Handling Cross-Dialect Syntactic Variation: a Theory-Driven Web Resource

Emanuela Li Destri[1], Marco Longhin[1], Gaia Sorge[2], Sofia Ferroni[3], Giovanni B. Matteazzi[4], Andrea Artioli[2], Lorenzo Carletti[2], Federico Motta[2], Giuseppe Longobardi[5], Cristina Guardiano[*1,6]

[1]Dipartimento di Comunicazione ed Economia, Università di Modena e Reggio Emilia
[2]Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio Emilia
[3]Dipartimento di Studi Linguistici e Letterari, Università di Padova
[4]Dipartimento di Matematica, Università di Padova
[5]Department of Language and Linguistic Science, University of York
[6]Scuola Universitaria Superiore IUSS, Pavia
[*]cristina.guardiano@unimore.it

Abstract

Cross-dialect syntactic variation tests the limits of comparative analysis, owing to the entanglement of inheritance and contact in dialect systems. Addressing this challenge requires analytical tools combining the theoretical depth of formal models of grammatical competence with quantitative taxonomic techniques. The Parametric Comparison Method (PCM) embodies this integration by quantifying structural similarity across grammars through the comparison of abstract syntactic rules. The method has been shown to achieve a good degree of resolution in dialectal domains, capturing subtle contrasts and yielding configurations aligning with phylogenetic expectations while remaining sensitive to contact-induced convergence. Fully assessing its effectiveness as a resource for the quantitative study of syntactic dialectology, however, requires an infrastructure that ensures systematic data collection, consistent parameter setting, and dedicated computer-based taxonomic analyses. The PCM Hub is a web-based resource designed for this purpose. It integrates guided elicitation, automated parameter-setting procedures, data management, and the computation of distances and automatic classifications within a unified environment. By standardizing the transition from raw linguistic observations to a structured, replicable empirical apparatus, the PCM Hub provides the practical and quantitative support necessary to test the power of the PCM across expanded comparative domains.

Keywords: Web Resources, Data Collection and Analysis, Dialect Variation, Syntactic Comparison

1. Challenges for Dialect Comparison

Comparative approaches to cross-dialect syntactic diversity (Kayne, 1996, 2005; Poletto, 2012; Ledgeway, 2000; Ledgeway et al., 2018, 2020, a.o.) have demonstrated the relevance of investigating closely related systems to understand the minimal mechanisms underlying grammatical variation and reconstruct pathways of change by disentangling phylogenetic inheritance from interference.

In the framework of the Parametric Comparison Method (PCM; Longobardi and Guardiano, 2009), these issues converge in what is termed the 'ultralocality' problem (Guardiano et al., 2021, 146). The PCM, conceived for purposes of long-term historical reconstruction and phylogenetic inference, quantifies relations between languages through the comparison of syntactic *parameters* (Chomsky, 1981). The core assumption is that the distribution of parametric variation carries historical significance.

A first point of inquiry concerns the level of resolution achieved by parametric systems. The question is whether a system capable to encode syntactic variation at the *universal* level (Guardiano et al.'s 2021 *globality* problem) is also able to capture minimal differences among languages whose degree of relatedness stems from both 'vertical' (inheritance) and 'horizontal' (contact) historical paths. If so, a further question is whether the patterns of divergence identified correlate with a historically meaningful distribution.

These two dimensions have been addressed along two distinct lines of research. A first line investigates aspects of diversity between varieties that share a history of contact and are related to varying degrees of historical depth. Dedicated work has examined the extent to which the variation observed within Italo-Romance (Silvestri, 2013, 2016, 2018, 2020; Guardiano et al., 2022; Guardiano, 2023; Guardiano and Stalfieri, 2026, a.o.), or between Greek and Romance, particularly in Italiot Greek (Guardiano and Stavrou, 2014, 2019a,b, 2020, 2021), reflects vertical inheritance or horizontal diffusion, and how this information can be encoded in terms of parameter settings. A second line has focused on the mathematical modeling of the observed diversity by measuring parameter distances and assessing their distribution (Guardiano et al., 2016, 2021). This line of research has shown that "the PCM, even when applied to groups with minimal internal differentiation, is able to discrim-

inate their articulations" (Guardiano et al., 2021, 172): parameters encode a signal that "is robustly tree-like, even in critical areas of patent language contact" (Guardiano et al., 2016, 152), while also capturing some degree of secondary convergence, whose quantitative assessment, however, still calls for methodological refinement.

2. Assessing Dialect Diversity Through the PCM

A radical point of departure of the PCM from attempts to lexically measure dialect diversity (Pellegrini, 1970; Goebl, 1982; Nerbonne and Kretzschmar, 2003; Nerbonne, 2009; Nerbonne and Heeringa, 2009) or even to use 'structural data' to infer language classifications lies in the nature and theoretical status of the linguistic material employed as taxonomic characters. In the PCM, the relevant taxonomic characters are parameter values rather than words, sounds, or observable structural patterns: languages are compared on the basis of the abstract *rules* that define their grammars rather than on the observable sequences of words or morphemes stemming from them (*manifestations*, Crisma et al., 2020).

In the PCM, the binary states of parameters are symbolized by [+] and [−]: [+] signals that the parameter is active in the language, i.e., that the language displays at least one of its manifestations; [−] signals that the parameter is absent from the grammar, i.e., none of its manifestations is observed.

Checking whether a parameter is active or not in a grammar (I-language, Chomsky, 1986) involves detecting, in the available data, the structures depending on that parameter: this can either be obtained from textual evidence (Crisma et al., 2025) or, as in classical formal practice, from grammaticality judgments of individual informants.

To this end, a parameter setting algorithm (Crisma et al., 2020) is required to identify the relevant triggering data, along with a practical tool to collect and systematically organize them.

Crucially, the parameter setting process does not require large quantities of data; it relies instead on carefully selected evidence. In this respect, a major challenge in dialectal settings is that speakers often find themselves in conditions of unbalanced bilingualism (Lambert et al., 1959), a situation common in endangered contexts (Polinsky, 1995; Grinevald, 2003; Bidese, 2017), which may make it difficult to assess the native status of their dialect competence and, consequently, the reliability of their judgments. Hence, assessing the speaker's competence becomes a particularly delicate task that must be handled with care (Cornips and Poletto, 2005; Bailey, 2017; Macaulay, 2017, a.o.). On these grounds, the internal architecture of the parametric framework offers a built-in control mechanism: data yielding parameter values inconsistent with the dependency structure of the system or with its theoretical predictions are immediately detectable, thus making it possible to isolate inaccurate information and neutralize its potential distortive effects.

