A CLDF-Compliant Lexical Database for Modern Greek Dialects: Resource Design and Dialectometric Analysis

Stavros Bompolas[1,2], Natalia Chousou-Polydouri[2], Manuela Genitsaridi[2,3], Danae Karatzanou[2,3], Georgios Kostopoulos[2,3], Elena Anagnostopoulou[2,3], Dimitra Melissaropoulou[2,4]

[1]Archimedes/Athena RC, [2]Institute for Mediterranean Studies/Foundation for Research and Technology, [3]University of Crete, [4]Aristotle University of Thessaloniki

[1]Artemidos 1, 15125, Marousi, Greece, [2]Nikiforou Foka 130, 74131, Rethymno, Greece, [3]Department of Philology, 74100, Rethymno, Greece, [4]School of Italian Language and Literature, 54124, Thessaloniki, Greece

Correspondence: s.bompolas@athenarc.gr

Abstract

This paper presents the first CLDF database that systematically documents lexical variation across 36 Modern Greek varieties (including Standard Modern Greek). The dataset aligns 14,378 lexical items over 345 concepts, links varieties to stable identifiers (Glottocodes), introduces a Greek-specific concept list, and maps meanings to standardized Concepticon concept sets, enabling interoperability and reproducible workflows. To assess whether the database preserves a meaningful dialectological signal, we conduct a dialectometric analysis by computing feature-sensitive string distances over IPA transcriptions and applying hierarchical clustering. The resulting similarity structure recovers major macro-divisions—most notably a broad Northern vs. Southern partition among Koine-descended mainland varieties—and isolates peripheral groups with distinct historical trajectories (e.g., Asia Minor, Italiot, Tsakonian). The database provides scalable infrastructure for quantitative dialectology, comparative Greek linguistics, and dialect-aware language technology.

Keywords: Modern Greek dialects, lexical variation, lexical database, CLDF, dialectometry

1. Introduction

Lexical databases constitute core infrastructure for research in linguistic typology, historical linguistics, and Natural Language Processing (NLP). The adoption of structured and interoperable standards—most prominently Cross-Linguistic Data Formats (CLDF)—has significantly improved the transparency, comparability, and long-term reusability of lexical data (Forkel et al., 2018b). These advances have enabled large-scale cross-linguistic comparison and computational analysis at an unprecedented scale. Yet this progress has been uneven: while variation *across* languages is increasingly well documented, systematic dialectal variation *within* languages remains substantially underrepresented. This imbalance reflects a persistent standard-centric bias in resource development, whereby non-standard and regional varieties are marginalized, normalized, or reduced to peripheral deviations from an assumed "main" language.

Modern Greek (MG) exemplifies this asymmetry. Although Standard Modern Greek (SMG) is included in several major lexical and typological databases, dialectal varieties receive only limited coverage. Large cross-linguistic lexical databases include at most a small number of Greek varieties (e.g., IE-Cor includes four MG varieties, Heggarty et al., 2023), while typological resources typically encode a single standard-based profile (e.g., Grambank includes only SMG, Skirgård et al., 2023; WALS includes SMG and Cypriot Greek, Dryer and Haspelmath, 2013). At the same time, Greek dialectology has produced extensive descriptive scholarship, including dialect dictionaries, grammars, and regional atlases. Yet digital infrastructures that systematize and integrate this material remain scarce. As a result, despite a rich dialectological tradition, there is currently no large-scale, concept-aligned, computationally reusable database that documents lexical variation across the breadth of the MG dialect continuum in a format compatible with contemporary lexical infrastructures and NLP workflows.

In this paper, we address this infrastructural gap by presenting a CLDF-compliant lexical database documenting dialectal variation across 36 MG varieties (including SMG), aligned over 345 concepts and comprising 14,378 lexical items. The dataset transforms dispersed dialect material into a structured, interoperable resource that renders intra-language lexical variation computationally tractable, while providing a foundation for future multi-level extensions (phonology, morphology, and syntax). Beyond introducing the resource, we evaluate whether it encodes a meaningful dialectological signal through dialectometry, applying feature-sensitive distance measures and hierarchical clus-

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 287–299

tering to test whether aggregate lexical–phonetic similarity recovers groupings anticipated in the dialectological literature. Concretely, the paper:

1. introduces the first CLDF-compliant lexical database covering a broad sample of MG dialects;
2. documents the methodological and modeling decisions required to represent a dialect continuum within an interoperable framework; and
3. offers a dialectometric evaluation that both validates the internal coherence of the dataset and provides a data-driven reassessment of traditional dialect classifications.

2. Related Work

Our work sits at the intersection of (a) cross-linguistic lexical infrastructures grounded in concept-list methodology, (b) emerging efforts to develop computationally reusable resources for dialect continua and closely related varieties, and (c) MG dialect lexicography, where lexical documentation remains fragmented and infrastructurally underdeveloped. We therefore review prior work along these three dimensions.

2.1. Cross-Linguistic Lexical Infrastructures

Linguists have long relied on standardized concept lists to compare the lexicons of different languages (e.g., Swadesh lists, Swadesh, 1955; the Leipzig–Jakarta list, Haspelmath and Tadmor, 2009a; among others). Building on this tradition, large-scale databases now provide comparative wordlists across hundreds or thousands of languages. For example, the Automated Similarity Judgment Program (ASJP) maintains 40-item wordlists in a unified phonetic encoding for most of the world's languages (Wichmann et al., 2025). The World Loanword Database (WOLD) provides "mini-dictionaries" of approximately 1,000–2,000 entries for 41 diverse languages, with each entry annotated for loanword versus inherited status (Haspelmath and Tadmor, 2009b). Like ASJP, WOLD relies on a fixed meaning list (1,460 concepts derived from the Intercontinental Dictionary Series; Key and Comrie, 2023), enabling systematic cross-language comparison of lexical structure and borrowing patterns. In addition to global comparative resources, there are family-specific lexical databases designed for particular lineages, such as IE-CoR for the Indo-European family (Heggarty et al., 2023) and LA80 for a set of North-West Bantu A80 languages (Vermeir et al., 2024).

