Variation in valency patterns across Romani dialects is primarily shaped by contact languages

Kirill Kozhanov, Sergey Say (Universität Potsdam)

Abstract: This paper examines the effects of language contact on valency patterns by comparing Romani dialects with their contact languages. Due to their wide dispersion and extensive contact with diverse European languages, Romani dialects provide an excellent testing ground for exploring the interplay of genealogical and areal factors in valency encoding. Using data from the Romani morpho-syntax (RMS) database and BivalTyp, a typological database of bivalent verbs, we analyzed valency patterns in 43 predicates across 119 Romani varieties and 18 contact languages. Despite their relatively recent divergence (600–700 years), Romani dialects exhibit greater variation in valency patterns than some genealogical groups with a 2000-year history. These patterns align more closely with current geographic distribution and contact languages than with traditional genealogical classifications. The findings suggest that language contact is the primary driver of rapid changes in Romani valency systems, as some dialects resemble their contact languages more than other Romani varieties.

Keywords: Romani, valency, language contact, areality, flagging, case marking, adpositions, transitivity

1. Introduction

1.1. Setting the stage

The question of whether the distribution of observed linguistic diversity is primarily shaped by genealogy—i.e., inherited from a common ancestor—or by geographic proximity—i.e., acquired through areal diffusion—remains one of the central issues in typological research (Nichols 1992; Murawaki & Yamauchi 2018; Gast & Koptjevskaja-Tamm 2022; Skirgård et

al. 2023). The influence of these factors may vary across linguistic features, as they show differing degrees of diachronic stability (Nichols 1995; Wichmann & Holman 2009; Dediu & Cysouw 2013; Wichmann 2015; Greenhill et al. 2017). A corollary to this is that unstable features, which are easily updated through processes such as language contact, display stronger geographical patterning compared to diachronically stable features, which tend to preserve genealogical signal. Valency,[1] in particular, is considered to be a diachronically unstable feature, highly susceptible to influence from language contact (Say 2014; Grossman & Witzlack-Makarevich 2019; Trips 2020; see also Seržant et al. 2022: 327 for the conservative role of language contact in the development of valency patterns and Michaelis 2019 for the emergence of valency patterns in creoles).

The instability of valency patterns can be illustrated by (1–2) from two varieties of Romani spoken in Lithuania and Poland respectively:

Lithuanian Romani (Lithuania)

(1) *uźakir-á* ***tút*** *paše* *khangirí*
wait-FUT.1SG 2SG.ACC near church-NOM.SG
'I will wait for you by the church.' (RMS, LT008)[2]

Polish Romani (Poland)

(2) *źakir-áva* ***pe*** *túte* *paś* *khangirý*
wait-FUT.1SG on 2SG.LOC near church-NOM.SG
'I will wait for you by the church.' (RMS, PL019)

[1] Our definition of valency (pattern) is given in Section 1.2. However, it should not be confused with *valency alteration*, a term that, in Romani linguistics, refers to morphological derivation of transitive and intransitive verbs (see, e.g., Matras 2002: 120–128).
[2] All examples in this paper are taken from either the RMS or BivalTyp databases and are accordingly tagged. For further details, see Section 2.

These two Romani varieties belong to the same North-Eastern dialect group, with an estimated time depth of approximately 300 years since their split-up. Although these examples describe the same situation and use the same lexical items, they employ different encoding devices: in Lithuanian Romani, ‘the person waited for’ is marked by the accusative case, while in Polish Romani, a preposition *pe* ‘on’ in combination with the locative case (the default case with most prepositions) is used instead. The transitive pattern exemplified in (1) is omnipresent across Romani dialects and arguably reflects the inherited option. In contrast, the valency pattern shown in (2) replicates the pattern found in Polish, as seen in (3), where ‘cousin’ is encoded by the preposition ‘on’:

Polish

(3) *Basi-a* *czek-a* ***na*** *kuzynk-ę*
PN-NOM.SG wait-PRS.3SG on cousin-ACC.SG
‘Basia is waiting for her cousin.’ (BivalTyp)

A single instance of pattern replication—i.e., the use of inherited resources to copy an external model (see Matras & Sakel 2007; Matras 2020: 260–264 for a discussion of this concept)—might be an incidental linguistic fact. However, Romani dialects have undergone extensive development through contact with various languages and thus present a rare opportunity to systematically investigate the variation of valency patterns. Using Romani data, we can explore the ways in which, to what extent, and how fast a recipient language can adopt the valency patterns of the source language, and, consequently, how similar the recipient and source languages can become. Another intriguing question concerns how dissimilar varieties of the same language can become if one variety comes into contact with a certain external

language, while the other interacts with another external language. In the remainder of the Introduction, we clarify our definitions (1.2), provide background information on Romani (1.3) and outline our research questions (1.4).

1.2. Valency patterns: definition and challenges

By a verb's valency pattern, we refer to the set of semantic arguments licensed by the verb's meaning, where each argument is associated with specific encoding devices, such as cases, adpositions, and person indexes (Malchukov et al. 2015). Thus, the valency pattern of the verb 'to wait', as exemplified in (3), links 'the person who waits' and 'the person waited for' to the morphosyntactic slots signaled by different encoding devices: the first argument is marked by the nominative case and also indexed on the verb, while the second argument is marked by the preposition *na* 'on' in combination with the accusative case.

The notion of valency pattern, sometimes defined in slightly different ways or under other terms (e.g., argument structure, case frame, diathesis etc.), has been central to most theoretical approaches to Syntax-Semantic mappings (Apresjan 1969, Dowty 1991, Croft 1998, Levin & Rappaport Hovav 2005), including typological studies (Tsunoda 2004; Witzlack-Makarevich 2011; Malchukov & Comrie 2015). At least since Fillmore's groundbreaking study (1968), there has been broad consensus that the choice of argument-encoding devices is primarily triggered by the semantic roles of the arguments (Nichols 1975; Hawkins 1985; Croft 1993; Lazard 1994; Malchukov 2005). Proving this claim, however, is a challenging task due to the lack of agreed-upon procedures for identifying semantic roles for a given verb (but see Bickel et al. 2014; Hartman et al. 2014). In what follows, we adopt a fully empirical approach to this problem: instead of identifying abstract and discrete semantic roles on *a priori* grounds, we analyze the lexical distributions of verbs into valency classes, defined by the use of observable

encoding devices. This approach allows us to detect both similarities and differences between valency class systems in individual varieties (languages and dialects).

1.3. Language background

Romani is an Indo-Aryan language that has been spoken in Europe since the Middle Ages (Matras 2002; Matras & Tenser 2020; Matras et al. 2022). Early Romani evolved in close contact with Greek in the Byzantine Empire during the 11–14th centuries. There is no consensus on whether Early Romani was linguistically homogenous: a certain degree of uniformity is suggested by common structural innovations and lexical loans (Matras 2002: 19; Elšík & Beníšek 2020: 409), though some authors argue that dialectal differences may have already existed in the pre-European period (Boretzky & Igla 2004; Boretzky 2007). Be that as it may, new innovations were subsequently acquired during the migrations of Roma to Europe, which began no later than the early 15th century. Matras (2005) argues that most of the dialectal differences observed today developed during the 16–17th centuries.

Currently, dozens of Romani varieties are spoken across vast territories (primarily in Europe, but also beyond). These varieties are traditionally referred to as "dialects", even though the difference between some of them can arguably lead to mutual incomprehensibility (Elšík & Beníšek 2020: 391). There exist several classifications of Romani dialects based on phonological, grammatical, and lexical criteria, with a varying level of detail (Miklosich 1873; Matras 2002; Boretzky & Igla 2004; Elšík & Beníšek 2020). The division into Northern, Central, Vlax, and Balkan major dialect groups is widely recognized in Romani dialectology. At least in part, this division is linked to distinct migrations and subsequent innovations in several areas of diffusion (Matras 2002, 2005). As a result of later migrations, dialects of different "genealogical" origin can be now spoken in the same territories (Elšík & Beníšek 2020: 392–393).

All varieties of Romani function in similar sociolinguistic environments: most Romani speakers are at least bilingual, meaning that their language is always influenced by contact languages (primarily Indo-European, but also Uralic and Turkic).

Like many European languages, Romani uses flags—i.e., case and adposition marking—to encode syntactic arguments, as seen in (4–5), where *-asa* is an instrumental case marker and *pe* is a preposition ‘on’.

Kalderash Romani (Russia)

(4) *o* *Múrš-a* *divin-íl* *la* *Marijk-**ásа***
ART.M.DIR.SG PN-NOM.SG speak-PRS.3SG ART.F.OBL PN-INS.SG
‘Mursha is speaking with Marijka.’ (BivalTyp)

(5) *o* *Múrš-a* *xoľáv-el* ***pe*** *Maríjk-a*
ART.M.DIR.SG PN-NOM.SG be_angry-PRS.3SG on PN-DIR.SG
‘Mursha is angry with Marijka.’ (BivalTyp)

1.4. Research questions

The general goal of this paper is to contribute to the understanding of how genealogical and areal factors, including those related to language contact, shape the valency class systems in individual varieties. Specifically, it focuses on the Romani dialects of Europe as a testing ground for potential broader generalizations. The specific research questions addressed are as follows:

i) To what extent are the valency class systems similar (stable) across the Romani dialects of Europe? How does the degree of diachronic (in)stability observed in these dialects compare to other genealogical taxa?

ii) How can we measure (dis)similarities in the organization of valency class systems in pairs of Romani dialects? Do such distances reveal strong genealogical patterns, or are they primarily areal patterns?

iii) How can we measure (dis)similarities in the organization of valency class systems in Romani dialects and their respective contact languages, especially when direct etymological equations of specific markers, such as case affixes and adpositions, are not possible? How can we identify the effects of language contact in the resulting dialect systems without access to specific contact-induced events in their histories? Finally, in a situation where a certain dialect of a language is in contact with a different language, should we expect its valency class system to display closer similarity with other dialects or with the contact language?

2. Data and methodology

2.1. The Romani morpho-syntax database (RMS)

The primary source of our data is the Romani morpho-syntax database (RMS), a questionnaire-based database of Romani dialects in Europe (Matras & Elšík 2001–2016). The questionnaire includes approximately 300 lexical questions and 700 sentences designed to elicit morphosyntactic information (Matras et al. 2009). The RMS data consist of translations obtained through elicitation, which may therefore replicate the patterns of the interview language more closely than in free narratives. While we acknowledge this potential issue, we believe the data still provide a reliable overall picture.

The freely available online version of the database contains transcribed answers to the questionnaire from 119 locations (to be precise, 118 locations in Europe and one in Mexico). Each dialect in the database has a reference name followed by an ID. The reference names are based on the self-designation of the speaker, while the ID consists of a country abbreviation and a three-digit number. For example, Kalajdži (BG014) refers to Kalajdži Romani spoken in

Bulgaria. The recordings made in Serbia, Bosnia and Herzegovina, and Kosovo are coded as YU in the database. Each entry in the RMS database includes a sentence ID (i.e., the number of the question in the questionnaire), a transcription of the answer in the respective dialect, and an English translation. For most dialects and sentences, audio files are also available. A typical entry in the database is illustrated in (6) from East Slovak Romani spoken in Slovakia:

East Slovak Romani (SK002)

(6) 1021 - mindri čhaj daral jagata

1021 - My daughter is scared of fire.

As seen in (6), the database does not provide any morphological annotation of the data. All valency patterns for this study were coded manually (see Section 2.3 for further details). For instance, using (6) we would code the valency pattern of the verb 'be afraid' in this dialect as NOM_ABL, since *čhaj* 'daughter' is in the nominative, and *jagata* 'from the fire' is in the ablative form.