The dependency structure is a distinctive feature of parameter systems (Baker, 2001; Fodor, 2001; Longobardi, 2003; Biberauer and Roberts, 2015, 2017; Roberts, 2019): one state of a given parameter may entail the irrelevance of another parameter, whose manifestations become predictable. Parameter dependencies are formalized as *implicational conditions*; neutralization due to a violated condition is encoded as [0] (Guardiano and Longobardi, 2017), which represents redundant information that should be excluded from comparison. A resource designed to manage a system of this kind must incorporate all of its defining properties and ensure that data collection and parameter setting are consistent with them. Such a tool should allow for linking data to the relevant parameter(s) and setting their values, while also taking into account the network of dependencies that may neutralize them.

The comparison of the resulting parameter lists requires additional tools to align them, measure distances and statistically assess their distribution, select subsets of languages and/or parameters to explore patterns of convergence and divergence in greater detail, derive classifications, and map the distribution of parametric distances and values across the geographical space.

The PCM Hub was developed to address these issues. This web-based application optimizes the conceptual architecture of the parameter system and the parameter setting algorithm by integrating them with a user-oriented data collection mechanism that meets accessibility standards, accommodates specific properties of different data types, and addresses the empirical and methodological requirements for automatic analyses. Designed to be flexible and expandable, this tool facilitates cumulative research, supports comprehensive parameter encoding and generates outputs suitable for statistical testing and phylogenetic reconstruction. Starting from a set of data supplied by a native speaker or language expert, each linked to the manifestation(s) of a given parameter, the system determines parameter values and evaluates the data against parameter dependencies. A resource that meets these requirements can, in principle, be used to collect, analyse, and process any type of linguistic data, including dialectal ones.

The following sections illustrate its architecture and core functionalities.

3. The PCM Hub

So far, cross-linguistic data collection for PCM investigations has been conducted in the absence of an infrastructure capable of systematically recording, managing and analysing linguistic material while incorporating the theoretical apparatus of the parameter system. The need to expand the comparative scope of the PCM in terms of typological range, genealogical depth, and empirical coverage has now made such an infrastructure indispensable. Recent efforts have therefore moved toward shareable environments, designed to uphold the PCM methodological standards while improving the efficiency of data acquisition, management, and sharing. A resource of this kind should ideally satisfy the following criteria:

1. Data must be recorded and shared through a user-oriented web-based interface.

2. A structured network of connections must be established among data, between data and their manifestations, between manifestations and parameters, and among parameters.

3. The consequences of parameter dependencies must be automatically computed in order to identify neutralized parameters.

4. The list of parameters and their properties must remain expandable and adaptable to revisions intrinsic to a framework that evolves alongside expanding empirical evidence and ongoing theoretical development.

5. Subsets of languages and/or parameters must be extractable, enabling systematic comparison and analysis.

6. Measures of dissimilarity and the resulting classifications must be automatically derived.

The scope of the PCM parameter system, as outlined above, is much wider than others in syntactic research, and has a deductive depth unparalleled in typology and quantitative phylogenetics. Consequently, no pre-existing software or modelling framework can be directly adapted to it.

The PCM Hub was conceived to be sufficiently rigorous to incorporate all structural aspects of the system, yet flexible enough to allow for integrations and revisions. Indeed, PCM research on ultralocality (Guardiano et al., 2016) has already shown that the analysis of microvariation may reveal aspects of variation not captured by the parameter set; this often requires the introduction of finer-grained options and a corresponding restructuring of the dataset (as shown in Fig. 1).

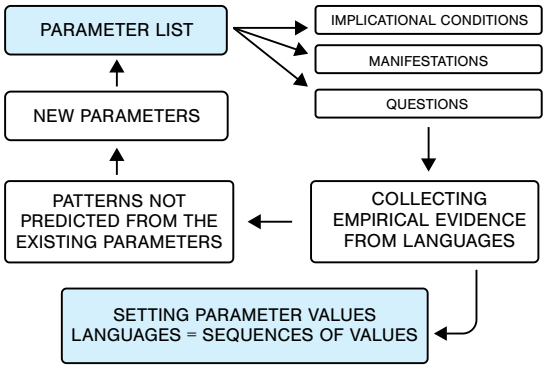

Figure 1: Parameter setting workflow

The infrastructure is inherently extensible: while the current parameter set reflects established theoretical knowledge, the platform is designed to seamlessly integrate new parameters as they emerge from the study of previously undocumented varieties or from further advancements in syntactic theory.

The platform has been designed with a view to future integration of phylogenetic inference, and will provide a usable and accessible tool to implement PCM analyses, while ensuring transparency of all methodological steps, in adherence to the FAIR principles (Wilkinson et al., 2016). In this sense, it is conceived primarily as a dedicated resource for PCM investigation, enabling researchers to carry out the workflow even in the absence of highly specialized expertise, while not positioning itself as a competitor to existing large-scale databases. Still, such resources, especially when grounded in rigorous principles of data collection and syntactic analysis – such as, in the Italo-Romance domain, the recent web platform based on Manzini and Savoia (2005) – may constitute a valuable source of empirical evidence. Interoperability with other databases, in line with standardization initiatives such as Forkel et al.'s (2018) Cross-Linguistic Data Formats (CLDF), is therefore envisaged as a potential feature of the infrastructure, although it is unlikely to provide a decisive solution for the full automation of data collection.

3.1. Interface overview

In its current version, the PCM Hub supports three types of activities: (a) data collection; (b) data analysis, including parameter setting, extraction of parameter sequences for statistical and phylogenetic processing, crosslinguistic and cross-parametric searches; and (c) retrieval of language information and procedural workflow. These activities correspond to three access modalities: (a) *User* (for informants, data collectors, language specialists),

(b) *Admin* (for PCM researchers), and (c) *Public Access* (for research replication). So far, development has primarily focused on the User and Admin interfaces, discussed below. Regarding Public Access, links to published PCM work (such as lists of languages with technical details and examples for each parameter, maps, dendrograms, comparative graphs and taxonomies) will be made accessible.

3.1.1. Data collection: the *User* interface

The data-entry platform is designed to allow the autonomous collection and entry of targeted evidence associated with the manifestations of each parameter by trained native speakers (or language experts). Since language experts may lack expertise in syntactic analysis, elicitation must be carefully guided (Koopman and Guardiano, 2022). To this end, a step-by-step procedure that presents the relevant syntactic environments needed to test the target structures was designed following the data collection protocol presented in Crisma et al. (2020) and Crisma et al. (2025). This procedure does not rely on standardized questionnaires. Instead, it offers a description of the structure(s) required to set the parameter, together with one or more examples from languages in which the relevant configuration is unambiguously attested. Doing so ensures that technical details are correctly interpreted by contributors. Language experts are then asked to provide their own examples, thus being free to make their own lexical choices and adapt the examples to language-specific needs. The system's support for IPA transcription makes the platform fully operational even for languages lacking a written standard. Finally, it is important to note that, although the User interface has been designed to make data entry as straightforward as possible, the resource is intended for expert use and does not accommodate untrained speakers, whose input therefore requires expert mediation and guidance.