More recently, efforts have focused on improving interoperability and standardization. The Cross-Linguistic Data Formats (CLDF) initiative proposes simple, tabular standards for lexical and structural datasets (Forkel et al., 2018b), along with associated tools for validation, processing, and reproducible dataset curation (Forkel et al., 2018a; Forkel and List, 2020; Kaiping et al., 2022). Built on CLDF, Lexibank aggregates dozens of published wordlist datasets into a unified, FAIR-compliant repository (List et al., 2022). In its current form, Lexibank covers over 2,000 language varieties drawn from roughly 100 source datasets. Crucially, it harmonizes transcription practices and links entries to standardized identifiers (e.g., Glottocodes for languages and Concepticon IDs for meanings), thereby ensuring interoperability and reproducibility. Complementing these efforts, Concepticon maps tens of thousands of concept labels from diverse published lists onto a few thousand standardized concept sets (List et al., 2025), facilitating cross-dataset comparison even when original glossing conventions differ.

Taken together, these infrastructures have transformed lexical typology and historical comparison into transparent and computationally tractable enterprises. However, their primary unit of organization remains the language. Typically, one representative lexical item per concept is recorded for each variety. As a result, while these resources excel at cross-language comparison, they are not designed to capture systematic microvariation within a single language, where multiple competing realizations per concept and fine-grained sociolinguistic metadata may be central.

2.2. Lexical Resources for Dialects

Resources specifically designed for closely related varieties (dialects, regional standards, or "similar languages") are comparatively rare. Traditional dialectology has produced extensive linguistic atlases (see Lameli, 2009; Kretzschmar, 2017) and regional glossaries or dialect dictionaries (see Moulin, 2009; Van Keymeulen, 2017), which document lexical and structural variation at fine geographical resolution. However, these works are typically published as printed maps, volumes, or static digital repositories. Even when digitized, data are often distributed across heterogeneous formats and lack standardized encoding. Coverage may be uneven, transcription practices inconsistent, and metadata incomplete. Consequently, although dialect atlases and glossaries are rich in empirical detail, they are rarely structured in a way that supports systematic, item-by-item comparison across many varieties or straightforward computational reuse.

This underrepresentation persists even within contemporary interoperable infrastructures. For example, in Lexibank—currently comprising approximately 75 curated lexical datasets—only 14 ex-

plicitly target dialectal or closely related varieties, covering just 87 varieties out of more than 2,029 included overall. Moreover, these datasets typically focus on relatively small and geographically restricted areas. This reflects a broader pattern: although current infrastructures technically support dialect-level encoding, dialect continua remain substantially underrepresented in practice.

The relative scarcity of such resources underscores the methodological and infrastructural gap that persists for dialect continua—including MG.

2.3. Lexical Resources for Modern Greek Dialects

MG dialectology has a long and productive tradition, including detailed monographs, dialectal dictionaries, and grammatical descriptions of dialects (Tzitzilis, 2000; Katsouda, 2024). Nevertheless, dialect documentation has typically proceeded on a per-variety basis, resulting in isolated lexica or grammatical studies rather than integrated, cross-dialect resources (for a relevant discussion, see Trudgill, 2003).

More recent digital initiatives focus on specific varieties or limited geographical areas—for example, work on Cappadocian Greek (Melissaropoulou et al., 2022; ILIK, 2024). Other efforts remain under development and are not yet released as openly-licensed, interoperable datasets suitable for systematic comparison and computational reuse (ILNE, 2026). As a result, despite the richness of dialectological scholarship, there is currently no large-scale, computationally reusable database that systematically documents variation across the breadth of MG dialects in a format suitable for structured comparison and NLP-oriented reuse.

Cross-linguistic lexical infrastructures occasionally include MG data, but typically restrict coverage to the standard variety. A partial exception is IE-CoR (Heggarty et al., 2023), which incorporates several Greek varieties, including Pontic, Cappadocian, Cypriot, Italiot, Peloponnese Tsakonian, and Propontis Tsakonian. Similarly, the Global Lexicostatistical Database (Starostin, 2011) documents SMG alongside Peloponnese Tsakonian, Pharasa Greek, and Cappadocian Greek (Aravan). In both cases, coverage remains limited and does not aim at systematic intra-language dialect comparison; moreover, some entries rely on abstracted or supra-dialectal forms—for instance, Cappadocian data in IE-CoR are derived from Dawkins' glossary (Dawkins, 1916), which presents normalized lemmas rather than systematically differentiated subvarieties.

In practice, MG dialectal data remain dispersed across independent projects and publications, without unified alignment, consistent metadata standards, or interoperable encoding. This gap constrains both dialectological research and computational applications.

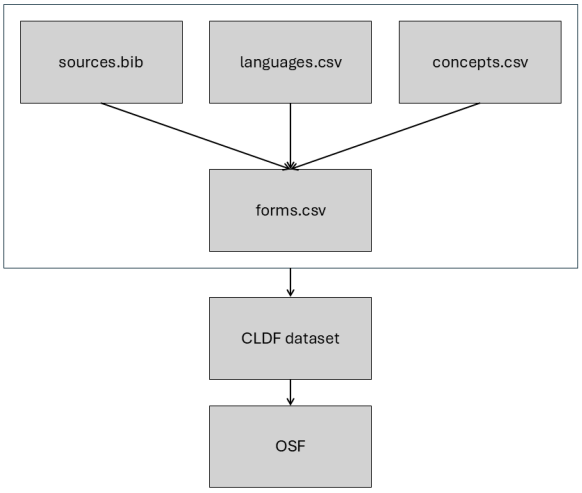

Figure 1: Structure of the CLDF dataset.

Our work addresses this gap by compiling a multi-dialectal MG lexical database in CLDF format, explicitly aligned by concept and designed for computational reuse. In doing so, we integrate Greek dialect data into contemporary lexical infrastructures and create a foundation for dialectometric analysis, NLP diagnostics, and future expansion toward multi-level linguistic documentation.

3. Resource Design

Our database conforms to the Cross-Linguistic Data Formats (CLDF) standard (Forkel et al., 2018b). The current release is structured around the standard CLDF Wordlist module and consists of four core files:

- concepts.csv
- forms.csv
- languages.csv
- metadata.json

In addition, a `sources.bib` file contains the bibliographic references associated with each entry.

The `metadata.json` file defines the dataset structure and explicitly specifies how the tables are interrelated (e.g., linking forms to concepts and language varieties). CLDF is modular and extensible: additional columns or new tables can be added without compromising compliance with the standard. This flexibility is essential for future extensions of the database beyond lexical data. All data are publicly available via OSF.

The core CSV files are described below, highlighting their principal columns and coding decisions. The overall CLDF architecture is presented in Figure 1.

3.1. Dialect Coverage and Selection Criteria (languages.csv)

The `languages.csv` table contains one row per variety (doculect), each identified by a unique ID. The table includes structured metadata such as:

- Glottocode (serves as a unique ID)
- Name
- Longitude
- Latitude
- Locality
- ISO code (if available)
- Phonetic conversion notes
- Comment (free-text)

The current release covers 36 MG varieties (including SMG) spoken in and outside Greece in the 20th century (Figure 2).