For each location, RMS provides additional information, including the name and geographical coordinates of the place where the recording was made, and, importantly, the contact languages (old, recent, and current) of the specific variety. In some cases, several contact languages are named.

Throughout the paper we use "dialect" as a technical term referring to a separate entry in our list of the 119 Romani varieties, without assuming any special linguistic status. This means that some "dialects" in our dataset are very similar and can essentially be treated as belonging to the same variety, whereas other "dialects" are quite dissimilar and may be mutually incomprehensible.

For this study, we used the version of RMS that was available on the University of Manchester website in February 2023. Currently, the database is no longer accessible at this address, but the same version is available on the University of Cologne website.

2.2. The BivalTyp database

The second source of our raw data is BivalTyp, an online typological database of bivalent verbs and their encoding frames (Say 2020–). BivalTyp is based on a questionnaire containing 130 predicates given in context, such as '(P. has to go out of the house, but there is a dog barking in the yard.). P. is afraid of the dog'. Most stimulus sentences in BivalTyp are sufficiently neutral and can be taken to represent the valency patterns associated with basic language-specific equivalents of the respective predicates, such as 'be afraid', 'kiss', or 'wait'.

Each BivalTyp entry contains a translation of a certain stimulus sentence into a target language (multiple translations are disallowed) and is annotated for the devices involved in the encoding of two pre-defined arguments, labeled X and Y. In the example above, 'P.' is X, and 'the dog' is Y. X is the argument that accumulates more agentive 'lexical entailments', in the sense of Dowty (1991), see also Bickel et al. (2014). In the predominantly dependent-marking languages of Europe, argument-encoding devices can be sufficiently characterized in terms of the cases and/or adpositions involved. The valency pattern in every entry is defined as the ordered combination of argument-encoding devices associated with X and Y. Thus, the valency pattern of the Slovenian equivalent of 'like' in (7) is schematically represented as "DAT_NOM".

Slovenian

(7)	*Petr-u*	*je*	*všeč*	*ta*	*srajc-a*
	PN-DAT.SG	be.PRS.3SG	pleasant	this.NOM.SG.F	shirt-NOM.SG

‘Peter likes this shirt.’ (BivalTyp)

For every language, the valency pattern associated with the predicate ‘kill’ is singled out as the ultimate transitive pattern (see also Haspelmath 2015: 136).[3] The language-specific argument-encoding devices involved in the transitive pattern are considered “core”. Non-transitive patterns are further classified based on whether one or both of the two arguments, X and Y, are encoded as non-core argument NPs. The pattern in (7), for example, displays the X-locus of non-transitivity, since its X-argument is encoded by the dative case, whereas Slovenian core arguments are encoded by the nominative or the accusative case.[4] The rationale behind the four-way classification in terms of the locus of (non-)transitivity — transitive, X-locus, Y-locus, and XY-locus — is two-fold. First, it is based on the idea that deviations from Hopper and Thompson’s (1980) transitivity prototype are usually encoded on the relevant constituent. In particular, verbs with non-volitional X’s tend to display X-locus, and verbs with non-affected Y’s tend to display Y-locus (Malchukov 2005, 2006). Second, this classification allows us to abstract from the language-specific details in the organization of case paradigms and alignment patterns, providing a tool for a broad cross-linguistic comparison of valency patterns associated with specific predicates (Say 2014: 142–148).

For this study, we used the latest development version of BivalTyp available in July 2023. Of the total sample encompassing 99 languages at that time, we used data from 18 languages that were tagged as primary contact languages for at least one of the Romani dialects covered in the RMS database (see Section 2.3).

[3] BivalTyp disregards differences in the encoding of arguments conditioned by intrinsic properties of arguments, such as animacy and definiteness. Languages with differential object marking are considered to display a single transitive pattern, even though the actual case forms observed in their transitive entries may vary.

[4] In the rare cases when a verb deviates from the transitive pattern without involving non-core devices, its locus is defined as the highest argument on the hierarchy X > Y that is encoded differently from the transitive pattern. Thus, ACC_NOM patterns in nominative-accusative languages are classified as patterns with an X-locus, and NOM_NOM patterns are classified as patterns with a Y-locus.

2.3. Preparing the dataset: selection of verbs and annotation

Our dataset contains data extracted from two databases — RMS and BivalTyp — and includes information about the valency patterns of 43 predicates in 119 Romani varieties and 18 contact languages. All data and code used in this study are available as Supplementary Materials at OSF (https://doi.org/10.17605/OSF.IO/389QM).

The list of 43 predicates used in this study includes the predicates that overlap between the two databases. To arrive at this list, we compared the questionnaires and selected those sentences from the RMS questionnaire that most closely correspond to the stimulus sentences in BivalTyp.

Data on 118 Romani dialects come from the RMS database (with one available location, GR002 (Romacilikanes), excluded due to excessive missing data points). Additionally, we included the only Romani variety covered in Bivaltyp: a Kalderash Romani variety spoken in the northwestern part of Russia, which we coded as RUS101.

Metadata on Romani dialects, available in the Supplementary materials, include the dialect ID, geographical coordinates of the location where the respective data were collected, dialect group names based on two alternative classifications, the country where the dialect is spoken, and its primary contact language. It is important to note that the dialect classifications used in this study are as close to capturing genealogical relationships as possible for Romani dialects. We further refer to dialect classification as "quasi-genealogical" because it closely approximates, but does not fully replicate, the genealogical classification of languages, which is based on their development from common ancestors. In the case of dialect groups, a distinct common ancestor cannot always be postulated. The quasi-genealogical groups in this study refer to earlier dialect clusters that may have emerged due migration, language contact, or areal innovations. Regardless of the factors behind the emergence of such clusters, our focus is on

the stability of their valency patterns over time — in other words, how (dis)similar contemporary dialects within each cluster are. We assigned dialects to groups based on available literature on Romani dialect classification (see also Section 1.3). Since dialect grouping can be subjective and multiple classifications have been proposed for Romani, we used two different schemes to assess whether these quasi-genealogical classifications align with differences observed in the argument-encoding data, and if so, which of the two performs better. The first classification, labeled "Dialect classification 1" in our metadata, includes eight dialect groups, following the reference grid commonly used in Romani dialectology (e.g., Matras 2005). The second classification, labeled "Dialect classification 2" in our metadata, represents the most recent classification and includes 12 dialect groups (Elšík & Beníšek 2020); however, our dataset covers only 10 of these groups. Importantly, neither classification is based on the distribution of valency patterns.

Apart from this, we identified primary contact languages based on the information provided in the RMS (see also Section 2.1). Out of 119 dialects in our dataset, 76 dialects (approximately 64%) have only one contact language, 34 dialects (about 29%) are described as having two contact languages, and 9 dialects have three contact languages. To make our data suitable for statistical analysis, we defined one primary contact language for each dialect. If the official language of the country was also the language used to collect the Romani data, it was usually considered the primary contact language for that variety. For instance, four Romani dialects in Bulgaria, which had two contact languages — Bulgarian and Turkish — were identified as having Bulgarian as their primary language. When the language of the interviews differed from the official language of the country, we took into account the sociolinguistic situation in the specific locations. In post-Soviet countries where data was collected in Russian, we defined Russian as the primary language for the Romani dialects of Lithuania and Ukraine, but not for those in Latvia, Estonia, and Moldova, where Russian, although widely used by the Romani

community, is arguably less prominent than the respective official languages. Additionally, we designated Hungarian as the primary contact languages for two Romani dialects in Transylvania and one dialect in Slovakia, reflecting the sociolinguistic situation in these varieties. Table 1 illustrates the structure of the metadata on Romani dialects used in our study.

Table 1. Examples of metadata on Romani dialects

Dialect ID	Latitude	Longitude	Dialect classification		Country	Contact language
			1	2		
AL001	40.73	19.56	BalkanSouth	BalkanSouth	Albania	Albanian
BG001	42.03	23.99	BalkanSouth	BalkanSouth	Bulgaria	Bulgarian
LV005	56.94	24.09	NorthEast	Northeastern	Latvia	Latvian
MX001	17.17	-97.09	VlaxNorth	Vlax	Mexico	Spanish
SK031	48.83	20.13	CentralNorth	CentralNorth	Slovakia	Slovak
YU002	45.46	19.21	VlaxSouth	Vlax	Serbia	Serbian

Finally, we compiled a list of 18 languages that are tagged as primary contact languages for at least one Romani dialect in the database. This list includes Spanish, Italian, Romanian, Albanian, Modern Greek, Bulgarian, Macedonian, Serbian, Croatian, Slovenian, Slovak, Czech, Polish, Russian, Latvian, Hungarian, Estonian and Finnish. The data on valency patterns in the contact languages were fully imported from the BivalTyp database. The only exception concerned the Finnish and Estonian patterns, where the X argument is in the nominative case, while the Y argument is in the partitive case. The Baltic Finnic patterns involving the partitive case are known to be theoretically challenging (Kiparsky 1998); however, for the practical purposes of our combined dataset, we regarded these patterns as

transitive, whereas in the original BivalTyp annotation, they were classified as patterns with the Y-locus.

We analyzed the Romani translations in the RMS and coded the observed valency patterns manually. When possible, we consulted audio files to verify the transcriptions provided in the database. When the RMS questionnaire had several entries for the same predicate (e.g., the predicate 'have' is attested in 30 sentences), we considered all entries for our coding. If we found variation in the valency patterns for a given predicate, either with the same verb lexeme or with different verb lexemes, we included all distinct patterns in the dataset (see the verb 'have' in Kalajdži Romani from Bulgaria (BG007) in Table 2 below). However, we disregarded distinct verbs corresponding to the same questionnaire sentence if they displayed the same valency pattern. The reason for that is that we are primarily interested in the valency patterns, not in the verbs themselves. Such an approach to variation puts a natural limitation on our data, as no variation could be captured for the predicates that appeared in the RMS questionnaire only once (e.g., 'call', 'sing' etc.).

Coding of the valency patterns with prepositions for the purposes of cross-dialect comparison followed the etymological criterion. For example, the various cognate forms of the Romani preposition 'on'—e.g., *uppe*, *pre*, *pe* etc., which all go back to the same adverb *opré* 'on the top, up'—would receive the same tag in the "Valency pattern" field of the final dataset. However, the preposition *na*, borrowed from Bulgarian, although having the same meaning 'on', would be coded separately. This distinction was particularly important for the coding of the predicates 'be afraid' and 'reach', as they employed various prepositions of different origin but supposedly similar semantics ('of, from' and 'to'/'in').

The entire dataset consists of 6067 data entries: 5293 entries for Romani dialects (43 predicates in 119 dialects, including cases with variation and 179 NAs, i.e., missing data) and 774 for contact languages (43 predicates in 18 languages; including 16 NAs). Each data entry

is a row in a spreadsheet, coded for the dialect, predicate, verb, its origin, encoding devices associated with the first argument (X), encoding devices associated with the second argument (Y), locus, and, finally, the valency pattern. Non-transitive valency patterns indicate the encoding of the two arguments, and transitive patterns are encoded as "TR". See Table 2 for some examples of our dataset entries.