The workflow is structured as follows. For each parameter, the interface presents a concise description alongside a series of YES/NO questions corresponding to its manifestations. After a positive answer is given, the User is required to submit a minimum of two examples, along with their transliteration (where necessary), gloss, and translation (see Figure 2). Conversely, selecting a negative answer requires the User to choose one of the predefined explanations accounting for that response (see Figure 3).

Each stage is supported by specific instructions, a glossary of technical terminology, and concise glossing guidelines. A dedicated Comment field allows for further observations/comments. This feature enables researchers to document gradient phenomena or cases that might, at first glance, be difficult to fit into a strict binary classification, ensuring that the quantitative abstraction is always supported by nuanced linguistic evidence.

Data reliability is monitored through a dual-color confidence system: a green button for complete and verified data, a red one for further review (Figure 4). Until final submission, entries remain editable. Users may download the data they contribute, which will be credited to them in relevant PCM publications.

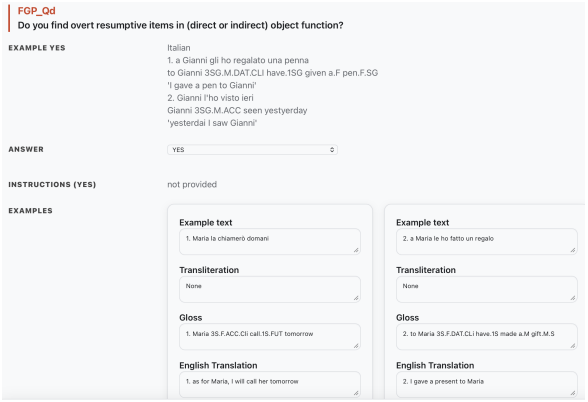

Figure 2: Data entry interface, answer YES

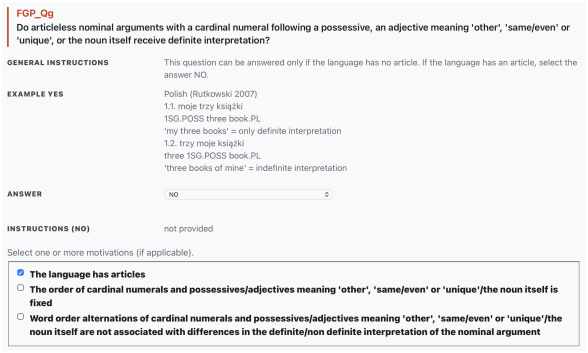

Figure 3: Data entry interface, answer NO

Figure 4: Data entry interface, confidence system

3.1.2. Data analysis: the *Admin* interface

Admins oversee access to the dataset and control data entry. They assign each language expert (User) to a specific language and enter relevant metadata, including standard identification codes (Glottocode and ISO code), genealogical classification (when available), and geographical coordi-

nates. These data can be exported and used for different types of analysis (Section 4.3).

Admins review the data to ensure that examples are consistent with the answer provided. If inconsistencies are detected, Users may be asked to modify or refine their entries. This process ensures quality control and uniformity.

The data collected as part of previous PCM publications are uploaded via structured files, which are automatically integrated into the platform.

Parameter values are computed through an algorithm that processes the answers in combination with the system of conditions formalizing cross-parametric dependencies. At the end of this process, each language is expressed as a sequence of symbols ([+], [−], and [0]).

Admins oversee database content: they can add, edit or deactivate parameters, link each parameter to its corresponding schema, type, set of manifestations, examples, and implicational conditions (Figure 5); edit the support material for Users. Parameters can be deactivated without loss of associated data, as the system ensures full preservation and traceability of all materials entered over time.

The platform supports cross-linguistic comparison by allowing Admins to select subsets of languages and/or parameters and to compute pairwise distances using dedicated algorithms. It enables multivariate techniques such as Principal Component Analysis (PCA) and hierarchical clustering visualized as dendrograms; it includes geographic visualization tools for generating maps such as those illustrated in Figure 8.

Finally, the system provides dedicated query tools for specific analysis of the distribution of parameter values and their manifestations.

All materials can be downloaded in multiple formats, including non-editable formats for dissemination (PDF) and machine-readable formats suitable for further processing through external scripts and computer-based tools not natively supported by the platform.

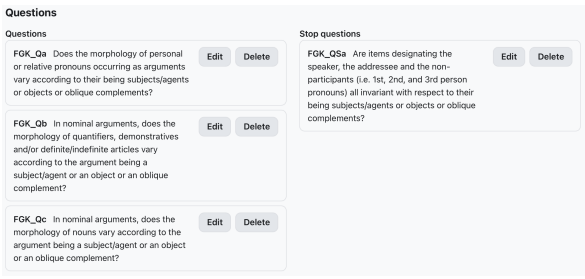

Figure 5: Question editing interface

3.2. Technical Specs

The PCM Hub is a scalable platform developed using the Django framework and a PostgreSQL backend. Its architecture follows a clear separation between data management, application logic, and user interface, ensuring stability, maintainability, and future extensibility. This approach allows the system to consistently handle the progressive increase in the number of languages, parameters, and operators. The interface is developed with responsive web technologies (HTML5, CSS, JavaScript). Each language and parameter is represented as a relational entity in the database, allowing for a structured and queryable organization of data. Parameter states are automatically assigned in response to the User's answers. Subsequently, a dedicated logical engine currently based on directed acyclic graphs (DAGs) assigns a [0] to those parameters whose implicational condition is not satisfied.

All software and technologies employed are open source; by contrast, the platform design, its structure and all of its content are protected by copyright.

4. Dialect Analysis and Comparison

This section illustrates some key functionalities of the PCM Hub in their application to ultralocality.

4.1. Languages

The current PCM dialect dataset consists of a selection of Romance and Greek dialects. The Romance dialects are all spoken in Italy, and are representative of six independently recognized groups in this domain, spanning across a north-south geographical space: (1) Ladin (Rocca Pietore); (2) Venetian (Trieste, Motta di Livenza, Vicenza, Chioggia, Porto Viro); (3) Northern Italy Gallo-Italic (Casalmaggiore, Parma, Reggio Emilia, Novellara, Correggio, Savignano sul Rubicone); (4) Upper Southern (Teramo, Santa Maria Capua Vetere, Frattamaggiore, Palma Campania, Felitto, Francavilla in Sinni, Verbicaro, Gargano, Barletta, Bari, Taranto); (5) Extreme Southern (Mesagne, Cellino San Marco, Botrugno, Cutro, Nicastro, Catanzaro, Reggio Calabria, San Filippo del Mela, Trapani, Ribera, Sant'Angelo Muxaro, Mussomeli, Catania, Ragusa); (6) Gallo-Italic of Sicily (Nicosia, Aidone). The Greek dialects instantiate three major groups, ideally representing the major compartments of the Greek-speaking world: (1) Italiot Greek (Grecìa Salentina, Bovesìa – two varieties); (2) Asia Minor Greek (Romeyka Pontic – Sitaridou, 2014; Schreiber, 2024; Neocleous and Sitaridou, 2025, Cappadocian – Karatsareas, 2009, 2013, 2016, Pharasiot – Bağrıaçık, 2017, 2018; Bağrıaçık and Danckaert, 2022); (3) 'Central' Greek (Standard Modern Greek, Cypriot).