A central design decision concerned the selection of varieties to be included in the database. Our sampling strategy aimed to balance three criteria:

1. **Geographical representativity**: coverage of the major dialect areas across the MG continuum, including mainland, insular, and Asia Minor varieties.
2. **Dialectological representativity**: inclusion of members from all traditionally recognized macro- and micro-groupings in Greek dialectology, ensuring comprehensive coverage of the established classificatory landscape.
3. **Documentary adequacy**: availability of sufficient reference materials to support reliable data extraction and future extensibility (including to other levels of linguistic analysis, thereby ensuring cross-level comparability).

Given these criteria, we prioritized varieties for which both lexical and grammatical documentation are available. Because the long-term goal is to expand the database to additional linguistic levels (e.g., phonology and morphosyntax), the inclusion of both types of documentation ensures structural comparability and facilitates future development. As a result, some historically important varieties that are lexically documented but lack sufficient grammatical description (e.g., the Greek of Constantinople) were not included in the present release.

Additionally, to account for intra-dialectal diversity, multiple varieties were included for certain groups (e.g., Cappadocian, Pontic, Italiot Greek). This serves two purposes: (a) to capture significant internal variation often resulting from contact-induced change; and (b) to avoid reliance on supra-dialectal or artificially normalized lemmas that obscure local variation.

Glottocodes. ISO 639-3 codes are limited in dialect coverage and often represent abstract language groupings. As part of this project, we systematically mapped, updated, or created Glottocodes for the varieties in our sample (Hammarström and Forkel, 2022). Each doculect is identified by a Glottocode and Glottolog name.

Many MG dialects are inconsistently represented—or entirely absent—from widely used language catalogues. This complicates dataset linking, reproducibility, and interoperability. Establishing stable identifiers provides:

- Unambiguous reference to dialect entities
- Improved dataset merging and comparison
- Durable cross-resource integration
- Support for cumulative work (e.g., adding grammatical features or linking corpora)

We mapped existing Glottocodes where possible, refined imprecise legacy entries, and introduced new identifiers for previously uncoded varieties.[1]

3.2. Definition of the Concept List (concepts.csv)

The database is organized around a concept list of 345 concepts covering core vocabulary across parts of speech (nouns, verbs, adjectives, and functionally salient items).

The `concepts.csv` file contains one concept per row with associated metadata:

- ID
- Concept
- Comment (precisions on the intended meaning and contexts)
- Concepticon ID
- Concepticon Name
- Concepticon Comment (precisions regarding the Concepticon mapping)

The concept list was constructed by combining and adapting established concept inventories widely used in historical linguistics, especially in research on the Indo-European family (e.g., Dyen et al., 1992; Ringe et al., 2002; Scarborough, 2019). Since there is, to our knowledge, no dedicated concept list specifically designed for the Greek branch or for systematic comparison across MG dialects, the present study also represents an initial step in that direction. In the absence of a Greek-specific benchmark, we used these earlier concept lists as a starting point and adapted them to the lexical and dialectological particularities of Greek as follows:

- merging overlapping concepts,
- resolving differences in semantic granularity,
- removing redundant, sparsely attested, or diagnostically uninformative items, and
- adding or refining concepts that are diagnostically relevant for Greek dialect differentiation.

[1] New and updated Glottocodes will be included in the upcoming Glottolog release (v5.3), but are already available in the Glottolog GitHub repository: https://github.com/glottolog/glottolog.

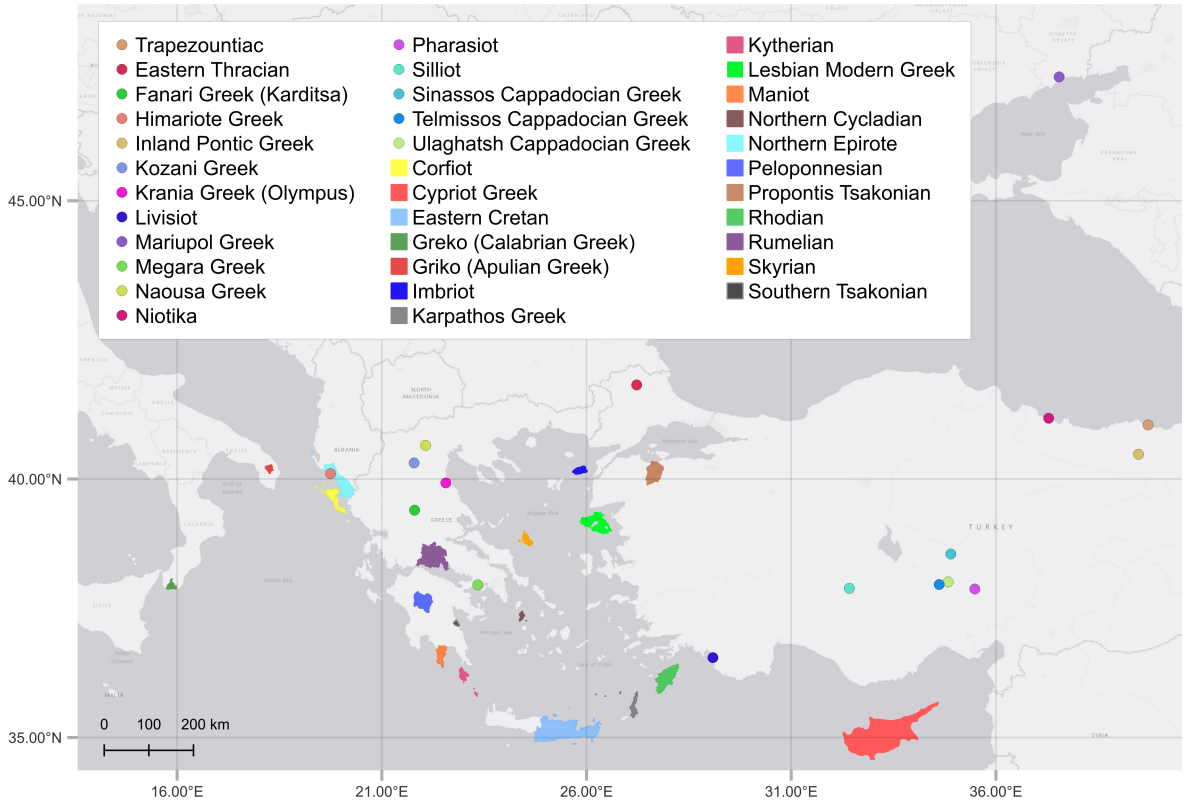

Figure 2: Geographic distribution of the sampled MG varieties represented in the database (excluding SMG).

