Evolution and areal expansion of differential object marking in Romani

Kirill Kozhanov | ORCID: 0000-0003-3852-6617
Institute of Slavic Studies, Potsdam Typology Lab, Universität Potsdam,
Potsdam, Germany
kozhanov@uni-potsdam.de

Ilja A. Seržant | ORCID: 0000-0002-8066-9251
Institute of Slavic Studies, Potsdam Typology Lab, Universität Potsdam,
Potsdam, Germany
serzant@uni-potsdam.de

Abstract

This article traces the evolution of differential object marking (DOM) in Romani, an Indo-Aryan language primarily spoken in Europe in contact with different languages. Drawing on dialectological data from 119 locations in Europe, we demonstrate that DOM in Romani dialects is generally a stable feature constrained by the pronoun vs. noun distinction, animacy, and—in some conservative varieties—definiteness. By comparing Romani with other Indo-Aryan languages, we propose a diachronic scenario. An indirect object case marker expands into marking direct objects, starting with pronouns, then definite animate nouns, and finally, encompassing all animate nouns. This developmental sequence is reflected in the distribution of DOM-constraining factors in contemporary dialects. By contrast, Romani dialects of Finland are undergoing the process of losing DOM, while Italian varieties have lost it completely by forfeiting nominal case inflection altogether. However, in the varieties of southern Italy this loss is compensated by a new prepositional DOM, a pattern replication from dialectal Italian.

Keywords

1   Introduction

Romani is an Indo-Aryan language that has been spoken in Europe since the Middle Ages (Matras, 2002; Matras and Tenser, 2020). As a result of numerous migrations, Romani dialects are widespread across vast territories in Europe, where they are in contact with languages from different groups and families, including Indo-European (Greek, Slavic, Romance, etc.), Uralic (Hungarian, Finnish), and Turkic (Turkish). These dialects may be mutually unintelligible and are characterized by structural influences from contact languages (Elšík and Beníšek, 2020). In this paper, we explore the variation of the systems of differential object marking (henceforth DOM) across contemporary Romani dialects and propose a diachronic scenario for its evolution.

Similarly to other New Indo-Aryan languages, the Romani case system morphologically consists of two layers (Zograf, 1976; Masica, 1991). The first layer of case markers distinguishes between direct and oblique forms, which are cumulatively coded with number and gender. The second layer is agglutinatively attached to the oblique form, as illustrated in Table 1. Direct forms of Romani nouns may be zero-marked, as in *manúš* 'person,' or marked with an affix, as in *bakr-ó* 'sheep.' This opposition is solely determined by the noun's inflectional class. Direct forms are typically used as the subjects of intransitive (S) and transitive (A) verbs as well as certain (mostly inanimate) objects of transitive verbs (P; see details below). Oblique forms, on the other hand, are used for some (typically animate) objects of transitive verbs, as well as for the possessor in the predicative possession construction and for the recipient of the verb *del* 'give.' The secondary cases generally include ablative, dative, instrumental, and locative forms, with some variation across dialects. For an overview of nominal inflection in Romani, see Elšík (2000, 2020: 163–165).

All contemporary Romani dialects exhibit accusative alignment; that is, they do not differentiate between the sole argument of intransitive verbs and the first argument of transitive verbs but may mark the more patient-like argument of transitive verbs differently, as shown in (1) and (2):[1]

---

1 Examples follow the Leipzig Glossing Rules. Abbreviations used: 1/2/3 first/second/third person; ART article; CL clitic; COMP complementizer; DAT dative; DEF definite; DIR direct; DOM differential object marker; ERG ergative; F feminine; GEN genitive; IMP imperative; IMPF inperfect; INDF indefinite; INF infinitive; INS instrumental; M masculine; NEG negation; NOM nominative; OBL oblique; PL plural; PROG progressive; PRS present; PST past; REL relative; SBJ subjunctive; SG singular.

TABLE 1   Case marking in Romani (on the example of Kalderash Romani)

		manúš 'person'		*bakró* 'sheep'	
	Case	**SG**	**PL**	**SG**	**PL**
I	Direct	*manúš*	*manúš*	*bakr-ó*	*bakr-é*
	Oblique	*manuš-és*	*manuš-én*	*bakr-és*	*bakr-én*
II	Ablative	*manuš-és-tar*	*manuš-én-dar*	*bakr-és-tar*	*bakr-én-dar*
	Dative	*manuš-és-ke*	*manuš-én-ge*	*bakr-és-ke*	*bakr-én-ge*
	...				

Lithuanian Romani (Kirill Kozhanov's fieldwork notes)

(1) *Mir-í   phén   naš-él.*
     my-DIR.SG.F sister.DIR.SG run-PRS.3SG
     'My sister runs.'

(2) *Mir-í   phén   dźin-él   do ćhav-és.*
     my-DIR.SG.F sister.DIR.SG know-PRS.3SG that guy-OBL.SG
     'My sister knows that guy.'

Romani has completely lost the tense-split ergativity of Middle Indo-Aryan, with its only trace arguably found in the origin of certain pronominal forms, such as Romani *me* 'I' (direct form) < Vedic *máyā* (INS.SG, used as A in an ergative construction; see Bubeník, 2000, for a discussion of ergativity in Proto-Romani).

At its core, DOM in Romani consists of an alternation formally realized as a contrast between direct and oblique case forms. Romani DOM crucially depends on inherent lexical properties of the input, such as word class and animacy (Matras, 2002: 86–87; Adamou and Matras, 2020: 94; for a typological overview, see Witzlack-Makarevich and Seržant, 2018). Here, animate nouns, as in (2), and pronouns, seen in (3), are typically marked by the oblique case, thus distinguishing P from A/S, whereas inanimate nouns remain in the direct case (P = A/S)—that is, they are mostly unmarked—as seen in (4):

Lithuanian Romani (Kirill Kozhanov's fieldwork notes)

(3) *Mir-í   phén   dźin-él   lés.*
     my-DIR.SG.F sister.DIR.SG know-PRS.3SG 3SG.M.OBL
     'My sister knows him/it.'

(4) *Mir-í   phén   dźin-él   da **láv**.*
     my-DIR.SG.F sister.DIR.SG know-PRS.3SG this word.DIR.SG
     'My sister knows this word.'

Most descriptions of Romani varieties state that DOM is governed solely by
animacy (for example, Wentzel, 1980, on Russian Romani; Pobożniak, 1964:
2, on Lovari in Poland; Wagner, 2012: 87–88, on Lovari in the Czech Repub-
lic; Igla, 1996: 93, 117, on the Vlax variety of Ajia Varvara in Greece; Boretzky,
1993: 22, on Bugurdži in Kosovo; Valtonen, 1968: 161–162, and Belugin, 1977:
263, on Finnish Romani; Gjerdman and Ljungberg, 1963: 60, § 62, on Kalderash
Romani in Sweden). However, upon closer inspection, the situation is more
complex. For example, Valtonen (1968: 161–162) notes that in Finnish Romani,
some "hypercorrect formations" with oblique-marked inanimate nouns are
also found. Boretzky (1993: 22) observes that definiteness plays a role in Bugur-
dži Romani (Kosovo and North Macedonia) with nouns that refer to animals,
and that the oblique marking may occasionally be absent with nouns that refer
to humans. This is further supported by Matras (2002: 86–87), who states that
the DOM systems of certain Romani dialects are also sensitive to definiteness,
that is, that only definite animate nouns receive the oblique marking; this can
be seen in the opposition between (5) and (6):

Kalderash Romani (Kirill Kozhanov's fieldwork notes)
(5) *mudar-dé   khə   **bal-ó***
     kill-PST.3PL ART.INDF pig-DIR.SG
     'they killed a pig'

(6) *mudar-dé   le   **bal-és***
     kill-PST.3PL ART.DEF.OBL.SG.M pig-OBL.SG
     'they killed the pig'

This suggests the need for a more fine-grained study of the predictors of DOM
in Romani, one that takes dialectal variation into account.
    In the following sections, we explore the effects of various predictors of DOM
across Romani dialects. We then offer a tentative diachronic explanation for
the rise and expansion of DOM in Romani. Our paper is structured as follows.
In Section 2, we introduce the data and methodology of our study. In Section 3,
we establish the main predictors that affect case selection in the DOM systems
of contemporary Romani dialects: the distinction between personal pronouns
and nouns, animacy (including humanness), and definiteness. We then discuss
whether the distribution of relevant predictors of DOM across contemporary

Romani dialects corresponds to the geographical location of dialects or their genealogical classification (3.1). We furthermore examine four areal case studies: Romani dialects of Romania (3.2), Croatia (3.3), Finland (3.4), and Italy (3.5). In Section 4, we propose a scenario for the emergence and evolution of DOM in Romani. Finally, in Section 5, we present our conclusions and discuss our findings in the broader context of diachronic developments in DOM systems and language change.

2 Data and methodology

Data for this study comes from the Romani Morpho-Syntax (RMS) database, a database of Romani dialects developed at the University of Manchester. The online version of the database contains questionnaire-based dialect samples from 118 locations across different parts of Europe and one location in Mexico.[2] Each dialect sample consists of responses to the questionnaire from a single speaker. The questionnaire, compiled by Yaron Matras and Viktor Elšík, includes approximately 300 lexical questions and 700 sentences aimed at eliciting morphosyntactic information (Matras et al., 2009).

All dialect samples from the RMS database were used in this study. The distribution of the dialects and their dialectal affiliations, based on the recent classification by Elšík and Beníšek (2020), is illustrated in Fig. 1.[3]

In this study, we treat answers to the RMS dialectological questionnaire as a parallel corpus, comparing the frequency of oblique vs. direct case marking of direct objects in each sample.[4] The RMS questionnaire contains a total of 260 transitive clauses. Since we expect greater variation in the marking of animate nouns, we extracted all clauses with animate nominal Ps (58 instances). We then coded the P argument in these sentences for the following variables: (a) case marking (direct vs. oblique); (b) definiteness (definite vs. indefinite); (c) humanness (human vs. nonhuman, that is animal); (d) number (singular vs. plural); and (e) origin (borrowed vs. inherited). The final variable reflects the degree of morphological integration: nouns with morphological paradigms typical of inherited and pre-European loans (referred to as athematic or oico-

---

2 Originally, we used the database version located at https://romani.humanities.manchester.ac .uk/rms/, which is no longer available. The same data can be currently accessed at https:// romani.dch.phil-fak.uni-koeln.de/.