This selection allows for testing ultralocality questions beyond those addressed in previous PCM studies. Once the tree-like signal identified in earlier work (namely, the distinction of the Romance cluster from Greek) is replicated, the next question is whether finer-grained internal structure can be detected within each cluster. For example, (i) Does the distribution of parametric diversity align with traditionally recognized classifications of the Romance dialects of Italy? (ii) Does it recover the tripartite structure of the Greek-speaking domain?

The data required to set each parameter were collected over the period 2010-2025, before the PCM Hub was implemented. These data are being stored in the PCM Hub through structured files, following the procedure described in Section 3.1.2.

As observed in Section 2, the lists of parameter values used for comparison are assumed to instantiate the competence of individual speakers, such that, in principle, each speaker is associated with a distinct list of values. Ideally, each list should therefore be constructed consistently on the basis of the judgments of a single individual speaker. In principle, judgments provided by speakers of the same language/dialect are expected to largely overlap. Accordingly, when judgments from different speakers of the same language/dialect converge, only one list of values is produced. By contrast, when different speakers provide judgments which lead to opposite settings for one or more parameters, multiple lists of values are produced, one for each speaker.

This idealized procedure, however, cannot always be implemented with full precision, as in practice, circumstances may arise in which it is not possible to obtain all the required judgments from a single speaker (for instance, when updates to the parameter system make it necessary to elicit additional judgments and the original speaker is no longer available). In such cases, corrective measures are adopted, including consultation of the relevant literature.

In the present dataset, for some dialects (including Gargano, Nicosia, Asia Minor Greek, and one variety of Italiot Greek from Bovesìa), due either to the unavailability of native speakers or to the impossibility of maintaining systematic contact with the same speaker, data were collected from written sources or from the existing literature, in accordance (when possible) with the protocol defined in Crisma et al. (2025).

The parameter dataset currently implemented in the platform consists of the 94 parameters described in Crisma et al. (2025), with the addition of 14 further points of variation tentatively formulated to account for aspects of microparametric diversity detected across the Romance sample and briefly presented in Guardiano and Stalfieri (2026).

The lists of parameter values corresponding to each dialect are analysed and compared using exploratory statistical analyses, clustering techniques, and mapping algorithms implemented within the platform through dedicated scripts. Concerning statistical and computer-assisted analyses, a range of approaches have been applied over the years to various PCM datasets (Bortolussi et al., 2011; Guardiano et al., 2016; Franzoi et al., 2019; Ceolin et al., 2021; Guardiano and Stalfieri, 2026; see Longhin et al. (in prep.) for a detailed methodology-based comparative discussion). These studies consistently show that the results are robust across methods, with minimal variation in the resulting taxonomic classifications. The three methods presented in what follows are intended to provide the clearest illustrative examples.

4.2. Principal Component Analysis

The Principal Component Analysis (PCA; Jolliffe, 2002; Jolliffe and Cadima, 2016; Gewers et al., 2021) is an exploratory statistical technique that extracts orthogonal components from a matrix of variables. Despite distortions introduced by binary encoding of parameter states, the PCA provides more interpretable representations than other multivariate techniques and has yielded the most transparent results across different language datasets. An additional advantage is that it enables the identification of the variables shaping the observed parametric diversity. However, more refined analytical approaches may prove even more effective in future applications.

The analysis begins by removing those variables (i.e., parameters) that behave as constants, i.e., whose values are identical across all taxonomic units (the dialects). On this basis, 61 out of 108 variables were excluded. The distribution of parameter values across the taxonomic units can therefore be analysed only on a reduced set of variables, as is to be expected in a dataset of genealogically closely related languages (Guardiano and Longobardi, 2005). Despite this reduction, the two principal components account for 38.69% of the total variance, and the distribution of the taxonomic units across the space defined by these two components (F1, 25.89%, and F2, 12.80%), as illustrated in the scatter plot in Figure 6, remains highly informative.

This distribution can first be interpreted based on F1 scores. The Greek varieties are associated with positive F1 scores, as opposed to the Romance dialects, all displaying negative F1 scores. This result provides empirical corroboration of the expectation that parameter values identify a Greek cluster separate from Romance. This supports the conclusion that, although effects of contact-induced parameter resettings are visible in the two domains (see Section 1), inherited patterns in parameter transmission

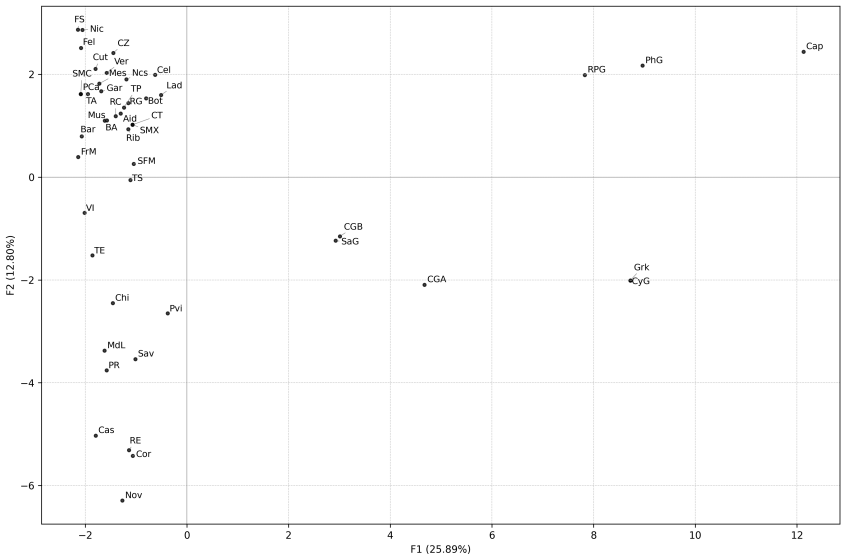

Figure 6: Scatter plot of the PCA

are not obscured by horizontal convergence.

It must also be noted that, while the Romance dialects are compressed within a narrow range of negative F1 scores, the Greek varieties are distributed across a much broader area, suggesting a sharper internal dialect differentiation.

The combination of F1 and F2 scores defines three Greek clusters, corresponding to Asia Minor Greek (highly positive F1 and F2 scores), Central Greek (highly positive F1 scores, moderately negative F2 scores), and Italiot Greek (moderately positive F1 scores, moderately negative F2 scores). This provides a positive answer to question (ii): the tripartite structure of the Greek-speaking domain is reflected in the distribution of parameter values.