Table 2. Examples of dataset entries

Dialect ID	Predicate	Verb	Origin	X	Y	Locus	Valency pattern
GR032	believe	pistinel	borrowed	NOM	ACC	TR	TR
UKR010	believe	patjal	inherited	NOM	DAT	Y	NOM_DAT
GR002	be angry	nevrijazi	borrowed	NOM	INS	Y	NOM_INS
CZ001	be angry	xojajel	inherited	NOM	pre	Y	NOM_opre
MK001	be angry	xolavol	inherited	NOM	DAT	Y	NOM_DAT
AL001	have	therel	inherited	NOM	ACC	TR	TR
BG007	have	si	inherited	ACC	NOM	X	ACC_NOM
BG007	have	si	inherited	DAT	NOM	X	DAT_NOM
LT005	have	sy	inherited	LOC	NOM	X	LOC_NOM

All statistical calculations were performed in R (R Core Team 2021). The R code used in this study is also available in the Supplementary materials. We used the following packages for the analysis and visualization of our data: `dendextend` (Galili 2015), `ggpubr` (Kassambara 2023), `smacof` (de Leeuw & Mair 2009), `vegan` (Oksanen et al. 2022).

3. Results

3.1. Variation in valency patterns across Romani dialects

How much variation, and hence dissimilarity, can we observe in contemporary Romani dialects when it comes to the valency patterns they employ? In this section, we describe the cross-dialectal diversity of valency patterns associated with a given (semantic) predicate in terms of richness (how many distinct patterns are attested for a given predicate across dialects) and variation ratio (how many patterns deviate from the most frequent one). Out of 43 predicates used for this study, 19 show no variation across Romani dialects at all: these predicates display only one pattern in each of the 119 dialects in our dataset, and it is the same pattern in all of them. Unsurprisingly, most of these predicates correspond to transitive verbs. For instance, with the verb *pjél* 'drink' (example 8), the same etymon is preserved in all dialects, and it is always a transitive verb.

Romacilikanes Romani (Greece)

(8) *áma pj-áva thúd bút, k-av-á bút zural-í*
if drink-PRS.1SG milk.ACC.SG a_lot FUT-be-1SG very strong-NOM.SG.F
'If I drink a lot of milk, I will be strong.' (RMS, GR002)

At the same time, more than half of the predicates in our dataset exhibit at least some variation in their valency patterns. Table 3 lists the ten most variable predicates in the dataset. Variation is captured here in terms of richness, i.e., the number of distinct valency patterns observed across the dialects, with both frequent and rare patterns being treated equally.

Table 3. Predicates with the highest number of distinct valency patterns

Predicate	Number of distinct patterns
'be afraid'	17

‘need’	10
‘remember’	9
‘play instrument’	8
‘be angry’	7
‘hate’	6
‘feel pain’	5
‘have’	5
‘have enough’	5
‘like’	5

As an example, let us take a closer look at the predicate ‘have’. In our dataset, this predicate displays five distinct valency patterns, which can be grouped into two macro-types in terms of their locus. In the vast majority of dialects, this predicate is used in patterns with X-locus: here, the possessee (Y) is in the nominative case, while the possessor (X) is encoded by some non-nominative device: accusative (9),[5] locative (10), or dative (11) case, or alternatively, the preposition *ke* ‘at’ (12). In a small minority of Romani dialects in the dataset, the predicate ‘have’ is used with a transitive pattern (13).

Kalajdži Romani (Bulgaria)

(9)*isí mán dúj phral-á*

be.PRS.3 1SG.ACC two brother-NOM.PL

‘I have two brothers.’ (RMS, BG013)

[5] For the sake of simplicity, the (relatively rare) ACC_NOM patterns were tagged as displaying X-locus in the dataset, although, technically speaking, accusative NPs belong to the core of the clause according to the definition adopted above.

Vlax Romani (Mexico)

(10) *lá-te sí jékh phrál*

3F.SG-LOC be.PRS.3 one brother.NOM.SG

'She has a brother.' (RMS, MX001)

Lotfitka Romani (Latvia)

(11) *lá-ke špál sý*

3F.SG-DAT brother.NOM.SG be.PRS.3

'She has a brother.' (RMS, LV006)

Plaščuna Romani (Ukraine)

(12) *ke lá-te jí phrál*

at 3F.SG-LOC be.PRS.3 brother.NOM.SG

'She has a brother.' (RMS, UKR019)

Mečkari Romani (Albania)

(13) *ther-áva duj-é phen'-én*

have-PRS.1SG two-OBL sister-ACC.PL

'I have two sisters.' (RMS, AL001)

However, the mere number of distinct patterns observed in individual dialects, termed 'richness' above, does not always allow for a reliable assessment of cross-dialectal variation associated with a given predicate. For instance, it is intuitively clear that if one specific pattern accounts for 95% of all entries associated with a given predicate, the predicate does not display any significant level of variation, even if the remaining 5% of entries differ across dialects.

This is the case for the predicate ‘remember’: as shown in Table 3, this exhibits 9 distinct etymological patterns across dialects, but in reality, the transitive pattern represents 86% of all entries for this predicate in our dataset, which makes the overall distribution only moderately variable. To better capture the degree of variation in the observed distributions, we calculated the ratio of patterns that deviate from the most frequent (and likely inherited) pattern associated with a given predicate. Table 4 lists all the predicates with the variation ratio of 0.32 or greater.

Table 4. Predicates with the highest variation ratio

Predicate	Most frequent valency pattern	Ratio of other attested patterns
‘play instrument’	NOM_opre[6]	0.62
‘need’	DAT_NOM	0.54
‘like’	TR	0.53
‘reach’	NOM_andre	0.52
‘help’	NOM_DAT	0.48
‘have’	ACC_NOM	0.47
‘be afraid’	NOM_ABL	0.44
‘believe’	NOM_DAT	0.42
‘feel pain’	ACC_NOM	0.37
‘have enough’	ACC_NOM	0.33
‘be angry’	NOM_opre	0.32

[6] The labels ‘NOM_opre’ and ‘NOM_andre’ correspond to valency patterns involving prepositions with the meanings ‘on’ and ‘in,’ respectively. These labels represent the prepositions in a generalized pan-Romani form.

To illustrate the meaning of figures shown in Table 4, let us take the predicate ʻhelp’ as an example. The majority of Romani varieties (52%) use the dative case for encoding the Y argument of this predicate, as in (14). However, the transitive pattern is nearly as frequent (47%) as the dative marking, illustrated in (15). Latsly, there is one dialect — a Lovari Romani dialect in Serbia — where the second argument is marked by the preposition *pe* ʻon’ (16).

Gurvari Romani (Hungary)

(14) *na* *bājin-áv* *te* *šegītin-áv* ***túke***
NEG mind-PRS.1SG SBJ help-SBJ.1SG 2SG.DAT
ʻI don’t mind helping you’ (RMS, HU007)

Kelderash Romani (Romania)

(15) *či* *dukh-ál* *ma* *te* *ažutí-u* ***tut***
NEG ache-PRS.3SG 1SG.ACC SBJ help-SBJ.1SG 2SG.ACC
ʻI don’t mind helping you’ (RMS, RO008)

Lovari Romani (Serbia)

(16) *šáj* *žutí-v* ***pe*** ***túte*** *te* *kam-és*
can help-PRS.1SG on 2SG.LOC SBJ want-SBJ.2SG
‘I don’t mind helping you’ (RMS, YU015)

The figures shown in Tables 3 and 4 are arrived at using different techniques, but they converge in that the following predicates display the highest degree of argument encoding variation: ʻbe afraid’, ʻbe angry’, ʻfeel pain’, ʻhave’, ʻhave enough’, ʻneed’, ʻlike’, and ‘play instrument’.

At this point, there is little doubt that Romani dialects display a considerable amount of variation in the choice of valency patterns they use to encode the same predicates. Intuitively, this level of variability is very high, especially given that the varieties in question are commonly viewed as mere 'dialects' of the same language, with a time depth of approximately 600 years. However, this intuition should be supported by objective data, as it cannot be ruled out on *a priori* grounds that other genealogical taxa display similar levels of variation. To this end, we analyzed the distribution of *transitivity prominence* scores across Romani dialects, comparing them to other genealogical taxa. Transitivity prominence is an intuitively unproblematic comparative concept that captures the ratio of transitive valency patterns in a given subset (Haspelmath 2015). For example, the Arli dialect of North Macedonia (MK002) has 32 transitive entries out of a total of 43 entries, which corresponds to a transitivity prominence score of 0.74. This value is close to the mean observed in the sample of Romani dialects. The transitivity prominence score varies in the range between 0.62 and 0.88 in our sample of 119 dialects, with an observed standard deviation of 0.046. In Table 5, we compare these variability metrics with those observed in three genealogical taxa that are sufficiently well covered in Bivaltyp: Slavic, Romance, and Turkic languages. When using the BivalTyp data, we considered only the 43 predicates that were also used for the Romani dialects.

Table 5. Variability of transitivity prominence scores in Romani dialects and other genealogical taxa

Taxon	Range	SD	Number of varieties
Romani	0.26	0.046	119
Slavic	0.14	0.044	11
Turkic	0.09	0.034	8
Romance	0.08	0.024	8

Measuring transitivity prominence is a simplistic instrument that only provides an aggregate characteristic of a given valency class system. That being said, the data in Table 5 clearly show that transitivity prominence scores vary greatly across Romani dialects. In fact, both variability metrics we used are higher in Romani dialects than in the sets of Slavic, Romance, or Turkic languages, that is, in genealogical taxa with a time depth of approximately 2000 years.[7] This finding substantially supports our first generalization: Romani dialects are highly divergent in their use of valency patterns.

3.2. Clustering of Romani dialects

The data discussed in the previous section unequivocally indicate that the set of Romani dialects displays a high level of variability in terms of valency patterns. This initial observation raises the following questions: how can we measure this variability, and do Romani dialects form any robust clusters—groups whose members are more similar to each other than to other dialects?

To measure the degree of variability in our sample of dialects, we employed a distance metric with a potential range between 0 and 1, where higher values correspond to a higher degree of dissimilarity in a given pair of dialects. There are various metrics for calculating such

[7] The statistics presented in Table 5 partially reflect differences in sample size. To account for this, we conducted subsampling by generating 100 random subsamples of Romani dialects (each containing eight dialects) and calculated the mean range and SD. The adjusted values are 0.14 for range and 0.044 for SD for Romani, compared to 0.13 and 0.043 for range and SD, respectively, for Slavic. Although these adjusted values are closer than the raw statistics in Table 5, Romani still exhibits greater variability than the other three taxa.

linguistic distances; in our case, we needed a metric that could handle the problem of within-dialect variation. As discussed in Section 2.3, there were cases where a certain dialect had multiple entries for the same predicate. To address this challenge, we began with a predicate-by-predicate comparison for each pair of dialects and used the Jaccard distance to compute the dissimilarity distance between the two dialects for a given predicate. The Jaccard distance is calculated according to the following formula: 1 – I/U, where I stands for Intersection and U stands for Union. In a simple case where the two dialects have one entry for a given predicate each, the Jaccard distance is either 0 if the two dialects use the same valency pattern, or 1 if they use different patterns. In this basic scenario, the Jaccard distance is equivalent to the simple matching distance. However, the Jaccard distance allows for some gradience in more complex scenarios. For example, if dialect 1 displays the set of patterns <a, b> for a given predicate, while dialect 2 displays the set of patterns <a, c> for the same predicate, the Jaccard distance between the two dialects for this predicate equals 2/3 (their Intersection contains one element, a, and their Union contains three elements, a, b, and c). To calculate the aggregate distance between a given pair of dialects, we averaged the Jaccard distance observed for individual predicates. This resulted in a distance matrix covering all 119 Romani dialects in our sample.

We then applied Multidimensional Scaling (MDS), a standard algorithm for dimensionality reduction and visualization, and hierarchical clustering to generate a dendrogram from the distance matrix.

The MDS algorithm was implemented in R using the `smacof` package (de Leeuw & Mair 2009). The resultingt two-dimensional visualization is shown in Figure 1.

