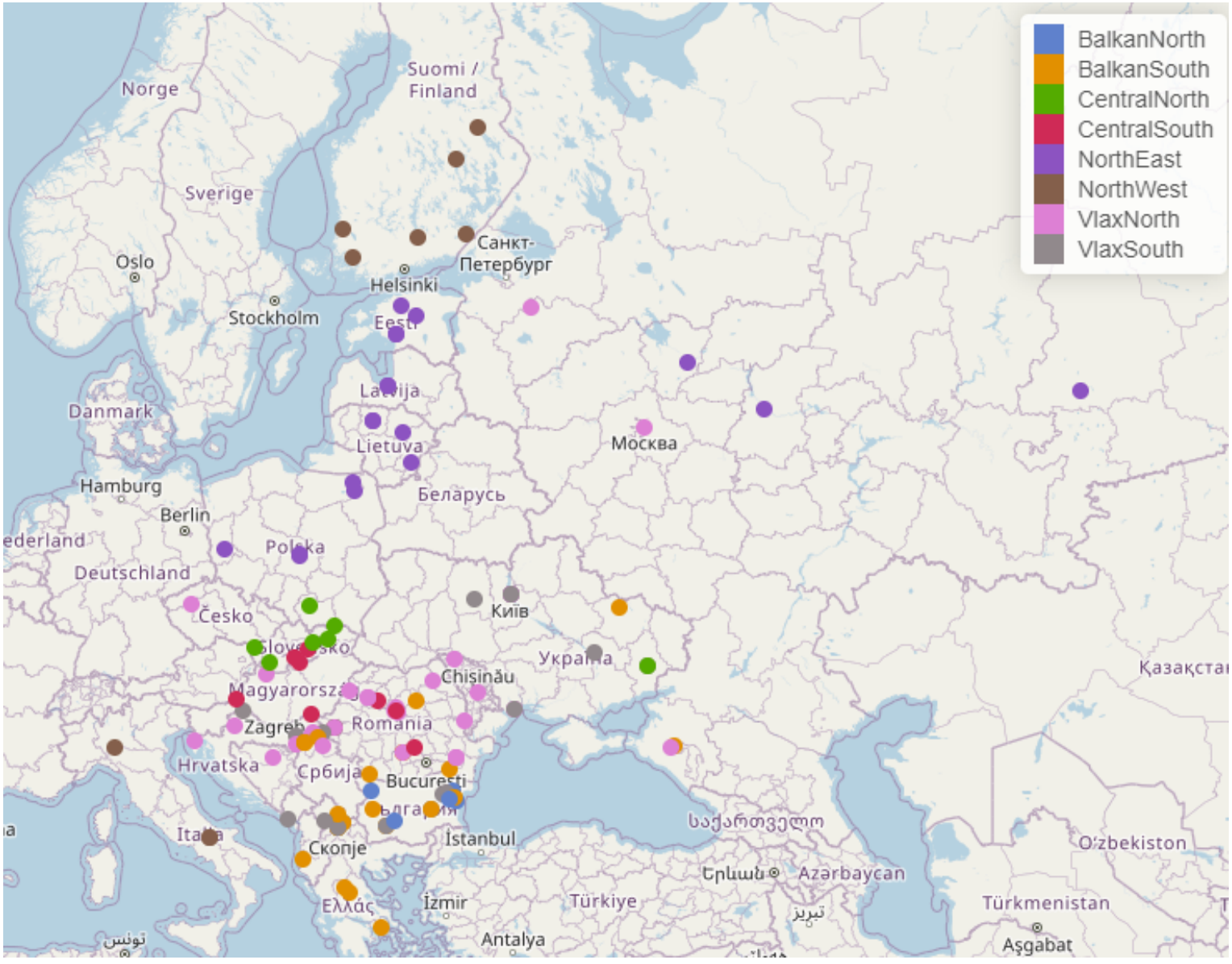
Data: predicates

‘feel_pain’	‘tell’	‘paint’	‘sing’	‘kiss’
‘be_afraid’	‘reach’	‘bite’	‘drink’	‘be_angry’
‘throw’	‘eat’	‘forfeit’	‘remember’	‘want’
‘have_enough’	‘wait’	‘break’	‘help’	
‘believe’	‘call’	‘wash’	‘understand’	
‘take’	‘know’	‘find’	‘speak’	
‘see’	‘play’	‘hate’	‘hear’	
‘encounter’	‘make’	‘like’	‘lose’	
‘drive’	‘have’	‘need’	‘kill’	
‘bend’	‘look_for’	‘open’	‘hit’	

Russian	22	Croatian	3
Romanian	19	Greek	3
Bulgarian	15	Italian	3
Serbian	13	Slovak	3
Hungarian	7	Latvian	2
Finnish	6	Albanian	1
Macedonian	6	Slovenian	1
Polish	6	Spanish	1
Estonian	4		
Czech	4		

Table 1: Number of Romani dialects by primary contact language

Romani dialect groups in the dataset



Dataset

- A dataset of 5965 entries:

	predicates	idioms	NAs
Romani dialects	43	119	189
Contact languages	43	18	16

- Each entry is coded for several variables:

dialect	predicate	predicate	argument_frame	verb	X	Y	locus	valency_patte	verb_origin
UKR010	1	feel_pain	X (have) Yache	dukhal	LOC	NOM	X	LOC_NOM	inherited
UKR010	3	be_afraid	X (be afraid) of Y	daral	NOM	ABL	Y	NOM_ABL	inherited
UKR010	4	throw	X (throw) Y	čhuvel	NOM	ACC	TR	TR	inherited

