languages

MDPI

Article

Understanding Dialectal Variation in Contact Scenarios Through Dialectometry: Insights from Inner Asia Minor Greek

Stavros Bompolas [1,*] and Dimitra Melissaropoulou [2]

[1] Archimedes, Athena Research Center, 15125 Athens, Greece
[2] School of Italian Language and Literature, Faculty of Philosophy, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; dmelissa@itl.auth.gr
* Correspondence: s.bompolas@athenarc.gr

Abstract: This study investigates the interplay between linguistic and extralinguistic factors in language contact scenarios, focusing on *inner Asia Minor Greek* (iAMGr), a dialect cluster influenced by Turkish and isolated from other Greek-speaking regions. Using dialectometric techniques, we quantified the dialect distances—encompassing both grammatical and lexical features, many of which reflect foreign interference—between nineteen iAMGr varieties. A regression analysis was then employed to evaluate the impact of geographic, demographic, and other macro-social factors on these distances. The results reveal distinct patterns. The grammatical features show a substantial divergence between communities, linked to structural borrowing and primarily influenced by the dominant group's population size and degree of contact (low- vs. high-contact variety types). In contrast, lexical features exhibit greater convergence, primarily influenced by geography, linked to the susceptibility of lexical borrowing to casual contact. Unlike previous dialectometric studies that report a strong correlation between geographic and dialect distances, our findings suggest that geography's influence varies by linguistic level, being more pronounced in lexical distances. Furthermore, the analysis reveals that certain dialect-specific factors previously identified in qualitative studies on iAMGr are statistically insignificant. The study concludes that, while geography remains relevant, macro-social factors often play a more critical role in language contact settings, particularly in shaping grammatical distances. These findings provide new insights into the determinants of dialect distances in such contexts.

Keywords: dialectometry; regression analysis; inner Asia Minor Greek; Cappadocian; Pharasiot; Silliot; language contact; grammatical variation; lexical variation

check for updates

Academic Editors: Alfred Lameli, Simonetta Montemagni and John Nerbonne

Received: 16 July 2024
Revised: 13 December 2024
Accepted: 3 January 2025
Published: 16 January 2025

Citation: Bompolas, S., & Melissaropoulou, D. (2025). Understanding Dialectal Variation in Contact Scenarios Through Dialectometry: Insights from Inner Asia Minor Greek. *Languages*, *10*(1), 13. https://doi.org/10.3390/languages10010013

1. Introduction

This paper explores the application of dialectometric and correlational techniques to investigate dialectal distances in language contact settings. We focus on *inner Asia Minor Greek* (henceforth iAMGr), a group of varieties that have developed under the influence of Turkish and in relative isolation from other Greek-speaking regions (Dawkins, 1916; Kontosopoulos, 1981/2008; Karatsareas, 2011; Manolessou, 2019). Employing iAMGr as a case study, this research aims to showcase the application of dialectometric techniques in analyzing dialectal variation through the lens of foreign interference. In particular, it seeks to empirically assess the impact of various extra-linguistic factors acknowledged in the relevant literature on this process.

Dialectometric approaches, utilizing computational and statistical methods, have revolutionized the field of dialectology by revealing hidden patterns in language variation

(Wieling & Nerbonne, 2015). However, limited research has focused on applying these techniques to understand the areal effects in language contact situations (e.g., Heeringa et al., 2000 for Dutch–German contact; Heeringa et al., 2010 for contact effects in Bulgarian; and Sousa & García, 2020 for Galician–Spanish contact) and, particularly, to an in-depth understanding of the influence of extra-linguistic factors (Kortmann, 2013). This is further compounded by a focus within the field on automating methods to identify lexical borrowing (Heeringa et al., 2010; List, 2019; Zhang et al., 2021). Still, a significant challenge persists in incorporating external factors beyond geography into the formal statistical models employed by correlational approaches (Wieling, 2012; Huisman et al., 2021). This study addresses this gap by examining how geography and social factors influence dialect distances in language contact settings by using the iAMGr varieties as a case study.

This study adopts Thomason and Kaufman's (1988, p. 9) framework for language variation and change, focusing on three key drivers: (i) drift, representing inherent tendencies within a language to evolve due to internal structural imbalances; (ii) dialect interference, encompassing both interactions between established dialects and the diffusion of changes across less distinct varieties; and (iii) foreign interference, arising from contact with another language.

Geography is recognized as a well-established factor that influences both drift and dialect interference, with spatial proximity often correlating with linguistic similarity (Chambers & Trudgill, 1998; Nerbonne & Heeringa, 2001). Nerbonne and Heeringa (2007, pp. 274, 291) even argue that the significant influence of geography as an explanatory factor for dialectal variation suggests that social variables are unlikely to have a greater impact. This notion is further supported by the *Fundamental Dialectological Postulate* (FDP)[1] (proposed by Nerbonne & Kleiweg, 2007, p. 154; cf. Bloomfield, 1933, p. 476, *principle of density*). However, despite the seeming universality of the FDP (e.g., Nerbonne, 2010; contra Szmrecsanyi, 2012), its influence is demonstrably contingent upon the specific linguistic level under investigation (Spruit et al., 2009; Scherrer & Stoeckle, 2016; Bompolas & Melissaropoulou, 2023a, among others), the geographical scale of the area examined (Stanford, 2012; Jeszenszky et al., 2017; Bompolas, 2023, Chapter 5), and the type (i.e., low vs. high contact) of varieties analyzed (Kortmann, 2013; Bompolas & Melissaropoulou, 2023b).

Furthermore, the FDP might not fully capture the complexities of foreign interference. Geographic distance, while a well-established factor, may not be the sole influence, particularly in situations where multilingual speaker groups interact within a single geographical area (Thomason, 2001). In such settings, additional factors at a more granular level, particularly the intensity of social and cultural interaction among speakers of different languages, can mediate the impact of geographic distance (e.g., Thomason, 2008; Yakpo, 2021). Kortmann (2013) reinforces this notion by demonstrating that geography holds less explanatory power for morphosyntactic variation in high-contact varieties (e.g., pidgins and creoles) compared to low-contact English dialects. Additionally, the intensity of contact itself is a crucial factor in foreign interference, with structural borrowing necessitating a higher level of contact intensity compared to lexical transfer (Thomason & Kaufman, 1988, pp. 74 ff.; Thomason, 2001, pp. 68 ff.).

Building on the preceding discussion, this study emphasizes the importance of a socially oriented perspective for a comprehensive understanding of dialectal distances in language contact settings. In this line, extra-linguistic macro-variables provide a robust framework for explaining the diverse outcomes observed in the dialect continua shaped by intense contact (Yakpo, 2021). To explore this further, we computationally measure dialect distances between nineteen varieties of iAMGr based on 279 grammatical and 423 lexical features extracted from the electronic version of the DiCaDLand ("Digitizing the Cappadocian Dialectal Landscape") atlas (Melissaropoulou, 2024) and dictionary (ILIK, 2024). Crucially, these features capture not only instances of drift and dialect interference but also foreign influence from Turkish, reflecting the unique dialectal context. Subsequently, through statistical analysis, we analyze linguistic distances and their correlations with extra-linguistic variables, revealing distinct patterns: (i) grammatical features show significant divergence linked to structural borrowing influenced by factors such as the dominant group's population size and the degree of contact (operationalized in terms of high-/low-contact variety types), with geography playing a secondary role, and (ii) lexical features exhibit convergence driven predominantly by geography, linked to the ease of lexical borrowing through casual contact. Notably, our findings challenge traditional assumptions regarding the relative importance of specific explanatory factors in shaping dialect distances, both in general and with particular reference to iAMGr. Moreover, our results reveal that the relative impact of these factors differs depending on the linguistic level under investigation. Overall, this study suggests that, while geography remains a significant factor for lexical distances, macro-social dynamics play a more pronounced role in shaping grammatical distances, highlighting the complex interplay between linguistic and extra-linguistic forces in dialect formation within contact settings.

2. Materials and Methods

2.1. The Inner Asia Minor Greek Dialect Continuum

The dialectal data examined in this study pertains to the Greek dialects of iAMGr (Kontosopoulos, 1981/2008, pp. 6–10). These dialects were historically spoken by Greek-Orthodox communities that inhabited the Cappadocian plateau in the southeastern region of Asia Minor, which is nowadays Central Turkey (Figure 1). The iAMGr dialect group consists of three closely related dialects (Dawkins, 1916): Cappadocian, predominantly spoken in various villages across Niğde, Nevşehir, and Kayseri; Pharasiot, used in Pharasa and five adjacent villages in southeastern Kayseri; and Silliot, used in the village of Silli near the town of Konya.[2]

At the beginning of the 20th century, the Greek-Orthodox population of inner Asia Minor amounted to approximately 44,792 inhabitants, as estimated by the Centre for Asia Minor Studies (Kitromilidis & Mourelos, 1982). Out of these, around 21,104 spoke a Greek variety (Cappadocian, Pharasiot, or Silliot), while the rest spoke Turkish (Figure 1). Within the larger Muslim population, the Christian minority was numerically small, and due to conquest and the need for social survival, the majority of them had also adopted Turkish as their language. In certain pockets of this Turkophone Christian society, the Greek language had managed to survive in local varieties (Kitromilidis & Mourelos, 1982, κς′–κζ′, fn. 2–3). By the beginning of the 20th century, iAMGr had been confined to a geographically limited area consisting of twenty-nine villages, primarily located in the rural regions between the Ottoman urban centers of Nevşehir, Kayseri, and Niğde.

In this particular context, the intra-dialectal differentiation within iAMGr is generally attributed to the linguistic Turkification process and the varying degrees of Turkish influence experienced by the remaining Greek-speaking communities (Dawkins, 1916; Thomason & Kaufman, 1988, p. 19; Thomason, 2001, pp. 66–67).[3] Notably, Cappadocian varieties are more influenced by Turkish compared to those of Silliot and Pharasiot, and even within Cappadocian, various extra-linguistic factors have shaped a non-uniform dialectal landscape. As a result, some iAMGr varieties have often been referred to as "an excellent example of heavy structural borrowing" (Thomason & Kaufman, 1988, p. 215). Drawing on Dawkins (1916, pp. 203, 209), Thomason and Kaufman (1988, pp. 215 ff.) enumerate a variety of lexical and grammatical innovations found in the three dialects, making a strong case for language contact in iAMGr so as to claim that, while most of the varieties "clearly retain enough inherited Greek material to count as Greek dialects in the full genetic sense"—the "Greek substratum" in terms of Dawkins (1916, p. 212)—"a few dialects may be close to or even over the border of nongenetic development" (see also Thomason, 2001, pp. 63–65, 74, 86).

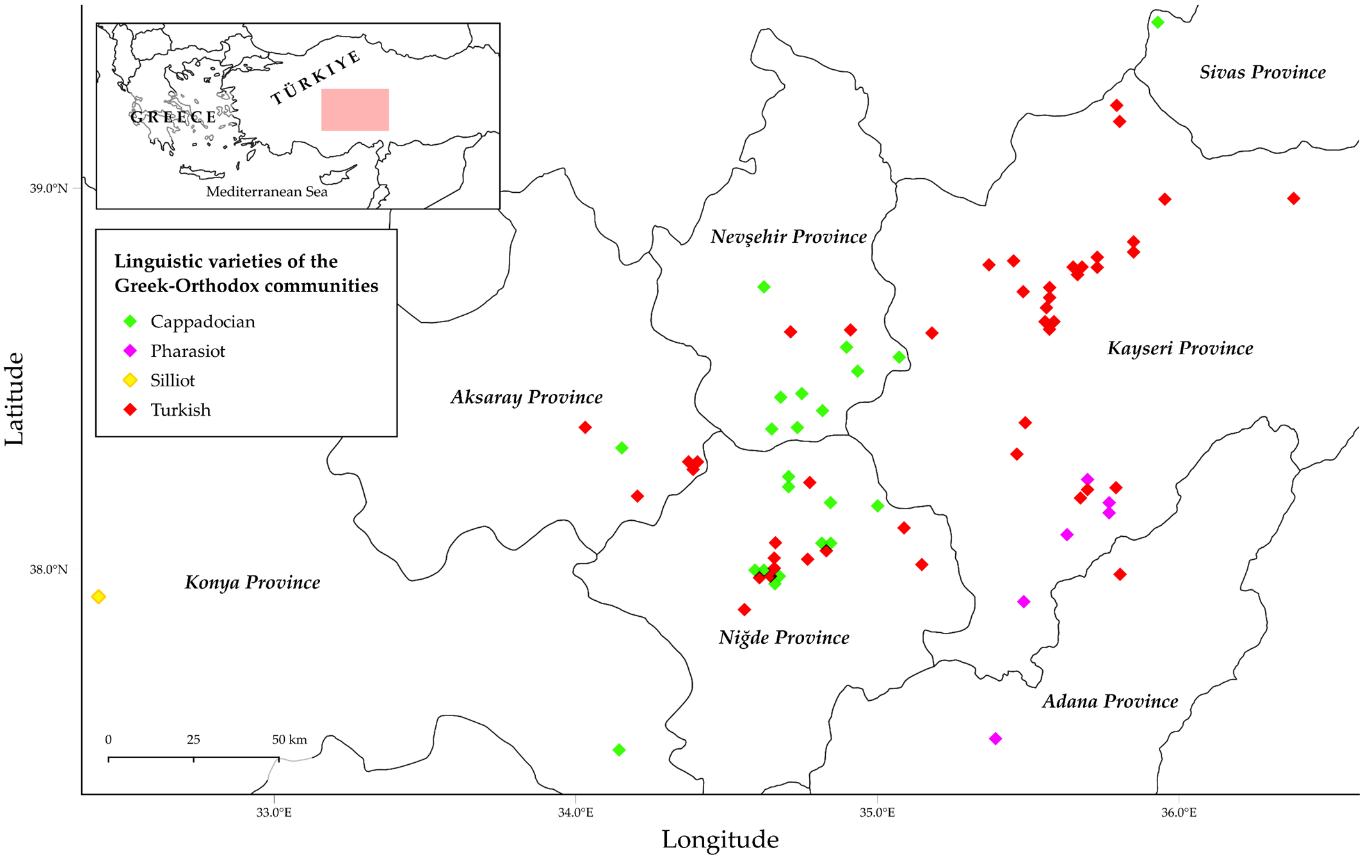


Figure 1. Language map of the Greek-Orthodox communities of inner Asia Minor at the beginning of the 20th century (Some places that are too close to each other have been displaced for better readability of the map).