3 The map does not show the location in Mexico.

4 Our entire dataset as well as the R code can be accessed in the supplementary materials at https://doi.org/10.17605/OSF.IO/7SC8X.
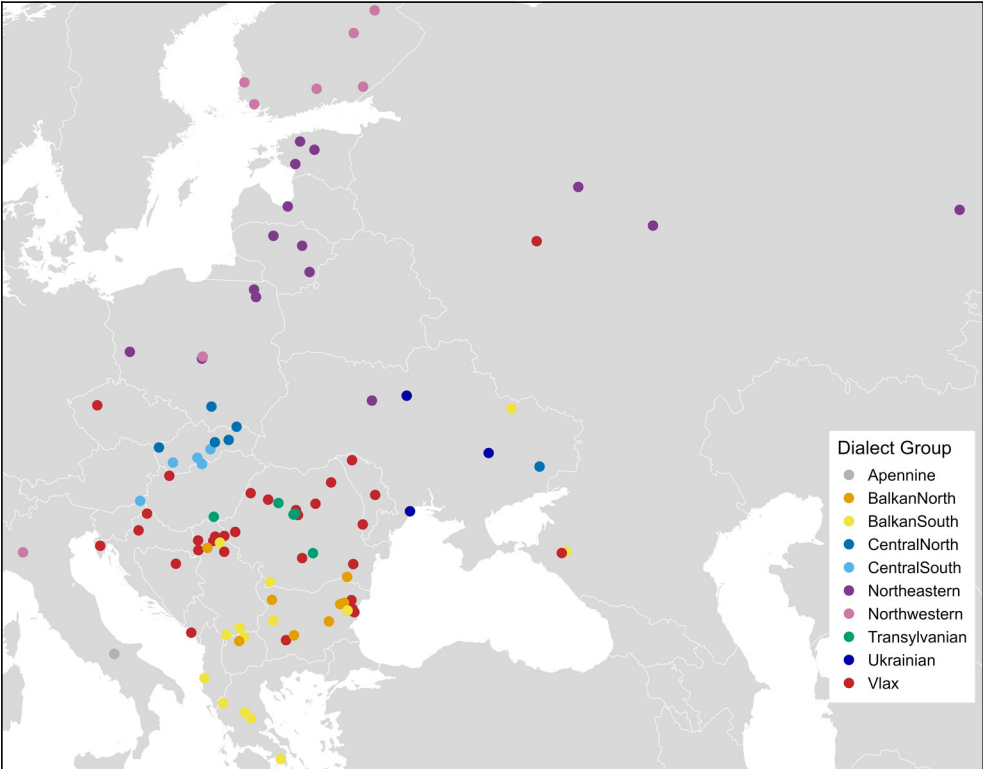
FIGURE 1   Geographical distribution of Romani dialects in the dataset, along with their genealogical
affiliation
Note: Maps in this study were created in R (R Core Team, 2024) using the packages *ggplot2*
(Wickham, 2016), *sf* (Pebesma and Bivand, 2023; Pebesma, 2018), *rnaturalearth* (Massicotte
and South, 2025), *rnaturalearthdata* (South, Michael and Massicotte, 2025), and *viridis* (Gar-
nier et al., 2024). The full reproducible code is available in Supplementary materials.

clitic morphology in Romani linguistics) are coded as inherited, while nouns
with inflection and stress patterns typical of European loans (referred to as
thematic or xenoclitic morphology) are tagged as borrowed (for details on this
distinction in Romani, see Matras, 2002: 20–25; Elšík, 2020: 168–170).

Table 2 illustrates the structure of our dataset for animate nominal direct
objects.

The first row in the table corresponds to the sentence *Kiss your sister!* (ID
number 368 in the questionnaire). The answer from a Romungro speaker in
Romania (ID RO-001), namely *čumide tra phenja*, contains the direct object
phenja, the oblique form of *phen* 'sister.' This word is animate, definite, singular,
human, and an inherited Indic noun. Its inherited status is determined by its
origin from Old Indo-Aryan *bhaginī* (F) 'sister' (Oslon and Kožanov, 2025) and

TABLE 2   Structure of the dataset (animate nouns)

Dialect ID	Sentence ID	Case marking	Definiteness	Humanness	Number	Origin
RO-001	368	oblique	definite	human	SG	inherited
UKR-001	584	oblique	definite	nonhuman	SG	inherited
MK-001	570	direct	indefinite	human	SG	inherited
MX-001	1001	direct	definite	nonhuman	SG	inherited

TABLE 3   Dataset on P marking in Romani

Type of P	No. of contexts	No. of varieties	No. of NA[a]	Total no. of examples
personal pronoun	7	119	53	833
inanimate noun	58	119	488	6902
animate noun	58	119	547	6902

[a]This column (NA) indicates how often the relevant element is missing in the questionnaire answers, either because of a skipped context or because a stimulus sentence is not fully translated.

its belonging to the paradigm of feminine nouns typical of inherited words. In this manner, we coded all instances of animate nominal direct objects in the RMS database.

As we do not expect variation in the marking of inanimate nominal Ps (which are always in the direct form) or Ps expressed by personal pronouns (which are always in the oblique form), we used only a subset of the available examples: 58 clauses with inanimate nominal Ps and 7 clauses with pronominal Ps. This part of the dataset was coded only for the case marking of Ps. The number of entries in the dataset is summarized in Table 3. Additional coding was also carried out for specific locations (see the discussion of Finnish Romani in Section 3.4).

Since the RMS database is not morphologically annotated, we manually tagged all examples. When possible, we consulted audio files to verify the transcriptions in the database. However, a disclaimer is warranted at this point. Questionnaire-based elicitation has inherent limitations, as the answers provided by speakers are translations of sentences without context. This limitation is especially prominent when dealing with semantic features such as definiteness or specificity. Some data samples were collected with a questionnaire mediated by languages that lack overt markers of, for example, definiteness

(e.g., most Slavic languages). In such cases, speakers might interpret the context differently (as definite or indefinite). Nevertheless, we expect that this shortcoming of the data does not alter the overall picture, which we present in the following section.

3   Predictors for the direct vs. oblique marking in DOM

Our data confirm that the two main predictors of oblique case marking on objects in Romani are the word class distinction between nouns and pronouns, and animacy; see Table 4.

All Romani dialects in the database consistently mark pronominal Ps with the oblique case as in (7), where *lén* is the oblique form, contrasting with the direct form *joné* 'they.' There are only a few exceptions to this; see (10) and (11) below.

Lithuanian Romani

(7) *jów*   **lén**   *na*   *dykh-já*
     3SG.M.DIR 3PL.OBL NEG see-PST.3SG
     'He didn't see them' (LT-005, 353g)

Some dialects have developed object clitics, typically originated from oblique forms of pronouns, as seen in (8), where the oblique form is double-marked by the clitic:

Laješa Romani (Moldova)

(8) *óv*   **mán**   *ní*   *dikh-ľá*   **ma**
     3SG.M.DIR 1SG.OBL NEG see-PST.3SG 1SG.CL
     'He didn't see me' (MD-001, 353a)

In total, our dataset contains 25 examples where a clitic is used in addition to an oblique form. All such examples are categorized under oblique marking in Table 3.

In some instances, other types of marking are observed for pronominal direct objects. For Arli, a South Balkan Romani dialect, Boretzky (1996) mentions the possibility of marking pronominal objects with the locative alongside the accusative. Our dataset did not contain such examples, but certain Balkan Romani varieties, for instance, employ third person object clitics, which may originate from oblique forms or represent remnants of archaic pronominal clitics (see Matras, 2004: 77–78, for further discussion). In (9), for instance, the

TABLE 4   P marking in the dataset

	Direct marking P=A/S	Oblique marking P≠A/S	Other marking P≠A/S	Total number of examples
personal pronouns	8 (1.0%)	759 (97.3%)	13 (1.7%)	780
inanimate nouns	6,407 (99.9%)	7 (0.1%)	0	6414
animate nouns	757 (11.9%)	5,566 (87.6%)	32 (0.6%)	6,355

object clitic *-os* '3SG.M,' although of pronominal origin, is not directly associated with the oblique form *olés-* of the pronoun *ová* 'he.'

Romacilikanes (Greece)

(9) *ój   daránd-il-i   kána dikh-ľás-os*
     3SG.F.DIR get.scared-PST-3SG.F when see-PST.3SG-3SG.M.CL
     'She became scared when she saw him' (GR-002, 394)

In nine examples, all from Finnish Romani, a dative form is used. This choice of marking is arguably due to partial language competence (see also Section 3.4). For instance, the form *tukke*, originally '2SG.DAT,' is used in the sample FIN-011 not only as a dative form but also as a general nominative form, as a direct object marker, and as a possessive pronoun 'your.' The third person clitic marking in Romacilikanes and the formally dative marking in Finnish Romani[5] are categorized under "other" marking in Table 3.

Regarding the eight examples of direct marking of personal pronouns mentioned in Table 3, three examples are in Finnish Romani (see Section 3.4 for further details on these dialects) and appear to be authentic, such as (10). In contrast, the remaining five examples, including (11), are attested in a single sample from Croatia (HR-001). The answers from this Croatian sample seem to be mistakes when translating the sentences word by word (other instances of pronominal Ps are consistently marked throughout the sample).

Finnish Romani

(10) *jou   na tikhe-tas   jou*
     3SG.M.DIR NEG see-PST.3SG 3SG.M.DIR
     'He didn't see him' (FIN-008, 353c)

---

5 There is also one example (353d) from the sample MD-007 that contains a dative-marked pronoun doubled by a clitic. We categorized it as "Other marking" in Table 3, but it appears to be more a slip of the tongue.

Gurbet Romani (Croatia)

(11) *óv   či   dikh-ľá   ój*
     3SG.M.DIR NEG see-PST.3SG 3SG.F.DIR
     'He didn't see her' (HR-001, 353d)

To conclude, personal pronouns in the P slot are nearly always marked by the oblique and so are consistently distinguished from A/S.

    Animacy is also a strong predictor of oblique marking in nouns. Animate nominal Ps receive oblique marking in most Romani dialects, accounting for 87.6% of all relevant cases in the dataset. Other marking of nominal direct objects (32 examples) is observed in Romani dialects of Italy, representing a contact-induced innovation (see Section 3.5). In contrast, inanimate Ps nearly always appear in the direct case (P = A/S), constituting more than 99% of all examples. Occasional instances of oblique marking in three dialects appear to be slips of the tongue. The only sample with a somewhat higher frequency of oblique marking for inanimate Ps (3 occurrences in FIN-002) comes from Finland. Earlier descriptions of Finnish Romani also report the use of the oblique case with inanimate nouns such as 'book,' which Valtonen (1968: 161, fn. 1) describes as a "hypercorrect formation." Compare (12), where the noun 'car' is used in the direct case, with (13), where the same noun appears in the oblique case.

Finnish Romani

(12) *lē   tukko   pēr-o   ta ja   rāt-es*
     take.IMP.2SG your(SG) car-DIR.SG and go.IMP.2SG drive-INF
     'Take your car and go for a ride!' (FIN-002, 1002)

(13) *jou   piha-tas   mān   piknav-es mo pēr-es   varo pereh*
     3SG.M.DIR say-PST.3SG 1SG.OBL sell-INF   my car-OBL.SG other year
     'he told me to sell my car next year' (FIN-002, 439)

The number of inanimate Ps marked by the oblique case in Finnish Romani is too small to allow for generalizations. However, a certain degree of lexicalization, that is, a tendency for specific nouns to appear in oblique form, may be at play.