The distribution of the Romance dialects is considerably less clear-cut, although some groupings can still be identified. Negative F2 scores separate a central-northern geographical area (except for Ladin) from southern dialects (except for Teramo), all of which display positive F2 scores. In this quadrant, no further clearly identifiable subdivisions emerge. By contrast, in the quadrant corresponding to negative F2 scores, the dialects are more dispersed. The most notable exception to the North-South divide captured by F2 scores is Ladin, a Rhaeto-Romance dialect group spoken in Northern Italy, whose classification and internal relations have been the subject of long-standing debate (Ascoli, 1873; Battisti, 1931; Haiman and Benincà, 2005). In the absence of closely related dialects such as Romansh or Friulian, Ladin constitutes an isolate in the sample. Hence, no strong a priori prediction can be made regarding its placement. The fact that it does not display any marked similarity with the other dialects of Northern Italy indicates that the distribution of parametric diversity cannot be reduced to a mere function of geographical proximity.

4.3. Hierarchical cluster analysis

Hierarchical agglomerative clustering methods (Everitt et al., 2011) take dissimilarity matrices as input and implement several linkage criteria, including: (1) Single linkage (minimum inter-cluster distance: nearest neighbor – Johnson, 1967; Sneath and Sokal, 1973); (2) Complete linkage (maximum inter-cluster distance: farthest neighbor – Johnson, 1967; Sneath and Sokal, 1973); (3) Average linkage (mean inter-cluster distance: UPGMA – Sokal and Michener, 1958; Lance and Williams, 1967); (4) Ward's method, which relies on Euclidean distances and is applied to the character matrix to minimize the increase in within-cluster variance at each agglomerative step (Ward, 1963).

These algorithms are implemented in the PCM Hub. Although comparable procedures are available in other web-based platforms for dialect data analysis (e.g., Gabmap, Nerbonne et al., 2011), their integration within the PCM Hub enables analyses to be conducted within a continuous workflow, eliminating the need to adapt PCM data formats to external platforms. This minimizes the risk of information loss. It also allows PCM researchers to operate on a set of algorithms directly suited to the analysis of parametric datasets.

To obtain a dissimilarity matrix from the lists of parameter values, pairwise parametric distances are computed using a formula traditionally adopted in PCM works (Ceolin et al., 2020), here simplified as *d/(i+d)*, where *d* denotes the number of differences in parameter states and *i* the number of identities. The latter can be calculated either by counting all

cases in which both languages display the same value ([+] or [−]) for a given parameter, or by counting only the cases in which both display [+]. The analyses presented here adopt the former criterion, although, at this level, both yield virtually identical results. Pairs in which one or both languages display a [0] are excluded from computations.

Figure 7 presents a dendrogram generated using the average linking method (UPGMA). The dendrogram reflects a taxonomic structure consistent with the PCA results: a clear primary split into two nodes corresponding to Greek and Romance, followed by an internal articulation of each cluster that largely mirrors the established internal structure of the groups. Within the Greek cluster, the outermost nodes separate Asia Minor Greek from a group including the central varieties and Italiot Greek. Within the latter, the two currently spoken varieties form a distinct cluster, as opposed to the conservative variety of Bovesìa (CGA), which is closer to Central Greek. This result is consistent with proposals suggesting that contact-induced effects from Romance represent relatively recent innovations (Guardiano and Stavrou, 2021). The articulation of Romance is largely consistent with independently established dialect classifications, though with some noteworthy divergences. Leaving aside the position of Ladin, already discussed in Section 4.2, two results are of particular interest. First, the Gallo-Italic dialects of Sicily are located within the cluster grouping the Extreme Southern varieties. This outcome is in line with recent work on the syntax of the Gallo-Italic dialects of Sicily (De Angelis, 2023; Guardiano and Stalfieri, 2026, a.o.). Second, the dialects spoken in central Calabria cluster with the Upper dialects, rather than with the Extreme Southern group. This configuration aligns with the literature identifying the corresponding area as a transitional zone of uncertain classification (Pellegrini, 1977; Trumper and Maddalon, 1988; Trumper, 1997, a.o.). Overall, these results provide a positive answer to question (i) in Section 4.1, as the only inconsistencies with traditional dialect classifications concern already controversial areas.

The clustering structure and its robustness were further tested using a second algorithm, K-means clustering (Forgy, 1965; Macqueen, 1967; Lloyd, 1982), which partitions taxonomic units on the basis of their parameter-value vectors by minimizing within-cluster variance (typically computed using Euclidean distance). These analyses yield a taxonomic structure closely comparable to that observed in Figure 7.

A dedicated script was developed to project the clustering results onto a geographical map. The script takes as input a table of geographical coordinates for each language, automatically generated within the PCM Hub. Once a predefined number of clusters is selected, the algorithm produces a map assigning each language to one of the clusters identified in the tree. Figure 8 was generated by setting the number of clusters to two. Increasing the number of clusters yields progressively finer partitions, which, as noted above, largely correspond to traditionally established dialect classification.

Algorithms for the areal visualization of cluster structures are standard in dialectometry (Goebl, 1982; Nerbonne and Kretzschmar, 2003; Nerbonne, 2009; Nerbonne and Heeringa, 2009, see also Dialektkarten.ch) and are implemented in web-based platforms such as Gabmap. These procedures model the spatial distribution of features, or feature bundles, across a territorially discretized space, represented through granular polygonal tessellations. While these tools have been shown to perform well even on sparse datasets, their resolution improves as geographical coverage becomes denser, since they rely on the cartographic projection of feature-based distance measures across finely articulated spatial grids. At present, the high degree of spatial granularity required for these techniques to be truly informative is as yet only partially available in the PCM. Additionally, this type of cartographic representation is methodologically better suited to examining the geographical diffusion of individual phenomena than to extracting historical signals from the distribution of multiple features. These are therefore more appropriate for mapping the distribution of individual parameter values, provided sufficiently fine-grained geographical coverage. Currently, PCM analyses use cartographic visualization to assess the extent to which parametric similarity is predicted by geographical proximity. For this purpose, a lightweight algorithm that directly projects cluster assignments onto a geographical map using information already stored within the PCM Hub offers a more efficient solution. At the same time, the PCM Hub does not position itself as an alternative to existing platforms. Owing to its data-export functionalities in multiple formats, it allows PCM datasets to be used in a wide range of external analytical environments. It should therefore be understood not as a replacement, but as a complementary theory-oriented infrastructure for quantitative dialect analysis.