Dawkins (1916, p. 204) proposed a dialectal continuum for the three primary dialect clusters, which reflects the level of dialectal differentiation within iAMGr in relation to the Turkish influence. According to this continuum, the Pharasiot dialect is least influenced by Turkish, while the Cappadocian dialect exhibits the greatest influence, with Silli occupying an intermediate position (Figure 2). Dawkins (1916, p. 203) identifies a diverse range of interference features in iAMGr from Turkish, which he proposes as key criteria for this classification:

1. borrowing of Turkish idioms;
2. use of Turkish word order;
3. effects of Turkish vowel harmony;
4. unvoiced final consonants;
5. unchanged velars in paradigms;
6. pronunciation of [ɣ] as [q];
7. failure to pronounce /θ/ and /ð/;
8. loss of genders;
9. partial disuse of the article;
10. use of the accusative ending -*ον*/-on/ only after the article and generalization using -ς/-s/;
11. agglutinative declension;
12. comparative of adjectives based on the Turkish model;

13. use of Turkish numerals;
14. use of Turkish derivative verbal suffixes in Greek;
15. addition of Turkish personal endings to the Greek verb;
16. agglutinative formation of the imperfect passive;
17. pluperfect on the Turkish model;
18. position of the enclitic substantive verb.

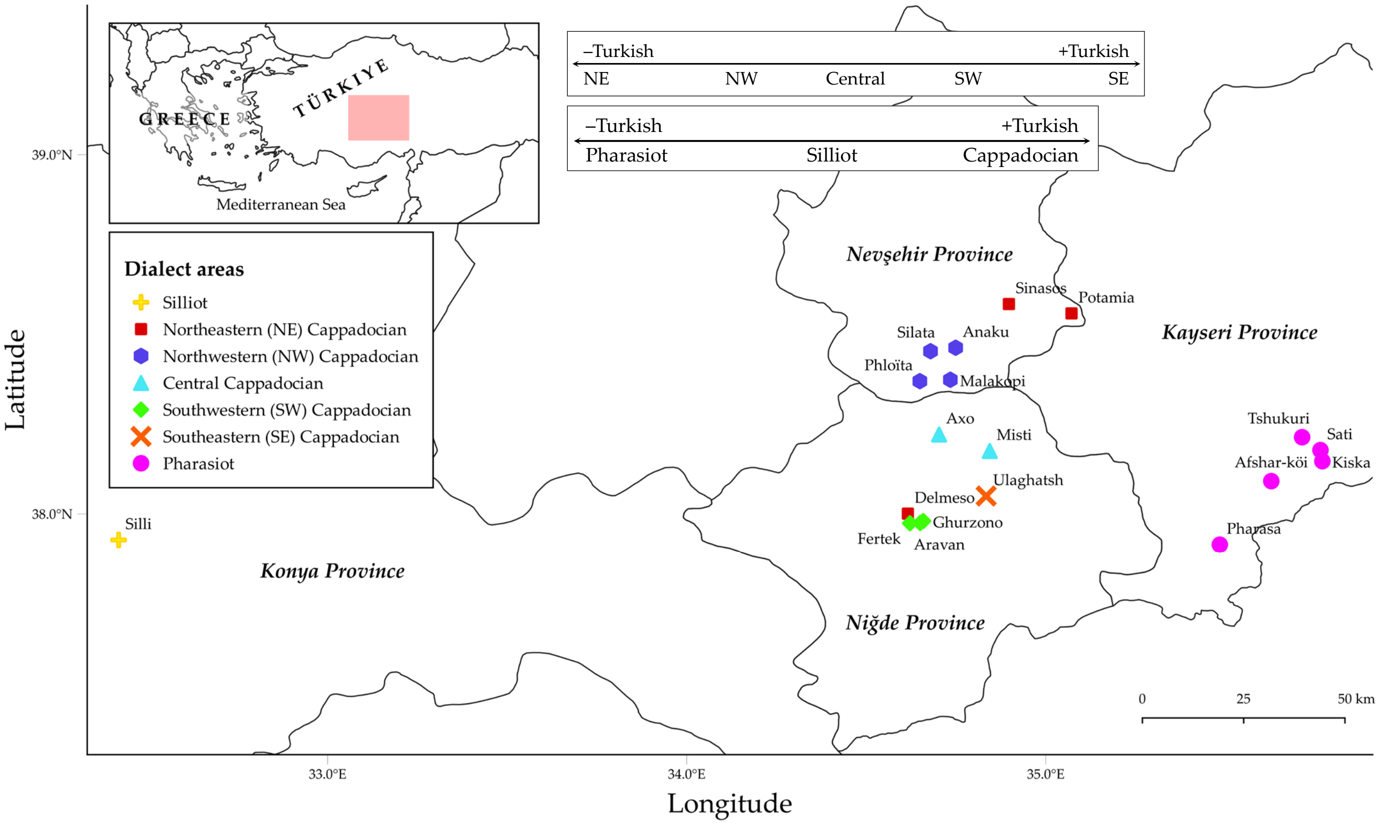


Figure 2. Dialect zones and the extent of Turkish influence on the Greek-speaking communities in inner Asia Minor. The map depicts the varieties studied here. Please refer to the text for more details.

Within the Cappadocian dialect cluster, diverse Greek communities also experienced varying levels of Turkish influence, resulting in a further classification of the Cappadocian varieties into five groups based on linguistic features, reflecting the depth and extent of this influence (Dawkins, 1916, p. 209). Specifically, in Dawkins' (1916, pp. 208–209, 211) classification scheme, Turkish influence is measured using the following linguistic features:

1. the preservation vs. loss of the Greek interdental fricatives /θ, ð/ and their replacement by dental stops;
2. the preservation vs. loss of traces of the Greek gender system;
3. the absence vs. presence of (Turkish) 'agglutinative' patterns in noun inflection; and
4. the degree of use of Turkish syntactic structures.

Based on these linguistic features, Cappadocian varieties are classified into Northeast Cappadocian (Sinasos, Potamia, and the geographically southwestern village Delmeso), Northwest Cappadocian (Silata, Phloita, Malakopi, and Anaku), Central Cappadocian (Misti and Axo), Southwest Cappadocian (Aravan, Ghurzono, and Fertek), and Southeast Cappadocian (Ulaghatsh and Semendere). Turkish was thought to have a stronger influence in the southern Cappadocian zone, while the villages in the northern zone were thought to be less influenced by Turkish. Figure 2 illustrates this isogloss-based partitioning of the iAMGr dialect landscape, along with the varying degrees of Turkish influence across the region.

Pharasiot also exhibits some intra-dialectal variation among its speech communities. However, the differences are not as extensive as in the case of Cappadocian dialects, are mainly observed between the central village (i.e., Pharasa) and the peripheral ones, and are more saliently observed in the respective lexical stocks (Bağrıaçık, 2018, p. 24), which are further attributed to Turkish influence (Anastasiadis, 1976, p. 7; see also Dawkins, 1916, p. 197).

Early classifications of the iAMGr dialects by Dawkins (1916) sought to establish a correlation between linguistic and extra-linguistic phenomena, suggesting a relationship between contact intensity and borrowing scales. Building upon this fundamental work, subsequent research has expanded upon these insights, examining factors such as the intensity and duration of contact, language status within the community, and speaker population size as potential influences on language change (Thomason & Kaufman, 1988, pp. 65 ff.; Thomason, 2001, pp. 66 ff.; Karantzola et al., 2021). Thomason and Kaufman (1988, pp. 216, 356) further highlighted the presence of Greek schools within the communities, proximity to Turkish-speaking urban centers, and migration patterns to Constantinople as additional factors influencing the contact situation in iAMGr.

Given the complex nature of the contact situation, involving diverse varieties and multifaceted social and cultural interactions both within and among communities, this study employs quantitative methods to investigate these factors, aligning with recent methodological trends in the field (Yakpo, 2021, pp. 138–139).

2.2. Linguistic Data

This study is based on data from the research project titled "Digitizing the Cappadocian Dialectal Landscape" (2018–2022), the main outcomes of which include two major reference works: the (electronic version of the) *Linguistic Atlas of the Dialectal Varieties of Cappadocia* (Melissaropoulou et al., 2022; Melissaropoulou & Bompolas, 2022; Melissaropoulou, 2024) and the (electronic version of the) *Historical Dictionary of the Cappadocian Dialects* (Karasimos et al., 2020; Manolessou et al., 2022; ILIK, 2024), both incorporating state-of-the-art methods and data. These two reference works served as the empirical testbed of this study.

Data for the atlas and dictionary were drawn based on all available sources, both primary (such as folktales and narratives) and secondary (such as grammatical descriptions, glossaries, and dictionaries). These sources primarily date back to the late 19th century till the early 20th century. The complete list of these materials can be found at http://cappadocian.upatras.gr/ (accessed on 2 January 2025).

2.2.1. Grammatical Data

The grammatical (and a selection of lexical) data for this study are derived from the *Dialectal Atlas of Cappadocian Dialects*. This comprehensive atlas encompasses 429 linguistic variables of (morpho-)phonological, morphological, and (morpho-)syntactic natures drawn from twenty distinct communities (Table 1). While dialects were historically spoken in a greater number of communities, the available data are limited to twenty, representing approximately two-thirds of all Greek-speaking communities in inner Asia Minor (Kitromilidis & Mourelos, 1982). The atlas is available in two formats: an open-access interactive atlas (http://cappadocian.upatras.gr/atlas/, accessed on 2 January 2025) and a printed version of five volumes (Melissaropoulou, 2024). For this paper, we utilized the digitized version of the atlas.

Table 1. Grammatical domains covered in the DiCaDLand Atlas.

Grammatical Domain		Type of Data	Sum of Features in Group	% of Total Features
Phonetics—(Morpho)phonology	Vowels	Categorical	19	4.43%
	Consonants		24	5.59%
	Stress		1	0.23%
	Morphophonology		5	1.17%
Morphology	Articles		27	6.29%
	Gender		16	3.73%
	Nominal Morphology		68	15.85%
	Pronouns	String	103	24.01%
	Numerals		34	7.93%
	Verbal Morphology	Categorical	61	14.22%
	Derivational Morphology		5	1.17%
(Morpho)syntax	Verb Morphosyntax		3	0.70%
	Noun Phrase		25	5.83%
	Verb Phrase		7	1.63%
	Adverbial Phrase		2	0.47%
	Comparative Structures		4	0.93%
	Sentence		25	5.83%

2.2.2. Lexical Data

The lexical data for this study are drawn from the *Historical Dictionary of the Cappadocian Dialects*, a comprehensive resource containing approximately 9000 entries collected from 29 distinct locations. This dictionary features a carefully organized microstructure, systematically detailing variations in meanings, etymology, and forms. Each variant is presented with orthographic and phonetic transcriptions and mapped to their specific geographical distribution, which is critical to the aims of this study. The *Historical Dictionary of the Cappadocian Dialects* will be available in two formats: an open-access interactive database and a three-volume printed edition (ILIK, 2024). For the purposes of this paper, the digitized version of the dictionary was used.

2.3. Parameterization of Linguistic Data

To address the lack of a standardized measurement that can effectively combine both categorical (nominal) and string (lexical) data (Heeringa & Prokić, 2017, pp. 342–343), this study treats each type of phenomenon separately. This approach facilitates an inclusive and comprehensive dialectometric analysis. The grammatical phenomena, encompassing phonological, morpho-phonological, morphological, morphosyntactic, and syntactic elements, are derived from the DiCaDLand atlas. These categorical variables are parameterized in such a way that permits binary comparisons, as implemented in Gabmap (Nerbonne et al., 2011; Leinonen et al., 2016). Conversely, the lexical phenomena, primarily sourced from the DiCaDLand dictionary, are parameterized to be compatible with string (edit) distance calculations, as implemented in LED-A (Heeringa et al., 2023).