    To identify additional predictors for the oblique case marking of animate nominal Ps in Romani, we fitted a mixed-effect logistic regression to the subset containing only animate noun Ps, excluding the two dialects of Italy with innovative marking. The model, fitted using the *lme4* package (Bates et al., 2015) in R, included dialect as a random effect and definiteness, number, humanness,
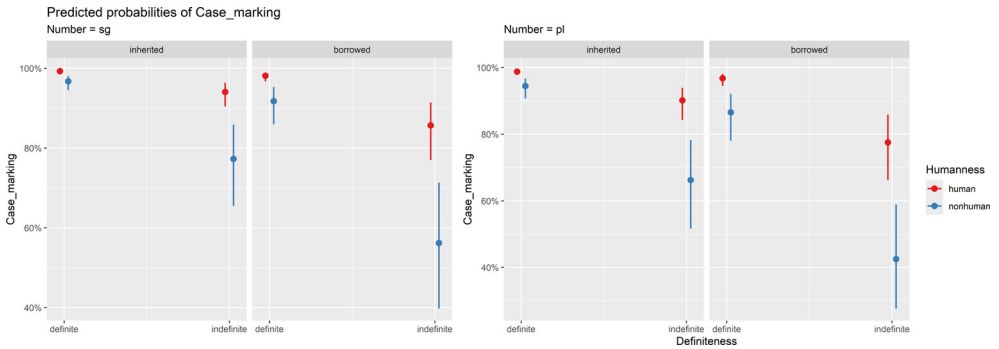
FIGURE 2   Predictors for the oblique case marking of animate nominal Ps in Romani dialects

and origin of the noun as fixed effects. The model's overall explanatory power is substantial (conditional R^2 = 0.686), though the part related to the fixed effects alone (marginal R^2) is of 0.096. The model fit parameters indicate high predictive power (concordance index C = 0.94; Somers' Dxy = 0.88).

The model's intercept, corresponding to definiteness = definite, humanness = human, origin = inherited, and number = singular, is 4.93 (95 % CI [4.35, 5.50], p < .001). The effect of definiteness [indefinite] is statistically significant and negative (β = –2.16, 95 % CI [–2.44, –1.89], p < .001). The effect of humanness [nonhuman] is statistically significant and negative (β = –1.54, 95 % CI [–1.83, –1.25], p < .001). The effect of origin [borrowed] is statistically significant and negative (β = –0.97, 95 % CI [–1.28, –0.67], p < .001). Finally, the effect of number [plural] is statistically significant and negative (β = –0.55, 95 % CI [–0.78, –0.31], p < .001). The effects of the model are visualized in Fig. 2.

The model indicates that indefinite nouns are less likely to receive oblique marking than definite ones, and nouns referring to nonhumans are less likely to be marked by the oblique than those referring to humans—these factors are well known crosslinguistically for conditioning DOM. Our data further show that plural nouns and borrowed lexemes are less likely to receive oblique marking than singular nouns and inherited lexemes, respectively. Singular nouns tend to be specific more often than plural ones, and perhaps it is precisely the effect of specificity—another factor that often plays a role in DOM systems— that we observe in our data. The nominative-accusative syncretism of nouns, particularly borrowed ones, in some Romani dialects results from the loss of shifting stress (cf. Elšík, 2000: 10–11; Elšík and Matras, 2006: 92–93). However, this alone is unlikely to explain why borrowed nouns in our data are less frequently marked. In fact, in dialects where a lower percentage of oblique-marked direct objects is attested, a mobile stress system is still retained (see, e.g., Section 3.3). This suggests that borrowed nouns exhibit a lower degree

of morphological integration, which may account for their lower likelihood of oblique marking.

To summarize the key predictors for P marking by the oblique case: personal pronouns are always marked while inanimate nouns are (nearly) never marked. Animate NPs, by contrast, are frequently marked, with definite animate NPs slightly more likely to receive oblique marking than indefinite animate NPs:

(14) Major predictors for oblique case marking in Romani DOM (effect decreasing from left to right)
      personal pronouns vs. nouns > animacy (> definiteness)

The distribution of DOM across Romani dialects correlates with geography. In the following sections, we examine this geographical factor in detail. First, we show how geography plays an important role in constraining frequency distributions and referential predictors of DOM (Section 3.1). Then, we focus on two geographical areas where oblique marking of animate nouns is less frequent than in the rest of the database (as shown in Fig. 3), namely Romania (Section 3.2) and Croatia (Section 3.3). Subsequently, we discuss two areas where the dialects are currently undergoing the demise of DOM, Finland (Section 3.4) and Italy (Section 3.5); some of the dialects spoken in the latter region are developing a new, preposition-based DOM.[6]

3.1   *Geographical distribution*

The frequency of oblique marking of animate nominal Ps is subject to cross-dialectal variation. Romani dialects in our dataset differ significantly in the ratio of animate Ps receiving oblique marking, as illustrated in Fig. 3. The ratio is calculated by dividing the number of oblique-marked Ps by the total number of analyzed animate nominal Ps—58 contexts per location, as explained in Section 2. Thus, the more frequently animate nominal Ps are marked with the oblique in a given Romani variety, the higher the ratio. Figure 4 shows the frequency of oblique marking across dialects in the dataset, as well as within groups defined by country and dialect classification.[7]

Figures 3 and 4 indicate that most dialects in Eastern (the Baltic states, Ukraine, Russia) and Central Europe (Hungary, Czechia, Slovakia, Poland)

---

6   We have excluded Mexico here, as the dataset contains only one variety from that country, and it is the result of recent migration (late nineteenth century) from Romania.

7   Note that two dialects in Molise (Italy) exhibit innovative prepositional marking which we did not consider for this map and following figures. They are discussed separately in Section 3.5.
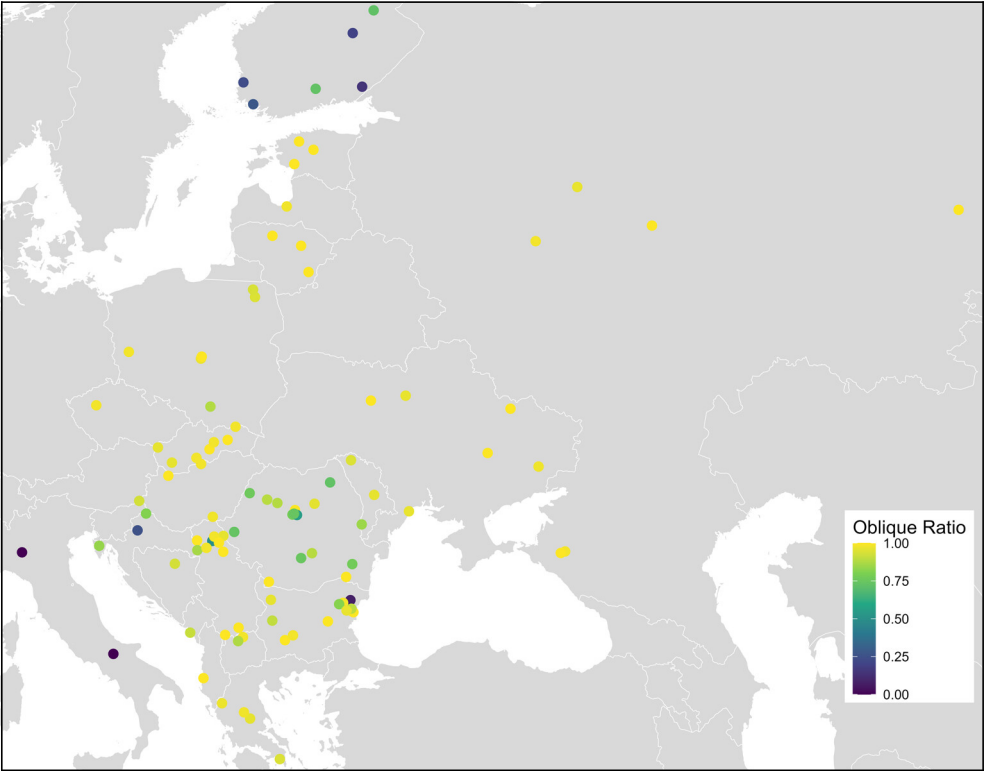
FIGURE 3   Ratio of oblique marking in animate nominal Ps across Romani dialects. A higher ratio corresponds to a lighter dot and indicates a higher frequency of oblique-marking

always mark animate nominal Ps with the oblique case (30% of the dialects in the dataset). The dialects spoken in the Balkans exhibit some variation (cf. North Macedonia, Serbia), but here too, animate nominal Ps predominantly receive oblique marking (particularly in Albania, Kosovo, and Greece). The Romani dialects of Romania demonstrate a somewhat lower degree of oblique marking, followed by the Romani dialects of Croatia and Finland (the latter two are discussed separately in Sections 3.3 and 3.4). The only Romani variety of Italy included in the subset for Figs. 3 and 4 contains no examples of oblique marking. We summarize this in (15):

(15) Frequency scale for the marking of animate nominal Ps in Romani dialects by geography (from no oblique marking on the left to total marking on the right of the scale)
Italy > Finland > Croatia > Romania > Balkans > Eastern and Central Europe
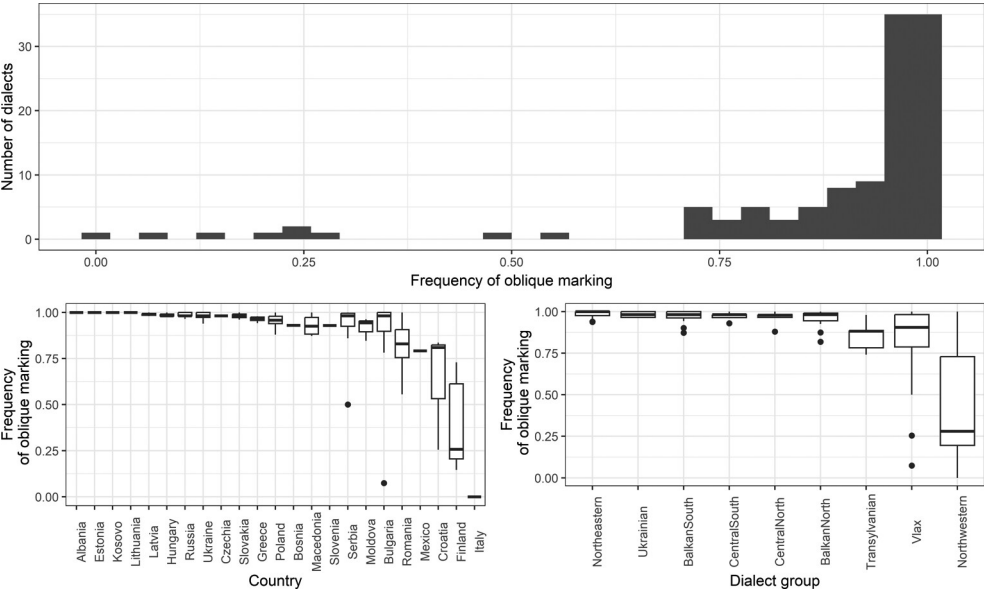
FIGURE 4   Frequency of oblique marking in animate nominal Ps across Romani varieties

In some cases, we observe outliers within dialects of the same region. For instance, among the 15 Romani dialects of Bulgaria, the majority consistently mark animate Ps with minimal variation, except for one dialect, namely BG-014, in which only 7% of animate Ps are marked with the oblique case. Unfortunately, we cannot account for such rare outliers without additional information about the sociolinguistic history of specific speakers.

