Towards a Greek Proverb Atlas: Computational Spatial Exploration and Attribution of Greek Proverbs

John Pavlopoulos[1,2], Panos Louridas[1], Panagiotis Filos[3]
[1] Athens University of Economics and Business, Greece
{annis,louridas}@aueb.gr
[2] Archimedes/Athena RC, Greece
[3] University of Ioannina, Greece
pfilos@uoi.gr

Abstract

Proverbs carry wisdom transferred orally from generation to generation. Based on the place they were recorded, this study introduces a publicly-available and machine-actionable dataset of more than one hundred thousand Greek proverb variants. By quantifying the spatial distribution of proverbs, we show that the most widespread proverbs come from the mainland while the least widespread proverbs come primarily from the islands. By focusing on the least dispersed proverbs, we present the most frequent tokens per location and undertake a benchmark in geographical attribution, using text classification and regression (text geocoding). Our results show that this is a challenging task for which specific locations can be attributed more successfully compared to others. The potential of our resource and benchmark is showcased by two novel applications. First, we extracted terms moving the regression prediction toward the four cardinal directions. Second, we leveraged conformal prediction to attribute 3,676 unregistered proverbs with statistically rigorous predictions of locations each of these proverbs was possibly registered in.

1 Introduction

A proverb (paroemia) is a popular saying that offers general advice or wisdom (Davis et al., 2021). Proverbs have not only been guiding social interactions of people for thousands of years (Hrisztova-Gotthardt and Varga, 2014), but they continue to do so today, as is evident from proverbs such as "Garbage In, Garbage Out" (GIGO) (Mieder, 2004), a popular concept in computer science. Recently, computational approaches have attempted to assist paroemiography (Baptista and Reis, 2022; Pimpalgaonkar et al., 2021), concerned with the collection and classification of proverbs. Also, paroemiology (Davis et al., 2021) addresses questions regarding the definition, form, structure, style, content, function, meaning, and value of proverbs.

1.1 Motivation

By occurring cross-linguistically/culturally or inherited from generation to generation (Mieder, 2008; Hrisztova-Gotthardt and Varga, 2014; Davis et al., 2021), proverbs are carriers of oral wisdom with cultural and historical value. Understanding their distribution is vital for insights into regional variations in Greek culture, but their propagation has hardly been tackled in literature (Villers, 2022). In this work, inspired by computational paroemiography that has already addressed the thematic classification of proverbs (Noah and Ismail, 2008; Baptista and Reis, 2022), we focus on their geolocation, i.e., classifying proverbs based on where they were registered in. Motivated by the unique linguistic identity of specific locations (Prokić and Nerbonne, 2008), we hypothesise that a proverb may be geolocated based on its text alone.

1.2 Contributions

New dataset and exploration: We introduce the first machine-actionable publicly available dataset of Greek proverbs,[1]comprising information about the location each has been collected from. Our data analysis revealed information about the dispersion of the most and least widespread proverbs, which was followed by clustering the linguistic alternations of the former, and by presenting the most distinctive character n-grams per location of the latter. This computational spatial approach is novel, especially due to its complementing to traditional methods nature in terms of scale and depth.
Attribution benchmark: We used our dataset to benchmark machine and deep learning classification and regression algorithms for the task of attributing the least widespread proverbs. Our findings show that (i) specific locations are classified with high accuracy while others are not; (ii) conventional machine learning classifiers outperform a

[1] https://github.com/greek-proverb-atlas/proverbs.gr

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11842–11854
November 12-16, 2024 ©2024 Association for Computational Linguistics

fine-tuned BERT classifier for most locations; (iii) text geocoding (Melo and Martins, 2017) can yield a mean absolute error of 1.31 (lat; 145km) and 1.85 (lon; 163km), and terms that push the regression prediction to specific cardinal directions.

Attributing unregistered proverbs: The location of 3,676 Greek proverbs is missing. To address this issue, we equipped multi-class text classification with conformal prediction (Sadinle et al., 2019), which quantifies the uncertainty of predictions made by machine learning models, providing a way to generate prediction sets that contain the true output with a specified probability. Paving that way, we provide all the possible locations each unregistered proverb could have been collected from, coming with mathematical guaranteed coverage.