In order to perform the dialectometric analysis, the data has been encoded in a format compatible with the web applications utilized in this study (Nerbonne et al., 2011, p. 73; Snoek, 2014, pp. 193–195; Leinonen et al., 2016, p. 72; Heeringa et al., 2022). Specifically, the data need to be entered into a tabular format that has the conceptual categories for the items across the horizontal axis and the varieties to be compared along the vertical axis. Missing data are indicated by empty cells. When multiple variants exist for one linguistic variable in one location, they are separated by a slash preceded and followed by a space.

2.3.1. Parameterization of Grammatical Data

The study draws upon the DiCaDLand atlas to extract grammatical phenomena. To facilitate the dialectometric analysis, the categorical variables within the atlas underwent a standardization process.

First, single variables (and their variants), originally distinguished in the atlas due to organizational or technical considerations (e.g., complementizers), were unified. Secondly, the analysis deliberately excluded systematic phonological changes (i.e., phonological allomorphs) at the morphological and syntactic levels. This approach aimed to maintain the independence of each linguistic level and minimize the potential confounding factors that arise from phonological variations (see Scherrer & Stoeckle, 2016, for a similar approach). Additionally, morphological maps exhibiting string data characteristics (i.e., variations in pronouns and numerals; Table 1) were omitted, as they did not conform to the established categorical distinctions (see also below). Finally, the data related to the community of Semendere were excluded due to insufficiently attested variants, particularly at the syntactic level.

The resulting data matrix consists of linguistic data from nineteen sites, comprising 279 standardized nominal/categorical variables and totaling 6751 variants (see Supplementary Materials).

2.3.2. Parameterization of Lexical Data

Parameterization of the lexical data involved obtaining data from the DiCaDLand dictionary, which were exported as a CSV file. Data-mining techniques were then utilized to extract a comprehensive list of lemmas (headwords), along with their usage labels, part-of-speech tags, IPA annotations for attested variants, definitions, usage areas, and etymologies. The dictionary data underwent a multi-step pre-processing methodology.

Initially, qualitative filtering criteria were applied to exclude variants with vague geographic references. Entries labeled as 'Cappadocian' and 'and elsewhere' were omitted due to their broad nature and potential bias in the analysis of dialect differences. Additionally, only citation forms were retained, excluding all inflected forms, including the first-person singular present tense for verbs and the singular nominative case for nouns. Following this, quantitative criteria were implemented, discarding entries lacking a minimum of two geographical variants and preserving only those with variants in at least half of the research areas, equivalent to ten or more, to ensure statistical validity. Lastly, we pre-processed certain digraphs (i.e., [ʤ], [ʥ], [ʦ], [ʧ]) as single entities. As these affricates were prevalent within the dataset, unitary symbols were used to represent individual phonological segments, ensuring the integrity of the data analysis.

The outcome of this procedure is a linguistic data matrix comprising 292 lexical variables with 4680 instances across nineteen locations. It should be noted that string data from the atlas, such as pronouns and numerals (Table 1), were also integrated. However, duplicates found in both the dictionary and atlas data were excluded to prevent redundancy. Following this merging process, the final dataset includes 423 lexical phenomena with a total of 9291 instances across nineteen sites (see Supplementary Materials).

2.4. Limitations of Linguistic Data

Bompolas (2023, Chapter 7) and Bompolas and Melissaropoulou (2023b) identified a significant proportion of atlas-based grammatical variables (162 variables with 3593 variants) that can be attributed to foreign interference. In contrast, the number of identified lexical borrowings is relatively limited and less diverse than other domains of variation, primarily due to the quantitative criteria used during the lexical data extraction. While validity analyses confirm the reliability of statistical results based on contact-induced lexical phenomena, the identified 47 lexical phenomena and their 878 variants may not fully represent the existing variation compared to the grammatical variables related to language contact. This discrepancy underscores the need for a careful interpretation of results based on lexical variables.

2.5. Extra-Linguistic Data

The refined linguistic datasets enable the testing of hypotheses regarding the influence of external factors on grammatical and lexical variation. It is imperative to consider social meanings specific to the iAMGr context rather than relying solely on general principles of dialectology that are applicable to larger regions (Stanford, 2012, p. 252). These factors encompass social/administrative organization, geographic and demographic distribution, and ideological superstructure within the iAMGr context. While dialectologists have acknowledged the potential impact of such language-external variables, their inclusion in analyses has often been post hoc in dialectometric studies rather than as formal explanatory variables within statistical models (Wieling, 2012; Wieling & Nerbonne, 2015, pp. 250–251; Nerbonne & Wieling, 2017; Huisman et al., 2021). By integrating quantitative ecology techniques into dialectometric methodology, robust explanatory models of dialectal variation can be developed (e.g., Honkola et al., 2018).

In line with Yakpo's (2021, pp. 131–132) work, this study emphasizes the significance of macro-level social factors, such as demography, in understanding language contact dynamics and outcomes. This approach diverges from variationist sociolinguistics, which predominantly focuses on micro- and meso-level social variables (e.g., gender, social class, or ethnicity) and their associated social meanings (e.g., Theodoridi, 2017; Karantzola et al., 2021 for iAMGr). While aligning with Karantzola et al.'s (2021, p. 25) perspective on incorporating external factors, the current methodology differs by integrating social, demographic, and other predictors, along with their interactions, into a single statistical model, as opposed to their qualitative approach.

In our analysis, we employed seven extra-linguistic factors that had been previously identified as potential predictors of linguistic variation and change in the iAMGr dialectal continuum. These factors were organized within Karantzola et al.'s (2021) framework of macro-variables, as shown in Table 2 (see also Supplementary Materials). This framework distinguishes between local, regional, national, and extra-national levels of external macro-variables based on Grenoble and Whaley's (1998) model. In the context of iAMGr, the local level encompasses Greek-speaking communities and nearby villages, the regional level includes administrative regions, the national level comprises the Ottoman Empire and the Ecumenical Patriarchate of Constantinople, and the extra-national level includes the Greek state and organized Greek diaspora.

However, it is essential to acknowledge the limitations in reconstructing the complete geographic, administrative, demographic, and social ecology of the iAMGr dialectal landscape due to the scarcity of data on the entirety of the Greek-speaking communities (Logotheti-Merlier, 1977, p. 43; Kitromilidis & Mourelos, 1982, λβ′–λδ′; Karantzola et al., 2021, pp. 42–43). This inherent complexity and the constraints imposed by the available data must be taken into account.[4]

Table 2. Extra-linguistic variables, based on the levels of macro-variables distinguished by Grenoble and Whaley (1998, pp. 38–41).

Setting Level	Extra-Linguistic Variable
Local	Number of Turkish-speaking people within the community
	Distance between communities
Regional	Variety type (high-/low-contact variety; based on Dawkins' dialectal division)
	Number of Turkish-speaking people within the province
	Distance between communities and urban centers within the province
National	Contact with/migration to Constantinople
Extra-national	Presence/absence of (semi-)organized Greek school

2.5.1. Geographic Data

The first factor investigated in this study is geographic distance, a well-established predictor of linguistic variation in dialectology and dialectometry. To incorporate this factor, a comprehensive geodatabase was meticulously constructed, encompassing precise geographic coordinates for various locations within the Cappadocian plateau. These locations were selected based on the identification of Greek-Orthodox communities by Kitromilidis and Mourelos (1982), as depicted in Figure 1. A key challenge was reconciling historical placenames from inner Asia Minor found in historical and dialectal sources with current placenames used in the geographic data and information systems. Fortunately, we greatly benefited from the toponymic research conducted within the framework of the DiCaDLand research program by Prof. Melissaropoulou and Dr. Manolessou. In particular, they compiled a thorough catalog of the exact geographic coordinates of all Cappadocian communities with the help of the Index Anatolicus (https://www.nisanyanyeradlari.com/, accessed 2 January 2025), an online database that documents over 56,000 geolocated toponyms across Anatolia (for details, see Melissaropoulou et al., 2022; Melissaropoulou & Bompolas, 2022; Melissaropoulou, 2024; see also the introduction of ILIK, 2024, for information on the toponymic research). These coordinates served as the foundation for calculating geographic distances, allowing for a subsequent analysis of their correlation with linguistic distances. The emphasis on precision in this aspect was crucial for the successful execution of the dialectometric study.

2.5.2. Demographic Data

Another factor influencing linguistic distances in iAMGr is population size (Dawkins, 1916; Thomason & Kaufman, 1988, p. 356; Thomason, 2001, p. 66; Karantzola et al., 2021, pp. 44–48). Trudgill's gravity model suggests that population size mediates the effect of geographic distance, with language changes diffusing first from larger cities to smaller locations due to the influence of numerically dominant groups (Trudgill, 1974; 1986, Chapter 3; Chambers & Trudgill, 1998, Chapters 11.7–11.8). Similarly, in language contact settings, the language of a smaller minority group is more likely to acquire features from a larger dominant group's language (Thomason, 2001, p. 66).

Reconstructing the demographic ecology of iAMGr presents challenges due to limited data availability for all Greek-speaking communities (Logotheti-Merlier, 1977, p. 43; Kitromilidis & Mourelos, 1982, λβ′–λδ′; Karantzola et al., 2021, pp. 42–43). Historical events, such as population exchanges between Greece and Turkey in the early 20th century, further complicate the demographic picture.[5] The following authors provided data for each community: Alektoridis (1883), Farasopoulos (1895), Sarantidis (1899), Kholopoulos (1905), Dawkins (1916), and Kitromilidis and Mourelos (1982). However, the main issue with demographic data coming from older sources is that they are estimations that may be considerably inflated, especially for Christians and possibly for the Turkish population, according to Dawkins (1916, pp. 11, 19, 26, 34). Therefore, we rely on the data of Kitromilidis and Mourelos (1982), which come from the official demographic results of the Exchange of Populations in the report of the Society of Nations, *L'établissement des réfugiés en Grèce* (Société des Nations, 1926). Although using demographic data from 100–150 years prior to the linguistic data would be ideal (Nerbonne & Heeringa, 2007, p. 279), relying on unreliable population figures would be statistically unsound.

Based on available population numbers, this study examines the magnitude and effects of Turkish population size, both within the studied communities and within the broader administrative provinces. Regarding the latter, our choice is in line with Logotheti-Merlier's (1977) observation that cultural and social interactions among iAMGr communities were more frequent within these administrative contexts, a phenomenon termed 'isolation by administrative history' (Honkola et al., 2018). For this purpose, Logotheti-Merlier's (1977, pp. 50, 52, 58–61, 65–66, 69) sociohistorical classification of Cappadocian communities, which considers the degree of isolation and contact among Greek speakers and their affiliation with economic and social centers (typically towns or cities), is utilized (Table 3).

Table 3. Division of the Greek-Orthodox communities of inner Asia Minor (based on Logotheti-Merlier, 1977). In italics are the linguistic varieties studied here.

Province	Center	Greek-Orthodox Communities	Language
Pharasa	Pharasa	*Pharasa*	Pharasiot
Pharasa colonies	Kiska	*Tshukuri*, Fkosi, *Sati*, *Ashar-köi*, *Kiska*	
		Kurumza (Ghariptsas), Beskardas, Tastsi, Khostsa	Turkish
Iconium	Iconium	*Silli*	Silliot
Neapoli	Neapoli	*Anaku*, Arabison, Dila, *Malakopi*, *Silata*, *Phloïta*	Cappadocian
		Neapoli	Turkish
Nigdi	Nigdi	*Axo*, *Aravan*, *Ghurzono*, *Misti*, *Ulaghatsh*, Semendere, T' Axenu to khorio, *Delmeso*, Trokho, Tsharakly, *Fertek*	Cappadocian
		Andaval, Enekhil, Iloson, Kitsaghats, Limna/Limnos, Matala, Nigdi, Poros, Sazaldza, Suludzova, Teneï	Turkish
Prokopi	Prokopi	*Potamia*, *Sinasos*, Zalela	Cappadocian
		Prokopi	Turkish

Population sizes were incorporated as an individual factor, rather than adopting a gravity-based approach to ascertain the independent strength of population size on linguistic variation. This decision was motivated by previous research indicating that, in certain instances, the primary influence attributed to gravity models may stem predominantly from distance alone (Nerbonne & Heeringa, 2007).

2.5.3. Proximity to Urban Centers

Thomason and Kaufman (1988, p. 216) suggest that proximity to the large Turkish(-speaking) urban centers influences the linguistic outcome of contact situations in iAMGr, a notion that has been further explored but not fully corroborated by Karantzola et al. (2021, pp. 48–49). Empirical evidence supports the idea that urban centers play a significant role in disseminating linguistic innovations, potentially leading to dialect convergence towards the language spoken in areas with common economic, political, and cultural dominance (Trudgill, 1974, p. 233; 1986, pp. 73–76; Chambers & Trudgill, 1998, pp. 172, 178 ff.). In line with this, we have calculated the distance of each location in the dataset from its associated urban center, as outlined in Table 3, to examine this factor's potential impact on dialect distances within the iAMGr context.

2.5.4. Variety Type

Building on Kortmann's (2013) findings, that variety type may account for more variance in linguistic patterns than geographic distance, this study incorporates the dialect areas proposed by Dawkins (1916) as a factor in its analysis. Kortmann's framework emphasizes the socio-historical conditions under which language varieties emerge, providing a valuable perspective on the dynamics of linguistic variation. While we do not adopt his classification of variety types (e.g., L1, L2, pidgins, and creoles) in its entirety, we draw on elements of his conceptualization to explore the factors influencing dialect distances in iAMGr.