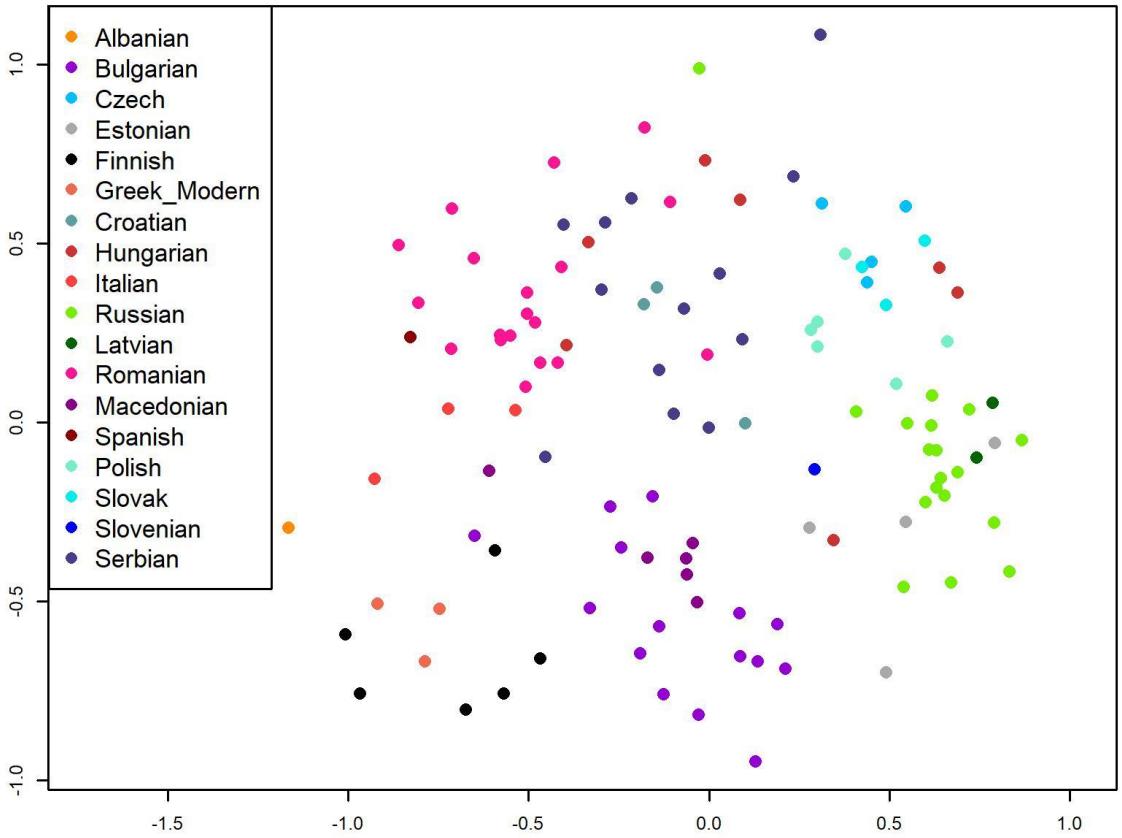


Figure 1. MDS of Romani dialects based on their valency patterns: colored by the primary contact language

Every point in Figure 1 represents a Romani dialect and is colored according to the respective dialect's primary contact language. The MDS plot in Figure 1 suggests that structural distances between Romani dialects are largely shaped by their primary contact language: visually, dialects sharing the same primary contact language tend to form contiguous zones.

However, it is not possible to identify specific clusters of dialects through intuitive visual inspection of Figure 1 alone. To determine specific clusters, we applied a hierarchical clustering algorithm implemented in R, using the `dendextend` package (Galili 2015). The resultant dendrogram is available in the Appendix. To classify the dialects, we set an arbitrary threshold

of eight clusters (the discrete classification of the dialects into these eight clusters is also provided in the Supplementary materials). The eight clusters include Romani dialects: (i) from Italy, Romania, Moldova, and Mexico; (ii) from Albania and Greece, and (iii) from Finland (together with IT007, i.e., Molise Romani); (iv) a group of dialects from Eastern Europe (Estonia, Russia and Ukraine); (v) a cluster formed by the dialects from Bulgaria; (vi) a mixed group of dialects from several countries (Hungary, Moldova, Romania, Russia, and Serbia); (vii) another mixed cluster of dialects spoken in Central and Eastern Europe (Croatia, Czechia, Hungary, Latvia, Lithuania, Poland, Slovakia, Slovenia, Russia, and Ukraine); and finally, (viii) a South-Eastern European cluster (Croatia, Macedonia, Serbia). The clustering algorithm makes the first split in the data by contrasting clusters (i–iii) with the rest.

These clusters include Romani varieties from different dialect groups, with the Vlax and South Balkan dialects being particularly diverse (represented in six and five clusters, respectively). In other words, similar to the multidimensional scaling, hierarchical clustering indicates that Romani dialect groupings are chiefly determined by geography and shared primary contact language. This is evident in the Bulgarian cluster (v) or the cluster consisting of the Romani dialects of Eastern Europe (iv). However, geography does not explain all groupings — cluster (vi) is a clear example of this. This cluster also includes two outliers in the MDS graph, namely RUS101 (Kalderash Romani, BivalTyp) and YU015 (Lovari Romani). Here, we oserve a genealogical signal, as all these dialects belong to the Northern Vlax dialect group, which was historically formed in the Romanian-speaking territories. Importantly, at least these two geographically divergent varieties from cluster (vi) are spoken in their current locations due to recent migrations in the last 150 years.

At this point, we can pose the crucial question: what is a better predictor for the structural distances between Romani varieties — geography or genealogy? Some intuitive evidence comes from MDS visualizations where the points correspond to the same dialects and are

located at the same positions as in Figure 1, but the colors correspond to quasi-genealogical dialect groups. As we mentioned in Section 2.3, we used two alternative classifications in this study. However, the resultant visualizations are predictably very similar to each other. For reasons discussed below, "dialect classification 1" yields a slightly clearer picture and is shown in Figure 2 below. The same visualization based on "dialect classification 2" is available in the Supplementary materials.

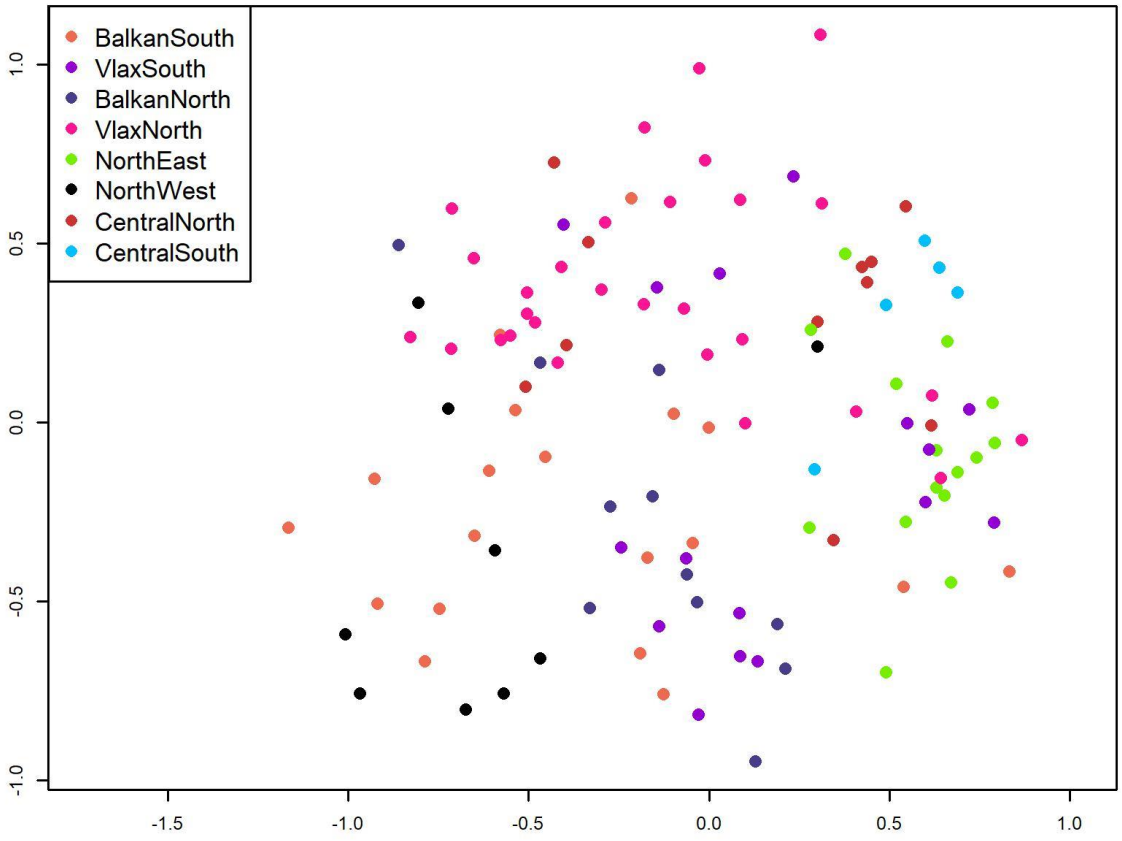


Figure 2. MDS of Romani dialects based on their valency patterns: colored by the dialect classification (1)

The visualization in Figure 2 seems to present a significantly more distorted picture than the visualization in Figure 1. Substantially, this suggests that a dialect's contact language is a better predictor for its valency behavior than its quasi-genealogical classification.

However, this hypothesis should not be based exclusively on intuitive visual judgement: it is also necessary to employ a rigorous statistical procedure to test it. To this end, we used the ANOSIM (analysis of similarities) algorithm, as implemented in R using the `vegan` package (Oksanen et al. 2022). This algorithm operates on a dissimilarity (distance) matrix and tests whether the variation within some pre-established groups is smaller than the variation between groups.

We tested four types of pre-established groupings in order to measure within- as opposed to between-groups variation. These groupings were based on i-ii) two quasi-genealogical dialect classifications, iii) primary contact language, and iv) country. The ANOSIM algorithm returns a useful statistic R: higher values of R are obtained if there is a greater difference between the variation observed between the pre-established groupings and the variation observed within these groupings. The results of implementing the ANOSIM algorithm are shown in Table 6, which contains both the values of R and the respective significance levels for the four types of groupings.

Table 6. Distances between Romani dialects

Grouping by	p-value	R
Contact language	0.001	0.804
Country	0.001	0.800
Dialect classification 1	0.001	0.374
Dialect classification 2	0.002	0.356

The results in Table 6 show that all four groupings are significant. However, the level of robustness varies across the four cases: the statistic R is higher for groups associated with areal predictors, namely country and contact language, than for the two dialect classifications.[8] Among the two quasi-genealogical classifications, "Dialect classification 1" with 8 groups performed slightly better than "Dialect classification 2" with 12 groups (this is the reason why we chose "Dialect classification 1" for the visualization in Figure 2 above). In any case, the data in Table 6 provide statistical support for the claim that geography is the main predictor of dissimilarity between Romani dialects in the domain of valency patterns.

3.3. Romani dialects and contact languages: (dis)similarities in valency class systems

3.3.1. The challenge

In the previous sections, we observed that argument-encoding systems in the Romani dialects of Europe exhibit a high level of variability (Section 3.1) and ascertained that the differences between these systems align with the classification of Romani dialects based on their primary contact language (Section 3.2). It is natural to surmise that both findings are ultimately attributable to intense contact with non-Romani languages of Europe. However, so far, we have only provided indirect evidence for this hypothesis, as we have not directly compared the valency patterns of the Romani dialects with those observed in their contact languages.