Items were excluded when they were poorly documented, overly culture-specific, or unlikely to contribute to meaningful cross-dialect contrast. For example, collective items such as GRAIN and CATTLE were removed due to inconsistent documentation across dialect sources. We also eliminated pairs that are etymologically related within the Greek branch and therefore unlikely to yield independent contrastive evidence (e.g., SISTER VS. BROTHER, HE VS. THEY, HE VS. THIS).

Conversely, we split concept pairs that are not etymologically related in Greek in order to avoid masking dialectal differentiation. Nominal and verbal pairs such as SLEEP VS. TO SLEEP and VOMIT VS. TO VOMIT were treated as distinct concepts. We further refined semantically broad items where dialects exhibit systematic lexical differentiation, including FIGHT (WITH SOMEONE) VS. FIGHT (IN WAR), male vs. female goat terminology, and SCRATCH (OF CAT) VS. SCRATCH (AN ITCH). In addition, OLD MAN and OLD WOMAN were introduced as distinct concepts, separate from OLD as an adjectival property applicable to objects.

This approach preserves comparability with established historical-linguistic practice while making it possible to capture distinctions that are especially relevant for dialect-level comparison within Greek. It prioritizes broad attestation across dialect sources, semantic stability, diagnostic value for dialect differentiation, and compatibility with cross-linguistic practice.

Crucially, the list is aligned with Concepticon, linking each entry to standardized concept sets. This alignment ensures cross-dataset comparability, reproducibility, and interoperability with other CLDF-based resources and typological datasets.

3.3. Lexical Entries (forms.csv)

Lexical entries were compiled from dialect dictionaries, glossaries, grammars, published texts, and verified digital databases (including IE-Cor). Where discrepancies were identified, forms were systematically checked against primary sources or dialect speakers when available.

The `forms.csv` table contains one row per dialectal form. Each entry links a specific dialect form to a concept and a language variety. Core fields include:

- ID: Unique identifier
- Language_ID: External reference to the *ID* column of the `languages.csv` table

- Concept_ID: External reference to the *ID* column of the `concepts.csv` table
- Orthographic_form: Attested spelling in the source
- Phonetic_form: IPA transcription
- Phonemic_form: Phonemic representation, where available
- Segments: Segmented IPA representation
- Gender: Grammatical gender (for nouns)
- Translation: Disambiguating gloss and additional meanings, as they appear in the source
- Variations: Attested intra-dialectal variants
- Source: Bibliographic reference, including page numbers
- Source_comment: Notes from the original source (e.g., etymology, usage restrictions, loan status)
- Curator_comment: Editorial comments added by the dataset curator

Orthographic transcriptions follow strictly the spelling conventions of the original sources, with no normalization applied, so that forms can be located directly using the cited references. The only systematic adaptation concerns the removal of polytonic diacritics, which do not encode dialect-specific phonetic information and mainly reflect older printing conventions rather than pronunciation.

The IPA layer required substantial editorial work. Greek dialects lack a unified orthographic standard, and no comprehensive orthography–IPA mapping exists for all varieties. Manolessou et al. (2012) provide the most systematic attempt toward such standardization and serve as a key reference point, particularly for older descriptions, although their coverage remains far from exhaustive. Moreover, some dialect sources deliberately avoid representing phonetic particularities and instead employ SMG orthography to emphasize lexical rather than phonetic differentiation. Therefore, we adopted a multi-step approach: (a) for varieties covered in Manolessou et al. (2012), we followed their transcription guidelines; (b) where IPA or IPA-like transcriptions were provided in the sources, these served as the primary basis; (c) in other cases, IPA forms were established on the basis of grammatical descriptions and general dialectological studies. In such cases, greater divergence between orthographic and phonetic representations may reflect the original editorial choices of the sources. Where necessary, transcription conventions specific to individual varieties are documented in `languages.csv`.

The `Segments` field is derived directly from the `Phonetic_form` field. In the current release, segmentation follows the IPA string and treats as single segments only those consonant clusters that are explicitly marked as unitary by means of a tie bar. At present, this applies only to the dialectal affricates [t͡ʃ] and [d͡ʒ], which in MG dialects typically derive from earlier /k/ and /g/, respectively, before /i, e/. By contrast, clusters such as [ts] and [dz] are currently segmented into two symbols. This choice is deliberate, since their status as single phonemes in MG remains disputed and available research on their phonological behavior is not conclusive enough to justify treating them uniformly as single segments in the dataset (for an overview, see Arvaniti, 2007, 114–117). One limitation of this approach is that dialectal sequences such as [ts] and [dz] may still be segmented into two symbols even in cases where they originate from a single phoneme. This issue will be addressed more systematically in future versions of the dataset.

3.4. Data Format and Interoperability

The database is released in CLDF, ensuring:

- Interoperability with existing lexical comparison and dialectometric tools
- Transparent separation of forms, concepts, and metadata
- Integration with infrastructures such as Concepticon and Lexibank
- Long-term extensibility as new linguistic layers are added

CLDF enables the resource to function not only as a standalone dataset but as an extensible component in a broader ecosystem of dialect-aware NLP and comparative linguistic workflows.

3.5. Automated Validation and CLDF Generation

Throughout the coding process, `pycldf` (Forkel et al., 2018a) and `lexedata` (Kaiping et al., 2022) were used to:

- Transform working spreadsheets into CLDF format
- Validate IPA transcriptions
- Check references consistency
- Detect structural errors and warnings
- Split the IPA representations into segments

After validation, the dataset was exported into its final CLDF form. All files are publicly available via OSF.