In this study, "variety type" refers to Dawkins' classification of iAMGr varieties, which categorizes dialects along a continuum reflecting varying degrees of contact with Turkish, ranging from low- to high-contact variety types (see Section 2.1 for details; see also Figure 2). By incorporating this variable, we aim to account for the role of contact-induced variation in the geographic distribution of iAMGr, as highlighted in previous research (Bompolas, 2023, Chapter 7; Bompolas & Melissaropoulou, 2023b).

To control for the potential effects of these dialect areas or variety types (see also Shackleton, 2005, 2007; Nerbonne, 2013), each location in the dataset was assigned to one of the following subgroups based on Dawkins' (1916) and Janse's (2008, p. 191) classifications: (i) the three core areas: Cappadocia, Pharasa, Silliot; and (ii) a finer subdivision of Cappadocian varieties into northwest/northeast/southwest/southeast and central, reflecting varying degrees of Turkish influence (see Figure 2).

2.5.5. Education

This factor was introduced due to the proposed dual impact of the spread of Greek education on local dialects through schooling. While it threatened the existence and led to the potential obsolescence of these dialects (Kitromilidis & Mourelos, 1982, λε′–λζ′; Karatsareas, 2011, pp. 18, 20), it also reduced Turkish influence on Greek in villages with established Greek schools compared to those without (Thomason & Kaufman, 1988, pp. 67, 216, 356). To investigate this, a variable was introduced that classifyied varieties based on the presence or absence of (semi-)organized Greek schooling (for which, see Dawkins, 1916, pp. 10–37; Logotheti-Merlier, 1977, p. 51).

2.5.6. Migration

Additionally, we have investigated the potential impact of migration on the sociolinguistic landscape, a factor previously proposed as influential (Dawkins, 1916; Thomason & Kaufman, 1988, p. 356; Karantzola et al., 2021, pp. 49–51). Our focus has been on migration to Constantinople, where migrants encountered both Turkish and Greek, specifically the official written form of Greek, Katharevousa, and a spoken koine (common language) that was prevalent within the Greek community at that time. This exposure to diverse linguistic forms could have influenced the migrants' native varieties of Greek upon their return or through ongoing contact with Constantinople.

2.6. Parameterization of Extra-Linguistic Data

To investigate the factors influencing linguistic variation, distinct data matrices were constructed, each representing a specific variable. Rows in the matrices corresponded to data collection sites, while columns represented the individual variables. In line with the study's goals, only numeric values were used. Numeric variables, like population sizes and geographic distances, were directly represented as absolute numbers. Non-numeric variables, including variety type and contact with/migration to Constantinople, were incorporated using dummy variables. This approach facilitated the inclusion of categorical effects in the analysis. Dummies indicate the presence (1) or absence (0) of a particular category. Non-numerical variables with multiple levels, like the level of organized schooling, were represented using multiple dummy variables (0 for absence, 0.5 for semi-organized, and 1 for organized). Missing values were denoted as N/A. Crucially, the data matrices adhered to the principle of allowing only one value per cell for numerical data. This ensured data consistency and streamlined the subsequent statistical analysis.

2.7. Computational Analysis

After collecting the data, the next step involved converting the qualitative data into a quantitative format (Goebl, 2017, p. 131). This conversion is crucial for dialectometric studies, as it allows for more rigorous statistical analysis. To do this, we utilized the linguistic, historical, cultural, demographic, and sociolinguistic attributes of each pair of locations to calculate their respective distances. These distance matrices, with each entry representing the level of difference between the paired locations, served as the foundation for our statistical analysis (see Supplementary Materials).

2.7.1. Linguistic Distances (Dependent Variables)

Dialectometry aims to objectively measure the linguistic distance between dialects (Séguy, 1971). Several measures of linguistic distance have been used in dialectometry to calculate how much two forms differ from each other (see Heeringa & Prokić, 2017, for a recent overview). Since there is no standardized measurement that can combine nominal (categorical) and lexical (string) data (Heeringa & Prokić, 2017, pp. 342–343), we approached each type of data differently. In this study, the measurement of dialectal data quantifies the distances between dialects based on the structures or loanwords they have acquired, as well as the native structures or lexicon they have retained (Sousa & García, 2020).

As detailed in Section 2.2.1, grammatical phenomena in this study are of a nominal or categorical type (see Supplementary Materials). To compute grammatical distances between all pairs of locations, the *relative distance value* (RDV) metric (see Heeringa & Prokić, 2017, p. 330), operationalized in Gabmap as a binary comparison, was employed (Nerbonne et al., 2011, p. 69; Snoek, 2014, p. 195; Leinonen et al., 2016, p. 75). In RDV, two strings are considered either identical (distance of zero) or different (distance of one). Thus, the overall linguistic distance is simply the sum of how many of the linguistic variables have different forms in two sites. In the case of competing forms, which were included in the analysis, Gabmap calculates the mean of the two distances when comparing—in the simplest case—one form at one site with two forms at another (Nerbonne & Kleiweg, 2003).

The lexical (string) data extracted from the *Historical Dictionary of the Cappadocian Dialects*, transcribed using a standardized IPA-based coding system, offer rich phonetic detail. Thus, to measure the lexical[6] and pronunciation distances between iAMGr varieties, the string edit (or Levenshtein) distance was employed (see Heeringa & Prokić, 2017, pp. 330–340). This computational technique is a well-established and robust method widely used in dialectometry to quantify the pronunciation differences between varieties. It calculates the minimum number of operations (insertions, deletions, or substitutions of single phones) needed to transform one string into another.

Among the various string edit distance variants, the *pointwise mutual information* (PMI) Levenshtein distance divided by the alignment length variant, as implemented in LED-A

(Heeringa, 2024), was utilized in this study. PMI Levenshtein evaluates the association strength between phones while enhancing alignment accuracy (Heeringa & Prokić, 2017, pp. 336–337). It employs an information-theoretic measure to assign smaller distances to frequently co-occurring sound segment pairs. While Zhang et al. (2021) have raised concerns regarding PMI Levenshtein for loanword detection, it has been shown to produce analyses that align with expert consensus (Heeringa et al., 2010, pp. 144–145).

2.7.2. Extra-Linguistic Distances (Independent Variables)

As the statistical analyses employed in this study rely on distance (or difference) matrices as the input data, all extra-linguistic variables underwent a transformation process to be represented in this format.

For geographic distances straight-line Euclidean distance in kilometers was calculated between each pair of locations. While Euclidean distance, the shortest path between two points, is a commonly used measure of geographic distance in dialectometry (e.g., Nerbonne, 2013), its accuracy as a predictor of dialectal variation can be influenced by topography and infrastructure. Consequently, alternative measures like least-cost (or travel time) distance, which account for barriers and surface resistance, have been employed in some studies (e.g., Gooskens, 2005; Nerbonne & Kleiweg, 2007; Jeszenszky et al., 2017, among others).

However, there is a dearth of historical travel time data for the iAMGr period in inner Asia Minor, with Logotheti-Merlier (1977) being the sole available source that offers limited information. To address this, Bompolas (2023, pp. 87–88) calculated least-cost distances between iAMGr communities using GIS techniques. The resulting distance matrix showed a strong positive correlation (r = 0.89, $p < 0.001$) with Logotheti-Merlier's limited travel time data, suggesting a reasonable approximation of the original least-cost distances. A statistical analysis further revealed that Euclidean and least-cost distances correlated perfectly (r = 0.99, $p < 0.001$).

The strong correlation between linear and least-cost distances can be attributed to the historical adaptation of iAMGr communities to the local terrain, allowing them to overcome major topographical obstacles (Logotheti-Merlier, 1977, p. 43). This finding is also supported by research in other mountainous regions like Japan, where hiking and modern travel distances also correlate strongly with straight-line distance (Jeszenszky et al., 2019).

Therefore, for the present study, Euclidean distances were deemed a suitable approximation of geographic distance. However, acknowledging Gooskens' (2005) findings that historical travel times can improve models for some linguistic areas, it is important to recognize the potential limitations of using Euclidean distance and the potential benefits of incorporating historical travel time data where available.

To investigate the impact of Turkish population size on linguistic variation, we calculated the absolute difference in population size between all location pairs in the database. This analysis encompassed both the Turkish population within each community (intra-community) and the Turkish population within the broader administrative province to which each community belonged (intra-provincial).

To assess the influence of nearby urban centers, we calculated the linear distance between each studied community and its designated urban center, as defined by Logotheti-Merlier's (1977) classification (Table 3). Subsequently, absolute differences in distance were computed for each pair of communities to facilitate a comparative analysis.

For variety type, in the analysis, each location was assigned to a specific (sub)group based on Dawkins' (1916) classification scheme. A binary variable was then constructed for each location pair, indicating whether both locations belonged to the same dialect area (coded as 0) or to different areas (coded as 1).

For education, to assess the potential impact of Greek schooling, several dummy variables were introduced to represent the presence of a formal Greek school (coded as 1), a less structured Greek school program (coded as 0.5), or the absence of any Greek school (coded as 0). Following this, the absolute difference between these coded values was calculated for all pairs of locations within the dataset.

For migration, a binary variable was created for each location pair, indicating whether both locations had documented histories of migration to Constantinople (coded as 1) or lacked such documented evidence (coded as 0).

2.8. Statistical Analysis

Dialectometry aims to quantify linguistic similarities and differences between dialects, as well as the relationship between aggregate linguistic distances and one or more extra-linguistic variables. The Salzburg school of dialectometry typically compares a matrix of geographic distances with a matrix of linguistic (dis)similarities (Goebl, 2006). Gabmap, developed within the Groningen school, follows this framework by separating geographic and linguistic information into distinct matrices (Nerbonne et al., 2011, p. 85). However, incorporating external factors beyond geography poses a challenge in such correlation-based approaches.

Goebl (2006, p. 421) suggests using Pearson correlation for linking linguistic and non-linguistic distances, but the Mantel test is more appropriate for assessing significance due to the non-independence of distance values (Mantel, 1967; Smouse et al., 1986; Guillot & Rousset, 2013). Despite its widespread use in dialectometry (Gooskens & Heeringa, 2004; Heeringa, 2004, pp. 74–75; Heeringa et al., 2006, p. 57; Nerbonne & Heeringa, 2007, p. 286; Prokić & Nerbonne, 2008, p. 161; Spruit et al., 2009, pp. 1636–1637; Stanford, 2012, pp. 274–275; Grieve, 2014, pp. 60–62; Scherrer & Stoeckle, 2016, pp. 103–104; Jeszenszky et al., 2017, p. 93; Huisman et al., 2021), recent reviews of statistical methods in the field surprisingly omit it (Wieling & Nerbonne, 2015; Nerbonne & Wieling, 2017). In this study, we utilize the Mantel test and its extension, the *multiple regression on distance matrices* (MRM), to conduct our analyses, drawing partially on the methodological pipeline established by Honkola et al. (2018) and Huisman et al. (2021). Both tests were performed in the R programming environment (R Core Team, 2024) using the *ecodist* package (Goslee & Urban, 2007; Goslee, 2010).

The Mantel test (Mantel, 1967) was applied to investigate the relation between a single linguistic distance matrix and an extra-linguistic (geographic, demographic, etc.) distance matrix based on a single extra-linguistic variable. This test accounts for non-independence by calculating the Pearson product–moment correlation coefficient and performing 10,000 permutations with 1000 bootstrap iterations with 95% confidence intervals. To examine the relationship between a linguistic distance matrix and multiple extra-linguistic distance matrices simultaneously, we employed MRM with 10,000 permutations (Lichstein, 2007). MRM is an extension of the (partial) Mantel test on two (or more) distance matrices. Essentially, the relationship between the Mantel test, the partial Mantel test, and MRM is similar to the relationship between analyses of correlations, partial correlations, and multiple regression. However, the MRM has an advantage over the Mantel test in that it allows for the inclusion of each explanatory factor separately, rather than combining them into a single distance matrix as in the partial Mantel test. This enables the assessment of their individual importance and provides flexibility in handling different types of data (e.g., binary, continuous). The MRM also provides estimates of explained variance and utilizes random permutations for significance testing to avoid overestimating correlations. Unlike previous studies that used generalized additive mixed-effects regression modeling to analyze the linguistic distances between observed points (dialects) and a reference point (standard language) (Wieling, 2012; Wieling et al., 2014; Wieling et al., 2018), this study utilizes the MRM to incorporate all pairwise distances without defining a reference point.[7]

For model building, as a first step, we excluded the extra-linguistic variables that did not correlate with linguistic differences or were correlated only due to geographical distance. This was conducted by conducting (partial) Mantel tests between the linguistic (grammatical/lexical) differences and each of the explanatory variables, with and without the effect of geographical distance taken into account.

The results indicated that both distance from urban centers and contact with Constantinople did not exhibit statistically significant correlations with linguistic distances. Therefore, two of the initial seven extra-linguistic variables were excluded from further analysis due to their lack of correlation with both grammatical and lexical distances, regardless of whether the effect of geographical distance was controlled for.