Measuring similarities and differences between Romani and non-Romani varieties is methodologically challenging, since valency classes of unrelated or remotely related languages cannot be directly equated (Comrie et al. 2015: 4–5). In Section 3.2, we equated individual valency-encoding devices across Romani dialects. This was possible due to the relatively shallow history of dialect divergence and, consequently, the transparent cognacy relationships

[8] Our data do not allow for a meaningful comparison between the two areal groupings we used, namely primary contact language and country. The results obtained from these two cases are almost identical, which is unsurprising given that the two groupings largely overlap.

between grammatical markers. For example, we assumed that verbs requiring ablative encoding of the Y argument in different dialects belong to the "same class", as the ablative case is a relatively well-preserved form across the Romani dialects.

However, such a straightforward equation of encoding devices is not applicable when comparing Romani dialects with non-Indic Indo-European languages, let alone with genealogically unrelated languages. Clearly, language-specific descriptive labels cannot serve as a *tertium comparationis*.[9] For example, Estonian, like the Romani dialects, has a case traditionally referred to as the 'ablative'. However, unlike the Romani 'ablative', the Estonian 'ablative' is rarely used in our dataset. Most Romani predicates taking objects in the 'ablative' correspond to Estonian predicates with objects in the so-called 'elative' case. Since neither the Estonian elative nor the ablative is historically related to the Romani ablative, it is impossible to directly equate valency classes in Estonian and Romani based on case labels and to calculate distances between them accordingly.

To overcome this challenge, we use two methods:[10] one based on transitivity prominence (Section 3.3.2) and the other on the locus of (non-)transitivity (Section 3.3.3). The core idea behind these solutions is the assumption that both transitivity and locus of (non-)transitivity are defined as comparative concepts (in the sense of Haspelmath 2010) and are thus independent of the genealogical and typological profiles of the languages compared.

3.3.2. Transitivity prominence

[9] In the literature, there have been attempts to equate oblique forms involved in encoding of objects in terms of some abstract notion related to case labels, such as 'dativity' (Van Belle & Langendonck 1996; Blume 1998). However, such an approach would involve a high degree of arbitrariness in assigning discrete annotations in cross-linguistic datasets.

[10] There are other robust techniques for comparing valency class systems in unrelated languages. For instance, Say (2014) uses Mutual Information to compare valency systems and ultimately views individual valency class systems as different solutions to the set-partition task. However, this technique requires relatively large samples of predicates and is not applicable to our dataset of only 43 predicates.

The use of transitivity prominence to compare Romani dialects with their contact languages is a natural extension of the discussion in Section 3.1, where we observed considerable variation in transitivity prominence values across Romani dialects. For the contact languages, we relied on data from BivalTyp. To make the values from RMS and BivalTyp comparable, we subsetted the BivalTyp data to include only the 43 predicates that are also covered in RMS. In this way, we identified the transitivity prominence values for the 18 languages that serve as primary contact languages for at least one of the Romani dialects in our sample. These values vary between 0.62 (for Czech) and 0.84 (for Finnish), which closely matches the range observed in the case of Romani dialects (see Section 3.1).[11] More importantly, there is a statistically significant positive correlation between the transitivity prominence observed in a given Romani dialect and the value observed in its primary contact language (Pearson's $R = 0.59$, $p < .001$). This finding is visualized in the scatterplot in Figure 3, where each point represents a Romani dialect. The y-axis shows the transitivity values for the Romani dialects, and x-axis represents the transitivity values observed in their respective contact languages. The linear regression line represents the correlation between the two variables.

[11] These values are significantly higher than the transitivity prominence values reported in the BivalTyp database (Say 2020–) and related publications (Say 2014, 2018). The discrepancy arises from the subsetting procedure: the 43 predicates used in this study generally belong to the highly transitive section of the original 130-predicate sample in BivalTyp.

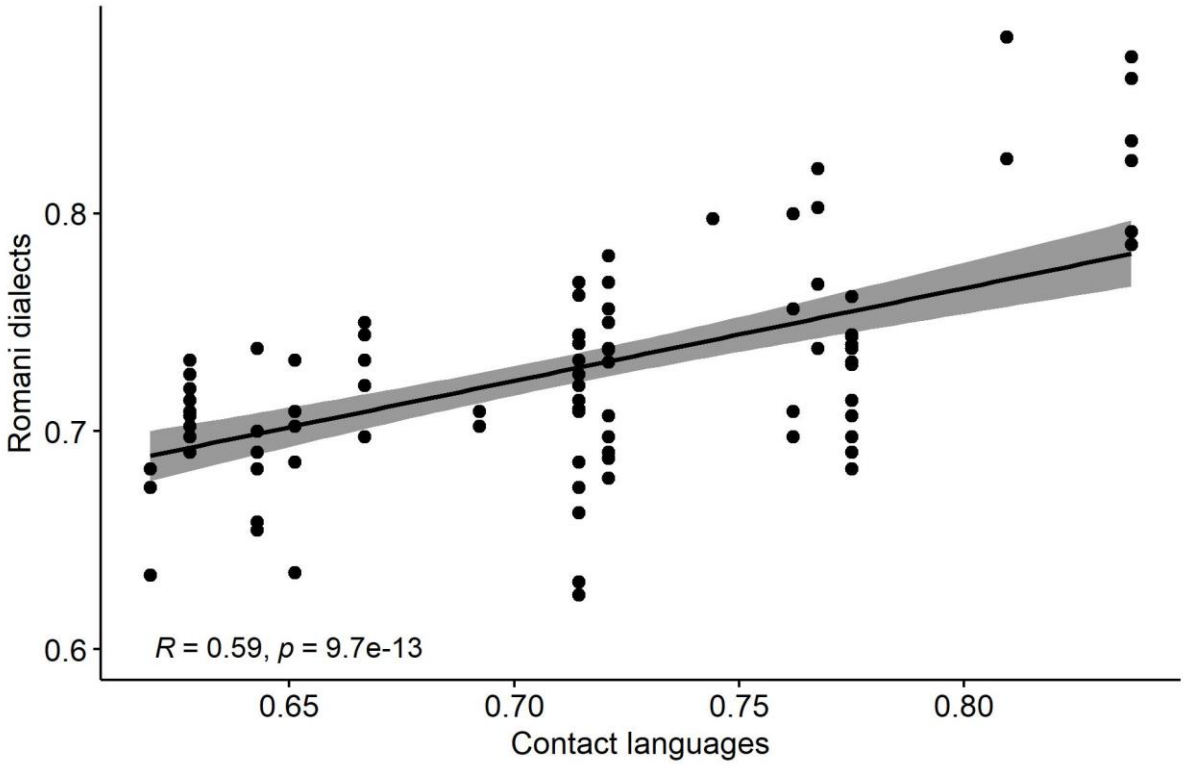


Figure 3. Transitivity prominence: Romani dialects vs. contact languages

The correlation visualized in Figure 3 clearly shows that Romani dialects in contact with highly transitive languages, such as Finnish[12] or Modern Greek (represented by the two rightmost vertical groups of points), tend to display higher transitivity prominence values than dialects in contact with low-transitivity languages, such as Czech or Russian (represented by the two leftmost vertical groups of points). This finding corroborates the idea that the Romani-internal variation in transitivity prominence, as discussed in Section 3.1, is largely shaped by language contact. From a geographical perspective, Romani dialects align with a broad areal trend, where low-transitivity languages of Eastern Europe, e.g., East Slavic and Baltic, are

[12] This holds true only if we treat Finnish and Estonian valency patterns with the partitive marking of the Y argument as transitive ones (see Section 2.3). Considering the NOM_PART pattern as non-transitive yields a more distorted picture.

flanked by areas with higher transitivity values, such as Western Europe and the Balkans (Lazard 2002: 153–154; Say 2014; Haspelmath 2015: 139–140).

3.3.3. Locus of (non-)intransitivity

The notion of transitivity prominence as a numeric typological variable is intuitively transparent, but it characterizes every language-specific valency class system in a very generalized way. While differences in transitivity prominence values always reflect substantial differences between varieties, any observed similarities can be fully fortuitous. For example, if the sets of transitive verbs in two varieties are largely divergent but happen to be of comparable size (as in the case of Latvian and Slovak), the similarities might not be meaningful. To arrive at finer generalizations, we propose a distance metric that considers the internal structure of the verbal lexicon and is based on the locus of (non-)transitivity.

As discussed in Section 2.2, the locus of (non-)transitivity is a four-way classification of valency patterns based on whether one, both, or neither of the two predefined arguments X and Y are encoded as non-core NPs, that is, "deviate" from the encoding devices employed in the transitive construction. At the level of individual entries in the dataset, the metric we use is discrete and binary: two equivalents of the same predicate in two varieties either display the same locus of (non-)transitivity, in which case the distance between them is 0, or they display different loci of (non-)transitivity, in which case the distance is 1. To illustrate this, let's consider the equivalents of 'be afraid' in Modern Greek, Albanian, and Croatian. The valency patterns associated with these equivalents are schematically represented as TR, NOM_*nga* and NOM_GEN, respectively. Although the individual argument-encoding devices of Albanian, including its preposition *nga* (≈ 'from'), cannot be directly equated with those in Croatian, the valency patterns observed in Albanian and Croatian are intuitively similar, as both encode the object feared in a non-core position (Y-locus). This is not the case in Modern Greek, where the

equivalent of ‘be afraid’ is a transitive verb. The metric we propose effectively captures these differences: the locus-based distance between Albanian and Croatian in the case of ‘be afraid’ equals 0, whereas the distances between either of these two languages and Modern Greek is 1.[13] We applied the same procedure to all the 43 predicates in the dataset, averaged the observed values, and eventually arrived at a locus-based distance matrix for all 137 varieties (119 Romani dialects and 18 contact languages) in our dataset.

The resultant distance matrix is empirically rich but difficult to inspect or visualize. To address this standard problem, we used Multi-Dimensional Scaling, one of the two techniques we introduced in Section 3.2. The algorithm was implemented in R using the `smacof` package (de Leeuw & Mair 2009). The results are visualized in Figure 4. In this visualization, labeled points represent contact languages, whereas unlabeled points represent Romani dialects. Dialect points are colored according to the color used for the respective contact language.

[13] This procedure could not be directly applied in situations where more than one entry corresponded to a certain predicate in a given variety. To adapt the metric to such cases, we again used the Jaccard distance (see the discussion in Section 3.2). We employed the same formula, 1-I/U, but this time I (Intersection) corresponded to the number of shared locus possibilities attested, and U (Union) corresponded to the total number of all attested loci among the equivalents of a certain predicate in the two varieties.