We now turn to an intricate question: is the observed variation in the marking of animate nominal Ps in Romani dialects better explained by genealogy or geography? To address this question, we fitted several mixed-effect regression models, using the variable of case marking as a response and the variable of dialect being nested within country (i.e., the geographical location of the dialect) or dialect group (i.e., its genealogical classification) as a random effect. Note that we use the term "genealogical" here in a broad sense, referring to a cluster of dialects with shared features, as classified in the literature on Romani dialectology. Traditional classification of Romani dialects are typically based on sound changes, loanwords, and, to some extent, morphosyntactic phenomena. DOM, however, has not been considered in such classifications. We remain agnostic about the factors that originally shaped these clusters. Including dialect as a random effect significantly improves an intercept-only baseline model. Significance was assessed using the Akaike information criterion (AIC), an estimator of model fit; the AIC of the model with a random effect

was much lower than that of the intercept-only model. A model likelihood ratio test confirmed that the model with a random effect explained significantly more variance. This is not surprising, as case marking of direct objects varies considerably across dialects. When we extend the minimal model with dialect as a random effect by adding another random effect—country or dialect group—both models show significant improvement. A comparison of model fits reveals that the model with country as a random effect provides a slightly better fit than the one with dialect group (AIC 2829.5 vs. 2835.5; conditional R^2 = 0.679 vs. 0.593, respectively).[8] The difference between the two models is small, likely because the distribution of dialect groups correlates with geography. Nevertheless, the areal predictor appears to explain more variation along the cline in (15). This, in turn, suggests that the referential predictors of Romani DOM are not particularly stable and are not consistently retained within genealogical clades (i.e., dialect groups). Instead, these predictors are subject to areal and contact influences, regardless of genealogical affiliation. This finding aligns with the analysis of another phenomenon—valency patterns across Romani dialects—which shows a clear areal distribution (see Kozhanov and Say, 2025).

3.2   *DOM in the dialects of Romania*

The largest portion of our dataset comes from Romania (19 locations). Within this subset, only three samples mark all animate Ps with the oblique case; that is, in these samples other factors such as definiteness do not play a role. The remaining varieties demonstrate some degree of variation, with the ratio of the oblique-marked animate Ps ranging from 54% to 96%.

To establish the factors conditioning the marking of direct objects in these dialects, we created a subset of the 19 samples from Romania and fitted a logistic mixed model to predict the use of the oblique form. The sample size was n = 1,016, with 847 observations featuring oblique marking and 169 observations featuring direct marking. The model included dialect as a random effect. We used the *gmulti* package to identify the model with the lowest Bayesian information criterion (BIC) among all possible models with the predictors definiteness, humanness, origin, and number. The best-fitting model included the random effect and the independent variables definiteness and humanness. The model's total explanatory power is substantial (conditional R^2 = 0.69), and

---

8   Although we report only the results for the classification based on Elšík and Beníšek (2020), we also tested alternative dialect classifications, namely, one based on a traditional reference grid of eight groups (see Matras, 2005, for details) and another similar to Elšík and Beníšek (2020) but maintaining the distinction between Northern and Southern Vlax. Out of the three classifications, the first provided the best fit (see supplementary materials).
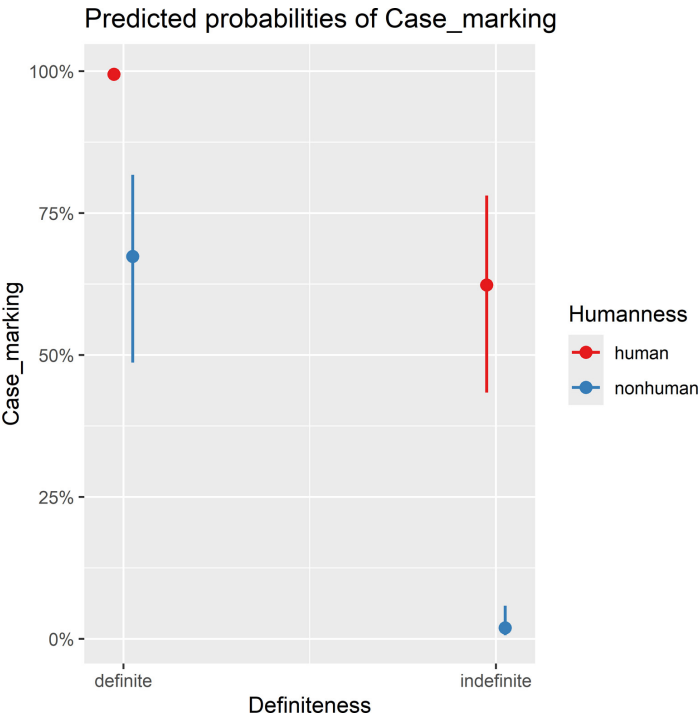
FIGURE 5   Predictors of oblique marking in the dialects of Romania

the part related to the fixed effects alone (marginal R^2) is of 0.47. The model fit parameters indicate high predictive power (concordance index C = 0.93; Somers' Dxy = 0.86).

The model's intercept, corresponding to definiteness = definite and humanness = human, is at 5.16 (95 % CI [4.11, 6.22], p < .001). The effect of definiteness [indefinite] is statistically significant and negative (β = –4.66, 95 % CI [–5.50, –3.82], p < .001). The effect of humanness [nonhuman] is also statistically significant and negative (β = –4.44, 95 % CI [–5.29, –3.60], p < .001). In other words, indefinite and nonhuman nouns have a significantly lower probability of being marked by the oblique compared to definite and human nouns.

The effects of the model based on the predicted values of oblique marking are visualized in Fig. 5. The figure shows that in the Romani dialects of Romania, the variables definiteness and humanness demonstrate significant association with oblique case marking of Ps. First, indefinite Ps, as in (16), receive oblique marking significantly less frequently than definite ones, as in (17). Second, definite Ps exhibit additional variation based on referent type: human referents, like in (17), are more likely to receive oblique marking than nonhuman referents, such as in (18).

Kaldaraš Romani (Romania)

(16) *dikh-lem* **murš-a**   *kaj beš-en-as*   *anglal o*
      see-PST.1SG man-DIR.PL REL sit-3PL-IMPF in.front ART.DEF
      magazin-o
      shop-DIR.SG
      'I saw men standing in front of the shop.' (RO-008, 573)

(17) *dikh-lem*   *sa kodo-le*   **murš-en**   *vi   arati*
      see-PST.1SG all that-OBL.PL man-OBL.PL and yesterday
      'I saw the same men there yesterday as well.' (RO-008, 574)

(18) *lesk-o*   *dad*   *kər-el*   *les*   *te*
      his-DIR.SG.M father.DIR.SG make-PRS.3SG 3SG.M.OBL COMP
      bikin-el   *o*   **grast**   *parpale*
      sell-SBJ.3SG ART.DEF horse.DIR.SG back
      'His father makes him sell the horse back.' (RO-008, 799)

The scale with the three predictors established above in (14) for all Romani dialects is calibrated for the dialects of Romania in (19):

(19) The main predictors for oblique case marking of Ps in Romani dialects of Romania (effect decreasing from left to right)
      personal pronouns > definite human nouns > definite animal nouns > indefinite animate nouns > inanimate nouns

Generally, DOM systems are sensitive to topic-like input types, that is, they are restricted by both definiteness and animacy simultaneously, at least in their earlier developmental stages (Iemmolo, 2010; Dalrymple and Nikolaeva, 2011). At some point in their development, animacy or definiteness (or specificity) may be lost as a predictor (e.g., in Slavic; see Eckhoff, 2015). We propose a similar scenario for Proto-Romani below (Section 4). Consequently, Romani dialects of Romania are the most conservative in retaining both predictors: animacy and definiteness. In contrast, Central and Eastern European dialects are more innovative in this respect, having generalized animacy as the sole predictor, possibly under the influence of neighboring Slavic languages.

3.3   *DOM in the dialects of Croatia*

Another subset of Romani dialects with a lower ratio of oblique marking comes from Croatia. The dataset includes responses to the questionnaire from three locations in Croatia. In two of them (HR-001, HR-003), approximately
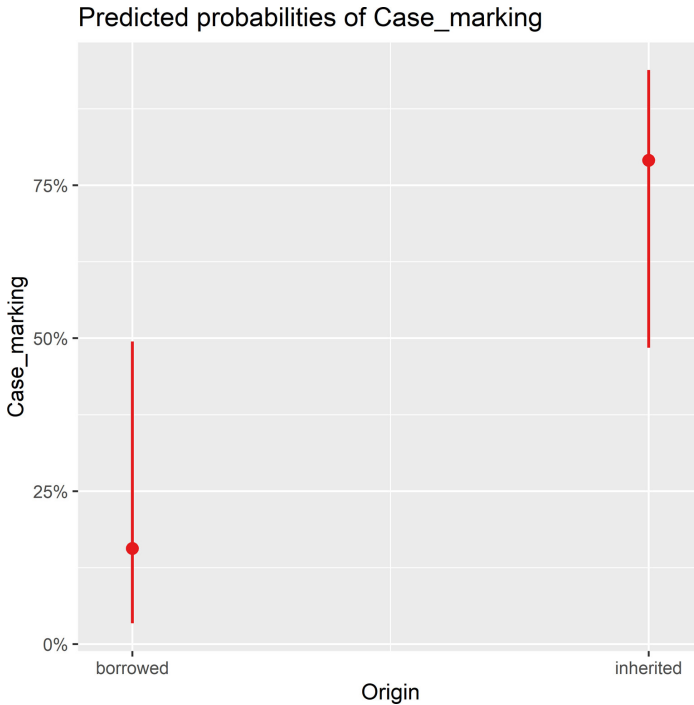
Predicted probabilities of Case_marking

FIGURE 6   Predictors of oblique marking in the dialects of Croatia

84% of all animate nouns are marked with the oblique case, whereas in HR-002, the ratio of oblique-marked animate nouns is significantly lower, at only 25%.

As in the previous analysis, for this subset, we fitted a logistic mixed model. The sample consisted of 168 observations, with 107 instances of oblique marking and 61 instances of direct marking. The best-fitting model included dialect as a random effect and origin as an independent variable. The model's total explanatory power is substantial (conditional R^2 = 0.47), while the part related to the fixed effects alone (marginal R^2) is 0.25. The model fit parameters indicate high predictive power (concordance index C = 0.86; Somers' Dxy = 0.73).