Additionally, we assessed the multicollinearity among the remaining extra-linguistic variables. Collinearity, a high correlation between two or more predictors, can impact the accuracy of the coefficient estimates in a multiple regression (Vittinghoff et al., 2012, p. 148). While a correlation matrix can detect high pairwise correlations, it may not reveal multicollinearity, where a set of variables exhibits collinearity despite low pairwise correlations (James et al., 2021, p. 102). To address this, we employed the *variance inflation factor* (VIF), a more robust measure of collinearity that quantifies the correlation of a variable with a group of other variables. A VIF of 1 indicates no collinearity, while values below 5 suggest low collinearity. VIF values between 5 and 10 signify moderate collinearity, and values above 10 indicate high collinearity, which may be problematic for the model. As Table 4 illustrates, all independent variables in our model exhibit VIF values lower than 5. Consequently, we conclude that multicollinearity is not a significant concern in this analysis.

Table 4. VIF values for each explanatory variable utilized in MRM analysis.

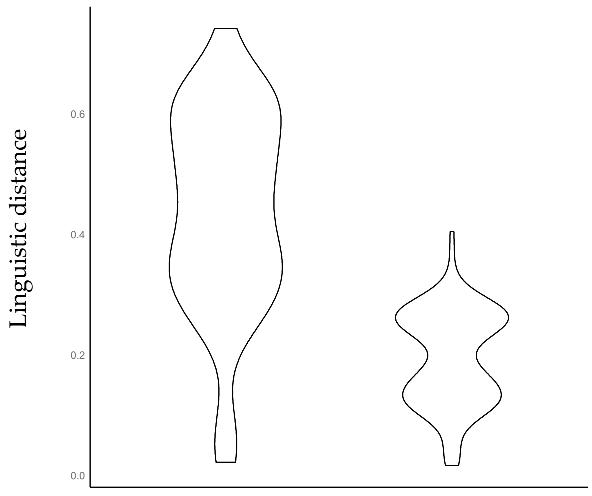
Variables	VIF
Distances (km)	4.33
Turkish community population	3.61
Variety type	1.32
Turkish regional population	1.88
Greek schooling	1.10

Finally, to ensure interpretability and facilitate meaningful comparisons within the MRM framework, all selected features underwent standardization. This process transformed the variables to have a mean of zero and a standard deviation of one. Standardization eliminates the influence of differing measurement scales on the regression coefficients, enabling a more accurate assessment of the relative contribution of each feature to the linguistic variation observed.

3. Results

3.1. Overview

Figure 3 is a violin plot of linguistic distances (as measured through the RDV and PMI Levenshtein) for all unique pairwise location-by-location comparisons across the two domains, excluding comparisons with the same location. The figure reveals several key differences between the distributions of grammatical and lexical distances. Both data sets share a minimum value of 0.02, indicating the presence of highly similar dialects within the sample. However, the maximum grammatical distance (0.74) is considerably higher than the maximum lexical distance (0.40). This suggests a wider range of variation in grammatical features compared to the lexical features across the analyzed dialects. This observation is further supported by the higher median and mean values for grammatical distances (0.46 and 0.44, respectively) compared to lexical distances (both at 0.19). Additionally, the standard deviation is slightly larger for grammatical distances (0.17) compared to lexical distances (0.08), indicating a greater spread of values around the mean for the grammatical data.



	Grammar	Lexicon
Minimum	0.02	0.02
Maximum	0.74	0.40
Median	0.46	0.19
Mean	0.44	0.19
SD	0.17	0.08

Figure 3. Distribution of linguistic distances across the two datasets.

These findings suggest a potentially greater influence of historical or external factors on grammatical structures compared to lexical items within the analyzed dialects. The wider range and higher central tendency of grammatical distances point toward a more dynamic and evolving nature of grammatical features across these communities. Conversely, lexical distances seem to be more tightly clustered, suggesting a higher degree of lexical convergence across dialects.

The aforementioned observations are corroborated by the two-dimensional plots of dialect distances (Figure 4). With respect to grammar, the phenomena reveal a continuum ranging from −Turkish to +Turkish, reflecting the social circumstances and levels of bilingualism within the communities discussed earlier. This continuum demonstrates a direct correlation between contact intensity and the borrowability scales outlined in the introduction (Thomason & Kaufman, 1988, pp. 74 ff.; Thomason, 2001, pp. 70 ff.). Notably, the dialectometric analysis aligns remarkably well with Dawkins' (1916, pp. 204, 209) non-dialectometric continua of Turkish influence (see Section 2.1. for details; see also Figure 2). Conversely, lexical phenomena exhibit a reverse effect, with reduced differences among subdialects and increased differences among the three main dialect groups. This observation is not unexpected, as lexical borrowing can occur even in the absence of bilingualism (Thomason & Kaufman, 1988, pp. 77 ff.), which may explain the divergent outcomes in the analyses for grammatical and lexical phenomena (see also below for more details).

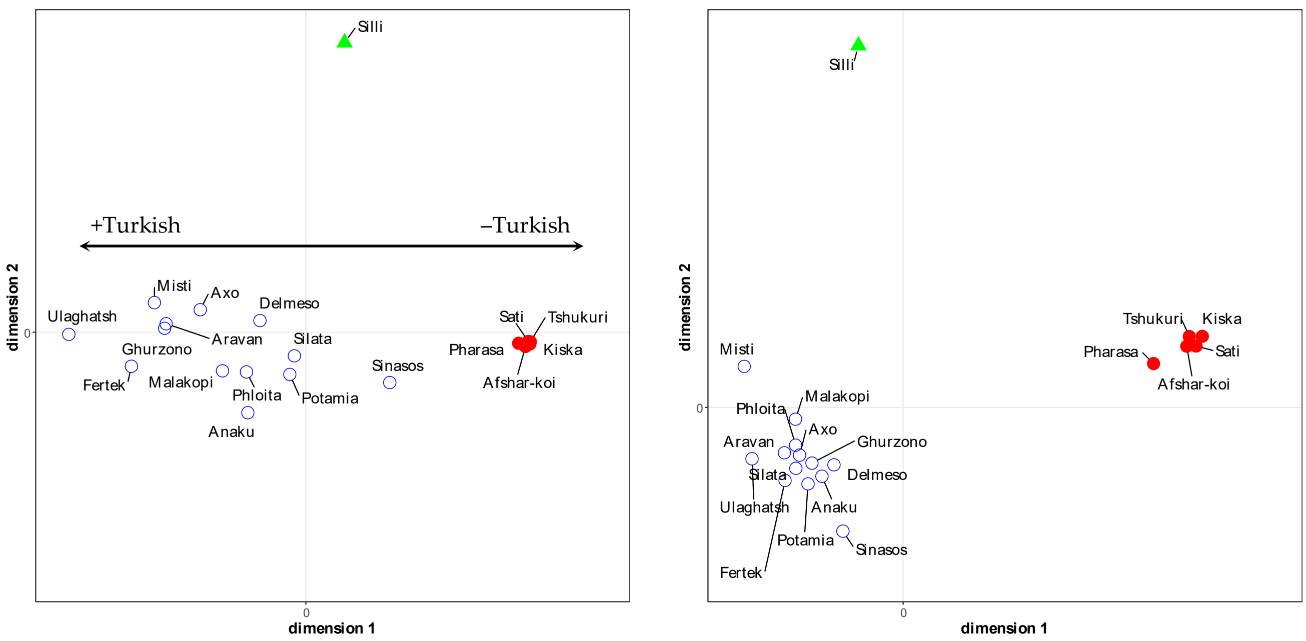


Figure 4. Two-dimensional plots for grammatical distances (left; variance explained by Kruskal's 2D MDS: 92.5%) and lexical distances (right; variance explained by Kruskal's 2D MDS: 86.7%).

3.2. Multiple Regression on Distance Matrices

In this section, an MRM analysis was conducted to investigate (i) the role of geographic, demographic, and sociolinguistic context in shaping the grammatical and lexical distances between the communities and (ii) whether the theoretically proposed extra-linguistic variables contribute equally to the observed grammatical and lexical distances. This approach allowed for a more nuanced understanding of the interplay between linguistic and extra-linguistic factors in the evolution of iAMGr dialects.

Table 5 presents a summary of the results of MRM for lexical and grammatical distances separately. It compares the estimated effect size (Estimate) and statistical significance (*p*-value < 0.05) of each explanatory variable. A higher absolute value of a coefficient indicates a stronger effect for the corresponding variable. The last row provides the R^2-values obtained through the MRM analyses for each dataset. Specifically, the R^2 measures the amount of variation in the dependent variable accounted for by the model.

Table 5. Standardized coefficients and R^2-values based on the MRM analyses across grammatical and lexical distances.

	Lexical Distances		**Grammatical Distances**	
	Estimate	***p***	**Estimate**	***p***
Intercept	0.00	<0.0001	0.00	<0.0001
Geographic Distance (km)	0.60	<0.0001	0.19	<0.0001
Variety Type	0.19	<0.0001	0.29	<0.0001
Turkish Population in the Communities	−0.10	0.16	0.14	0.02
Turkish Population in the Provinces	0.34	<0.0001	0.60	<0.0001
Greek Schooling	0.12	<0.0001	0.02	0.57
$\mathbf{R^2}$	0.78		0.84	

Starting with geographic distance, it has a statistically significant positive effect on both the lexical and grammatical distances. However, the effect size is considerably larger for lexical distances (0.60) compared to grammatical distances (0.19). This suggests that geography plays a more prominent role in driving lexical divergence than grammatical divergence.

The reverse effect is observed in the case of the Turkish population at the provincial level (Turkish population in the provinces). The effect size of this factor is larger for grammatical distances (0.60) compared to lexical distances (0.34). This finding suggests potential language contact effects, with a greater divergence in provinces with larger Turkish populations, potentially influencing grammatical structures to a greater extent than vocabulary.

However, the effect of the Turkish population within the communities shows a contrasting pattern and does not align with population differences at the provincial level.[8] This factor has a negative and statistically non-significant effect on lexical distances (p-value = 0.16) and a statistically significant but weaker effect on grammatical distances.

Variety type has a statistically significant positive effect on both lexical and grammatical distances. While the effect sizes are of similar magnitude (0.19 for lexical and 0.29 for grammatical), they differ across levels, with the effect on grammatical distances being stronger and ranking as the second most important factor. Moreover, variety type predicts divergence in grammatical structures more effectively than other factors, including geography.

Greek schooling has a statistically significant positive effect on lexical distances but not on grammatical distances (p-value = 0.57). This implies that exposure to standardized education might lead to a convergence of lexical features (e.g., vocabulary usage), but it does not significantly influence grammatical variation.

Figure 5 depicts the relative effects of significant extra-linguistic factors on grammatical and lexical distances. For grammatical distances, the Turkish population in the province exhibits the highest positive estimate among significant factors, suggesting that larger Turkish populations correlate with increased grammatical divergence, potentially due to language contact and borrowing. This finding aligns with the importance of variety type (low vs. high contact) as the second most influential factor, although its estimate is lower than that of the provincial Turkish population. Geographic distance and the presence of Turkish populations in the communities have the lowest effect estimates among the significant factors influencing grammatical distances.

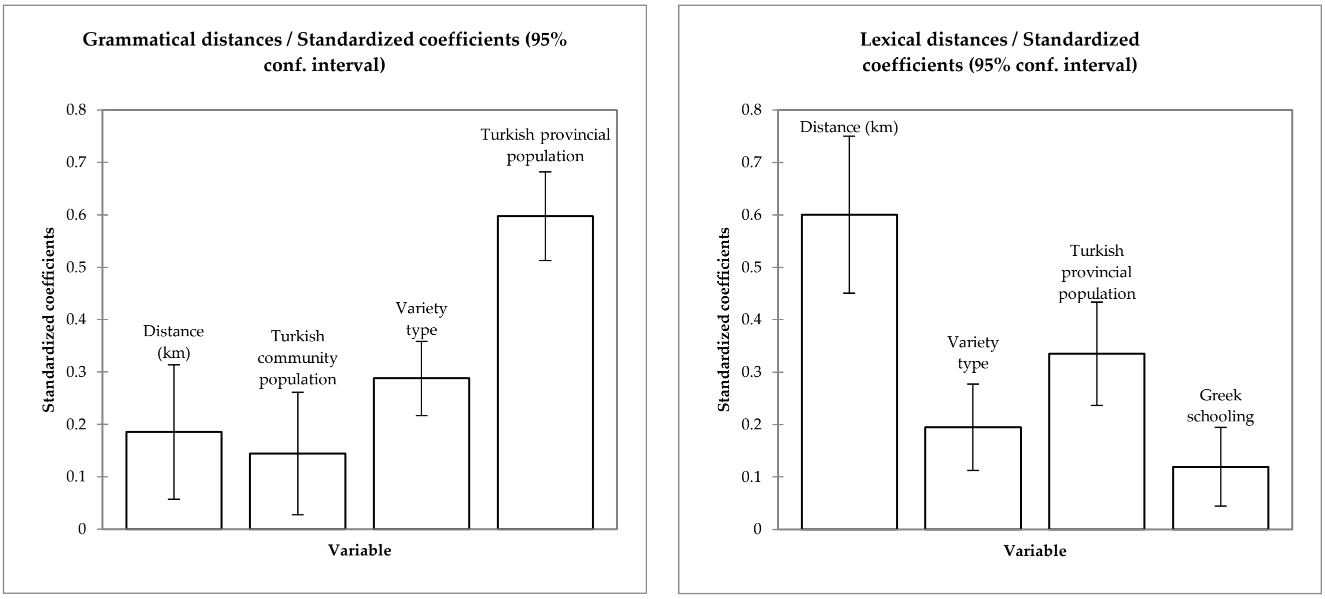


Figure 5. Chart of standardized coefficients for comparing the relative influence of the significant predictors on grammatical (**right**) and lexical (**left**) distances (based on Table 5).