Results

Table of Contents

Variation in argument encoding

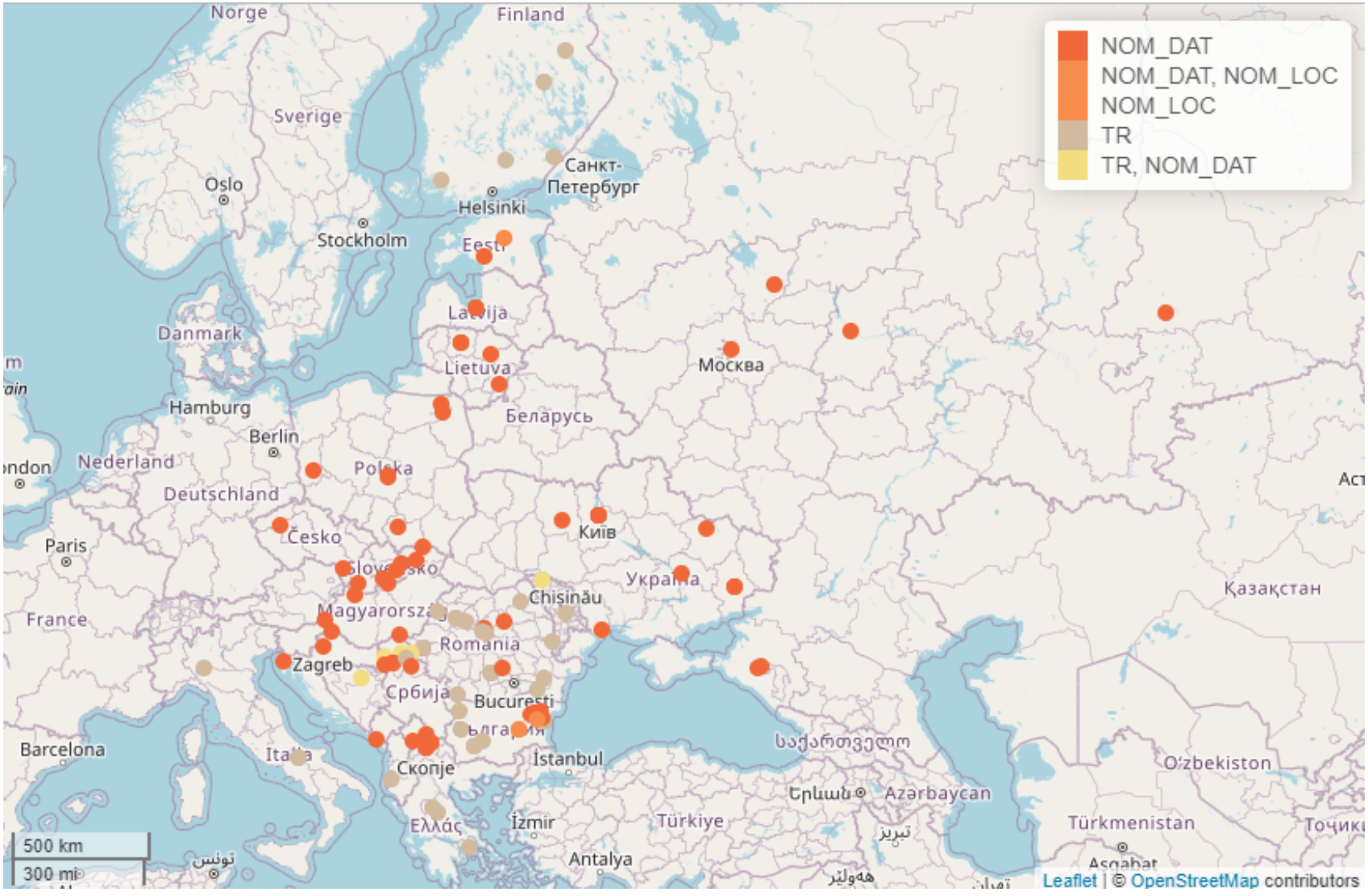
- In Romani data, 12 predicates display considerable variation in valency patterns (at least 27% of entries deviate from the most common valency pattern).

‘be_afraid’	‘have_enough’	‘reach’
‘be_angry’	‘help’	‘wait’
‘believe’	‘like’	
‘feel_pain’	‘need’	
‘have’	‘play_instrument’	