3.6. Dataset Coverage and Descriptive Statistics

The dataset's completeness varies by variety, reflecting differences in documentary depth. Across the 36 doculects, concept coverage ranges from 53/345 (15.4%) at the low end to 344/345 (99.7%) at the high end, with a median coverage of 85.4% (294.5 concepts). The overall distribution of concept coverage across varieties is shown in Figure 3. In addition to coverage, the database captures substantial intra-dialectal lexical variation: the mean

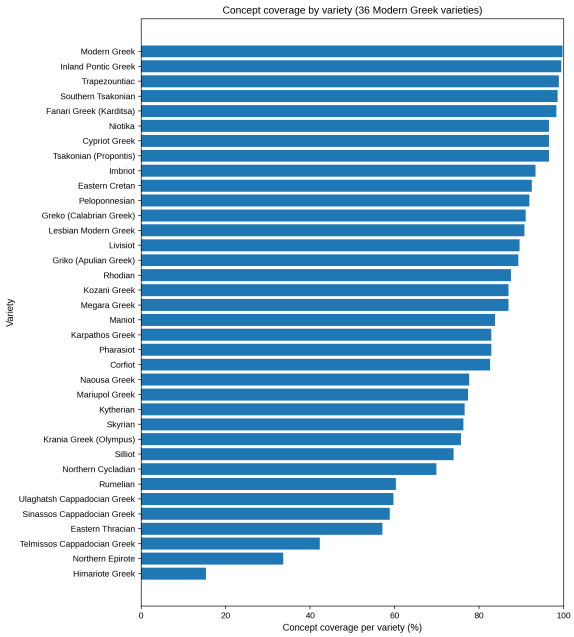

Figure 3: Concept coverage by variety.

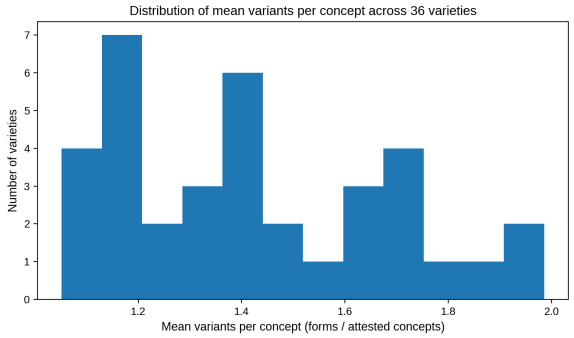

Figure 4: Mean lexical items per concept.

number of lexical items per attested concept ranges from 1.05 to 1.99 across varieties (median 1.37), indicating that many doculects record multiple competing realizations for the same concept. The distribution of this "variation density" measure is shown in Figure 4. Together, these statistics quantify both (a) the dataset's broad comparability across the dialect continuum and (b) its ability to represent microvariation within individual varieties.

4. Dialectometric Analysis

We use the lexical database introduced above to (a) evaluate whether it encodes a dialectologically plausible signal and (b) explore aggregate, data-driven (Nerbonne, 2009) groupings of MG varieties that can be compared to established dialectological classifications.

While dialectometry has substantially advanced modern dialectology (cf. Wieling and Nerbonne, 2015; Goebl, 2017; Heeringa and Prokić, 2017; Nerbonne and Wieling, 2017; Nerbonne et al., 2021; for an overview of computational dialectology with Greek-oriented examples, see Bompolas, 2023, Ch. 2), MG dialectology has only partially benefited from these developments. Existing computational studies have typically focused on restricted geographic areas (e.g., for inner Asia Minor Greek, see Bompolas, 2023; for the Pontic–Azov area, see Kisilier and Solliec, in press) or include only a small number of varieties (cf. Chatzikyriakidis et al., 2026), largely due to infrastructural constraints: the absence of a comprehensive dialect atlas covering the full dialect continuum and the limited availability of large, digitized dialectal corpora or machine-readable dictionaries. Notably, many of the challenges identified by Trudgill (2003, 46–48) for the classification of MG dialects remain relevant, as resource fragmentation and digitization gaps continue to limit quantitative work across broad samples.

Against this background, we provide a dialectometric evaluation grounded in our concept-aligned lexical database spanning a wide range of MG varieties. To our knowledge, no previous dialectometric study has combined this breadth of MG varieties with a concept-aligned, computationally-reusable lexical dataset. We next outline our workflow, which comprises (a) data extraction and parameterization from the CLDF release, (b) computation of lexical–phonetic distances, and (c) exploratory statistical analysis via clustering.

4.1. Data Preparation

For dialectometric analysis, we use the LED-A web application (Heeringa et al., 2023). The `forms.csv` table of the CLDF database was converted into the tabular format required by LED-A. Concretely, we constructed a data matrix in which (a) columns correspond to the 345 concepts, (b) rows correspond to the 36 varieties, and (c) cells contain IPA transcriptions. To maximize comparability, we rely exclusively on the IPA layer of the database, which represents attested forms in a consistent manner and thus provides a reliable basis for measuring dialect distances. When more than one lexical item (either variant or synonym) is attested for a concept in a given variety, all items are encoded within the same cell and separated by a slash. Missing data are left blank.

Since concept coverage varies across varieties (cf. Figure 3), distance computations are necessarily based on the subset of concepts attested for each pair of varieties. Consequently, comparisons involving sparsely documented varieties rely on fewer shared data points and may therefore yield less stable results than comparisons involving better documented varieties.

4.2. Lexical–phonetic Distance Measures

Dialectometric analysis transforms qualitative observations (lexical items, in our case) into a symmetric distance matrix in which each cell expresses the degree of linguistic difference between a pair of varieties. LED-A implements several string-based distance measures. We selected the distance measure that performed best under *local incoherence*, a diagnostic introduced by Nerbonne and Kleiweg (2007): geographically proximate varieties are generally expected to be linguistically more similar, and lower local incoherence indicates a distance measure that better fits this dialectological expectation (see also Heeringa et al., 2006).[2]

Among the available LED-A options, the `A&B sub. ≤ 1, indel ≤ 0.5` setting yielded the lowest local incoherence for our dataset. This corresponds to a feature-sensitive Levenshtein-style distance in which substitution costs depend on phonetic feature similarity (cf. Heeringa and Braun, 2003; Heeringa, 2004). Insertions and deletions are penalized with lower maximum cost than substitutions.[3] The output is a symmetric site × site distance matrix (36×36), which serves as input for clustering analysis.

4.3. Hierarchical Clustering

To explore structure in the distance matrix, we apply hierarchical clustering and visualize the results as a dendrogram (for its usage in dialectometry, see Prokić and Nerbonne, 2008; see also Nerbonne and Wieling, 2017, 401–403). We tested the linkage methods implemented in LED-A and selected the one that accounted for the most variance in our data. In our setting, UPGMA achieved the highest explained variance (71.1%).

4.4. Results and Comparison to Dialectological Classifications

Figure 5 presents the UPGMA dendrogram derived from feature-sensitive string distances. For interpretability, we highlight eight major clusters (color-coded in the figure).