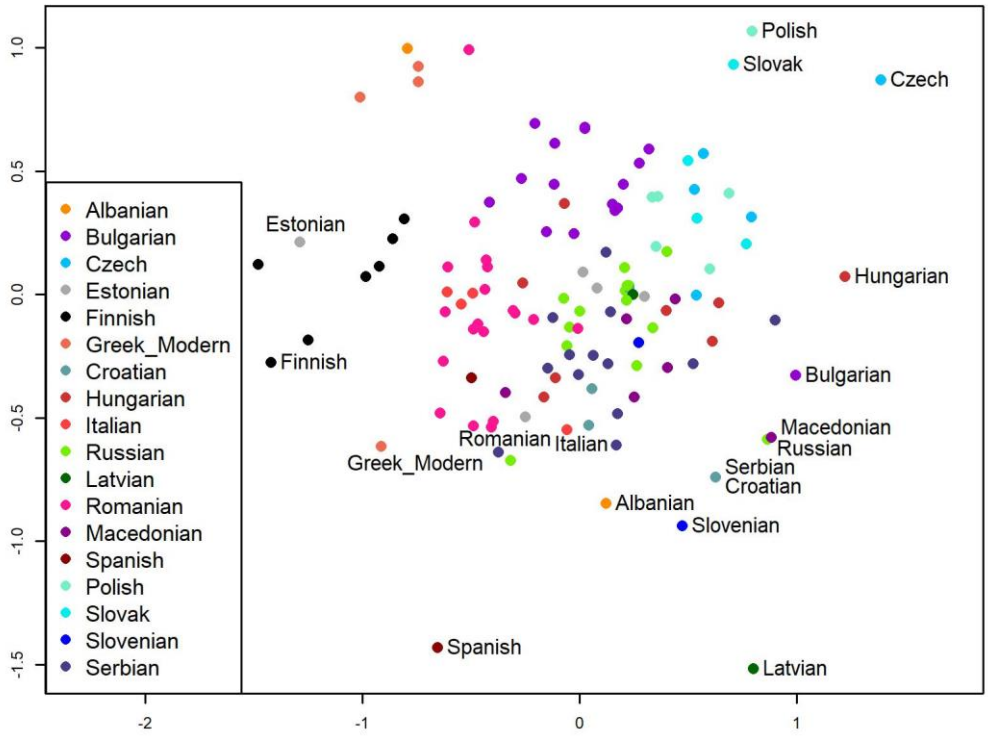


Figure 4. Romani dialects and primary contact languages: locus-based distances (MDS-visualization)

It should be borne in mind that the MDS-visualization in Figure 4 distorts the original distances. The degree of such distortion, called "stress", is relatively high, with a value of 0.21, which is close to 0.2, typically considered the threshold between 'poor' and 'fair' MDS-analyses (Dexter et al. 2018: 434). Nonetheless, a visual inspection of Figure 4 allows several generalizations, all of which are supported by the original data, i.e., by the unmodified distance matrix.

First, the Romani dialects form a large cluster that is not interspersed with non-Romani languages. Essentially, this means that Romani dialects retain a certain degree of unity. However, the distances between some Romani dialects are often larger than the distances between varieties that are considered separate languages. In fact, Romani dialects are scattered

over greater distances than, for instance, West or South Slavic languages (as shown in Figure 4 and the underlying distance matrix).

Second, dialects with the same primary contact language tend to cluster together in the visualization. More importantly, these dialect clusters seem to gravitate towards their respective contact languages. For example, among all the Romani dialects, those closest to Finnish are the dialects whose primary contact language is Finnish (represented by black points in the visualization). Similar patterns can be observed for dialects whose primary contact languages are West Slavic (light bluish points) or Romanian (pink points).

However, visual inspection of Figure 4 is not sufficient to draw a robust empirical generalization. To test the hypothesis statistically, we went back to the actual distances and averaged them by contact language. Specifically, for each contact language L, we calculated i) its mean distance to the Romani dialects that have L as their primary contact language, and ii) its mean distance to the Romani dialects that do not have L as their primary contact language. The actual data strongly support the generalization that the values of distances in i) are significantly smaller than the values in ii) (paired Student's t-test, $t = -4.42$, $p < .001$). The two types of averaged distances for each contact language are shown in Figure 5. What is significant here is that the majority of points lie above the $y = x$ line, indicating that dialects in contact with a particular language are closer to that language than the dialects with different primary contact languages.[14]

[14] Albanian is the only language in the dataset for which the reverse is observed. However, this exception can be attributed to the fact that only one Romani dialect in the RMS database has Albanian as its primary contact language: the Mečkari dialect, encoded as AL001. The locus-based distance between this dialect and Albanian is 0.20, a high value compared to other distances in the matrix. However, this dialect displays very low distances to Romani dialects spoken in Greece (0.01–0.06) and a moderate distance to Modern Greek (0.13). Importantly, although Albanian is identified as the main current contact language for the Mečkari dialect, Greek is also mentioned as a recent contact language. Evidently, in terms of its valency behaviour, the dialect in question has been more strongly influenced by Greek than by Albanian.

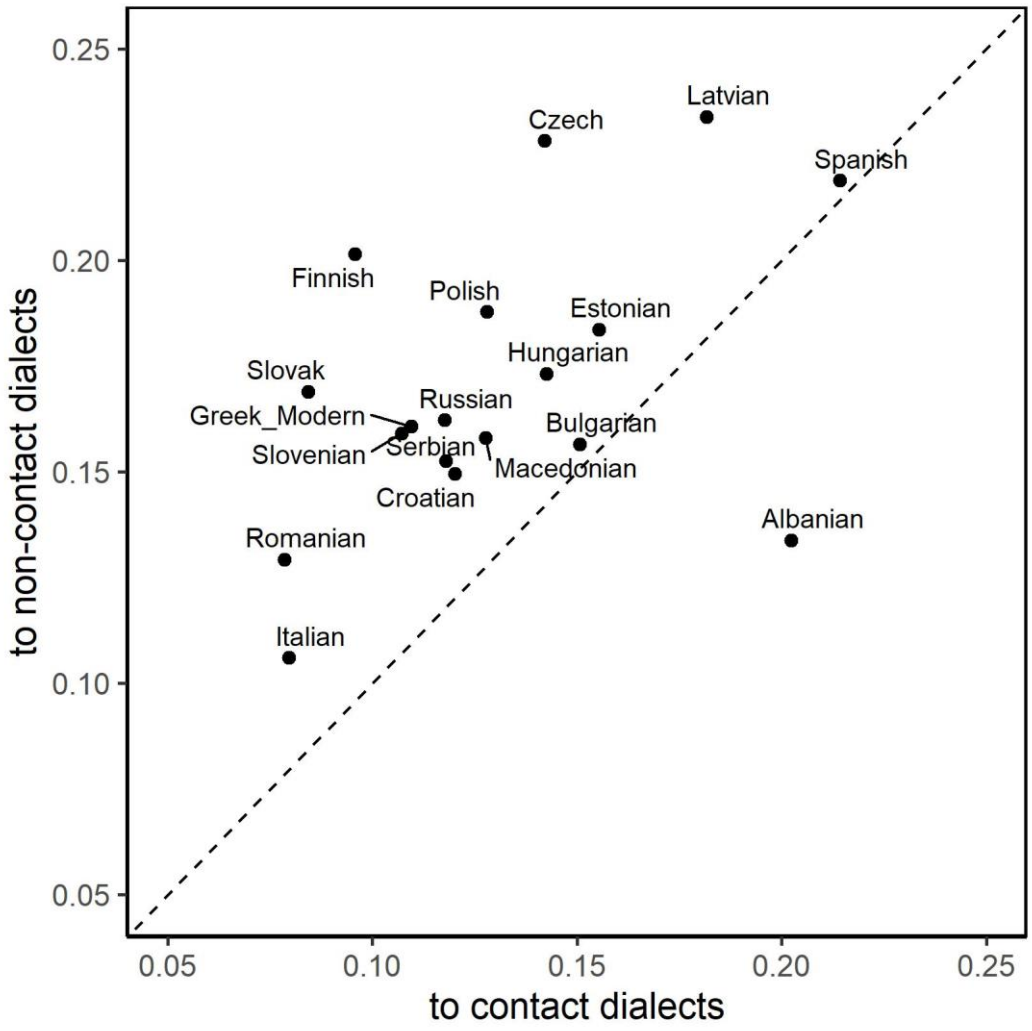


Figure 5. Average locus-based distances between 18 languages of Europe and contacting vs. non-contacting Romani dialects

The generalization discussed above offers new insights into the findings reported in Section 3.2, where we observed that Romani dialects sharing the same primary contact language tend to display significant similarities in the distribution of their valency frames across the lexicon. While the influence of the contact language was a very likely explanation, it was not the only possible one. An alternative scenario is the one where Romani dialects in contact with the same non-Romani language are also geographically close to each other and can develop or preserve common features through inter-dialect contacts (more complex scenarios are also possible).

However, a direct comparison between Romani dialects and their contact languages provides strong evidence against the alternative explanations. In the light of the data discussed in this section (see Figures 4 and 5), it is clear that the impact of contact languages is the main factor driving the observed divergence in valency patterns across Romani dialects.

Although this finding is important for our objectives, it was largely predictable: Romani dialects in contact with a certain language L could be *a priori* expected to display more similarities with L than those not in contact with it. However, a more challenging question arises: if a Romani dialect D's primary contact language is L, is D more likely to display closer structural similarity with L or with Romani dialects that are not in contact with L? This question essentially operationalizes the central problem of our study, i.e., evaluating the relative weight of genealogical and areal factors in shaping valency class systems.

To answer this question, we used the same locus-based distance matrix as the starting point. For every Romani dialect, we calculated i) the distance to its primary contact language and ii) the average distance to Romani dialects that have a different primary contact language. The next step was to calculate the average values for groups of dialects sharing the same primary contact language. The resultant scatterplot is shown in Figure 6, where individual points represent groups of Romani dialects sharing the same contact language, the x-axis displays averaged distances to the contact language, and the y-axis represents averaged distances to Romani dialects with different contact languages.

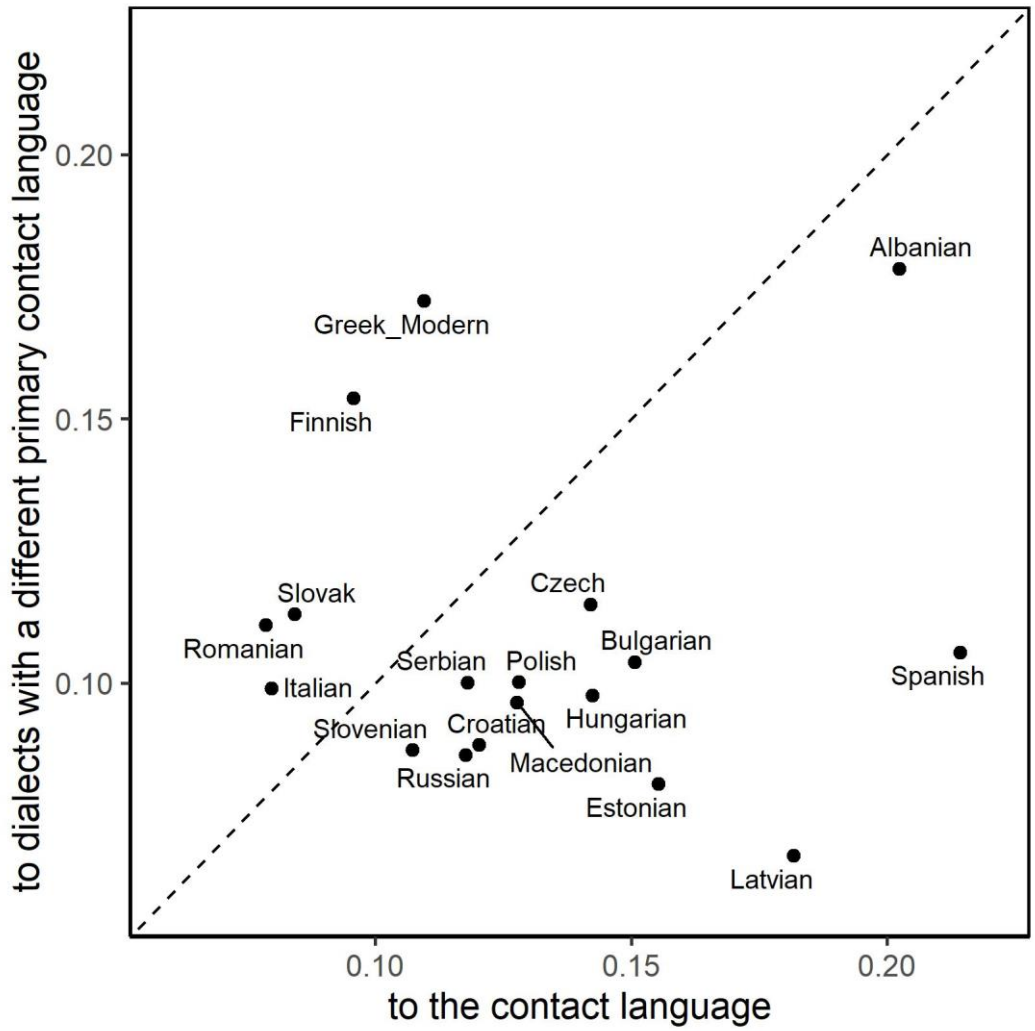


Figure 6. Average locus-based distances between Romani dialects and contact languages vs. dialects with a different primary contact language

A few generalizations emerge from the data shown in Figure 6. First, the two distances are very close to each other: most observations for *both* values fall in the range between 0.08 and 0.18. Distances between contact languages themselves (not shown in the visualization) sometimes fall outside this range, being either very small (e.g., 0.05 between Albanian and Romanian) or very large (e.g., 0.33 between Czech and Finnish). Substantially, this suggests that Romani dialects tend to be equidistant from both their contact language and remote

Romani dialects. This general conclusion is further corroborated by the fact that there are points both above and below the y = x line.