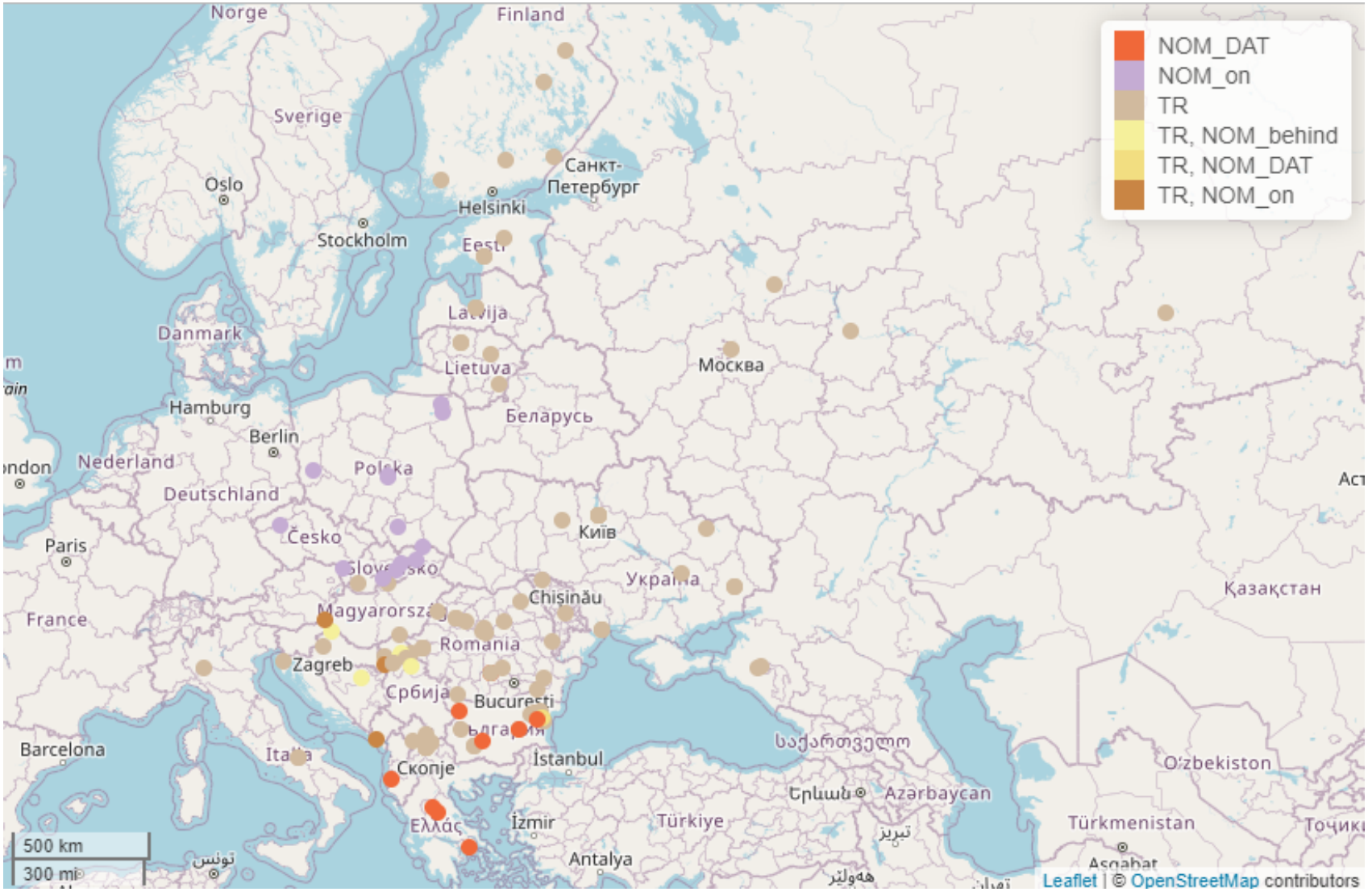
Variation in argument encoding

- Valency patterns in Romani dialects display areal distribution.

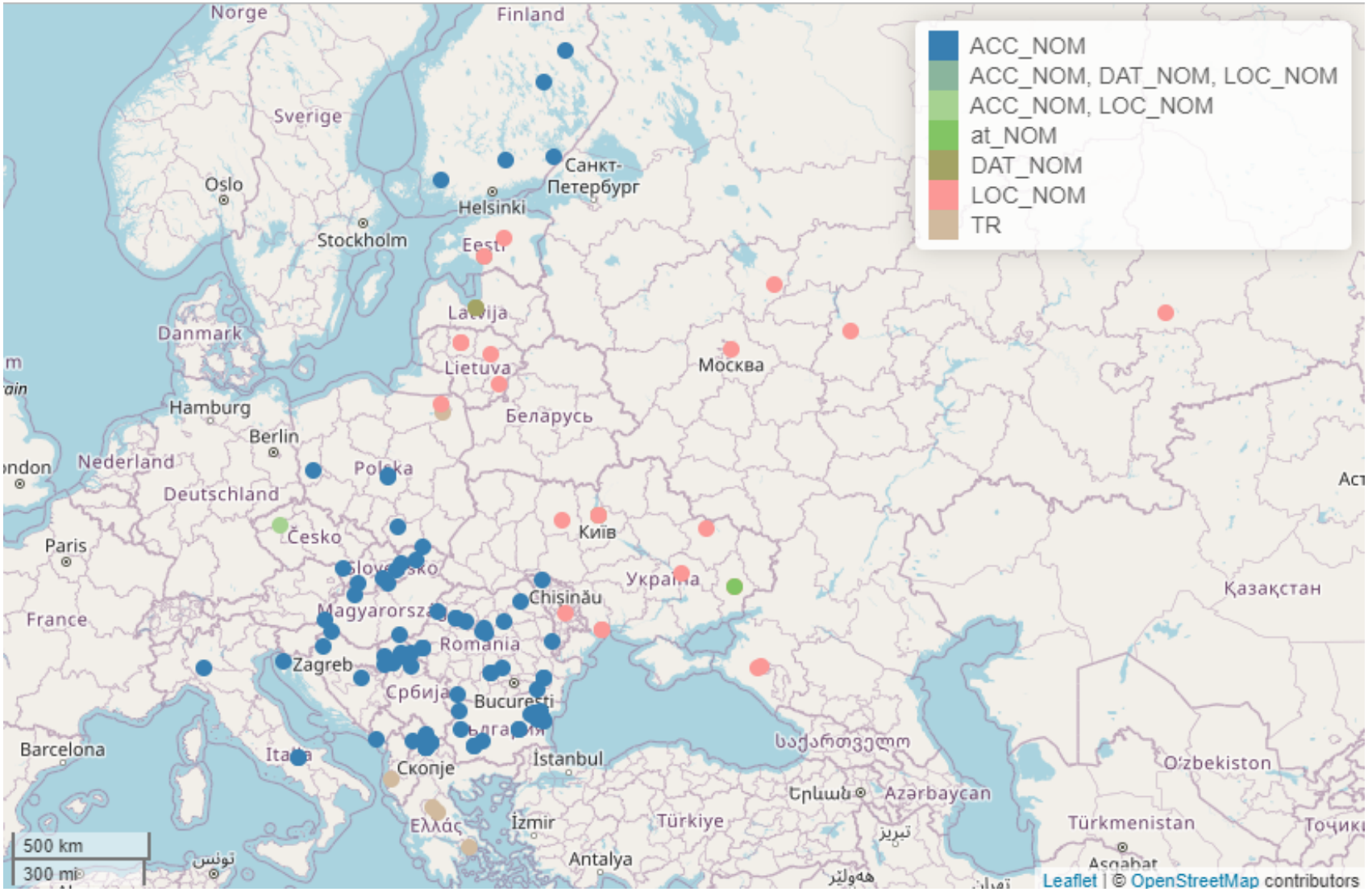
Argument encoding of the predicate ‘believe’