In the remainder of this work, we commence with an exploratory analysis (§2) that includes a spatial distribution of proverbs. We then transitioned to a benchmarking phase for the prediction of the location of proverbs (§3). This sequential approach is fundamental in comprehensively understanding the dataset and unlocking two potential applications (presented in §4). A discussion section (§5) is followed by our conclusions.

2 The Greek Proverb Atlas Dataset

2.1 The primary source

We collected 134,493 proverbs (and variants) from the Hellenic Folklore Research Centre.[2] Recorded by various contributors since 1807 (Appendix A), these proverbs exist in a digital repository.[3] Although accessible online, currently, they cannot serve any data exploratory analysis or machine learning purposes. To address this, we crawled this repository to extract the proverb, the collector (more details in Appendix B), and the location the proverb was registered in.

We removed 384 that were noisy (e.g., comprising a translation of the proverb or simply defining a word). In 3,698 proverbs no information about the place of collection was present, while in 14,835 proverbs the collector was anonymous. Out of the rest, 108,410 were unique. The location contained both the broader location and the specific place where the proverb was recorded; we kept the broader location by tokenising and keeping the first token (e.g., ‘Ioannina’ in ‘Ioannina, Chouliarades’). This process yielded 134 unique locations, which were geocoded,[4] in order to link each one to a latitude and longitude.

2.2 Proverb dispersion

We focused on 3,204 proverbs (i.e., their types, excl. duplicates) that were registered by named collectors and which were found in exactly the same form in 104 different Greek locations (i.e., dispersed). Figure 1 depicts a heatmap of the number of proverbs that occur at least 25 times in two different locations. It can be seen that the branching factor per location (i.e., the number of non-zero cells per row, reflected visually in the heatmap by rows in light colours) varied. The highest branching factor, indicating a location whose proverbs exist in many other locations, was that of Asia Minor and Thrace, followed by Epirus and other regions (more information is added in Appendix C).

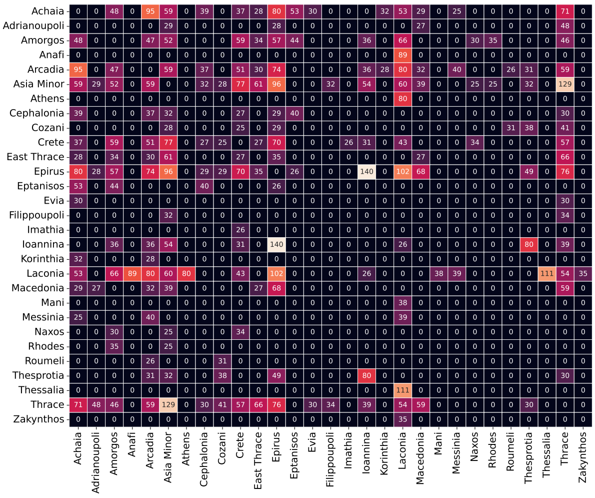


Figure 1: Heatmap of dispersed Greek proverbs.

The picture is different for proverbs existing in a single location, i.e. not shared with other locations (not dispersed). As can be seen in Figure 2, the locations with the most non-dispersed proverbs are islands. Specifically, 7,962 proverbs were collected from the Ionian Islands, and 7,020 from Amorgos, in the Aegean sea, followed by Achaia (6,743) in the Peloponnese (not an island) and Cyprus (4,910).