At the highest level, the clustering recovers a broad *Northern* vs. *Southern* partition within the mainland Koine-descended varieties, consistent with accounts that emphasize the phonological

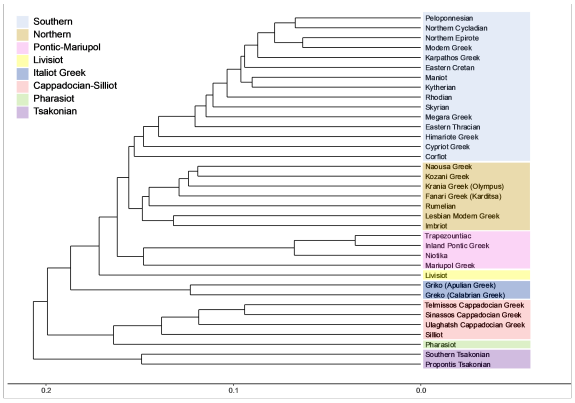

Figure 5: UPGMA dendrogram showing dialect similarity across 36 MG varieties; eight clusters are highlighted.

criterion of Northern vocalism (i.e., raising of unstressed /e, o/ and deletion of unstressed /i, u/) as a major diagnostic, originally proposed by Hadzidakis (1892) (see also Trudgill, 2003, 49–50 and references therein).

Several deviations from a purely geographic North–South split are also informative. In our sample, the geographically northern varieties Skyros and the Eastern Thracian variety (Saranta Ekklisies) pattern closer to the Southern cluster. Both have been described as *semi-northern* or as lacking the full set of Northern vocalism diagnostics (e.g., Psaltis, 1905, 12; Papadopoulos, 1926, 12,14; Katsouda, 2021, 53). Under a distance measure that captures cumulative lexical–phonetic similarity, their aggregate profiles align more closely with the Southern than with the Northern cluster; this is compatible with independent observations that they also share properties with southern varieties (e.g., Katsouda, 2021, 299–300 for Skyrian). Another exception to this geographically aligned partition concerns the varieties of southern Albania (Northern Epirote and Himariote), which do not exhibit Northern vocalism (Kyriazis and Spyrou, 2011, 179). For the remaining Northern varieties, the clustering broadly aligns with traditional expectations, and geographically proximate varieties tend to cluster together.

Within the Southern group, finer-grained subgroupings proposed in the literature are not consistently recovered (see Trudgill, 2003, 49–50 and references therein). This may reflect (a) the fact that the present analysis is based on aggregate lexical–phonetic distances rather than a small set of diagnostic features (e.g. Trudgill's classification relies on six phonological features), and (b) differences in how features are represented in the sources used here. For instance, some subgroup descriptions may be too coarse: Trudgill (2003)

[2] A limitation is that SMG is represented by the coordinates of Athens (as a proxy for its early-20th-century sociolinguistic center), which may slightly distort local incoherence for this single data point. Since this affects only one variety, we consider the diagnostic still informative for model selection.

[3] See the LED-A documentation for the implementation of feature string comparison: `https://www.led-a.org/docs/A&B.pdf`.

notes that Peloponnesian does not exhibit any of the six features used in his classification (pp. 58–59) and that Northern Cycladian is characterized only by tsitakism (p. 60). In our sources, however, both varieties also show velar palatalization, which plausibly contributes to their proximity in our aggregate distance-based clustering. In addition, lexical borrowing can influence similarity structure: Corfiot, for example, appears relatively peripheral within the Southern cluster, plausibly due to extensive Italo-Venetian lexical influence in the data.

Modern Greek (i.e., SMG) likewise patterns clearly with the Southern cluster. This placement is historically plausible, given the widely acknowledged contribution of the Ionian Islands and the Peloponnese to the formation of the Standard (following Horrocks, 2010; see also Trudgill, 2003, 58–59), though the Peloponnesian input has been debated (Pantelidis, 2001). In the dendrogram, SMG clusters closely with Peloponnesian but does not form an immediate sister grouping, reflecting phonetic differences in the dataset.

Beyond the mainland core, several peripheral clusters emerge in line with expectations about contact, relative isolation, and distinct historical trajectories. Within the Pontic—Mariupol Greek cluster, the dendrogram yields a coherent internal subdivision that accords with traditional groupings (e.g. the West vs. East distinction and further Inland/Trapezountiac differentiation; Triandaphyllidis, 1938, 288), while Mariupol Greek clusters closely with Pontic, in line with prior quantitative work suggesting their affinity (Kisilier and Solliec, in press). A particularly instructive case is Livisiot, which emerges as a peripheral branch rather than clustering tightly with either southern varieties or the inner Asia Minor group, contrary to some earlier proposals (see Andriotis, 1961, 11–15 for an overview). Italiot Greek forms a compact pair, reflecting shared contact history and long-term divergence from mainland varieties (cf. Horrocks, 2010, Ch. 14.2.3). Within inner Asia Minor, Cappadocian/Silliot and Pharasiot show a subdivision compatible with qualitative accounts (Manolessou, 2019, 30 and references therein) and earlier computational results (Bompolas and Melissaropoulou, 2023) based on shared phonological traits. Finally, Tsakonian (Propontis and Southern) forms the most divergent cluster, as expected given its distinct genealogical status within Greek and its well-known structural distance from Koine-descended varieties (Horrocks, 2010, Ch. 4.4.3).

Overall, the dialectometric evaluation supports two conclusions. First, the database encodes a clear dialectological signal: major macro-divisions and established subgroupings emerge from aggregate lexical–phonetic distances without being hard-coded into the dataset. Second, mismatches with traditional classifications are interpretable rather than arbitrary, highlighting cases where (a) dialect labels obscure internal heterogeneity, (b) a variety occupies an intermediate position with respect to canonical diagnostics, or (c) borrowing and contact influence similarity structure. These results demonstrate the value of concept-aligned lexical resources for computational dialectology and provide a baseline for future analyses incorporating additional linguistic levels as the resource expands.

5. Conclusion and Future Work

This paper introduced the first systematically aligned, CLDF-compliant lexical database designed to capture lexical variation across the MG dialectal continuum. The current release covers 36 MG varieties (including SMG), aligned over 345 concepts, and includes 14,378 lexical items, encoded with stable identifiers, explicit metadata, and a harmonized IPA layer to support interoperability and reproducible reuse. Alongside the resource description and the modeling decisions required to represent a dialect continuum in an interoperable framework, we evaluated whether the dataset preserves a meaningful dialectological signal. A dialectometric analysis based on feature-sensitive string distances and hierarchical clustering recovers major macro-divisions—most notably the broad Northern vs. Southern partition among Koine-descended mainland varieties—and isolates historically divergent peripheral groups.

Overall, the database provides infrastructure for quantitative dialectology, comparative Greek linguistics, and dialect-aware language technology, establishing a baseline for research on Greek microvariation. It can also support downstream computational tasks such as automatic cognate (List, 2012) and loan (List and Forkel, 2022) detection, as well as dialect-aware NLP (e.g., multi-dialect training and transfer learning, Lin et al., 2019; Faisal and Anastasopoulos, 2022).