Points above the line (representing dialects in contact with Modern Greek, Finnish, Slovak, Romanian, and Italian) are particularly noteworthy. These dialects are closer to their primary contact languages than to other Romani dialects, indicating strong contact influence that these dialects must have undergone in their relatively recent history. These cases effectively demonstrate that, in the context of intense language contact, the usage of construction types defined in terms of locus within a particular variety can be drastically reshaped within a few centuries.

The points below the main diagonal in Figure 6 do not undermine this conclusion. When y-values being higher than x-values, language contact is the only plausible cause of observed similarities between Romani dialects and contact languages. However, there are many potential reasons why a certain dialect group can display x-values that are higher than y-values. One possibility is that these dialects have not been strongly affected by language contact and remained relatively conservative. There are further possibilities. For example, since y-values were obtained by averaging distances to all dialects with different primary contact languages, this method could have yielded lower values for dialects in dense areas with mutually related contact languages (such as Serbian, Croatian, and Slovenian). Another scenario is when some dialects undergo significant contact-induced impact from a language that is not treated as their "primary" contact language, see footnote 14 for a possible candidate, or Mexican Vlax (MX001), which has Spanish as its primary contact language, but is a recent migrant in Mexico (ca. 100–150 years ago) and had previously developed for centuries under Romanian influence. In short, the methodology used for calculating the values shown in Figure 6 is sufficient to demonstrate that, at least in some cases, relatively recent areal effects override inherited

similarities between dialects. However, it cannot prove the opposite claim, i.e., the observed similarities between dialects are due to inheritance.

4. Conclusion and discussion

Romani serves as a compelling example of how unstable valency patterns can be, particularly under the influence of contact languages. Over a time span of approximately 600 years since their split, Romani dialects developed striking variation in their valency patterns. This degree of variation is typologically unusual: when compared to transitivity prominence variation observed in older genealogical taxa such as Slavic, Romance, and Turkic languages, Romani dialects stand out for their exceptional diversity.

Where does this variation come from? Has it been shaped at some intermediate phylogenetic stages, or is it the result of prolonged evolvement in geographically distant areas under contact with other languages? Crevels & Bakker (2011) were not able to find any areal patterning or genealogical inheritance in the observed variation of external possession constructions across Romani dialects. We answer this question by comparing linguistic distances between Romani dialects on the basis of their valency patterns. Overall, geography and language contact are stronger predictors of dialect (dis)similarity than genealogy. However, in some cases, we observe a genealogical signal as well; for example, several North Vlax dialects spoken in different countries (Serbia, Russian, Mexico, Hungary) gravitate towards the Romani dialects spoken in Romania. The best explanation seems to be that it is exactly the historical development of these dialects in the Romanian-speaking territories that has shaped the current valency patterns systems in these dialects. It remains unclear how fast the reshaping of valency patterns can occur, as several factors play a role. The North Vlax example shows that the genealogical signal can still be detectable after 100–150 years of contact with a new language.

If we consider only contact-induced changes, there are two basic scenarios in which valency patterns can be reshaped. In the first scenario, a new verb is borrowed from a contact language, along with its original valency pattern, which is then calqued into the recipient language. In the second scenario, an inherited verb in the recipient language replicates the valency pattern of the corresponding verb in the source language (Grossman & Witzlack-Makarevich 2019). These two scenarios can be captured in terms of matter vs. pattern borrowing (Matras & Sakel 2007; Sakel 2007; Gardani 2020). Although the interplay between lexical and syntactic dimensions in Romani valency patterns are to be discussed elsewhere, our data clearly indicate that: i) the borrowing of verbs does not necessarily lead to a change in valency patterns, and, more importantly, ii) most of the variation we observe is the result of pattern borrowing unaccompanied by lexical borrowing.[15] In any case, we can affirm that in the situations of language contact, valency patterns can change rapidly, extensively, and relatively independent of matter borrowing. Specifically, we have demonstrated that, in some cases, Romani dialects become more similar to their contact languages in terms of valency patterns than to other Romani dialects.

Our main conclusion is that the similarities and dissimilarities in the valency class systems of Romani dialects are primarily shaped by language contact. There is little doubt that the calquing of lexically-specific valency patterns—whether or not accompanied by verb borrowing—drives the variation in valency patterns across Romani dialects, contributing to both dialect divergence and convergence.

Funding

[15] The evidence for these two claims is based on our observations of "pattern preservation rates", calculated as the likelihood of observing the same valency pattern for a given verb meaning across different dialects. We compared pattern preservation rates for i) pairs where both verbs involved are inherited and ii) pairs where one verb is inherited and the other is borrowed. For most verb meanings, the pattern preservation rates are approximately the same under both conditions, although for some verb meanings, the rates are slightly higher in the case of cognate inherited verbs. The relevant data and a scatterplot visualization are provided in the Supplementary materials.

Sergey Say's contribution to this study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 317633480 – SFB 1287.

Acknowledgements

We are grateful to Daria Alfimova, Aigul Zakirova, and other participants of the PoSla Typology Lab at the University of Potsdam, as well as three anonymous reviewers, for their valuable comments on an earlier version of this paper. All remaining errors are our own responsibility.

Abbreviations

1, 2, 3 – 1st, 2nd, 3rd person; ACC – accusative; ART– article; DAT – dative; DIR – direct; F – feminine; FUT – future; INS – instrumental; LOC – locative; M – masculine; NEG – negation; NOM – nominative; OBL – oblique; PL – plural; PN – person name; PRS – present tense; SBJ – subjunctive; SG – singular.

References

Apresjan, Jurij D. 1967. *Eksperimental'noe issledovanie semantiki russkogo glagola* [Experimental study of the Russian verb]. Moscow: Nauka.

Bickel, Balthasar, Taras Zakharko, Lennart Bierkandt & Alena Witzlack-Makarevich. 2014. An empirical assessment of semantic role types in non-default case assignment. *Studies in Language* 38(3). 485–511. https://doi.org/10.1075/sl.38.3.03bic

Blume, Kerstin. 1998. A contrastive analysis of interaction verbs with dative complements. *Linguistics* 36(2). 253–280. https://doi.org/10.1515/ling.1998.36.2.253

Boretzky, Norbert. 2007. The differentiation of the Romani dialects. *STUF - Language Typology and Universals* 60(4). 314–336. https://doi.org/10.1524/stuf.2007.60.4.314

Boretzky, Norbert & Birgit Igla. 2004. *Kommentierter Dialektatlas des Romani*. *T. 1–2*. Wiesbaden: Otto Harrassowitz.

Comrie, Bernard, Iren Hartmann, Martin Haspelmath, Andrej Malchukov & Søren Wichmann. 2015. Introduction. In Andrej Malchukov & Bernard Comrie (eds.), *Valency classes in the world's languages*. Vol. 1, 3–26. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110338812-004

Crevels, Mily & Peter Bakker. 2011. External possession in Romani. In Viktor Elšík & Yaron Matras (eds.), *Grammatical relations in Romani: The noun phrase*, 151–185. Amsterdam & Philadelphia: John Benjamins. https://doi.org/10.1075/cilt.211.09cre

Croft, William. 1993. Case marking and the semantics of mental verbs. In James Pustejovsky (ed.), *Semantics and the Lexicon*, 55–72. Dordrecht: Springer. https://doi.org/10.1007/978-94-011-1972-6_5

Croft, William. 1998. Event structure in argument linking. In Miriam Butt & Wilhelm Geuder (eds.), *The projection of arguments: Lexical and compositional factors*, 21–63. Stanford: CSLI Publications.

de Leeuw, Jan & Patrick Mair. 2009. Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software* 31(3). 1–30. https://doi.org/10.18637/jss.v031.i03

Dediu, Dan & Michael Cysouw. 2013. Some structural aspects of language are more stable than others: A comparison of seven methods. *PLoS One* 8(1). e55009. https://doi.org/10.1371/journal.pone.0055009

Dexter, Eric, Gretchen Rollwagen-Bollens & Stephen M. Bollens. 2018. The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling. *Limnology and Oceanography: Methods* 16. 434–443. https://doi.org/10.1002/lom3.10257

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67. 547–619. https://doi.org/10.2307/415037

Elšík, Viktor & Michael Beníšek. 2020. Romani dialectology. In Yaron Matras & Anton Tenser (eds.), *The Palgrave handbook of Romani language and linguistics*, 389–427. London: Palgrave Macmillan.

Fillmore, Charles J. 1968. The case for case. In Emmon Bach & Robert T. Harms (eds.), *Universals in linguistic theory*, 1–88. New York: Holt, Rinehart and Winston.

Galili, Tal. 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. *Bioinformatics* 31(22). 3718–3720. doi:10.1093/bioinformatics/btv428

Gardani, Francesco. 2020. Borrowing matter and pattern in morphology. An overview. *Morphology* 30. 263–282. https://doi.org/10.1007/s11525-020-09371-5

Gast, Volker & Maria Koptjevskaja-Tamm. 2022. Patterns of persistence and diffusibility in the European lexicon. *Linguistic Typology* 26(2). 403–438. https://doi.org/10.1515/lingty-2021-2086

Greenhill, Simon J., Chieh-Hsi Wu, Xia Hua, Micahel Dunn, Stephen C. Levinson & Russell D. Gray. 2017. Evolutionary dynamics of language systems. *Proceedings of the National Academy of Sciences of the USA* 114. E8822–E8829. https://doi.org/10.1073/pnas.1700388114

Grossman, Eitan & Alena Witzlack-Makarevich. 2019. Valency and transitivity in contact: An overview. *Journal of Language Contact* 12(1). 1–26. https://doi.org/10.1163/19552629-01201001

Hartman, Iren, Martin Haspelmath & Michael Cysouw. 2014. Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language* 38(3). 463–484. https://doi.org/10.1075/bct.88.02har

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Haspelmath, Martin. 2015. Transitivity prominence. In Andrej Malchukov & Bernard Comrie (eds.), *Valency classes in the world's languages*. Vol. 1, 131–147. Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110338812-008

Hawkins, John A. 1985. *A comparative typology of English and German: Unifying the contrasts*. Austin: University of Texas Press.

Hopper, Paul J. & Sandra A. Thompson. 1980. Transitivity in grammar and discourse. *Language* 56(2). 251–350.

Kassambara, Alboukadel. 2023. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.6.0. https://CRAN.R-project.org/package=ggpubr.

Kiparsky, Paul. 1998. Partitive case and aspect. In Miriam Butt & Willem Geuder (eds.), *The projection of arguments*, 265–307. Stanford: CSLI Publications.