The model's intercept, corresponding to origin = borrowed, is –1.69 (95% CI [–3.36, –0.02], p = 0.047). The effect of origin [inherited] is statistically significant and positive (β = 3.02, 95% CI [1.87, 4.17], p < .001). The effects of the predictors are illustrated in Fig. 6.

The statistical analysis of all three dialects, as visualized in Fig. 6, demonstrates that most of the variation within animate nouns is determined by lexical choice. In contrast to inherited nouns, borrowed nouns typically do not inflect for the oblique case in these dialects; see examples (20) and (21) with the

borrowed noun *muškarco* 'man' (from Croatian *muškarac* 'man'). This paradigmatic restriction consequently reduces the influence of animacy as a predictor.

Gurbet Romani (Croatia)

(20) *trin* **muškarc-urja** *a-en*   *katar e*   *pijac-a*
     three man-DIR.PL   come-PRS.3PL from ART.DEF.F market-DIR.SG
     'Three men are walking home from the market.' (HR-001, 541)

(21) *me*   *rod-av*   *trin* **muškarc-urja** *e*   *vordon-enca*
     1SG.DIR look.for-PRS.1SG three man-DIR.PL   ART cart-INS.PL
     'I am looking for three young men with a cart.' (HR-001, 498)

3.4   *Loss of DOM in the dialects of Finland*

All six varieties from Finland in the dataset are in Finnish Romani, a highly endangered dialect of the Northwestern dialect group. Most speakers of this dialect belong to the older generation. While younger Finnish Roma may acquire the language later in adolescence, Finnish remains their primary language of communication (Borin, 2000: 75). The following data should thus be considered in the context of language shift and partial language competence.

    In Finnish Romani, animate nouns can receive oblique case marking, as in (22), but they often remain unmarked, as in (23). Bourgeois (1910–1911: 546) similarly notes that oblique marking of animate nouns is not consistently observed.

Finnish Romani

(22) *me*   *pinsivā*   *tola stār saijeha* **däija**
     1SG.DIR know.PRS.1SG that four girl.INS mother.OBL.SG
     'I know the mother of those four girls.' (FIN-005, 502)

(23) *startto tīves me*   *jeijom*   *tikjom*   *mango* **däi**
     fourth day 1SG.DIR go.PST.1SG see.PST.1SG my   mother.DIR.SG
     'On Friday I went to see my mother.' (FIN-005, 427)

Similar to the previous analysis, we also fitted a mixed logistic model for this subset. Even though all samples from Finland in the dataset represent the same dialect, the inclusion of dialect (which, in this case, corresponds to individual speakers) as a random effect improved the model. This suggests significant variation among speakers of this dialect.

    Since Finnish Romani sporadically marks inanimate nouns with the oblique (see Section 3.1), we tagged all inanimate nominal direct objects available for
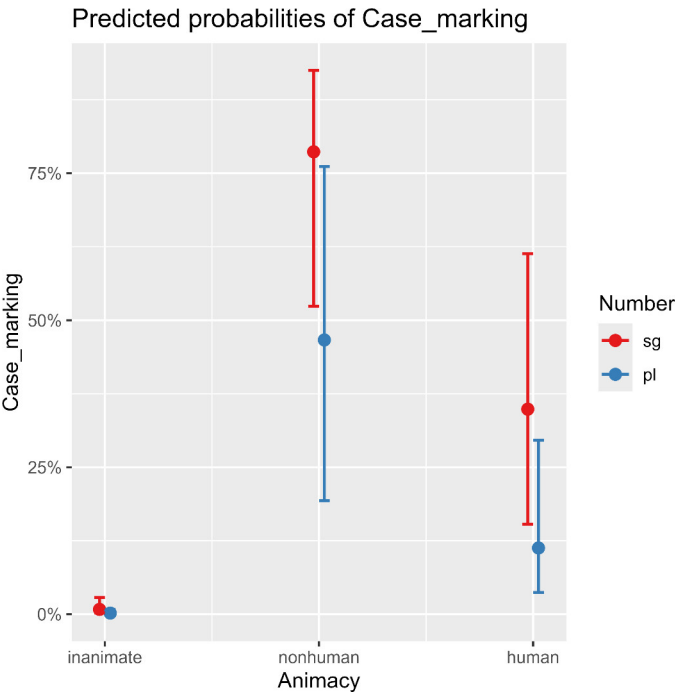
FIGURE 7　　Predictors of oblique marking in the dialects of Finland

that variety in the full RMS database. We thus created a new column "animacy", with the following values: inanimate, nonhuman, and human. The sample size was $n = 882$, with 115 observations exhibiting oblique marking and 767 observations exhibiting direct marking. The best-fitting model included the random effect dialect and the independent variables animacy and number. The model's total explanatory power is substantial (conditional $R^2 = 0.67$), while the part related to the fixed effects alone (marginal R^2) is 0.5. The model fit parameters indicate high predictive power (concordance index $C = 0.94$; Somers' Dxy = 0.88).

The model's intercept, corresponding to animacy = human and number = plural, is at −2.06 (95 % CI [−3.26, −0.87], $p < .001$). The effect of animacy [inanimate] is statistically significant and negative ($β = −4.16$, 95 % CI [−4.99, −3.34], $p < .001$). The effect of animacy [nonhuman] is statistically significant and positive ($β = 1.93$, 95 % CI [1.18, 2.68], $p < .001$). The effect of number [singular] is statistically significant and positive ($β= 1.44$, 95 % CI [0.75, 2.13], $p < .001$).

The effects of the predictors are visualized in Fig. 7. As observed in that figure, animacy remains the most significant predictor constraining DOM in Finnish Romani: only a handful of inanimate nouns receive oblique marking,

whereas the ratio of oblique-marked animate nouns is much higher. However, the marking of animate nouns varies across the responses to the questionnaire: in two response samples, 72%–73% of animate Ps are marked by the oblique, whereas in the remaining four samples, the percentage ranges between 15% and 28%. Somewhat counterintuitively, the model suggests that oblique marking is more likely to appear with nonhuman Ps, and among human referents, oblique marking occurs somewhat more frequently in the singular.

The overall low frequency of oblique marking, along with the higher ratio of nonhuman Ps receiving oblique marking, indicates a general tendency toward the loss of inflectional DOM in Finnish Romani (suggested previously in the literature by, inter alia, Vuorela and Borin, 1998: 70). Additionally, there are lexical biases in this process. For instance, Belugin (1977: 264) suggests that some verbs—whose objects regularly take the partitive case in Finnish—require the direct case in Finnish Romani as a way to adapt to the Finnish pattern.

Moreover, in our dataset, nonhuman referents appear in 11 clauses, and eight of them contain the lexeme 'horse.' A closer examination of the Finnish Romani data reveals that the oblique form of the lexeme 'horse' may have been reanalyzed as the default form—that is, the oblique form *gres* 'horse' is provided instead of the direct form *grai* in FIN-008 and FIN-011.

3.5   *Contact-induced emergence of a new DOM in the dialects of southern Italy*

Our dataset contains three questionnaire responses from Italy, representing two genealogically distinct dialects—Lombard Sinti and Abruzzian Romani. Lombard Sinti (IT-011), recorded in Piacenza, is related to Sinti dialects that developed in German-speaking territories and belongs to the Northwestern dialect group (Fig. 1 above). The two samples recorded in Molise (IT-007, IT-010) represent Abruzzian (or Apennine) Romani, a variety spoken by Roma populations that settled in Italy around the sixteenth century (Ascoli, 1865; Soravia, 1977; Scala, 2018). This variety is sometimes classified as a separate dialect group (Elšík and Beníšek, 2020: 401–402). Despite their distinct genealogical origins, both Lombard Sinti and Abruzzian Romani have undergone the loss of nominal case inflection, resulting in the complete absence of inflectional DOM in these dialects.

It remains unclear whether the loss of case inflection in these dialects is a result of language contact with Italian, which lacks nominal case inflection. Notably, a similar pattern is observed in Molise Slavic—genealogically a Croatian dialect spoken in Molise since the sixteenth century—which is gradually losing its old Slavic, inflectional DOM system (based on the alternation between the genitive and zero-marked nominative/accusative; Breu, 2017: 23). In Early

Slavic, as well as in modern Croatian, the genitive was obligatory for animate nouns in the object position. However, in Molise Slavic, this distinction has become optional, and animate nouns are not always distinguished from the zero-marked nominative by the genitive case (Breu, personal communication). Given this parallel development, it is likely that Italian has influenced the DOM systems of both Romani and Molise Slavic.

The loss of nominal case inflection is not the only reason why Romani varieties of Italy are relevant to our discussion. Of particular importance is the emergence of a new DOM system in Abruzzian Romani, which is based on the preposition *ki/ku* 'at,' as illustrated in examples (24) and (25). The direct object remains unmarked in (24), but is introduced by the preposition *ku* in (25).

Abruzzian Romani (Italy)

(24) *me   kamma   ni   grašt parn-o*
1SG.DIR want.PST.1SG ART.INDF.SG horse white-SG.M
'I want a white horse.' (IT-007, 1008)

(25) *u   dad   marja   **ku** grašt*
ART.DEF father beat.PST.3SG at horse
'His father killed the horse.' (IT-007, 776)

The new prepositional DOM was already present in the mid-nineteenth century, as evidenced by G.J. Ascoli's descriptions of Romani dialects in this area. This is illustrated in example (26), which Ascoli cites with some surprise, noting that the preposition appears "wo man den reinen accusativ erwarten könnte" ('where one might expect the simple accusative').

Abruzzian Romani (Italy)
(26) *kanán dekáv   k'o rom*
now   see.PRS.1SG at Rom
'now I see the Rom' (Ascoli, 1865: 140)

The emergence of prepositional DOM in Abruzzian Romani is undoubtedly a pattern replication from Southern Italian dialects, which employ the preposition *a* 'at, to' to mark animate direct objects (Rohlfs, 1969: 7; Iemmolo, 2010; Ledgeway et al., 2019; Ledgeway, 2022). Pattern replication refers to the reproduction of a certain structure in one language using functionally matching means of another language, a phenomenon well attested in the context of language contact (for an overview, see Matras, 2020: 254–264). In this case, the

Italian preposition *a* 'at, to' is replicated by the Romani preposition *ki* with similar spatial semantics.

Southern Italian

(27) *stamatina   vitti   a   Maria*
     this.morning see.PST.3SG DOM Maria
     'This morning he saw Maria' (Gioiosa Ionica; Ledgeway et al., 2019: 12)

For Abruzzian Romani, we examined all transitive clauses in the RMS database (see Section 2 for our initial dataset), including all examples with pronominal and nominal inanimate direct objects. We found no inanimate nominal direct objects marked by the preposition, indicating that animacy remains the primary predictor of prepositional DOM, just as it was for the old oblique-based DOM. In Abruzzian Romani, pronominal objects are typically expressed through verbal clitics, as in *dikátta* 'I see you (SG),' *dikkétta* 'he/she sees you,' and *dikkə́nata* 'they see you'; or *dikálla* 'I see him,' *dikkésala* 'you see him,' and *dikkə́nala* 'they see him' (Soravia, 1977: 88). Curiously, pronominal objects may sometimes be expressed both on the verb by clitics and by the locative-marked pronominal forms, as in *jov na dikjenč amende* 'he didn't see us' (IT-007, 353e), which is a calque of the Italian *a* marking. However, such examples are rare in our dataset (IT-007, 353a,e, 773, 958; IT-010, 353e), and further examples are needed to explain this phenomenon.