In contrast, for lexical distances, geographic distance emerges as the strongest driver of divergence in this analysis, followed by the Turkish population in the provinces. Preexisting dialectal boundaries (i.e., variety type) and Greek schooling also contribute, but to a lesser extent.

The R-squared values provide valuable insights into the explanatory power of each model. The model for lexical distances, with an R-squared of 0.78, indicates that the chosen factors account for approximately 78% of the observed variance in lexical distances. The model for grammatical distances boasts an even higher R-squared value of 0.84, explaining nearly 84% of the variance in grammatical distances. These results demonstrate the effectiveness of dialectometry in accounting for a substantial portion of observed variation. Notably, these models surpass the explanatory power of geographic distance alone.

4. Discussion

This study explored the impact of extra-linguistic factors on grammatical and lexical distances within the iAMGr dialect continuum in Asia Minor. In line with previous research that challenges the significance of geographic distance in language contact situations (Kortmann, 2013), we focused on seven extra-linguistic factors, including geographic distance between the studied communities, which have been examined as potential predictors of linguistic distance in the iAMGr dialectal continuum (Dawkins, 1916; Thomason & Kaufman, 1988; Thomason, 2008; Karantzola et al., 2021). Using regression analysis, we created a statistical model to test the influence of these factors on grammatical and lexical distances. Our findings support Kortmann's (2013) observation that factors beyond geography can be more effective predictors of dialectal distances in language contact situations. However, the influence of these factors appears to be dependent on the linguistic level (grammatical vs. lexical) under scrutiny.

Starting with an overview of the results for grammatical and lexical distances, our analysis reveals different patterns. Grammatical distances show a wider range of variation compared to lexical distances, as reflected in the higher maximum, median, and mean values. The larger standard deviation for grammatical distances indicates a greater spread of values around the mean. This suggests that grammatical features are more dynamic and evolve across the communities. On the other hand, lexical distances appear to be more tightly clustered, as further illustrated using MDS, suggesting a higher degree of convergence across the dialects. These results are consistent with Adamou's (2016, p. 162) empirical finding that changes in grammar and lexicon in language contact settings occur somewhat independently. We further examined this using a regression analysis.

Before conducting the MRM analysis, we performed a (partial) Mantel test to include only those variables that showed a significant correlation with linguistic differences. Surprisingly, proximity to urban centers and contact with Constantinople did not significantly correlate with linguistic distances for either grammar or lexicon, regardless of controlling for geographical distance.

Contrary to previous theoretical assumptions (Dawkins, 1916, p. 13; Thomason & Kaufman, 1988, p. 216), proximity to Turkish-speaking urban centers did not significantly predict linguistic distances. Karantzola et al. (2021, pp. 48–49) suggest that, for communities in Cappadocia, this factor is mediated by other factors, such as isolation, which limit its influence. It is worth noting that previous quantitative studies using gravity models have observed similar diffusion patterns. Trudgill (1974; 1986, Chapter 3) and Chambers and Trudgill (1998, Chapters 11.7–11.8) have observed that linguistic changes tend to skip from one urban center to another, leaving rural areas unaffected until the late stages of the change (e.g., Britain, 2012, pp. 2035–2036; Wolfram & Schilling-Estes, 2017, pp. 725–726 and the references therein). Considering that most varieties in our dataset have only a few hundred speakers and that the influence of urban centers is highly dependent upon population density, it is reasonable to assume that the effect of urban centers is naturally limited.

Additionally, the recent historical timeframe of Constantinople's influence on iAMGr varieties weakens its association with dialect distances. Historically, the migration of predominantly male populations within Ottoman territory facilitated linguistic contact between speakers of different Greek dialects with Turkish and puristic/common Greek in Constantinople (Karantzola et al., 2021, pp. 49–51). However, migration increased notably in the 18th and 19th centuries due to economic hardships and insecurity in rural areas (Dawkins, 1916, pp. 14, 23; Karatsareas, 2011, p. 18; Manolessou, 2019, p. 27). Consequently, this factor is unlikely to have a significant direct impact on the observed linguistic distances, given that it has occurred relatively recently in the linguistic history of the iAMGr varieties.

Turning to the results of the MRM analysis, we tested the relative influence of the remaining extra-linguistic factors on linguistic distances. The influence of Greek schooling had no significant effect on grammatical distances, and it had the least impact on lexical distances. Exposure to common Greek through education intensified for Asia Minor enclaves after the formation of the first Greek state and its reconnection to the Greek-speaking communities of Cappadocia in the 1830s (Karatsareas, 2011, p. 20 and the references therein). Furthermore, although Greek schools were established in many communities during this period, it remains uncertain whether they taught Greek in its puristic form (the language of literacy) or in common Greek. Regardless, the linguistic distance between the two varieties and iAMGr was considerable, despite the continued use of the puristic form in church texts (Karatsareas, 2011, p. 20). Thus, Greek schools did not significantly affect linguistic distances, refuting the proposition of Thomason and Kaufman (1988, pp. 216, 356). Although Dawkins (1916, pp. 27–28) observed a gradual replacement of the variety of Sinasos by common Greek, which was widely spoken and taught in its schools (see also Logotheti-Merlier, 1977, p. 51; Karantzola et al., 2021, p. 55), the interaction between common/puristic Greek and iAMGr dialects (the language of domestic and social domains) was predominantly competitive in other communities (Theodoridi, 2017, pp. 553–559; Theodoridi & Karantzola, 2019).

Similarly, what our analyses revealed is that the presence of a Turkish population in the communities did not significantly impact lexical distances, while it had the least influence on grammatical distances. The effect of population sizes within iAMGr communities has been previously analyzed in studies conducted by Dawkins (1916, p. 209), Theodoridi (2017, p. 534), and Karantzola et al. (2021, p. 48). These qualitative approaches have also revealed an intriguing 'paradox' that becomes apparent in certain communities, such as Axo and Misti, where a certain degree of Turkish influence is evident despite the absence of a Turkish-speaking population. In fact, our findings indicate that the size of the Turkish population is the most significant predictor of grammatical variation in iAMGr, especially when considering the regional context. This is not surprising, considering that this regional context can be viewed as a politically significant construct (Grenoble & Whaley, 1998, p. 39) or a relatively closed social network (Nerbonne, 2013, p. 237), where cultural and socioeconomic pressures primarily impact Greek-speaking communities due to their politically and numerically dominant Turkish population (Thomason, 2001, p. 66). Consequently, the size of the Turkish-speaking population in the entire region has a greater impact.

Figure 6 illustrates the close correspondence between the relative size of the Turkish population at the provincial level and the proposed degrees of Turkish influence. This finding suggests that the demographic dominance of Turkish speakers within a province plays a crucial role in shaping the extent of linguistic influence on smaller communities. This pattern also explains paradoxical cases like Misti and Axo, as they are administratively located within provinces with the highest Turkish populations. Our findings align with the previous research demonstrating that language contact is a multifaceted phenomenon, influenced by a combination of factors beyond the mere coexistence of two linguistic groups within a single community. As Foley (2010, p. 796 and the references therein) has noted, the size of the minority group and its proximity to other communities speaking a different language are key determinants. In the context of our study, a larger Turkish population dispersed across a province increases opportunities for interaction and communication with smaller minority groups, fostering bilingualism and facilitating the transfer of linguistic elements. The case of Delmeso presents an important deviation from this pattern, highlighting the influence of geographical factors on language contact. Delmeso's geographical isolation likely limits opportunities for interaction with the dominant Turkish population, thus mitigating the expected degree of linguistic influence (Dawkins, 1916, p. 13; Theodoridi, 2017, p. 184; Karantzola et al., 2021, p. 47).

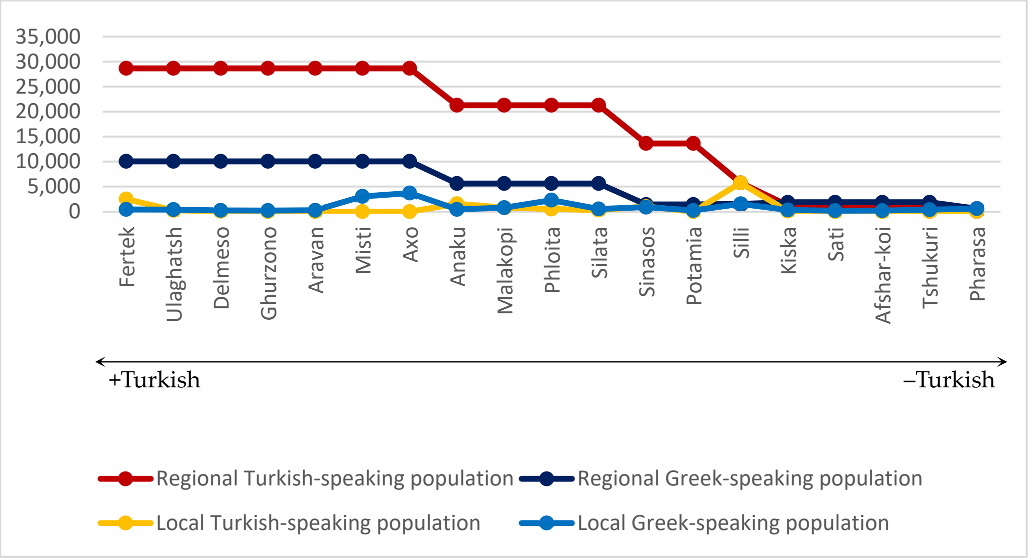


Figure 6. Community and provincial population sizes per site in relation to the assumed degree of Turkish influence in iAMGr.

In contrast to grammatical distances, the influence of the provincial Turkish population is less evident in the lexical distances. This divergence likely stems from the inherent differences in borrowability between structural and lexical elements, as highlighted by Thomason (2001, pp. 69, 70–71). Lexical borrowing, notably, can occur even without bilingualism (Thomason & Kaufman, 1988, pp. 77 ff.), making it prevalent in both isolated communities and those not experiencing a language shift.

This observation explains why contact intensity (i.e., variety type) emerged as the second most influential factor for grammatical distances (consistent with Kortmann, 2013) but not for lexical distances. The weaker correlation with contact intensity for lexical distances is likely due to the feasibility of borrowing without extensive bilingualism, aligning with established borrowability scales (Thomason & Kaufman, 1988, pp. 50, 74–76; Thomason, 2001, pp. 70–71). As a result, lexical differences are more pronounced between the three main dialect clusters, whose distinctions primarily lie in loanword integration and language-internal mechanisms (Melissaropoulou, 2016a, 2016b). Consequently, lexical distances are significantly influenced by geographical distance, due to their strong dependence on language-internal mechanisms and/or dialect interference. Conversely, geography exerts a weaker influence on grammatical distances, as these often reflect instances, in terms of foreign interference, resulting from direct contact with the dominant Turkish language.

Finally, the high R^2 values obtained in our models emphasize the significance of considering various factors, in addition to geographic distance, in order to gain a deeper understanding of the intricate dynamics of dialect variation, especially in situations involving contact. Importantly, our findings question the idea that geography is the only or main factor influencing linguistic distances, especially in cases of foreign interference. In such situations, it becomes crucial to take into account factors that reflect the social meanings and dynamics of the area being studied (Stanford, 2012).

5. Conclusions

Our findings highlight the intricate relationship between sociolinguistic and geo-demographic factors that shape linguistic variation in language contact scenarios, as exemplified by the iAMGr dialect continuum. While geographic distance remains significant, our analysis reveals that population size and contact intensity have a stronger influence, particularly on grammatical distances. This is likely due to the association between the increasing Turkish population and increased contact intensity, which leads to structural borrowing. On the other hand, lexical distances seem to be more influenced by geographic distance, as lexical borrowing may occur independently of contact intensity. The observed dissociation between grammatical and lexical distances emphasizes the need to consider domain-specific effects in dialectometric analyses, as the impact of language contact and other extra-linguistic variables may vary depending on the linguistic level being studied (Spruit et al., 2009; Scherrer & Stoeckle, 2016; Bompolas & Melissaropoulou, 2023a; Bompolas, 2023).

The R^2 values, ranging from 0.78 to 0.84, indicate that the models explain a substantial proportion of the variance in the observed data. This suggests that the factors included in the study strongly contribute to the shaping of the observed patterns, highlighting the effectiveness of dialectometry in accounting for dialectal variation. It is worth noting that these models explain a greater proportion of the variance compared to geographic distance alone, underscoring the importance of incorporating additional sociolinguistic and demographic variables. However, it is important to acknowledge that the remaining unexplained variance may be attributed to other factors not examined in this study. Previous theoretical research has identified micro- and meso-level variables as potential contributors to linguistic variation in Cappadocian and Pharasiot (Karantzola et al., 2021; Theodoridi, 2017). Although this study did not directly examine these factors due to data limitations, they offer promising avenues for future research to comprehensively investigate their influence on the observed variation.