Lazard, Gilbert. 1994. *L'actance*. Paris: Presses Universitaires de France.

Lazard, Gilbert. 2002. Transitivity revisited as an example of a more strict approach in typological research. *Folia Linguistica* 36(3–4). 141–190. https://doi.org/10.1515/flin.2002.36.3-4.141

Levin, Beth & Malka Rappaport Hovav. 2005. *Argument realization*. Cambridge: Cambridge University Press.

Malchukov, Andrej L. 2005. Case pattern splits, verb types and construction competition. In Mengistu Amberber & Helen de Hoop (eds.), *Competition and variation in natural languages: The case for case*, 73–117. Oxford: Elsevier.

Malchukov, Andrej. 2006. Transitivity parameters and transitivity alternations: constraining co-variation. In Leonid Kulikov, Andrej Malchukov & Peter de Swart (eds.), *Case, valency and transitivity*, 175–190. Amsterdam, Philadelphia: John Benjamins.

Malchukov, Andrej & Bernard Comrie (eds.). 2015. Valency classes in the world's languages. Vol. 1–2. Berlin, Boston: De Gruyter Mouton.

Malchukov, Andrej. 2015. Valency classes and alternations parameters of variation. In Andrej Malchukov & Bernard Comrie (eds.), *Valency classes in the world's languages. Vol. 1. Introducing the framework, and case studies from Africa and Eurasia*, 73–130. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110338812-007

Matras, Yaron. 2002. *Romani: A linguistic introduction*. Cambridge: Cambridge University Press.

Matras, Yaron. 2005. The classification of Romani dialects: A geographic-historical perspective. In Dieter W. Halwachs, Barbara Schrammel & Gerd Ambrosch (eds.), *General and applied Romani linguistics*, 7–26. Munich: Lincom Europa.

Matras, Yaron. 2020[2009]. *Language contact*. 2nd edn. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108333955

Matras, Yaron & Viktor Elšík. 2001–2016. *Romani Morpho-Syntax Database*. University of Manchester. Available at https://romani.dch.phil-fak.uni-koeln.de/ (last access 22 August 2023)

Matras, Yaron & Jeanette Sakel. 2007. Introduction. In Yaron Matras & Jeanette Sakel (eds.), *Grammatical borrowing in cross-linguistic perspective*, 1–13. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110199192.1

Matras, Yaron, Christopher White & Viktor Elšík. 2008. The Romani Morpho-Syntax (RMS) database. In Martin Everaert, Simon Musgrave & Alexis Dimitriadis (eds.), *The use of databases in cross-linguistic studies*, 329–362. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110198744.329

Matras, Yaron & Anton Tenser (eds.). 2020. *The Palgrave handbook of Romani language and linguistics*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-030-28105-2

Matras, Yaron, Márton A. Baló, Kirill Kozhanov, Daniele Viktor Leggio & Jakob Wiedner. 2022. Romani. In Lenore Grenoble, Pia Lane & Unn Røyneland (eds.), *Linguistic minorities*

in Europe online. Berlin & Boston: De Gruyter Mouton. https://doi.org/10.1515/lme.18104356.

Michaelis, Susanne Maria. 2019. World-wide comparative evidence for calquing of valency patterns in creoles. *Journal of Language Contact* 12 (1). 191–231. https://doi.org/10.1163/19552629-20190001

Murawaki, Yugo & Kenji Yamauchi. 2018. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution* 3(1). 13–25. https://doi.org/10.1093/jole/lzx022

Nichols, Johanna. 1975. Verbal semantics and sentence construction. *Proceedings of the First Annual Meeting of the Berkeley Linguistics Society*, 343–353. https://doi.org/10.3765/bls.v1i0.2355

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago & London: The University of Chicago Press.

Nichols, Johanna. 1995. Diachronically stable structural features. In Henning Andersen (ed.), *Historical Linguistics, 1993: Selected Papers from the 11th International Conference on Historical Linguistics*, 337–356. Amsterdam, Philadelphia: John Benjamins.

Oksanen J., Simpson G., Blanchet F., Kindt R., Legendre P., Minchin P., O'Hara R., Solymos P., Stevens M., Szoecs E., Wagner H., Barbour M., Bedward M., Bolker B., Borcard D., Carvalho G., Chirico M., De Caceres M., Durand S., Evangelista H., FitzJohn R., Friendly M., Furneaux B., Hannigan G., Hill M., Lahti L., McGlinn D., Ouellette M., Ribeiro Cunha E., Smith T., Stier A., Ter Braak C., Weedon J. 2022. _vegan: Community Ecology Package_. R package version 2.6-4.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Say, Sergey. 2014. Bivalent verb classes in the languages of Europe: A quantitative typological study. *Language Dynamics and Change* 4(1). 116–166. https://doi.org/10.1163/22105832-00401003

Say, Sergey. 2018. Markirovanie aktantov dvuxmestnyx predikatov: predvaritel'nye itogi tipologicheskogo issledovanija. In Sergey Say (ed.), *Valentnostnye klassy dvuxmestnyx predikatov v raznostrukturnyx jazykax*, 557–616. St. Petersburg: ILI RAN.

Say, Sergey. (ed.). 2020–. *BivalTyp: Typological Database of Bivalent Verbs and Their Encoding Frames*. St. Petersburg: Institute for Linguistic Studies, RAS. Available at https://www.bivaltyp.info (last access 22 August 2023).

Sakel, Jeanette. 2007. Types of loan: Matter and pattern. In Yaron Matras & Jeanette Sakel (eds.), *Grammatical borrowing in cross-linguistic perspective*, 15–29. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110199192.15

Seržant, Ilja A., Björn Wiemer, Eleni Bužarovska, Martina Ivanová, Maxim Makartsev, Stefan Savić, Dmitri Sitchinava, Karolína Skwarska & Mladen Uhlik. 2022. Areal and diachronic trends in argument flagging across Slavic. In Eystein Dahl (ed.), *Alignment and Alignment Change in the Indo-European Family*, 300–327. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198857907.003.0010

Skirgård, Hedvig et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Sci. Adv.* 9. eadg6175. DOI:10.1126/sciadv.adg6175

Trips, Carola. 2020. Copying of argument structure: A gap in borrowing scales and a new approach to model contact-induced change. In Bridget Drinka (ed.), *Historical Linguistics 2017. Selected papers from the 23rd International Conference on Historical Linguistics, San Antonio, Texas, 31 July – 4 August 2017*, 413–434. Amsterdam: John Benjamins. DOI:10.1075/cilt.350.19tri

Tsunoda, Tasaku. 2004. Issues in case-marking. In Peri Bhaskararao & Karumuri Venkata Subbarao (eds.), *Non-nominative subjects*. Vol. 2, 197–208. Amsterdam: Benjamins. https://doi.org/10.1075/tsl.61.11tsu

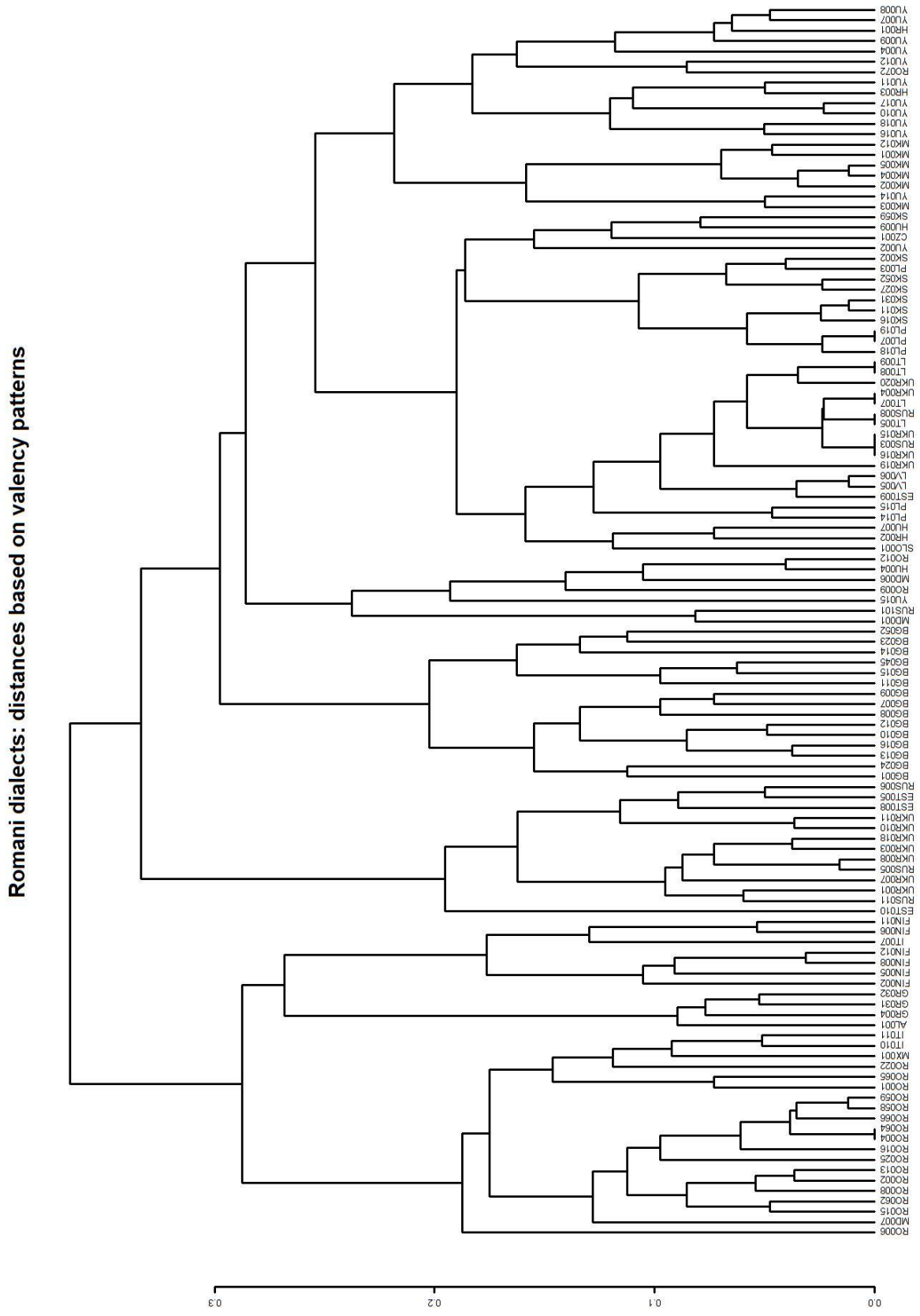
Van Belle, William & Willy van Langendonck (eds.). 1996. *The dative*. Amsterdam, Philadelphia: John Benjamins.

Wichmann, Søren & Eric W. Holman. 2009. *Temporal stability of linguistic typological features*. München: LINCOM Europa.

Wichmann, Søren. 2015. Diachronic stability and typology. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge handbook of historical linguistics*, 212–224. London: Routledge.

Witzlack-Makarevich, Alena. 2011. *Typological variation in grammatical relations*. Leipzig: University of Leipzig Ph.D. dissertation.

Appendix. Dendrogram of Romani dialects based on their valency patterns



Address for correspondence

Kirill Kozhanov

Institut für Slavistik,

Universität Potsdam

Am Neuen Palais 10, Haus 01, D-14469 Potsdam

Deutschland

kozhanov@uni-potsdam.de

https://orcid.org/0000-0003-3852-6617

Co-author information

Sergey Say

Universität Potsdam

Am Neuen Palais 10, Haus 01, D-14469 Potsdam

Deutschland

serjozhka@yahoo.com

https://orcid.org/0000-0001-8066-9166