Further evidence challenging the possibility that parametric similarity can be entirely reduced to geographical proximity is provided by Mantel tests (Mantel, 1967; Sokal and Rohlf, 1995; Guillot and François, 2013) correlating parametric and geographical distance matrices. The correlation coefficient (Pearson's r=0.5904, p=0.0010; r=0.5157, p=0.0010 when limited to the Italian peninsula; r=0.59, p=0.0010 if restricted to Romance) indicates a statistically significant and relatively strong

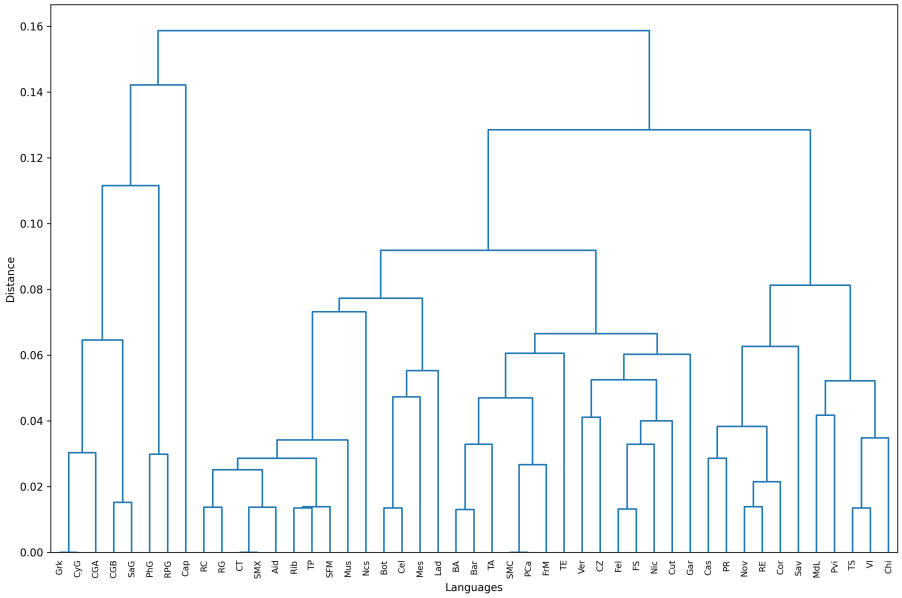

Figure 7: Hierarchical clustering dendrogram

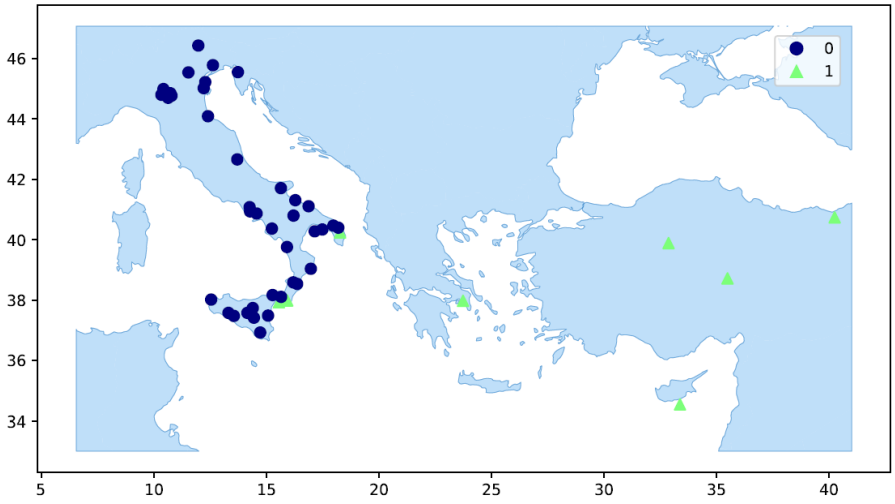

Figure 8: K-means clustering projected onto a map

association, yet not one sufficient to support direct predictability. Even within a geographically restricted and contact-intensive domain, geographical distance does not fully account for linguistic proximity, suggesting that the signal encoded in syntactic parameters remains robust enough to preserve genealogical differentiation despite areal interference.

5. Conclusion

The PCM approach to ultralocality demonstrates that dialect syntactic diversity can be measured in a formally constrained, reproducible and meaningful way. It introduces a quantitative perspective on ultralocal syntactic variation that, while building on dialectometric traditions, geolinguistic models, and formal treatments of syntactic microvariation, extends beyond them. The applications illustrated in this contribution suggest that the transmission of syntactic parameters, while sensitive to areal convergence, preserves a clear inheritance structure even in ultralocal domains. A more granular investigation of vertical transmission and horizontal diffusion in these domains is therefore poised to contribute to an integrated model of local variation, while also identifying patterns of syntactic transmission applicable to a deeper historical scale. In this light, the PCM Hub emerges as a key resource ensuring a theory-oriented empirical infrastructure to explore the parametric diversity of dialects.

6. Acknowledgements

Part of this research was funded by the Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN), project "Measuring the power of parameter setting theory on historical corpora – PARTHICO", Grant Assignment Decree n. 901, 21/06/2023, Italian Ministry of University and Research – MUR, and by the Progetto di Ricerca Interdisciplinare FAR Unimore 2024 Modeling crosslinguistic diversity in Differential Object Marking through the Parametric Comparison Method (POM; prot. n. 0323715 del 28/11/2024, Rep. n. 1307/2024).

Most of the data from the Romance dialects used in this work were collected by Vincenzo Stalfieri as part of the project n. Prin2017-2017K3NHHY, Models of language variation and change: new evidence from language contact, Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN), and analysed with the advice of Franco Fanciullo.

The data from the Greek domain were collected and analysed as part of collaborative work with Melita Stavrou.

The structure, content and design of the web platform were discussed many times with Paola Crisma, who provided essential advice.

Access to the list of informants, geographical coordinates and data for each dialect will be accessible through the website www.parametricomparison.unimore.it, where the link to the PCM Hub, along with detailed instructions to access the platform, will be made available.

All scripts and technical tools used to generate the results presented in this study are publicly available at https://github.com/GaiaSorge/PCM_Hub.

Author contribution. Conceptualization and methodology: CG and GL. Software development: ML and GBM. Data collection and analysis: SF, ELD, ML, CG, GS. Scripts and computer-assisted analyses: AA, ML, LC and FM. Writing - first draft: ELD, GS and ML. Writing - review and editing: CG, SF and GL (all authors, final version). Supervision, Funding acquisition, Project administration: CG.

7. References

Graziadio Isaia Ascoli. 1873. Saggi ladini. In Graziadio Isaia Ascoli, editor, *Archivio glottologico italiano*, volume 1, pages 1–537. Loescher, Roma/Torino/Firenze.

Guy Bailey. 2017. Field interviews in dialectology. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 284–299. Wiley, Hoboken.

Mark C. Baker. 2001. *The atoms of language. The mind's hidden rules of grammar*. Oxford University Press, Oxford.

Carlo Battisti. 1931. *Popoli e lingue nell'Alto Adige. Studi sulla latinità altoatesina*. Bemporad, Firenze.