Argument encoding of the predicate ‘wait’



Argument encoding of the predicate ‘have’

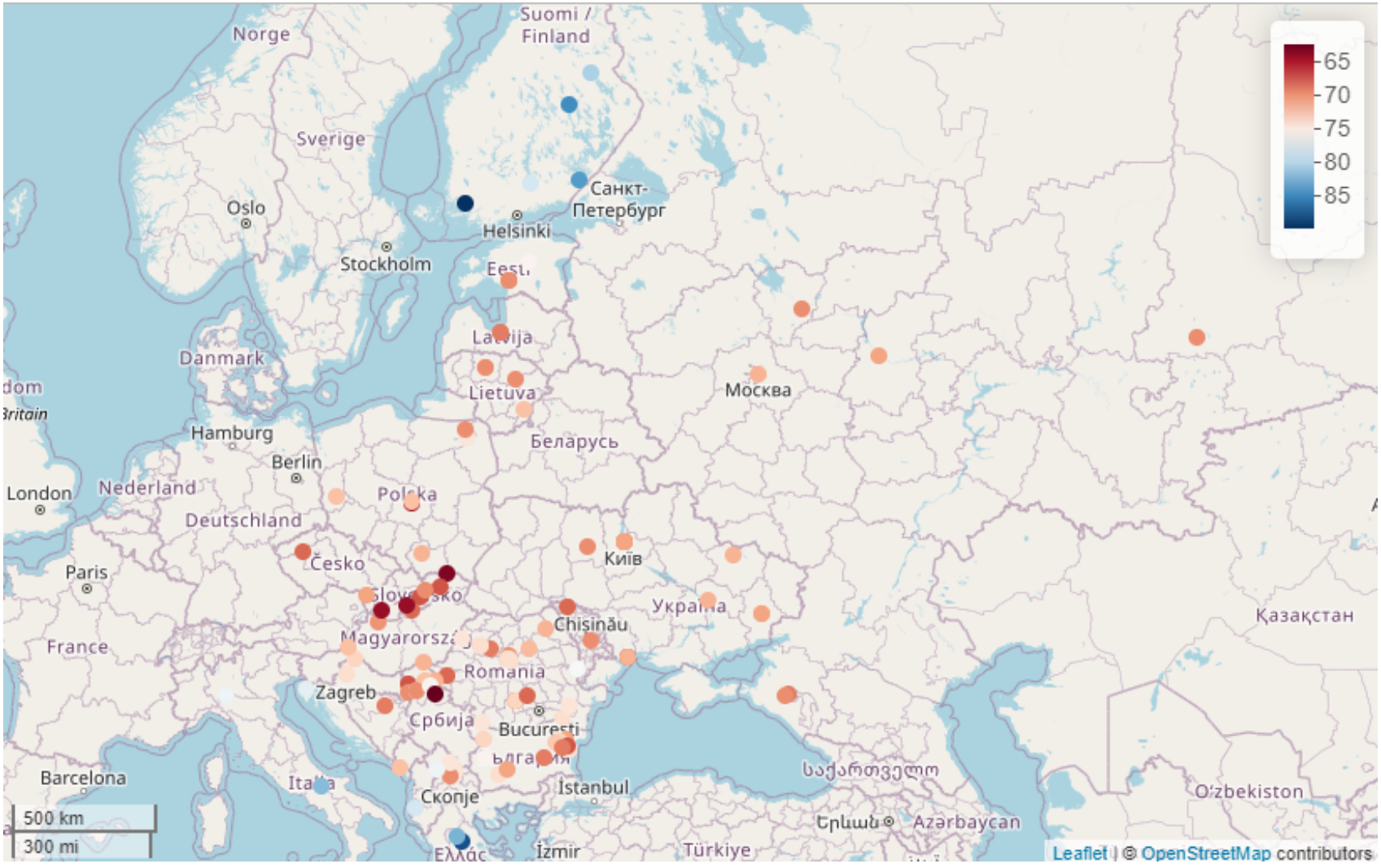


- How can we capture the range of variation across Romani dialects?

Transitivity prominence

- Transitivity prominence is the ratio of transitive valency patterns in a given subset (Haspelmath 2015).
- Transitivity prominence varies greatly across Romani dialects and ranges between 0.64 and 0.89.

Transitivity prominence across Romani dialects



- The range of transitivity ratio in Romani dialects is higher than in some genealogical taxons with the time depth of ca. 2000 years.

Taxon	Range	SD	Number of idioms
Romani	0.25	0.046	119
Slavic	0.14	0.044	11
Romance	0.08	0.024	8
Turkic	0.09	0.034	8

Table of Contents

Establishing shared patterns

- Due to recent common ancestry, it is possible to equate valency patterns across Romani dialects (despite minor differences in shapes).