We foresee four main directions for further development. First, we will extend the database beyond the lexical layer by adding grammatical features, moving toward a multi-level resource that captures phonological, morphological, and syntactic variation across MG varieties. Second, we plan to incorporate earlier stages of Greek to enable diachronic analyses linking historical change to present-day dialect differentiation. Third, we will enhance accessibility through a web-based interface (e.g., CLLD) alongside the archival CLDF release. Fourth, we will broaden the analytical framework beyond exploratory clustering to include additional dialectometric and phylogeny-oriented methods, systematically comparing quantitative results with established classifications.

6. Data Availability

The CLDF dataset and the corresponding matrix used for the dialectometric analysis are available on OSF: https://osf.io/t5avq/overview.

7. Acknowledgements

This work has been funded by the European Union (ERC, PhylProGramm, 101096554). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

We also express our sincere gratitude to Harald Hammarström for his invaluable assistance with the systematic assignment and revision of Glottocodes and for his support in ensuring their integration into Glottolog.

8. Bibliographical References

Nikolaos P. Andriotis. 1961. *The Idiom of Livisi in Lycia [in Greek]*. Centre for Asia Minor Studies, Athens.

Amalia Arvaniti. 2007. Greek Phonetics. The State of the Art. *Journal of Greek Linguistics*, 8(1):97–208.

Stavros Bompolas. 2023. *Computational dialectology in the linguistic varieties of Cappadocian, Pharasiot, and Silliot*. Ph.D. thesis, University of Patras, School of Humanities and Social Sciences, Department of Philology, Linguistics Section.

Stavros Bompolas and Dimitra Melissaropoulou. 2023. A dialectometric approach to inner Asia Minor Greek: Comparisons and associations between linguistic levels. *Digital Scholarship in the Humanities*, 38(4):1389–1403.

Stergios Chatzikyriakidis, Erofili Psaltaki, Dimitrios Papadakis, Erik Henriksson, and Veronika Laippala. 2026. Perplexity as a Metric for Dialectal Distance: A Computational Study of Greek Varieties. In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 101–112, Rabat, Morocco. Association for Computational Linguistics.

Richard MacGillivray Dawkins. 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, Cambridge.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo. Type: Data set.

Isidore Dyen, Joseph B. Kruskal, and Paul Black. 1992. An Indoeuropean Classification: A Lexicostatistical Experiment. *Transactions of the American Philosophical Society*, 82(5):iii–132.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-Inspired Adaptation of Multilingual Models to New Languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Robert Forkel, Simon J. Greenhill, and Christoph Rzymski. 2018a. cldf/pycldf: pycldf. Zenodo.

Robert Forkel and Johann-Mattis List. 2020. CLDF-Bench: Give your cross-linguistic data a lift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6995–7002, Marseille, France. European Language Resources Association.

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. 2018b. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1):180205.

Hans Goebl. 2017. Dialectometry. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 123–142. Wiley.

Georg Hadzidakis. 1892. *Einleitung in die Neugriechische Grammatik*. Breitkopf & Härtel, Leipzig.

Harald Hammarström and Robert Forkel. 2022. Glottocodes: Identifiers Linking Families, Languages and Dialects to Comprehensive Reference Information. *Semantic Web Journal*, 13(6):917–924.

Martin Haspelmath and Uri Tadmor. 2009a. *Loanwords in the world's Languages: a comparative handbook*. Mouton De Gruyter, Berlin.

Martin Haspelmath and Uri Tadmor, editors. 2009b. *WOLD*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at https://wold.clld.org, Accessed on 2026-02-26.

Wilbert Heeringa and Angelika Braun. 2003. The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances. *Computers and the Humanities*, 37(3):257–271.

Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In *Proceedings of the Workshop on Linguistic Distances*, pages 51–62, Sydney, Australia. Association for Computational Linguistics.

Wilbert Heeringa and Jelena Prokić. 2017. Computational Dialectology. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 330–347. Wiley.

Wilbert Heeringa, Vincent Van Heuven, and Hans Van de Velde. 2023. LED-A: Levenshtein Edit Distance App [computer program]. Available online at https://www.led-a.org/, Accessed on 2026-02-26.

Wilbert Jan Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance (Groningen Dissertations in Linguistics 46)*. Ph.D. thesis, University of Groningen, Groningen.

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, Henrik Liljegren, Richard F. Strand, Geoffrey Haig, Martin Macák, Ronald I. Kim, Erik Anonby, Tijmen Pronk, Oleg Belyaev, Tonya Kim Dewey-Findell, Matthew Boutilier, Cassandra Freiberg, Robert Tegethoff, Matilde Serangeli, Nikos Liosis, Krzysztof Stroński, Kim Schulte, Ganesh Kumar Gupta, Wolfgang Haak, Johannes Krause, Quentin D. Atkinson, Simon J. Greenhill, Denise Kühnert, and Russell D. Gray. 2023. Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages. *Science*, 381(6656):eabg0818.

Geoffrey C. Horrocks. 2010. *Greek: a history of the language and its speakers*, 2nd ed edition. Wiley-Blackwell, Oxford ; Malden, Mass. OCLC: ocn416717812.

ILIK. 2024. *Historical Dictionary of the Cappadocian Dialects [in Greek]*. Athens Academy, Athens. Available online at https://ilik.academyofathens.gr, Accessed on 2026-02-26.

ILNE. 2026. The Historical Lexicon of Modern Greek [in Greek]. Thesaurus Linguae Graecae Project, University of California, Irvine. Available online at https://stephanus.tlg.uci.edu/ilne, Accessed on 2026-02-26.

Gereon A. Kaiping, Melvin S. Steiger, and Natalia Chousou-Polydouri. 2022. Lexedata: A toolbox to edit CLDF lexicaldatasets. *Journal of Open Source Software*, 7(72):4140.

Georgia Katsouda. 2024. Dialect Lexicography: Greek Bibliography (II) [in Greek]. *Lexicographic Bulletin*, 27:355–379.

Georgia Th. Katsouda. 2021. *Stsyriana: A Synchronic Description and Analysis of the Linguistic Idiom of Skyros [in Greek]*. Stamoulis Publications, Athens.

Mary Ritchie Key and Bernard Comrie. 2023. The Intercontinental Dictionary Series. Max Planck Institute for Evolutionary Anthropology. Available online at https://ids.clld.org, Accessed on 2026-02-26.