Similar to the previous analysis, we applied a mixed logistic model for this subset. However, differently from Finnish Romani, Molise Romani did not show a significant improvement of the model by including dialect as a random effect. Subsequently, we proceeded by fitting a fixed-effect logistic regression model, following a manual stepwise step-up procedure, that is, predictors were added to the model if they significantly improved model fit. The only significant predictor turned out to be definiteness. The sample size was n = 103, with 32 observations showing new prepositional marking and 71 observations showing direct marking. While the model clearly predicts the prepositional marking for definite animate nouns (coefficient = –2.5; 5% CI [–5.41, –0.85], p = 0.017), its explanatory power is low (Nagelkerke R^2 = 0.141), and model fit parameters are only moderate (concordance index C = 0.625; Somers' Dxy = 0.25). This is likely due to the small number of direct objects receiving prepositional marking, with the majority remaining unmarked. The effects of the model are visualized in Fig. 8.

It is noteworthy that in Italian, likewise, animacy plays the primary role in conditioning the presence of the DOM marker *a* 'to' (Rohlfs, 1969: 7). Further research is required to systematically compare data from Abruzzian Romani with Italian dialectal data.

Predicted probabilities of Case_marking
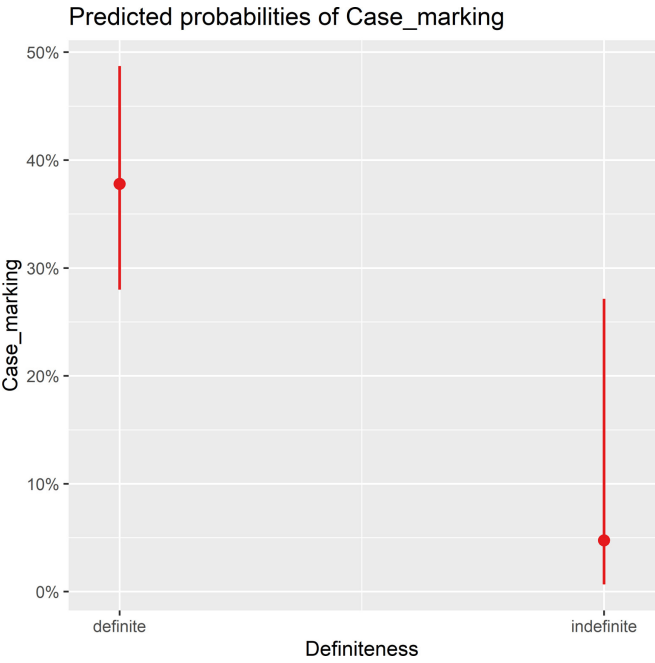
FIGURE 8   Predictors of prepositional marking in Abruzzian Romani

4   Historical development of the inflectional DOM of Romani

In what follows, we outline the historical development of DOM in Romani. We begin by discussing potential sources of DOM (Section 4.1) and then reconstruct its expansion across different input types (Section 4.2).

4.1   *Emergence of DOM*

The oblique affixes that Romani employs to mark Ps originate from Middle Indo-Aryan dative-genitive markers, as illustrated in Table 4 (Matras, 2002: 43; Beníšek, 2009). The dative and genitive were conflated into a single case already in Middle Prakrits (Bubeník, 1998: 65–66), and in Apabhraṃśa (a late Prakrit, timewise a transitional stage between Middle and Modern Indo-Aryan), these two cases are always marked the same (Tagare, 1948: 128–129). The morphological endings themselves stem from the Old Indo-Aryan genitive case (Beníšek, 2020: 30).

In Romani, the oblique case still retains some typical "dative" functions, such as marking the (non-internal) possessor in the existential predicative possession construction, as seen in (28), or the recipient of the verb 'give,' as in (29) (Matras, 2002: 85–87).

TABLE 5   Romani oblique endings and their sources in Middle and Old Indo-Aryan

Romani	Middle Indo-Aryan	Old Indo-Aryan
-es (M.SG)	DAT/GEN *-assa*	GEN.SG *-asya*
-a (F.SG)	DAT/GEN *-āya, -āa, -āe*	GEN.SG *-āyāḥ* or DAT.SG *-ā-yai*
-en (PL)	DAT/GEN *-āṇaṃ*	GEN.PL *-ānām*

Kalderash Romani (Kirill Kozhanov's fieldwork notes)

(28) *sý   **les**   dúj   śav-é*
       be.3.PRS 3SG.M.OBL two.DIR son-DIR.PL
       'He has two sons'

(29) *dé   **les**   lov-é*
       give.IMP.2SG 3SG.M.OBL money-DIR.PL
       'Give him some money!'

Note that Romani oblique forms of personal pronouns (see Table 5) are in part suppletive formations, originating from different Old Indo-Aryan pronominal stems. Some of these may be direct developments of Old Indo-Aryan genitive forms, whereas others may represent innovations based on nominal oblique endings (see Oslon and Kožanov, 2025).

   The oblique markers of the same origin as the Romani nominal ones exhibit a similar range of functions, including the P-marking, in the closely related language Domaaki, as well as in Kashmiri, another Northwestern Indo-Aryan language (Masica, 1991: 239–240; Weinreich, 2008: 302; Wali and Koul, 1997: 151, 155–156):

Domaaki (Weinreich, 2011: 175)

(30) *ma-s   paani   de*
       1SG-OBL.SG water.NOM.SG give.IMP.SG
       'Give me water!'

Kashmiri (Wali and Koul, 1997: 82)

(31) *ra:ma-s   chu maka:nɨ*
       Ram-DAT is   house
       'Ram has a house'

We suggest that it is the recipient function of the oblique case that serves as the source for its expansion from more topical input types (animate, definite)

TABLE 6   Romani personal pronouns in
             direct and oblique forms

	SG		PL	
	DIR	**OBL**	**DIR**	**OBL**
1	*me*	*man*	*ame*	*amen*
2	*tu*	*tut*	*tume*	*tumen*
3M	*ov*	*les*	*on(e)*	*len*
3F	*oj*	*la*		

to those that are less topical (indefinite and/or inanimate), thereby contribut-
ing to the evolution of DOM in Romani. The development of recipient markers
into DOM markers is not an infrequent phenomenon (Lehmann, 1995: 97; Heine
and Kuteva, 2002; König, 2008: 43; Dalrymple and Nikolaeva, 2011: 207). Recip-
ients are indeed semantically biased for topics, as they are usually animate
and definite, and often coded by personal pronouns (Bossong, 1991: 157; Ais-
sen, 2003: 447 fn. 10; Lazard, 2001: 875; Dalrymple and Nikolaeva, 2011: 207–211).
This gradual expansion of new marking can be described in terms of conven-
tionalization (or "grammaticalization" in Dalrymple and Nikolaeva, 2011: 18)
of the most frequent input types that topics take: topical referents are much
more frequently animate than inanimate because most of human conversation
is centered around living beings who can act; topical referents are also often
given and thus definite. The association between personal pronouns and topics
is even stronger, as personal pronouns are more frequent topics than animate
and definite nouns. Drawing on Dalrymple and Nikolaeva (2011), this develop-
ment is summarized in (32):

(32) Expansion scale for marking of topic-like NPs
      Secondary topics are marked > topic-like NPs (animacy/definiteness/
      specificity) are marked regardless of their information-structural role in
      the utterance

However, the development of recipient markers (i.e., "datives") into DOM mark-
ers has been questioned for some languages, such as Romance or Persian, in
view of historical evidence.[9] For other languages like English (with historically

---

9   The modern DOM marker -*rā(y)* appears in Middle Persian at the same time in the recipient

dative forms like *me, her, him,* etc.), alternative explanations have been proposed (Seržant, 2017). In English, for example, the explanation involves case syncretism resulting from various sound changes. This process began with specific nominal inflectional classes which are indistinguishable in terms of animacy or topic-likelihood, and which include many inanimate nouns. For Persian (Seržant, 2017) and Romance (Pensado, 1995; Seržant, 2017; Melis, 2021), the origin of DOM markers has been attributed to topicalization markers. Furthermore, the non-distinction of direct and indirect objects is the second most frequent strategy crosslinguistically (22% of the sample in Haspelmath, 2013), following languages with a dedicated recipient marker (49%). Therefore, the conflation and non-distinction of direct and indirect objects is generally a likely phenomenon which need not follow the specific pathway via topichood outlined in (32).

For Romani, this implies that the expansion of the oblique case can itself be independent from the referential patterns of recipients. Instead, it is more plausible that oblique marking expanded to Ps (details in Section 4.2 below), specifically targeting those NP types that were highly likely to be misinterpreted as As. Disambiguation is considered the primary driving force behind the expansion of DOM in Seržant (2019, in press). Alternatively, and relatedly, the expansion may have been driven by the role-reference association universal proposed by Haspelmath (2021a, 2021b), which constrained the expansion of the oblique case in Proto-Romani such that unexpected P types were marked by the oblique case.[10] The disambiguation-driven expansion of the genitive case onto direct objects, leading to DOM, is well attested in Slavic (Eckhoff, 2015). This development in Slavic is somewhat similar to the one found in Romani, as it is not an adposition that gives rise to DOM, but an old inflectional case going back to the Proto-Indo-European genitive case. Finally, a diachronic relation to topic marking would only seem to be motivated if the oblique marker itself could be shown to have emerged from a dedicated topic construction. We are unaware of any evidence in favor of topic-related functions of the oblique case in Romani.

---

and direct object functions, and the direct object function seems even to be more frequent than the recipient function (Seržant, 2017).

10   "Deviations from usual associations of role rank and referential prominence tend to be coded by longer grammatical forms if the coding is asymmetric" (Haspelmath, 2021b: 125). Note that in Romani, oblique markers are longer than direct makers for the majority of nouns. Only for some feminine nouns, specifically those with the direct marker *-i,* does the oblique marker have the same length.

Given the absence of diachronic data, it remains challenging to precisely determine the source construction, the pathway, and the main factors constraining the conflation of direct and indirect objects, and thus the expansion of the oblique case on direct objects in Romani.