In conclusion, our findings support the idea that geographic distance can be complemented or even replaced by more historically, culturally, and sociolinguistically specific variables for understanding dialectal variation (Stanford, 2012, p. 252). For smaller regions, it is essential to examine highly localized social meanings instead of simply applying general dialectological principles. However, to enhance the generalizability of these findings, future research should investigate the impact of similar extra-linguistic factors, such as population size and contact intensity, in other language contact settings. This would help determine the extent to which the patterns observed in iAMGr are applicable across diverse contexts, thus contributing to a more comprehensive understanding of the complex interplay between linguistic and extra-linguistic factors in dialect distances (Nerbonne, 2010, p. 3828).

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/languages10010013/s1.

Author Contributions: Conceptualization, S.B. and D.M.; methodology, S.B.; software, S.B.; validation, S.B.; formal analysis, S.B.; investigation, S.B.; resources, D.M.; data curation, D.M. and S.B.; writing—original draft preparation, S.B. and D.M.; writing—review and editing, S.B. and D.M.; visualization, S.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the 'Andreas Mentzelopoulos Scholarships for postgraduate studies at the University of Patras' (2019–2023) [to S.B.]; dialectal resources were provided by the project 'Digitizing the Cappadocian Dialectal Landscape', supported by the General Secretariat for Research and Technology and the Hellenic Foundation for Research and Innovation (2018–2022) [to D.M.].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are the same as those provided in the Supplementary Materials, with additional details available in the main text.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

[1] The FDP states that "geographically proximate varieties tend to be more similar than distant ones" (Nerbonne & Kleiweg, 2007, p. 154).

[2] This paper investigates the Cappadocian, Pharasiot, and Silliot (sub)varieties of the iAMGr subgroup within the broader Asia Minor Greek dialect continuum (cf. Manolessou, 2019). The term 'inner Asia Minor Greek' (iAMGr) is used as a convenient cover term for these varieties in line with Kontosopoulos' (1981/2008, pp. 6–10) without assuming a specific genetic affiliation among them in the exclusion of Pontic (Manolessou, 2019, pp. 20–21, 29–40), although belonging to the iAMGr group, which is not the focus of this study.

[3] The underlying causes of language change in iAMGr have been a subject of debate among scholars. Some argue that language change is driven by inherent linguistic forces (e.g., Karatsareas, 2011). Others argue that language change is primarily the result of language contact (e.g., Thomason & Kaufman, 1988; Janse, 2001, 2019; Melissaropoulou, 2012; Melissaropoulou, 2016a, 2016b). According to this perspective, the main distinction lies in the source of the change. In contact-induced change, the source is the influence of another language, whereas, for language-internal change, the source is the structural asymmetries within a single linguistic system (Thomason, 2001, p. 86). The objective of this paper is not to take a stance in this debate but rather to examine the intricate interactions between the extralinguistic factors and the dialect distances.

[4] Given this lack of evidence, more recent studies draw material from refugees' interviews carried out between 1930 and 1975 in Greece (Theodoridi, 2017; Theodoridi & Karantzola, 2019; Karantzola et al., 2021).

[5] The Greco-Turkish War (1920–1922) led to the end of the Greek presence in Asia Minor. In the aftermath of the war, the *Convention Concerning the Exchange of Greek and Turkish Populations* was signed by the Greek and Turkish governments in Lausanne (Switzerland) on 30 January 1923. This convention mandated a compulsory population exchange based on the religious affiliation of Turkish nationals of the Greek Orthodox faith residing in Turkey and Greek nationals of the Muslim faith residing in Greece. These individuals were prohibited from returning to live in Turkey or Greece, respectively, without authorization from their respective governments. Consequently, the Greek speakers of Asia Minor were displaced from their eastern homelands and relocated mainly to the newly acquired northern parts of Greece as refugees.

[6] *Lexical distances* in this study refer to the variation observed at the level of lexical items, encompassing, among others, phonetic variants, morphophonological adaptations, and the integration of borrowings, while also including comparisons of both cognate and non-cognate items. This approach goes beyond simple measures of vocabulary overlap, such as Hamming distance, which is limited to binary comparisons based on specific lexical forms. By employing PMI Levenshtein distance, we effectively capture these variations by quantifying differences in the phonetic and morphophonological structure of lexical (and sub-lexical) items. This provides a precise measure of lexical variation that reflects the linguistic complexities of iAMGr.

[7] As argued by Karatsareas (2020, pp. 184–185) present-day Modern Greek and Turkish, especially their standard forms, may not be the suitable reference varieties to compare when attempting to determine the causes of change observed in iAMG. What is more, Greek and Turkish were not the only languages constituting the so-called '(Graeco-)Anatolian Sprachbund' (for an overview, see Donabedian & Sitaridou, 2021). Therefore, one would ideally want to compare the iAMGr data with data derived from varieties of Greek and Turkish that are closer to iAMGr from a historical and/or geographical perspective. However, the almost complete absence of texts written in iAMGr in dialectal Greek or Turkish in the period before the 19th century makes this type of comparison unfeasible.

[8] We addressed the possibility of collinearity between the two population variables by employing the Mantel test to assess the correlation between their matrices. The results ($r = -0.0005$, $p = 0.9871$, $\alpha = 0.05$) indicate no significant collinearity, affirming the independence of the variables.

References

Adamou, E. (2016). *A corpus-driven approach to language contact: Endangered languages in a comparative perspective*. De Gruyter. [CrossRef]

Alektoridis, A. S. (1883). *Λεξιλόγιον τοῦ Ἐν Φερτακαίνοις τῆς Καππαδοκίας γλωσσικοῦ Ἰδιώματος* [Vocabulary of the linguistic variety spoken in Fertakena of Cappadocia]. Deltion Istorikis Ethnologikis Etaireias.

Anastasiadis, V. K. (1976). *Ἡ σύνταξη στὸ Φαρασιώτικο Ἰδίωμα τῆς Καππαδοκίας σὲ σύγκριση πρὸς τὰ Ὑπόλοιπα Ἰδιώματα τῆς Μικρᾶς Ἀσίας, καθὼς καὶ πρὸς τὴν Ἀρχαία, τὴ Μεσαιωνικὴ καὶ τὴ Νέα Ἑλληνικὴ γλῶσσα* [Syntax in the Pharasiot variety of Cappadocia in comparison to the other varieties of Asia Minor, as well as to Ancient, Medieval, and Modern Greek language]. University of Ioannina.

Bağrıaçık, M. (2018). *Pharasiot Greek: Word order and clause structure* [Ph.D. thesis, Ghent University].

Bloomfield, L. (1933). *Language*. Holt, Rinehart & Winston.

Bompolas, S. (2023). *Computational dialectology in the linguistic varieties of Cappadocian, Pharasiot, and Silliot* [Ph.D. thesis, University of Patras]. [CrossRef]

Bompolas, S., & Melissaropoulou, D. (2023a). A dialectometric approach to inner Asia Minor Greek: Comparisons and associations between linguistic levels. *Digital Scholarship in the Humanities*, *38*(4), 1389–1403. [CrossRef]

Bompolas, S., & Melissaropoulou, D. (2023b). A first dialectometric approach to contact-induced vs language-internal variation in inner Asia Minor Greek. *Studies in Greek Linguistics*, *42*, 51–62.

Britain, D. (2012). Varieties of English: Diffusion. In A. Bergs, & L. Brinton (Eds.), *Handbücher zur sprach- und kommunikationswissenschaft* [Handbooks of linguistics and communication science 34.2] (pp. 2031–2043). De Gruyter. [CrossRef]

Chambers, J. K., & Trudgill, P. (1998). *Dialectology* (2nd ed.). Cambridge Textbooks in Linguistics. Cambridge University Press.

Dawkins, R. M. (1916). *Modern Greek in Asia minor: A study of the dialects of Sílli, Cappadocia and Phárasa with grammar, texts, translations and glossary*. Cambridge University Press.

Donabedian, A., & Sitaridou, I. (2021). Anatolia. In E. Adamou, & Y. Matras (Eds.), *The Routledge handbook of language contact* (1st ed., pp. 404–433). Series: Routledge handbooks in linguistics: Routledge. Routledge. [CrossRef]

Farasopoulos, S. (1895). *Τὰ Σύλατα. Μελέτη Τοῦ Νομοῦ Ἰκονίου Ὑπὸ Γεωγραφικήν, Φιλολογικὴν Καὶ Ἐθνολογικὴν Ἔποψιν* [Sylata. A study of the region of Ikonio from a geographical, philological, and ethnological perspective]. Deligianni's and Kalergis' Print Shop.

Foley, W. A. (2010). Language contact in the New Guinea region. In R. Hickey (Ed.), *The handbook of language contact* (1st ed., pp. 795–813). Wiley. [CrossRef]

Goebl, H. (2006). Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing*, *21*(4), 411–435. [CrossRef]

Goebl, H. (2017). Dialectometry. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *The handbook of dialectology* (1st ed., pp. 123–142). Wiley. [CrossRef]

Gooskens, C. (2005). Travel time as a predictor of linguistic distance. *Dig*, *2005*(13), 38–62. [CrossRef]

Gooskens, C., & Heeringa, W. (2004). Perceptive evaluation of levenshtein dialect distance measurements using norwegian dialect data. *Language Variation and Change*, *3*, 189–207. [CrossRef]

Goslee, S. C. (2010). Correlation analysis of dissimilarity matrices. *Plant Ecology*, *206*(2), 279–286. [CrossRef]

Goslee, S. C., & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, *22*(7), i07. [CrossRef]

Grenoble, L. A., & Whaley, L. J. (1998). Toward a typology of language endangerment. In L. A. Grenoble, & L. J. Whaley (Eds.), *Endangered languages* (1st ed., pp. 22–54). Cambridge University Press. [CrossRef]

Grieve, J. (2014). A comparison of statistical methods for the aggregation of regional linguistic variation. In B. Szmrecsanyi, & B. Wälchli (Eds.), *Aggregating dialectology, typology, and register analysis* (pp. 53–88). De Gruyter. [CrossRef]

Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. Edited by Luke Harmon. *Methods in Ecology and Evolution*, *4*(4), 336–344. [CrossRef]

Heeringa, W. (2024). *PMI Levenshtein distance*. Available online: https://www.led-a.org/docs/PMI.pdf (accessed on 2 January 2025).

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance* [Ph.D. thesis, University of Groningen].

Heeringa, W., Kleiweg, P., Gooskens, C., & Nerbonne, J. (2006). Evaluation of string distance algorithms for dialectology. In *Proceedings of the workshop on linguistic distances* (pp. 51–62). Association for Computational Linguistics. Available online: https://aclanthology.org/W06-1108 (accessed on 2 January 2025).

Heeringa, W., Nerbonne, J., Niebaum, H., Nieuweboer, R., & Kleiweg, P. (2000). Dutch-German Contact in and around Bentheim. *Studies in Slavic and General Linguistics*, *28*, 145–156.

Heeringa, W., Nerbonne, J., & Osenova, P. (2010). Detecting contact effects in pronunciation. In M. Norde, B. de Jonge, & C. Hasselblatt (Eds.), *IMPACT: Studies in language and society* (Vol. 28, pp. 131–154). John Benjamins Publishing Company. [CrossRef]

Heeringa, W., & Prokić, J. (2017). Computational Dialectology. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *The handbook of dialectology* (1st ed., pp. 330–347). Wiley. [CrossRef]

Heeringa, W., van Heuven, V., & Van de Velde, H. (2022, August 1–5). *LED-A: A web app for measuring distances in the sound components among local dialects*. Seventeenth International Conference on Methods in Dialectology (Methods XVII), Johannes Gutenberg-University Mainz, Mainz, Germany. Available online: https://www.led-a.org/slides.pdf (accessed on 2 January 2025).

Heeringa, W., Van Heuven, V., & Van de Velde, H. (2023). *LED-A: Levenshtein edit distance app* [Computer Program]. Available online: https://www.led-a.org/ (accessed on 2 January 2025).

Honkola, T., Ruokolainen, K., Syrjänen, K. J. J., Leino, U.-P., Tammi, I., Wahlberg, N., & Vesakoski, O. (2018). Evolution within a language: Environmental differences contribute to divergence of dialect groups. *BMC Evolutionary Biology*, *18*(1), 132. [CrossRef]

Huisman, J. L. A., Franco, K., & van Hout, R. (2021). Linking linguistic and geographic distance in four semantic domains: Computational geo-analyses of internal and external factors in a dialect continuum. *Frontiers in Artificial Intelligence*, *4*, 668035. [CrossRef] [PubMed]

ILIK. (2024). *Ιστορικό Λεξικό Των Ιδιωμάτων Της Καππαδοκίας* [Historical dictionary of Cappadocian varieties]. Athens Academy.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* [Springer texts in statistics]. Springer. [CrossRef]

Janse, M. (2001). Morphological borrowing in Asia Minor. In *International conference on Greek Linguistics, 4th, proceedings* (pp. 473–479). University Studio Press.