Maxim L. Kisilier and Tanguy Solliec. in press. Is Azov Greek a variety of Pontic? Preliminary remarks. *Greek Around the World: Papers from the 9 February 2024 Workshop*. HAL Id: hal-04921554.

William A. Kretzschmar. 2017. Linguistic Atlases. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 57–72. Wiley.

Doris Kyriazis and Aristotelis Spyrou. 2011. The Greek Linguistic Idioms of Albania [in Greek]. *Modern Greek Dialectology*, 6:175–199.

Alfred Lameli. 2009. Linguistic atlases - traditional and modern. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space: Theories and Methods*, pages 567–592. Walter de Gruyter.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125, Avignon, France. Association for Computational Linguistics.

Johann-Mattis List and Robert Forkel. 2022. Automated identification of borrowings in multilingual wordlists. *Open Research Europe*, 1:79.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

Johann Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon Greenhill, and Robert Forkel, editors. 2025. *CLLD Concepticon 3.4.0*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Available online at https://concepticon.clld.org, Accessed on 2026-02-26.

Io Manolessou. 2019. The Historical Background of the Asia Minor Dialects. In Angela Ralli, editor, *The Morphology of Asia Minor Greek*, pages 20–65. Brill.

Io Manolessou, Stamatis Beis, and Christina Basea-Bezantakou. 2012. The phonetic transcription of the Modern Greek dialects [in Greek]. *Lexikografikon Deltion*, 26:161–222.

Dimitra Melissaropoulou, Stavros Bompolas, and Charalampos Tsimpouris. 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.

Claudine Moulin. 2009. Dialect dictionaries - traditional and modern. In Peter Auer and Jürgen Erich Schmidt, editors, *Language and Space: Theories and Methods*, pages 592–612. Walter de Gruyter.

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

John Nerbonne and Peter Kleiweg. 2007. Toward a dialectological yardstick*. *Journal of Quantitative Linguistics*, 14(2-3):148–166. Number: 2-3.

John Nerbonne, Jelena Prokić, and Martijn Wieling. 2021. Dialectology for Computational Linguists. In Marcos Zampieri and Preslav Nakov, editors, *Similar Languages, Varieties, and Dialects. A Computational Perspective*, 1 edition, pages 96–118. Cambridge University Press.

John Nerbonne and Martijn Wieling. 2017. Statistics for Aggregate Variationist Analyses. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 400–414. Wiley.

Nikolaos Pantelidis. 2001. Peloponnesian dialectal accent and Standard Modern Greek [in Greek]. In *Proceedings of the Fourth International Conference of Greek Linguistics*, pages 480–486, Thessaloniki. University Studio Press.

Anthimos Papadopoulos. 1926. *Grammar of the Northern Idioms of the Modern Greek Language [in Greek]*. P. D. Sakellariou Press, Athens.

Jelena Prokić and John Nerbonne. 2008. Recognising Groups among Dialects. *International Journal of Humanities and Arts Computing*, 2(1-2):153–172. Number: 1-2.

Stamatios B. Psaltis. 1905. *Thracian Studies or A Study of the Linguistic Idiom of the City of Saranta Ekklisies [in Greek]*. P. D. Sakellariou Press, Athens.

Don Ringe, Tandy Warnow, and Ann Taylor. 2002. Indo-European and Computational Cladistics. *Transactions of the Philological Society*, 100(1):59–129.

Matthew Scarborough. 2019. Cognacy and Computational Cladistics: Issues in Determining Lexical Cognacy for Indo-European Cladistic Research. In *Dispersals and Diversification: Linguistic and Archaeological Perspectives on the Early Stages of Indo-European*, pages 179 – 208. Brill, Leiden, The Netherlands.

Hedvig Skirgård, Hannah J. Haynie, Harald Hammarström, Russell Barlow, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Johannes Englisch, Angela Chira, Annemarie Verkerk, Russell Dinnage, Luke Maurits, Sam Passmore, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browen, Hoju Cha , Lennart Chevallier, Swintha Danielsen, Hugo De Vos, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Grace Ephraums, Giada Falcone, Anna-Maria Fehn, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Samuel Griggs, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Manuel Rüdisühli, Alexandra Marley,

Tânia R. A. Martins, Marvin Leonard Martiny, Celia Mata German, Suzanne Van Der Meer, Jacob Menschel, Jaime Montoya, Michael Müller, Saliha Muradoglu, HunterGatherer, David Nash, Kelsey Neely, Johanna Nickel, Miina Norvik, Olga Olina, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Leah Pappas, Jesse Peacock, India O.C. Pearey, Naomi Peck, Jana Peter, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Tihomir Rangelov, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Kristian Roncero Toledo, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Dineke Schokkin, Jeff Siegel, Jane Simpson, Amalia Skilton, Hilário De Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Daniel Wikalier Smith, Nicholas Williams, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye , Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Olena Shcherbakova, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank v1.0. Zenodo.

George Starostin. 2011. The Global Lexicostatistical Database. Higher School of Economics and Santa Fe Institute. Available online at https://starlingdb.org/new100/, Accessed on 2026-02-26.

Morris Swadesh. 1955. Towards Greater Accuracy in Lexicostatistic Dating. *International Journal of American Linguistics*, 21(2):121–137.

Manolis Triandaphyllidis. 1938. *Modern Greek Grammar. Volume A: Historical Introduction [in Greek]*. Manolis Triandaphyllidis Foundation, Thessaloniki.

Peter Trudgill. 2003. Modern Greek dialects: A preliminary classification. *Journal of Greek Linguistics*, 4(1):45–63.

Christos Tzitzilis. 2000. Modern Greek dialects and Modern Greek dialectology [in Greek]. In A.-Ph Christidis and others, editors, *The Greek Language and its Dialects [in Greek]*, pages 15–22. Ministry of National Education and Religious Affairs & Centre for the Greek Language, Athens.

Jacques Van Keymeulen. 2017. The Dialect Dictionary. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, 1 edition, pages 39–56. Wiley.

Tessa Y. Vermeir, Marc Allassonnière-Tang, and Guillaume Segerer. 2024. LA80: A Lexical Database of 10 Bantu A80 Languages. *Journal of Open Humanities Data*, 10:42.

Søren Wichmann, Eric W. Holman, Cecil H. Brown, Matthew S. Dryer, and Qibin Ran. 2025. The ASJP Database (version 21). Available online at https://asjp.clld.org, Accessed on 2026-02-26.

Martijn Wieling and John Nerbonne. 2015. Advances in Dialectometry. *Annual Review of Linguistics*, 1(1):243–264.