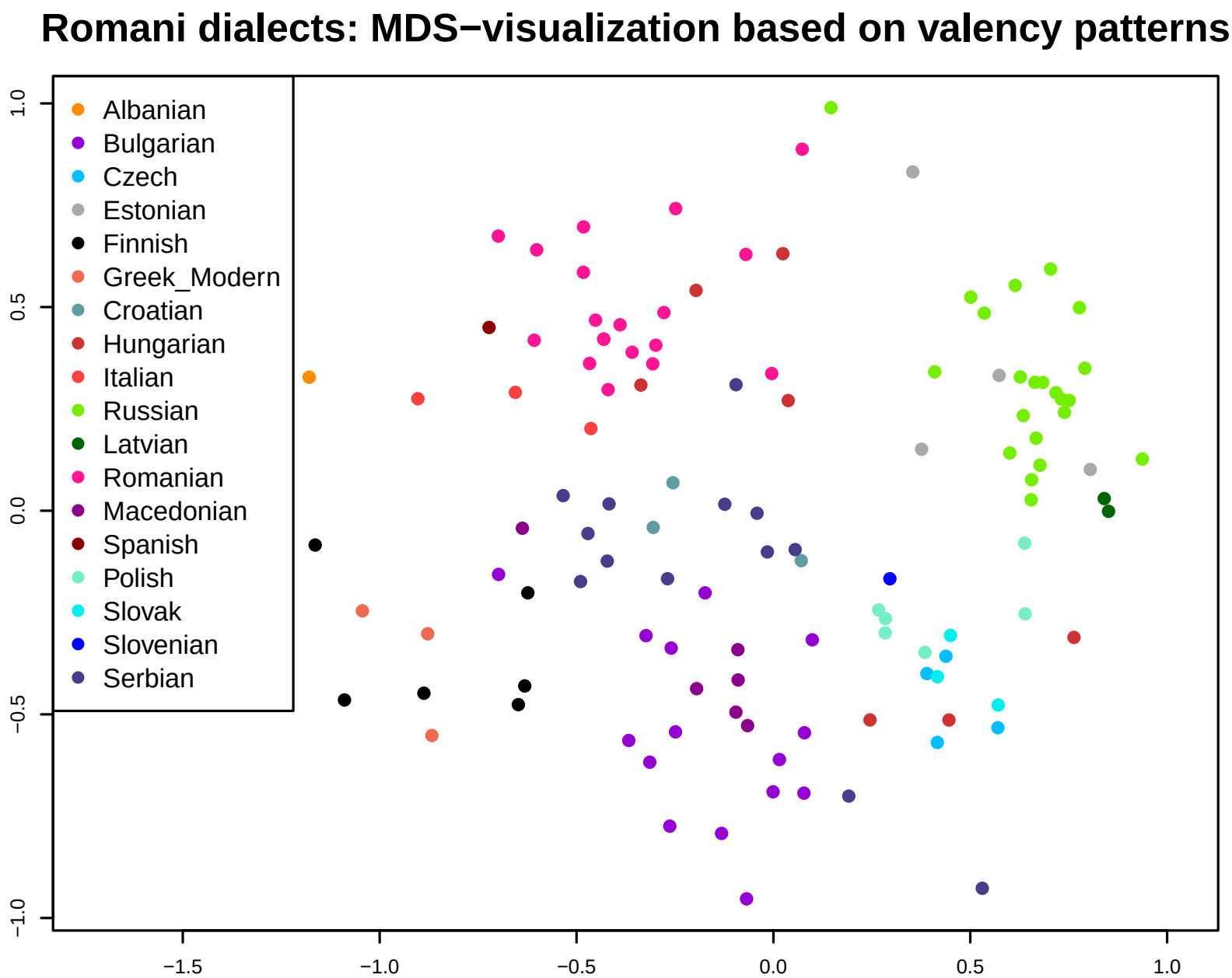
'get angry'

dialect	verb	X	Y	pattern
YU014	xol'anel	NOM	opri	NOM_on
CZ001	xojajel	NOM	pre	NOM_on
UKR010	xol'avel	NOM	pe	NOM_on
RO022	vel xojmen	NOM	an	NOM_in
MK001	xolavol	NOM	DAT	NOM_DAT