Metin Bağrıaçık. 2017. Representing discourse in clausal syntax. The *ki* particle in Pharasiot Greek. *Journal of Greek Linguistics*, 17:141–189.

Metin Bağrıaçık. 2018. *Pharasiot Greek: word order and clause structure*. Ph.D. thesis, Ghent University, Ghent.

Metin Bağrıaçık and Lieven Danckaert. 2022. Raising and matching in Pharasiot Greek relative clauses. A diachronic reconstruction. *Journal of Linguistics*, 58(3):495–533.

Theresa Biberauer and Ian Roberts. 2015. The clausal hierarchy, features and parameters. In Ur Shlonsky, editor, *Beyond Functional Sequence*, pages 295–313. Oxford University Press, Oxford.

Theresa Biberauer and Ian Roberts. 2017. Parameter setting. In Adam Ledgeway and Ian Roberts, editors, *The Cambridge Handbook of Historical Syntax*, pages 134–162. Cambridge University Press, Cambridge.

Ermenegildo Bidese. 2017. The correlation between unbalanced bilingualism and language decay in small language minorities. The current status of research and future perspectives. *Bollettino dell'Atlante linguistico italiano*, 41:95–107.

Luca Bortolussi, Andrea Sgarro, Giuseppe Longobardi, and Cristina Guardiano. 2011. How many possible languages are there? In Gemma Bel-Enguix, Veronica Dahl, and M. Dolores Jiménez-López, editors, *Biology, computation and linguistics*, pages 168–179. IOS Press, Amsterdam.

Andrea Ceolin, Cristina Guardiano, Giuseppe Longobardi, and Monica Alexandrina Irimia. 2020. Formal syntax and deep history. *Frontiers in psychology*, 11.

Andrea Ceolin, Cristina Guardiano, Giuseppe Longobardi, Monica Alexandrina Irimia, Luca Bortolussi, and Andrea Sgarro. 2021. At the boundaries of syntactic prehistory. *Philosophical Transactions of the Royal Society B*, 376:20200197.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

Noam Chomsky. 1986. *Knowledge of language. Its nature, origin, and use*. Praeger, New York.

Leonie Cornips and Cecilia Poletto. 2005. On standardising syntactic elicitation techniques (part 1). *Lingua*, 115(7):939–957.

Paola Crisma, Giulia Fabbris, Giuseppe Longobardi, and Cristina Guardiano. 2025. What are your values? Default and asymmetry in parameter states. *Journal of Historical Syntax*, 9:1–26.

Paola Crisma, Cristina Guardiano, and Giuseppe Longobardi. 2020. Syntactic diversity and language learnability. *Studi e Saggi Linguistici*, 58(2):99–130.

Alessandro De Angelis. 2023. The strange case of the Gallo-Italic dialects of Sicily. Preservation and innovation in contact-induced change. *Languages*, 8(3):163.

Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster analysis*. John Wiley and Sons, Chichester.

Janet Dean Fodor. 2001. Setting syntactic parameters. In Mark Baltin and Chris Collins, editors, *The Handbook of Contemporary Syntactic Theory*, pages 730–767. Blackwell, Oxford.

Edward W. Forgy. 1965. Cluster analysis of multivariate data. Efficiency versus interpretability of classifications. *Biometrics*, 21:768–780.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5:180205.

Laura Franzoi, Andrea Sgarro, Anca Dinu, and Liviu P. Dinu. 2019. Linguistic classification: dealing jointly with irrelevance and inconsistency. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 345–352, Varna. INCOMA.

Felipe L. Gewers, Gustavo R. Ferreira, Henrique F. De Arruda, Filipi N. Silva, Cesar H. Comin, Diego R. Amancio, and Luciano Da F. Costa. 2021. Principal component analysis. A natural approach to data exploration. *ACM Computing Surveys*, 54(4):1–34.

Hans Goebl. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*. Verlag der Österreichischen Akademie der Wissenschaften, Wien.

Colette Grinevald. 2003. Speakers and documentation of endangered languages. *Language documentation and description*, 1:52–72.

Cristina Guardiano. 2023. Differential object marking in a dialect of Sicily. In Monica Alexandrina Irimia and Alexandru Mardale, editors, *Differential Object Marking in Romance. Towards microvariation*, pages 192–231. John Benjamins, Amsterdam.

Cristina Guardiano, Michela Cambria, and Vincenzo Stalfieri. 2022. Number morphology and bare nouns in some Romance dialects of Italy. *Languages*, 7(4):255.

Cristina Guardiano and Giuseppe Longobardi. 2005. Parametric comparison and language taxonomy. In Montserrat Batllori, Maria-Lluïsa Hernanz, Carme Picallo, and Francesc Roca, editors, *Grammaticalization and Parametric Variation*, pages 149–174. Oxford University Press, Oxford.

Cristina Guardiano and Giuseppe Longobardi. 2017. Parameter theory and parametric comparison. In Ian Roberts, editor, *The Oxford Handbook of Universal Grammar*, pages 377–398. Oxford University Press, Oxford.

Cristina Guardiano, Giuseppe Longobardi, Guido Cordoni, and Paola Crisma. 2021. Formal syntax as a phylogenetic method. In Richard D. Janda, Brian D. Joseph, and Barbara S. Vance, editors, *The Handbook of Historical Linguistics*, volume 2, pages 145–182. Wiley-Blackwell, Malden.

Cristina Guardiano, Dimitris Michelioudakis, Andrea Ceolin, Monica Irimia, Giuseppe Longobardi, Nina Radkevich, Giuseppina Silvestri, and Ioanna Sitaridou. 2016. South by Southeast. A syntactic approach to Greek and Romance microvariation. *L'Italia dialettale*, LXXVII:95–166.

Cristina Guardiano and Vincenzo Stalfieri. 2026. Comparazione parametrica e strutture nominali. Alcune note dalla Sicilia galloitalica. In Elvira Assenza, Angela Castiglione, Alessandro De Angelis, and Salvatore Menza, editors, *I dialetti galloitalici di Sicilia tra resistenza e assimilazione*, pages 163–192. Centro Studi Filologici Linguistici Siciliani, Palermo.

Cristina Guardiano and Melita Stavrou. 2014. Greek and Romance in Southern Italy. History and contact in nominal structures. *L'Italia dialettale*, LXXV:121–147.

Cristina Guardiano and Melita Stavrou. 2019a. Adjective-noun combinations in Romance and Greek of Southern Italy. Polydefiniteness revisited. *Journal of Greek Linguistics*, 19:3–57.

Cristina Guardiano and Melita Stavrou. 2019b. Comparing patterns of adjectival modification in Greek. A diachronic approach. *Quaderni di Linguistica e Studi Orientali*, 5:135–173.

Cristina Guardiano and Melita Stavrou. 2020. Dialectal syntax between persistence and change. The case of Greek demonstratives. *L'Italia dialettale*, 81:119–155.