Janse, M. (2008). Clitic Doubling from ancient to Asia Minor Greek. In D. Kallulli, & L. Tasmowski (Eds.), *Linguistik aktuell/Linguistics today* (Vol. 130, pp. 165–202). John Benjamins Publishing Company. [CrossRef]

Janse, M. (2019). Agglutinative noun inflection in Cappadocian. In A. Ralli (Ed.), *The morphology of Asia Minor Greek* (pp. 66–115). Brill. [CrossRef]

Jeszenszky, P., Hikosaka, Y., Imamura, S., & Yano, K. (2019). Japanese lexical variation explained by spatial contact patterns. *ISPRS International Journal of Geo-Information*, *8*(9), 400. [CrossRef]

Jeszenszky, P., Stoeckle, P., Glaser, E., & Weibel, R. (2017). Exploring global and local patterns in the correlation of geographic distances and morphosyntactic variation in Swiss German. *Journal of Linguistic Geography*, *5*(2), 86–108. [CrossRef]

Karantzola, T., Theodoridi, A., & Sampanis, K. (2021). The interplay of external and sociolinguistic factors in contact-induced language change: Cappadocian Greek as a case study. *Mediterranean Language Review*, *28*, 21. [CrossRef]

Karasimos, A., Manolessou, I., & Melissaropoulou, D. (2020). Creating a DTD template for Greek dialectal lexicography: The case of the historical dictionary of the Cappadocian dialect. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Lexicography for inclusion: Proceedings of the 19th EURALEX international congress, 7–9 September 2021, Alexandro upolis* (Vol. 1, pp. 305–314). Democritus University of Thrace. Available online: https://www.euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_Vol1-p305-314.pdf (accessed on 2 January 2025).

Karatsareas, P. (2011). *A study of Cappadocian Greek nominal morphology from a diachronic and dialectological perspective* [Ph.D. thesis, University of Cambridge]. [CrossRef]

Karatsareas, P. (2020). The development, preservation and loss of differential case marking in inner Asia Minor Greek. *Journal of Language Contact*, *13*(1), 177–226. [CrossRef]

Kholopoulos, S. (1905). Μονογραφική Ἱστορία Ζήλης ἢ Σύλατας [Monographic history of Zili or Sylata]. *Xenophánis*, *II*, 92–96.

Kitromilidis, P., & Mourelos, G. (Eds.). (1982). *Η Έξοδος, Τόμος Β. Μαρτυρίες Από Τις Επαρχίες Της Κεντρικής Και Νότιας Μικρασίας* [The Exodus, volume b. Testimonies from the provinces of Central and Southern Asia Minor]. Kentro Mikrasiatikōn Spoudōn.

Kontosopoulos, N. (2008). Διάλεκτοι *Καὶ* Ἰδιώματα *Τῆς Νέας* Ἑλληνικῆς [Dialects and idioms of Modern Greek]. Grigoris. (Original work published 1981).

Kortmann, B. (2013). How powerful is geography as an explanatory factor in morphosyntactic variation? Areal features in the anglophone world. In P. Auer, M. Hilpert, A. Stukenbrock, & B. Szmrecsanyi (Eds.), *Space in language and linguistics* (pp. 165–194). De Gruyter. [CrossRef]

Leinonen, T., Çöltekin, Ç., & Nerbonne, J. (2016). Using Gabmap. *Lingua*, *178*, 71–83. [CrossRef]

Lichstein, J. W. (2007). Multiple regression on distance matrices: A multivariate spatial analysis tool. *Plant Ecology*, *188*(2), 117–131. [CrossRef]

List, J.-M. (2019). Automated Methods for the Investigation of Language Contact, with a Focus on Lexical Borrowing. *Language and Linguistics Compass*, *13*(10), e12355. [CrossRef]

Logotheti-Merlier, M. (1977). Οι Ελληνικές Κοινότητες Στη Σύγχρονη Καππαδοκία [The Greek communities in Modern Cappadocia]. *Bulletin of the Centre for Asia Minor Studies*, *1*, 29–74. [CrossRef]

Manolessou, I. (2019). The Historical Background of the Asia Minor Dialects. In A. Ralli (Ed.), *The morphology of Asia Minor Greek* (pp. 20–65). Brill. [CrossRef]

Manolessou, I., Karasimos, A., & Katsouda, G. (2022). Retro-digitization in Greek dialectology and lexicography: Challenges of morpho-phonetic representation of the Cappadocian dialect. *Modern Greek Dialects and Linguistics Theory*, *9*, 322–336.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*(2), 209–220. [PubMed]

Melissaropoulou, D. (2012). Reorganization of grammar in the light of the language contact factor: A case study on Grico and Cappadocian. *Modern Greek Dialects and Linguistics Theory*, *5*, 311–334. [CrossRef]

Melissaropoulou, D. (2016a). Loanwords integration as evidence for the realization of gender and inflection class: Greek in Asia Minor. In A. Ralli (Ed.), *Contact morphology in modern Greek dialects* (pp. 145–177). Cambridge Scholars Publishing.

Melissaropoulou, D. (2016b). Variation in word formation in situations of language contact: The case of Cappadocian Greek. *Language Sciences*, *55*, 55–67. [CrossRef]

Melissaropoulou, D. (2024). *Γλωσσικός Άτλαντας Των Διαλεκτικών Ποικιλιών Της Καππαδοκίας (5 Τόμ.)* [Linguistic atlas of the dialectal varieties of Cappadocia] (Vol. 5). Academy of Athens.

Melissaropoulou, D., & Bompolas, S. (2022). Μια Υπολογιστική Γεωγλωσσολογική Προσέγγιση Στις Γλωσσικές Ποικιλίες Της Καππαδοκίας [A computational geolinguistic approach to the linguistic varieties of Cappadocia]. *Modern Greek Dialects and Linguistics Theory*, *9*, 221–244. [CrossRef]

Melissaropoulou, D., Bompolas, S., & Tsimpouris, C. (2022). Digital cartography in the service of preservation of cultural linguistic heritage: Implementing the electronic dialectal atlas of Cappadocian Greek. *Scientific Culture*, *8*(2), 135–146. [CrossRef]

Nerbonne, J. (2010). Measuring the diffusion of linguistic change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1559), 3821–3828. [CrossRef]

Nerbonne, J. (2013). How much does geography influence language variation? In P. Auer, M. Hilpert, A. Stukenbrock, & B. Szmrecsanyi (Eds.), *Space in language and linguistics* (pp. 222–239). De Gruyter. [CrossRef]

Nerbonne, J., & Heeringa, W. (2001). Computational comparison and classification of dialects. *Dig*, *2001*(9), 69–84. [CrossRef]

Nerbonne, J., & Heeringa, W. (2007). Geographic distributions of linguistic variation reflect dynamics of differentiation. *Studies in Generative Grammar*, *96*, 267–298.

Nerbonne, J., & Kleiweg, P. (2003). Lexical distance in LAMSAS. *Computers and the Humanities*, *37*(3), 339–357. [CrossRef]

Nerbonne, J., & Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics*, *14*(2–3), 148–166. [CrossRef]

Nerbonne, J., & Wieling, M. (2017). Statistics for aggregate variationist analyses. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *The Handbook of dialectology* (1st ed., pp. 400–414). Wiley. [CrossRef]

Nerbonne, J., Colen, R., Gooskens, C. S., Leinonen, T., & Kleiweg, P. (2011). Gabmap—A web application for dialectology. *Dialectologia*, *SI II*, 65–89.

Prokić, J., & Nerbonne, J. (2008). Recognising groups among dialects. *International Journal of Humanities and Arts Computing*, *2*(1–2), 153–172. [CrossRef]

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available online: https://www.R-project.org/ (accessed on 2 January 2025).

Sarantidis, A. I. (1899). *Ἡ Σινασός, Ἤτοι Θέσις, Ἱστορία, Ἠθικὴ Καὶ Διανοητικὴ Κατάστασις, Ἤθη, Ἔθιμα Καὶ Γλῶσσα Τῆς Ἐν Καππαδοκίᾳ Κωμοπόλεως Σινασοῦ. Ἐν Ἐπιμέτρῳ Δὲ Καὶ Σύντομος Περιγραφὴ Τῶν Ἐν Ταῖς Ἐπαρχίαις Καισαρείας Καὶ Ἰκονίου Ἑλληνικῶν Κοινοτήτων Ὡς Καὶ Τῶν Ἐν Αὐταῖς Σῳζομένων Ἑλληνικῶν Διαλέκτων Ἐν Σχέσει Πρὸς Τὴν Ἐν Σινασῷ Λαλουμένην* [Sinasos, or the position, history, ethical and intellectual state, customs, traditions, and language of the town of Sinasos in Cappadocia. Additionally, a brief description of the Greek communities in the provinces of Caesarea and Iconium, as well as the preserved Greek dialects spoken in them in relation to the dialect spoken in Sinasos]. Ioannis Nikolaidis.

Scherrer, Y., & Stoeckle, P. (2016). A quantitative approach to Swiss German—Dialectometric Analyses and comparisons of linguistic levels. *Dialectologia et Geolinguistica*, *24*(1), 92–125. [CrossRef]

Séguy, J. (1971). La relation entre la distance spatiale et la distance lexicale. *Imprimerie Protat Frères*. [CrossRef]

Shackleton, R. G. (2005). English-American speech relationships: A quantitative approach. *Journal of English Linguistics*, *33*(2), 99–160. [CrossRef]

Shackleton, R. G. (2007). Phonetic variation in the traditional English dialects: A computational analysis. *Journal of English Linguistics*, *35*(1), 30–102. [CrossRef]

Smouse, P. E., Long, J. C., & Sokal, R. R. (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, *35*(4), 627. [CrossRef]

Snoek, C. (2014). Review of Gabmap: Doing dialect analysis on the web. *Language Documentation & Conservation*, *8*, 192–208.

Sousa, X., & García, F. D. (2020). Measuring language contact in geographical space—Sprachkontakt im geographischen raum messen. *Zeitschrift Für Dialektologie Und Linguistik*, *87*(2), 285–306. [CrossRef]

Spruit, M. R., Heeringa, W., & Nerbonne, J. (2009). Associations among linguistic levels. *Lingua*, *119*(11), 1624–1642. [CrossRef]

Stanford, J. N. (2012). One size fits all? Dialectometry in a small clan-based indigenous society. *Language Variation and Change*, *24*(2), 247–278. [CrossRef]

Szmrecsanyi, B. (2012). Geography is overrated. In S. Hansen, C. Schwarz, P. Stoeckle, & T. Streck (Eds.), *Dialectological and folk dialectological concepts of space* (pp. 215–231). De Gruyter. [CrossRef]

Theodoridi, A. (2017). *Καππαδοκικές Διάλεκτοι και Φαρασιωτική: Κοινωνιογλωσσικά και Δομικά Στοιχεία της Γλωσσικής Επαφής τους με την Τουρκική* [Cappadocian dialects and Pharasiot: Sociolinguistic and structural elements of their linguistic contact with turkish] [Ph.D. thesis, University of Aegean]. [CrossRef]

Theodoridi, A., & Karantzola, E. (2019). Επαφή Των Καππαδοκικών Διαλέκτων Με Την Τουρκική: Τα Πεδία Της Εκπαίδευσης Και Της Θρησκείας [Contact of Cappadocian dialects with Turkish: The fields of education and religion]. In C. Tzitzilis, & G. Papanastasiou (Eds.), *Language contact in the Balkans and Asia Minor* (Vol. 1, pp. 92–110). Institute of Modern Greek Studies [Manolis Triandaphyllidis Foundation].

Thomason, S. (2008). Social and linguistic factors as predictors of contact-induced change. *Journal of Language Contact*, *2*(1), 42–56. [CrossRef]

Thomason, S. G. (2001). *Language contact*. Edinburgh Univ. Press.

Thomason, S. G., & Kaufman, T. (1988). *Language contact, creolization, and genetic linguistics*. University of California Press.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society*, *3*(2), 215–246. [CrossRef]

Trudgill, P. (1986). *Dialects in contact*. B. Blackwell.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2012). *Regression methods in biostatistics: Linear, logistic, survival, and repeated measures models* (Statistics for biology and health). Springer. [CrossRef]

Wieling, M. B. (2012). *A quantitative approach to social and geographical dialect variation* [Ph.D. thesis, University of Groningen].

Wieling, M., & Nerbonne, J. (2015). Advances in dialectometry. *Annual Review of Linguistics*, *1*(1), 243–264. [CrossRef]

Wieling, M., Montemagni, S., Nerbonne, J., & Baayen, R. H. (2014). Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, *90*(3), 669–692. [CrossRef]

Wieling, M., Valls, E., Baayen, R. H., & Nerbonne, J. (2018). Border effects among Catalan dialects. In D. Speelman, K. Heylen, & D. Geeraerts (Eds.), *Mixed-effects regression models in linguistics* (pp. 71–97). Quantitative methods in the humanities and social sciences. Springer International Publishing. [CrossRef]

Wolfram, W., & Schilling-Estes, N. (2017). Dialectology and linguistic diffusion. In B. D. Joseph, & R. D. Janda (Eds.), *The handbook of historical linguistics* (pp. 713–735). Blackwell Publishing Ltd. [CrossRef]

Yakpo, K. (2021). Social factors. In E. Adamou, & Y. Matras (Eds.), *The Routledge handbook of language contact* (1st ed., pp. 129–146). Series: Routledge handbooks in linguistics: Routledge. Routledge. [CrossRef]

Zhang, L., Fabri, R., Nerbonne, J., & Nerbonne, J. (2021). Detecting loan words computationally. In E. O. Aboh, & C. B. Vigouroux (Eds.), *Contact language library* (Vol. 59, pp. 269–288). John Benjamins Publishing Company. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.