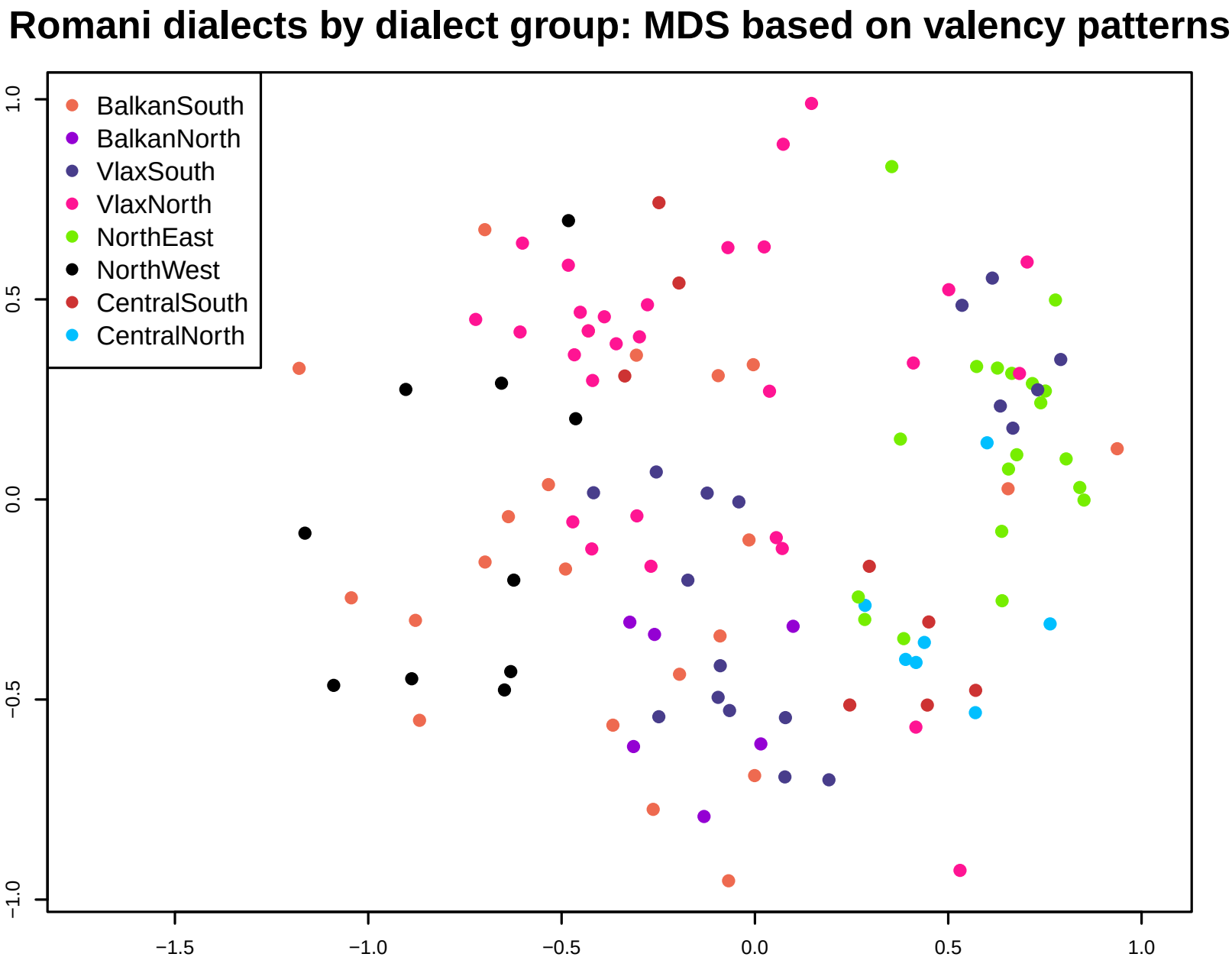
Dissimilarity between Romani dialects

- Dissimilarity distances between Romani idioms are based on the argument encoding of 43 predicates.
- The distance between two idioms for a given predicate was calculated as the Jaccard distance:
  - The formula is 1-I/U, where I is a number of shared patterns (Intersection) and U is the total of all attested patterns (Union).
- The distances are represented using the Multidimensional scaling algorithm implemented in R using the `smacof` package (de Leeuw, Mair 2009).

MDS plot by contact language



MDS plot by dialect groups



Clustering of dialects

- Based on their valency patterns, Romani dialects seem to cluster areally rather than genealogically.
- How can we test this hypothesis statistically?

Analysis of similarities

- Analysis of similarities (ANOSIM) operates on a dissimilarity matrix and tests whether the variation within some pre-established groups is smaller than between groups.
- ANOSIM was implemented in R using the `vegan` package (Oksanen et al. 2022).
- We tested the groupings based on:
  - (genealogical) dialect classification
  - primary contact language
  - country

- All three groupings prove to be significant.
- Higher R indicates larger dissimilarity between the groups.
- The statistic R is higher for groupings associated with areal effects: country and contact language.

Grouping by	p-value	R
Contact language	0.001	**0.815**
Country	0.001	**0.82**
Dialect groups	0.001	0.327

Table of Contents

Romani dialects vs. contact languages

- Romani dialects have similar valency patterns if they are located in the same area (defined by country or contact language).
- How can we compare valency patterns in Romani dialects and their respective contact languages? It is not possible to directly equate Romani and non-Romani valency patterns.
- We explored two alternatives:
  - transitivity prominence
  - locus of (non-)transitivity