With these observations in mind, we suggest the following scenario. Initially, Indo-Aryan languages showed an ergative-accusative split based on tense-aspect. This likely applied to Proto-Romani as well (Bubeník, 2000). However, over time, the ergative and the accusative clause types merged, resulting in neutral alignment, with some traces of the former ergative alignment. For instance, certain morphological forms of the direct case were historically ergative, for example, *me* '1SG.DIR' from '1SG.ERG'; compare *mae* '1SG.INS' in Apabhraṃśa (Bubeník, 2000: 211) and *mée* '1SG.ERG' in Domaaki. In Apabhraṃśa, all three macroroles (A, S, and P) were coded alike, that is, by the direct case (Bubeník, 1998: 65). The use of the oblique forms to mark Ps is a Middle Indo-Aryan innovation, which likely began in Proto-Romani at a time when it was still spoken among other Indo-Aryan languages. Given the comparable use of Middle Indo-Aryan genitive-dative markers found in Domaaki or Kashmiri, we suggest that this development may have been specific to Northwestern Indo-Aryan. This hypothesis is indirectly supported by phonetic features and arguably some morphological innovations that point to a Northwestern stage in the historical development of Romani (Turner, 1926; Beníšek, 2020: 23–24).[11]

In the following section, we reconstruct the expansion of the DOM system based on the alternation between direct (P = A/S) and oblique (P ≠ A/S) marking of P.

4.2   *Expansion of DOM*

We reconstruct the relative chronology of the expansion of DOM by comparing Romani evidence with data from other related Indo-Aryan languages that retain an oblique marker originating from Middle Indo-Aryan dative-genitive markers, namely Domari, Domaaki, and Kashmiri. This comparison does not assume a common protolanguage for these languages, but rather examines how a marker of the same origin is used across them.

We propose that DOM initially affected pronouns. In Romani, at least some oblique forms of personal pronouns originate from a different pronominal stem than the direct forms, though some may be derived with the nominal

---

11   The development of new DOM systems based on originally adpositional dative marking, such as the one found in Hindi, is unrelated to the processes discussed here and belongs to a more recent historical context. Montaut (2018) traces the development of DOM in modern Hindi back to the seventeenth century.

oblique markers. The special position of pronouns is also evident in Domaaki data, where only personal pronouns are regularly marked with the oblique case in the P role, as for example *m-as* '1SG-OBL' and *tu-s* '2SG-OBL' (Lorimer, 1939: 51). In Domari, demonstrative pronouns are marked with the oblique case in the P role, while personal pronouns occur as object indexes on the verb (Matras, 2012: 216, 219). This suggests that personal pronouns were likely the first type of NP affected by the expansion of the oblique into the domain of P in Proto-Romani as well.

In the following expansion of the oblique marking onto animate nouns in Proto-Romani, two predictors—animacy and definiteness—played an important role. In comparison, Lorimer (1939: 51–52) could identify just a handful of examples with oblique marking of nominal Ps (possibly only definite) in Domaaki, while Domari (as well as Kashmiri) marks only definite Ps with the oblique while indefinite Ps are left unmarked with respect to case and may contain the enclitic indefinite article -*ak* (Matras, 2012: 108, 113–115):

Domari (Matras, 2012: 113)
(33) *ama piyami   guld-**as***
     1SG   drink.1SG.PROG tea-**OBL.M**
     'I am drinking my tea' (with situational reference to a particular cup of tea)

(34) *ama inmangame'   piyam   gulda*
     1SG   NEG.ask.1SG.NEG drink.1SG.SBJ tea
     'I don't wish to drink any tea'

As demonstrated above, in some Romani varieties, definiteness still serves as a (weaker) predictor alongside animacy. Therefore, we suggest that during the second stage, the oblique marking expanded onto definite animate (human) nouns, in addition to the oblique marking of personal pronouns in the first stage. This is how DOM was likely constrained in Late Proto-Romani, when it was spoken in the Byzantine Empire prior to the fifteenth century.[12]

In the next stage, the oblique marking extended further to animate nouns across various Romani dialects, resulting in a generalization of DOM for all animate nouns, both definite and indefinite, in most dialects in Central and Eastern Europe; see (15) above. For convenience, we repeat the distribution of

---

12   It is noteworthy that the most important contact language of Romani at the time, namely Medieval Greek, did not have DOM (Holton et al., 2019).

TABLE 7   Diachronic development of DOM in Romani in idealized diachronic steps

Step 1	Proto-Romani	DOM on pronouns only
Step 2	Proto-Romani	DOM on pronouns and usually on definite human nouns
Step 3	Late Proto-Romani	DOM on pronouns and usually on definite animate nouns and sometimes on indefinite animate nouns
Step 4	most Romani dialects in Eastern and Central Europe	DOM on pronouns and animate nouns (all, definite and indefinite)
Step 5a	Abruzzian Romani	loss of nominal DOM (development of a new prepositional DOM)
Step 5b	Finnish Romani	gradual loss of DOM

dialects on this diachronic scale in (35), excluding the dialects of Finland and Italy where the demise of animacy constraint was historically secondary (see Sections 3.4 and 3.5 above).

(35) Expansion scale along animacy preference across Romani dialects by geography (from less preferred oblique marking on the left to total marking on the right)
Romania > Balkans (including Croatia) > Eastern and Central Europe

Definiteness is no longer a predictor in most dialects of Eastern and Central Europe.

Finally, in Section 3.5 we have discussed the complete loss of nominal inflection DOM and the development of new prepositional DOM in Abruzzian Romani recorded in Molise in Italy.

Table 7 summarizes the diachronic developmental steps.

5   Discussion and conclusions

In this paper, we analyzed the areal distribution of the two main referential predictors—animacy (the strongest predictor) and definiteness—that constrain DOM in Romani dialects.

The more conservative dialects primarily mark animate nouns that are definite, reflecting the original state of affairs. We argued that these predictors are better preserved in the dialects of Romania (Section 3.2). In contrast, most

dialects of Eastern and Central Europe expanded DOM along the definiteness scale, extending it to indefinite animate nouns. Consequently, definiteness is no longer a predictor in these dialects, since all animate Ps are marked. The most innovative dialects are Romani varieties in Italy, which have completely lost the inflectional nominal DOM. Finnish Romani dialects are currently undergoing a similar loss (Section 3.4). Abruzzian Romani in Italy developed a new prepositional DOM replicating the South Italian pattern (Section 3.5).

We further argued that the distribution of marked animate nominal Ps across Romani dialects in our dataset is not random, but can be explained by their geography or their dialect classification. Our analysis shows that the geographical predictor (proxied by the country of the dialect) explains more variation in the proportion of marked animate nouns than belonging to a certain dialect group. The preponderance of the geographical predictor over the genealogical classification indicates that the inheritance of referential predictors—and in the case of Italian and Finnish dialects, of the entire phenomenon of DOM—is not stable. Instead, inheritance is overridden by local areal pressures, which is why the referential predictors and DOM are not faithfully retained within the (sub)clades. This suggests that the observed variation in the marking of animate nominal Ps is relatively recent, developing after the initial dialect split in the fifteenth century. Indeed, it has been suggested that DOM systems based on semantic-pragmatic properties, like other similar grammatical patterns, can be unstable (Sinnemäki, 2014: 300), and thus more susceptible to the effects of language contact (Mardale and Karatsareas, 2020). At the same time, the Romani data demonstrates that over the course of the 600–700 years since the initial dialect split, most Romani dialects, regardless of their contact languages, have preserved a DOM system characterized by pronoun vs. noun and animate vs. inanimate distinctions. In other words, the mere existence of a DOM system based on referential properties can itself be stable.

Finally, we proposed a diachronic scenario for the evolution of DOM in Romani, considering diachronic typological evidence (Iemmolo, 2010; Dalrymple and Nikolaeva, 2011) and comparing it with genealogically related Indo-Aryan languages like Domaaki, Domari, and Kashmiri, which also retain case markers of the same origin (Section 4). We suggest that DOM initially applied only to pronouns and later expanded to definite nouns referring to animates or humans (Table 6). While the lack of historical records prevents a precise reconstruction of the details, we argue that the conflation of the P and recipient roles is a typologically frequent development. Additionally, we propose that the expansion of the oblique case onto the P role, conditioned by animacy and definiteness, was likely driven by disambiguation (Seržant, 2019, in

press) and/or by the role-reference association universal proposed by Haspelmath (2021a, 2021b). Our diachronic scenario summarized in Table 6 aligns with this explanation but does not entirely exclude the alternatives discussed in Section 4.1.

Acknowledgments

References

Adamou, Evangelia and Yaron Matras. 2020. Romani syntactic typology. In Yaron Matras and Anton Tenser (eds.), *The Palgrave Handbook of Romani Language and Linguistics*, 187–227. London: Palgrave Macmillan.

Aissen, Judith. 2003. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory* 213: 435–448.

Ascoli, G.J. 1865. *Zigeunerisches*. Halle: Heynemann.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Belugin, A.D. 1977. Sopostavitel'nyj analiz padežnyx sistem dialekta finskix cygan i finskogo jazyka. *Sovetskoe finno-ugrovedenie* 13: 261–271.

Beníšek, Michael. 2009. Middle Indo-Aryan ablative and locative markers in Romani. *Indo-Iranian Journal* 52: 335–362. https://doi.org/10.1163/001972409X445951.

Beníšek, Michael. 2020. The historical origins of Romani. In Yaron Matras and Anton Tenser (eds.), *The Palgrave Handbook of Romani Language and Linguistics*, 13–47. London: Palgrave Macmillan.

Boretzky, Norbert. 1993. *Bugurdži: Deskriptiver und historischer Abriß eines Romani-Dialekts*. Wiesbaden: Harrassowitz.

Boretzky, Norbert. 1996. Arli: Materialien zu einem südbalkanischen Romani-Dialekt. *Grazer Linguistische Studien* 46: 1–31.

Borin, Lars. 2000. A corpus of written Finnish Romani texts. In Donncha Ó Cróinín (ed.), *LREC 2000, Workshop Proceedings: Developing language resources for minority languages: Reusability and strategic priorities*, 75–82. Athens: ELRA.

Bossong, Georg. 1991. Differential object marking in Romance and beyond. In Dieter

Wanner and Douglas A. Kibbee (eds.), *New Analyses in Romance Linguistics*, 143–167. Amsterdam: John Benjamins.

Bourgeois, Henri. 1910–1911. Esquisse d'une grammaire du Romani finlandais. *Atti della Reale Accademia delle Scienze di Torino* 46: 541–554.

Breu, Walter. 2017. Einleitung. In Walter Breu (ed.), *Slavische Mikrosprachen im absoluten Sprachkontakt: Glossierte und interpretierte Sprachaufnahmen aus Italien, Deutschland, Österreich und Griechenland*, part 1, *Moliseslavische Texte aus Acquaviva Collecroce, Montemitro und San Felice del Molise*, 11–72. Wiesbaden: Harrassowitz.

Bubeník, Vít. 1998. *A Historical Syntax of Late Middle Indo-Aryan (Apabhraṃśa)*. Amsterdam: John Benjamins.

Bubeník, Vít. 2000. Was Proto-Romani an ergative language? In Viktor Elšík and Yaron Matras (eds.), *Grammatical Relations in Romani*: *The Noun Phrase*, 205–227. Amsterdam: Benjamins.