Cristina Guardiano and Melita Stavrou. 2021. Modeling syntactic change under contact. The case of Italiot Greek. *Languages*, 6(2):74.

Gilles Guillot and Rousset François. 2013. Dismantling the Mantel tests. *Methods in ecology and evolution*, 4:336–344.

John Haiman and Paola Benincà, editors. 2005. *The Rhaeto-Romance languages*. Routledge, London/New York.

Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Ian T. Jolliffe. 2002. *Principal Component Analysis*. Springer-Verlag, New York.

Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis. A review and recent developments. *Philosophical Transactions of the Royal Society A*, 374:20150202.

Petros Karatsareas. 2009. The loss of grammatical gender in Cappadocian Greek. *Transactions of the Philological Society*, 107(2):196–230.

Petros Karatsareas. 2013. Understanding diachronic change in Cappadocian Greek. The dialectological perspective. *Journal of Historical Linguistics*, 3(2):192–229.

Petros Karatsareas. 2016. Convergence in word structure. Revisiting agglutinative noun inflection in Cappadocian Greek. *Diachronica*, 33(1):31–66.

Richard S. Kayne. 1996. Microparametric syntax. Some introductory remarks. In James R. Black and Virginia Motapanyane, editors, *Microparametric Syntax and Dialect Variation*, volume 139, pages ix–xviii. John Benjamins, Amsterdam.

Richard S. Kayne. 2005. Some notes on comparative syntax. With special reference to English and French. In *Movement and Silence*, pages 277–334. Oxford University Press, Oxford.

Hilda Koopman and Cristina Guardiano. 2022. Managing data in TerraLing, a large-scale cross-linguistic database of morphological, syntactic, and semantic patterns. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, editors, *The Open Handbook of Linguistic Data Management*, pages 617–630. The MIT Press, Cambridge, MA.

Wallace E. Lambert, Jelena Havelka, and Richard C. Gardner. 1959. Linguistic manifestations of bilingualism. *The American Journal of Psychology*, 72(1):77–82.

Godfrey N. Lance and William T. Williams. 1967. A general theory of classificatory sorting strategies. 1. Hierarchical systems. *The Computer Journal*, 9(4):373–380.

Adam Ledgeway. 2000. *A comparative syntax of the dialects of southern Italy. A minimalist approach*. Blackwell, Oxford.

Adam Ledgeway, Norma Schifano, and Giuseppina Silvestri. 2018. Il contatto tra il greco e le varietà romanze nella Calabria meridionale. *Lingue antiche e moderne*, 7:96–133.

Adam Ledgeway, Norma Schifano, and Giuseppina Silvestri. 2020. Microvariation in dative-marking in the Romance and Greek varieties of Southern Italy. In Anna Pineda and Jaume Mateu, editors, *Dative constructions in Romance and beyond*, pages 311–342. Open Generative Syntax, Language Science Press.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

Marco Longhin, Isabella Morlini, Gaia Sorge, Monica A. Irimia, and Cristina Guardiano. in prep. Parametric comparison and phylogenetic testing. Ms. UniMoRe.

Giuseppe Longobardi. 2003. Methods in parametric linguistics and cognitive history. *Linguistic Variation Yearbook*, 3(1):101–138.

Giuseppe Longobardi and Cristina Guardiano. 2009. Evidence for syntax as a signal of historical relatedness. *Lingua*, 119:1679–1706.

Ronald Macaulay. 2017. Dialect sampling methods. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pages 241–252. Wiley, Hoboken.

James Macqueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkley. University of California Press.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2):209–220.

Maria Rita Manzini and Leonardo Maria Savoia. 2005. *I dialetti italiani e romanci. Morfosintassi generativa*. Edizioni dell'Orso, Alessandria.

Nicolaos Neocleous and Ioanna Sitaridou. 2025. Word order and information structure in Romeyka. *Frontiers in Psychology*, 16:1337962.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap – a web application for dialectology. *Dialectologia*, Special issue II:65–89.

John Nerbonne and Wilbert Heeringa. 2009. Measuring dialect differences. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space: Theories and Methods*, pages 550–567. De Gruyter Mouton, Berlin/New York.

John Nerbonne and William Kretzschmar. 2003. Introducing computational techniques in dialectometry. *Computers and the Humanities*, 37:245–255.

Giovan Battista Pellegrini. 1970. La classificazione delle lingue romanze e i dialetti italiani. *Forum Italicum: A Journal of Italian Studies*, 4(2):211–237.

Giovan Battista Pellegrini. 1977. *Carta dei dialetti d'Italia*. Pacini, Pisa.

Cecilia Poletto. 2012. Contrastive linguistics and micro-variation. The role of dialectology. *Languages in contrast*, 12(1):47–68.

Maria Polinsky. 1995. Cross-linguistic parallels in language loss. *Southwest journal of linguistics*, 14(1-2):87–123.

Ian Roberts. 2019. *Parameter Hierarchies and Universal Grammar*. Oxford University Press, Oxford.

Laurentia Schreiber. 2024. *A (contact-)grammar of Romeyka*. Ph.D. thesis, Otto-Friedrich-Universität, Bamberg.

Giuseppina Silvestri. 2013. *The nature of genitive case*. Ph.D. thesis, Università di Pisa., Pisa.

Giuseppina Silvestri. 2016. Possessivi e partitivi nei dialetti italo-romanzi dell'Area Lausberg. *La lingua italiana. Storia, struttura, testi*, XII:127–142.

Giuseppina Silvestri. 2018. Word-final schwa licensed by prosody and syntax. Evidence from Southern Italian dialects. *Rivista Di Grammatica Generativa*, 3:1–27.

Giuseppina Silvestri. 2020. Possessives in indefinite nominal phrases. A comparison between Italo-Romance and Daco-Romance. *Moderna språk*, 3(114):161–197.

Ioanna Sitaridou. 2014. The Romeyka infinitive. Continuity, contact and change in the Hellenic varieties of Pontus. *Diachronica*, 31(1):23–73.

Peter H. A. Sneath and Robert R. Sokal. 1973. *Numerical taxonomy. The principles and practice of numerical classification*. W. H. Freeman, San Francisco.

Robert Sokal and James Rohlf. 1995. *Biometry*. Macmillan.

Robert R. Sokal and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Scientific Bulletin*, 38:1409–1438.

John Bassett Trumper. 1997. Calabria and Southern Basilicata. In Martin Maiden and Mair Parry, editors, *The dialects of Italy*, pages 355–364. Routledge.

John Bassett Trumper and Marta Maddalon. 1988. Converging divergence and diverging convergence. The Dialect-Language Conflict and Contrasting Evolutionary Trends in Modern Italy. In Peter Auer and Aldo di Luzio, editors, *Variation and Convergence. Studies in Social Dialectology*, pages 217–259. De Gruyter.

Joe H. Ward. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 'T Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene Van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan Van Der Lei, Erik Van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.