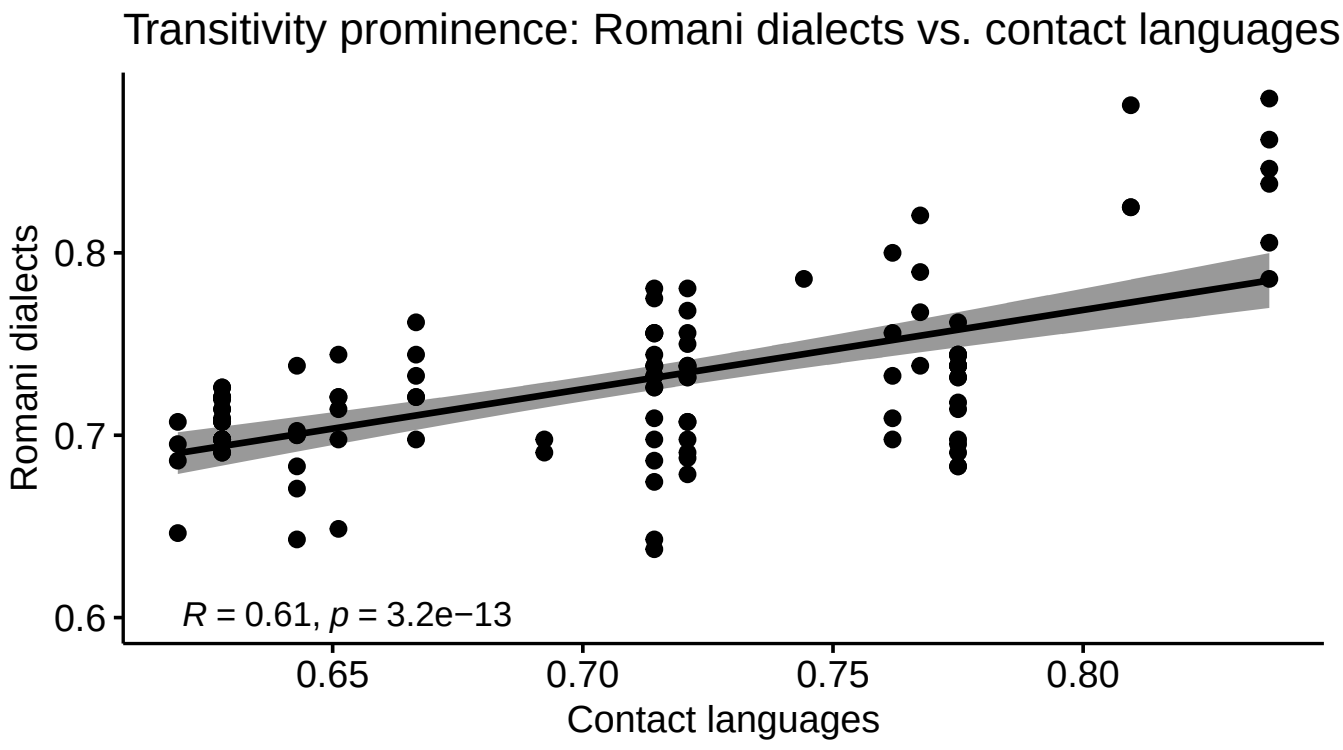


Figure 1: Strong positive correlation between the transitivity ratio in Romani dialects and their respective contact languages

Locus of (non-)transitivity

- Locus of (non-)transitivity is a four-way classification of valency patterns based on whether one, both or neither of the two arguments X and Y are encoded as non-core argument NPs.

Locus TR

(3) *abór* *aftokínit-a* *ther-él* *óv?*
how.many car-ACC.PL have-PRS.3SG 3M.NOM.SG
'How many cars does he have?' (Romacilikanes)

Locus X

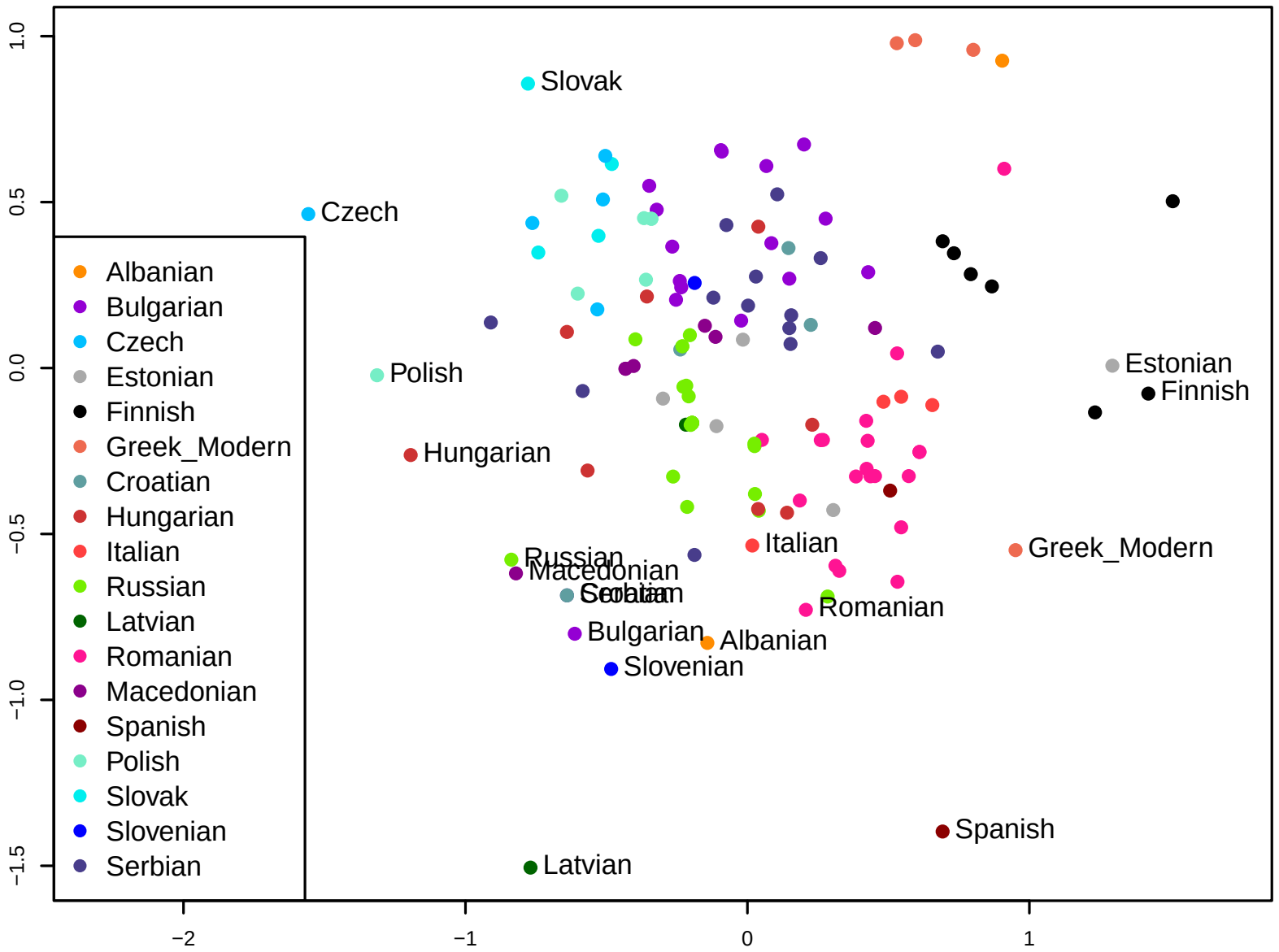
(4) *cík* ***léste*** *sí* *mašin-i?*
how.many 3M.LOC.SG be.PRS.3 car-NOM.PL
'How many cars does he have?' (Lotfitka Romani)

Locus Y

(5) *jóv* *xolisálij-as* ***pre mánde...***
3M.NOM.SG get.angry-PST.3SG on 1SG.LOC
'He got angry with me...' (Plaščuna Romani)

- Distances based on loci of (non-)transitivity can be measured in pairs of idioms regardless of their genealogical relatedness.