Dalrymple, Mary and Irina Nikolaeva. 2011. *Objects and Information Structure*. Cambridge: Cambridge University Press.

Eckhoff, Hanne M. 2015. Animacy and differential object marking in Old Church Slavonic. *Russian Linguistics* 39: 233–254.

Elšík, Viktor. 2000. Romani nominal paradigms: Their structure, diversity and development. In Viktor Elšík and Yaron Matras (eds.), *Grammatical Relations in Romani: The Noun Phrase*, 9–30. Amsterdam: John Benjamins.

Elšík, Viktor. 2020. Romani morphology. In Yaron Matras and Anton Tenser (eds.), *The Palgrave Handbook of Romani Language and Linguistics*, 155–186. London: Palgrave Macmillan.

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

Elšík, Viktor and Yaron Matras. 2006. *Markedness and Language Change: The Romani Sample*. Berlin: Mouton de Gruyter.

Garnier, Simon, Noam Ross, Robert Rudis, Pedro A. Camargo, Marco Sciaini, and Cédric Scherer. 2024. viridis(Lite)—Colorblind-Friendly Color Maps for R. https://doi.org/10.5281/zenodo.4679423, viridis package version 0.6.5, https://sjmgarnier.github.io/viridis/.

Gjerdman, Olof and Erik Ljungberg. 1963. *The Language of the Swedish Coppersmith Gipsy, Johan Dimitri Taikon: Grammar, Texts, Vocabulary, and English Word-Index*. Uppsala: A.-B. Lundequistska Bokhandeln & Copenhagen: Ejnar Munksgaard.

Haspelmath, Martin. 2013. Ditransitive constructions: The verb 'give.' In Matthew S. Dryer and Martin Haspelmath (eds.), WALS *Online* (v2020.3) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (accessed November 22, 2023).

Haspelmath, Martin. 2021a. Explaining grammatical coding asymmetries: Form-

frequency correspondences and predictability. *Journal of Linguistics* 57(3): 605–633. https://doi.org/10.1017/S0022226720000535.

Haspelmath, Martin. 2021b. Role-reference associations and the explanation of argument coding splits. *Linguistics* 59(1): 123–174. https://doi.org/10.1515/ling-2020-0252.

Heine, Bernd and Tania Kuteva. 2002. *World Lexicon of Grammaticalization*. Cambridge: Cambridge University Press.

Holton, David, Geoffrey Horrocks, Marjolijne Janssen, Tina Lendari, Io Manolessou, and Notis Toufexis. 2019. *The Cambridge Grammar of Medieval and Early Modern Greek*, vol. 4, *Syntax*. Cambridge: Cambridge University Press.

Iemmolo, Giorgio. 2010. Topicality and differential object marking: Evidence from Romance and beyond. *Studies in Language* 34(2): 239–272. https://doi.org/10.1075/sl.34.2.01iem.

Igla, Birgit. 1996. *Das Romani von Ajia Varvara: Deskriptive und historisch-vergleichende Darstellung eines Zigeunerdialekts*. Wiesbaden: Otto Harrassowitz.

Kozhanov, Kirill and Sergey Say. 2025. Variation in valency patterns across Romani dialects is primarily shaped by contact languages. *Studies in Language*. Advance online publication. https://doi.org/10.1075/sl.23067.koz.

König, Christa. 2008. *Case in Africa*. Oxford: Oxford University Press.

Lazard, Gilbert. 2001. Le marquage différentiel de l'objet. In Martin Haspelmath, Ekkehard König, Wulf Österreicher and Wolfgang Raible (eds.), *Language Universals and Language Typology: An International Handbook*, 873–885. Berlin: Mouton de Gruyter.

Ledgeway, Adam. 2022. Differential object marking in the dialects of southern Italy. *Caplletra* 74: 183–212. https://doi.org/10.7203/Caplletra.74.22958.

Ledgeway, Adam, Norma Schifano and Giuseppina Silvestri. 2019. Differential object marking and the properties of D in the dialects of the extreme south of Italy. *Glossa: A Journal of General Linguistics* 4(1). https://doi.org/https://doi.org/10.5334/gjgl.569.

Lehmann, Christian. 1995. *Thoughts on grammaticalization*. Unterschleissheim: Lincom Europa.

Lorimer, David L.R. 1939. *The Ḍumāki Language: Outlines of the Speech of the Ḍoma, of Bērīcho, of Hunza*. Nijmegen: Dekker & van de Vegt.

Mardale, Alexandru and Petros Karatsareas. 2020. Differential object marking and language contact. *Journal of Language Contact* 13(1). https://doi.org/10.1163/19552629-bja10001.

Masica, Colin P. 1991. *Indo-Aryan Languages*. Cambridge: Cambridge University Press.

Massicotte, Philippe and Andy South. 2025. rnaturalearth: World Map Data from Natural Earth. R package version 1.1.0, https://docs.ropensci.org/rnaturalearth/.

Matras, Yaron. 2002. *Romani: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Matras, Yaron. 2004. Romacilikanes: The Romani dialect of Parakalamos. *Romani Studies* 14(1): 59–109.

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

Matras, Yaron. 2012. *A Grammar of Domari*. Berlin: De Gruyter Mouton.

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

Matras, Yaron and Anton Tenser (eds.). 2020. *The Palgrave Handbook of Romani Language and Linguistics*. Cham: Palgrave Macmillan.

Matras, Yaron, Christopher White, and Viktor Elšík. 2009. The Romani Morpho-Syntax (RMS) database. In Martin Everaert, Simon Musgrave, and Alexis Dimitriadis (eds.), *The Use of Databases in Cross-Linguistic Studies*, 329–362. Berlin: Mouton de Gruyter.

Melis, Chantal. 2021. From topic marking to definite object marking: Focusing on the beginnings of Spanish DOM. In Johannes Kabatek, Philipp Obrist, and Albert Wall (eds.), *Differential Object Marking in Romance: The Third Wave*, 39–64. Berlin: De Gruyter.

Montaut, Annie. 2018. The rise of differential object marking in Hindi and related languages. In Ilja A. Seržant and Alena Witzlack-Makarevich (eds.), *Diachrony of Differential Argument Marking*, 281–313. Berlin: Language Science Press.

Oslon, Mixail and Kirill Kožanov. 2025. Ėtimologičeskij slovar' cyganskogo jazyka (obščecyganskaja leksika). 16 May 2025. https://rromanes.org/pub/%D0%AD%D0%A1%D0%A6%D0%AF.pdf (accessed May 27, 2025).

Pebesma, Edzer and Roger Bivand. 2023. Spatial Data Science: With applications in R. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016, https://r-spatial.org/book/.

Pebesma, Edzer. 2018. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* 10(1): 439–446. https://doi.org/10.32614/RJ-2018-009.

Pensado, Carmen. 1995. *La creación del complemento directo preposicional y la flexión de los pronombres personales en las lenguas románicas*. In Carmen Pensado (ed.), *El complemento directo preposicional*, 179–233. Madrid: Visor.

Pobożniak, Tadeusz. 1964. *Grammar of the Lovari Dialect*. Kraków: Panstwowe Wydawnictwo Naukowe.

R Core Team. 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rohlfs, Gerhard. 1969. *Grammatica storica della lingua italiana e dei suoi dialetti*, vol. 3, *Sintassi eformazione delle parole*. Torino: Einaudi.

Scala, Andrea. 2018. Italo-Romance phonological rules and Indo-Aryan lexicon: The case of Abruzzian Romani. In Roberta D'Alessandro and Diego Pescarini (eds.), *Advances in Italian Dialectology: Sketches of Italo-Romance Grammars*. Leiden: Brill.

Seržant, Ilja A. 2017. Differential object markers and datives. 50th Annual Meeting of the Societas Linguistica Europaea, University of Zurich, September 10–13.

Seržant, Ilja A. 2019. Weak universal forces: The discriminatory function of case in differential object marking systems. In Karsten Schmidtke-Bode, Natalia Levshina, Susanne M. Michaelis, and Ilja A. Seržant (eds.), *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the Evidence*, 149–178. Berlin: Language Science Press.

Seržant, Ilja A. In press. Diachronic and typological evidence from argument marking confirms generic audience-design effects and strong effects of ambiguity avoidance in the domain. *Studies in Language*.

Sinnemäki, Kaius. 2014. A typological perspective on differential object marking. *Linguistics* 52(2): 281–313. https://doi.org/10.1515/ling-2013-0063.

South, Andy, Schramm Michael, and Philippe Massicotte. 2025. rnaturalearthdata: World Vector Map Data from Natural Earth Used in 'rnaturalearth'. R package version 1.0.0.9000, https://docs.ropensci.org/rnaturalearthdata/.

Soravia, Giulio. 1977. *Dialetti degli zingari italiani*. Pisa: Pacini.

Tagare, Ganesh V. 1948. *Historical Grammar of Apabhramsa*. Poona: Deccan College.

Turner, Ralph Lilley. 1926. The position of Romani in Indo-Aryan. *Journal of Gypsy Lore Society* 5(4): 145–189.

Valtonen, Pertti. 1968. *Suomen mustalaiskielen kehitys eri aikoina tehtyjen muistiinpanojen valossa* [The development of the Finnish Romani language in the light of records made at different times]. MA dissertation, University of Helsinki.

Vuorela, Katri and Lars Borin. 1998. Finnish Romani. In A. Ó Corráin and S. Mac Mathúna (eds.), *Minority Languages in Scandinavia, Britain and Ireland*, 51–76. Acta Universitatis Upsaliensis: Studia Celtica Upsaliensia 3. Amsterdam: IFOTT.

Wagner, Peter. 2012. *A Grammar of North West Lovari Romani*. PhD dissertation, Univerzita Karlova v Praze.

Wali, Kashi and Omkar N. Koul. 1997. *Kashmiri: A Cognitive-Descriptive Grammar*. London: Routledge.

Weinreich, Matthias. 2008. Two varieties of Ḍomaakí. *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 158: 299–316.

Weinreich, Matthias. 2011. Domaaki jazyk [Domaaki grammar outline]. In Tatiana I. Oranskaya, Yulia V. Mazurova, Andrej A. Kibrik, Leonid I. Kulikov, and Aleksandr Y. Rusakov (eds.), *Jazyki mira: Novye indoarijskie jazyki*, 165–194. Moskva: Academia.

Wentzel, Tatiana W. 1980. *Die Zigeuner-Sprache* (*Nordrussischer Dialekt*). Moscow: Akademija Nauk SSSR.

Wickham, Hadley. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer.

Witzlack-Makarevich, Alena and Ilja A. Seržant. 2018. Differential argument marking:

Patterns of variation. In Ilja A. Seržant and Alena Witzlack-Makarevich (eds.), *Diachrony of Differential Argument Marking*, 1–40. Berlin: Language Science Press.

Zograf, Georgij A. 1976. *Morfologicheskij stroj novyx indoarijskix jazykov*. Moscow: Nauka.