MDS plot colored by contact languages

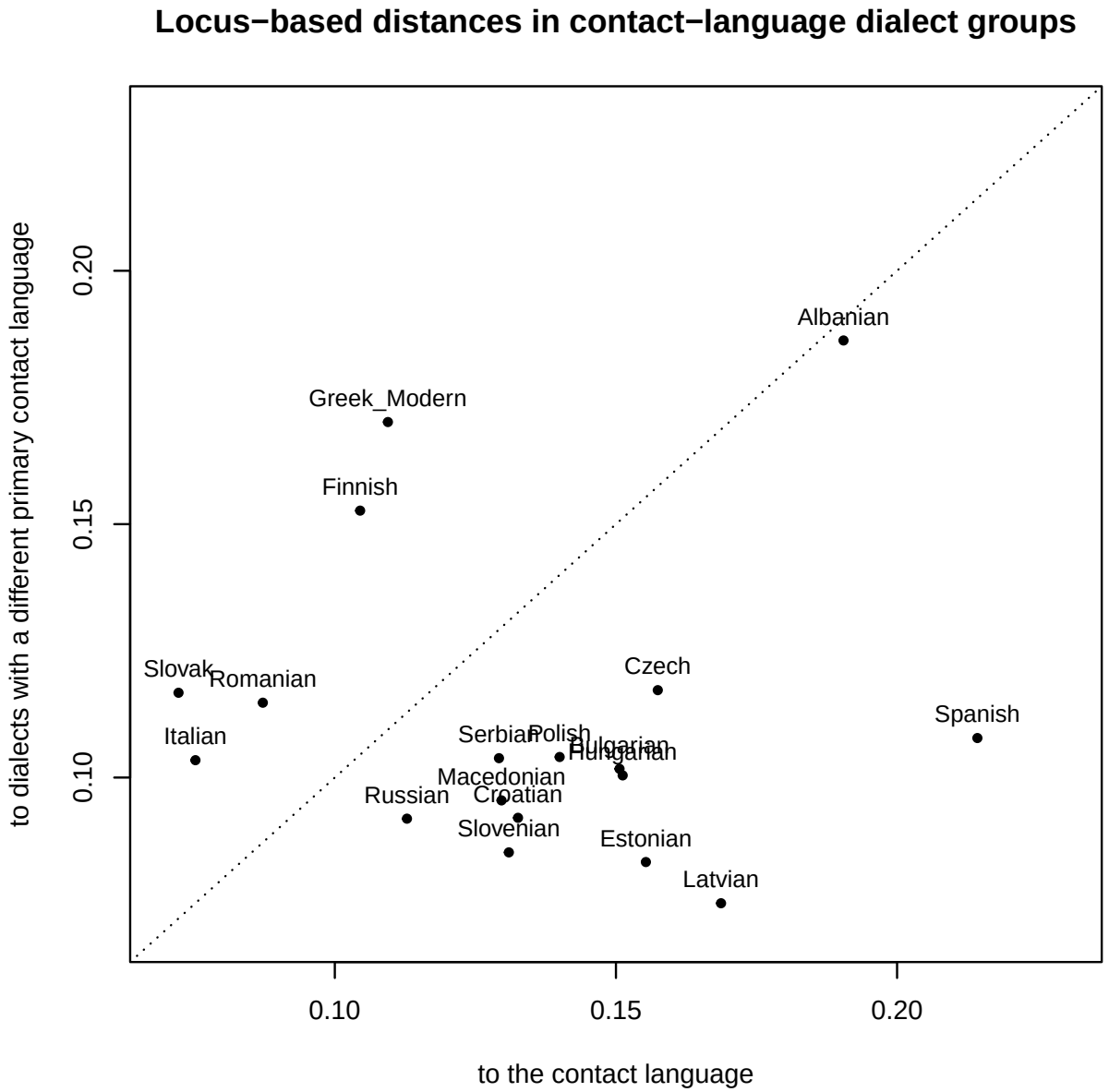


Locus of (non-)transitivity

- Romani dialects form a huge cluster with distances often greater than those between separate languages.
- Romani dialects do cluster together based on their primary contact language.
- There is a tendency for Romani dialects to be closer to their contact language, but they are not necessarily very similar.

- How can we assess the distance between Romani dialects and their contact languages vs. other Romani dialects?

Locus of (non-)transitivity



- Some Romani dialects are closer to their contact languages than to other Romani dialects: dialects in contact with Finnish, Greek, Italian, Romanian, and Slovak.
- If the distance to a contact language is smaller than to other Romani dialects, it is probably due to language contact.

Conclusions

Conclusions

- Argument encoding belongs to diachronically unstable phenomena and is easily reshaped via language contact (hence areal effects).
- Based on valency patterns, Romani dialects cluster areally rather than genealogically (traditional dialect classification).
- To a great extent, this is due to language contact. There is significant correlation between Romani dialects and their contact languages in regards to their transitivity prominence and locus of (non-)transitivity.