2.3 Linguistic alternations

We observe that linguistic alternations exist in our data, increasing the challenging nature of our work. By focusing on proverbs that are already widespread identically across locations, we assessed if they also exist in altered, non-duplicate

[2] http://kentrolaografias.gr

[3] https://creativecommons.org/licenses/by-nc-nd/4.0/

[4] https://github.com/geopy/geopy

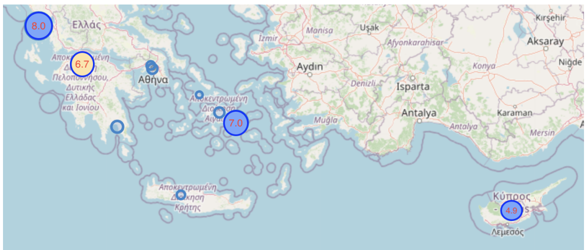


Figure 2: Distribution of proverbs existing in one location alone. Number of proverbs (in thousands) is shown for the four largest circles. Circles on islands are coloured (inside) in blue.

versions in other locations as well.

2.3.1 Widespread proverbs

The three most widespread proverbs are shown in Table 1. We measured the Levenshtein distance between each of these and any other (non-duplicate) proverbs respectively in our corpus (i.e., with a distance of five or less character edits).

Tongue doesn't have bones In the first case, we found six more locations where slightly altered versions of this proverb were collected. In Rhodes, a dialectal verbal form τσακά (tsaka) occurs instead of the far more common τσακίζει (tsakizei). On Skyros, on the other hand, και (kai) > τσαι (tsai) and τσακίζει (tsakizei) > τσατσίζει (tsatsizei), showing features of the geographically far most widespread palatalisation (Trudgill, 2003). We also observe that in locations where the proverb is already reported as a duplicate, other variants exist. In Thrace, for example, we observe a version with an (unaccented) high vowel loss, where τσακίζει (tsakizei) > τσακίζ' (tsakiz').

Knock on a deaf man's door In the second case, we found 19 more locations, which make this proverb the most widely distributed, i.e., over 39 locations in total. Cyprus was one of these, with final /n/ retention being observed: όσον (oson), not όσο (oso). The same feature was observed on the islands of Karpathos and Chios, both of which can be considered places showing retention of an ancient final nasal (Trudgill, 2003).[5]

Easy come, easy go In the third case, we found 17 more locations. Asia Minor and Adrianoupoli (Hadrianopolis) raise the unstressed /e, o/ to /i, u/ respectively, presenting the pronunciation ανεμομαζώματα (anemomazomata) > ανιμουμαζώματα (animoumazomata), which is considered a feature of Extreme Northern dialectism (Trudgill, 2003). We also observe an apparent deletion of dental nasal stop /n/ in δαιμο(ν)οσκορπίσματα (daimo(n)oskorpismata). This is a known feature of some Aegean dialects but evidence for its presence in other locations is insufficient (Trudgill, 2003).

2.3.2 Visualising linguistic paths

The linguistic alternations of a given target proverb can be clustered, in order to visualise possible linguistic "paths". For example, neighbouring locations may have used alternations of the same proverb, or a proverb may have travelled along a trade route. To mine the data for linguistic paths, we opted for hierarchical clustering. We used agglomerative clustering with single linkage and term-frequency-inverse-document-frequency (TFIDF) text representations with n-grams ($2 \leq n \leq 5$). In Appendix E, we present the clustering of the most widespread Greek proverb (Table 1 a) as an example. We note, however, that hierarchical clustering is only one of the possible ways to mine linguistic paths, and a more thorough benchmark of data mining methods is considered as a promising future research direction.

Our investigation of the clusters shows that Crete and Thrace both employ τσακεί (tsakei) rather than common τσακίζει (tsakizei), probably due to paradigmatic remodelling. A similar study on "Knock on a deaf man's door", the proverb with the most alternations across locations, we found that Kos, Samos and Epirus share the same alternation for "(the) door" (acc. sg.): (την)πόρτα ((tin)porta) > (τη)μπόρτα ((ti)mporta)) (pronounced /ti(m)borta/). A different alternation is shared by Chios and Adrianoupoli, with the final nasal retained: όσον (oson), not όσο (oso).

2.4 Distinctive local terms

Besides linguistic alternations, our data also comprise distinctive local terms, which could explain the linguistic identity of specific locations (Prokić and Nerbonne, 2008). A limitation of previous proverb classification studies concerns the insufficient confidence in the knowledge of the geographical incidence of specific linguistic features. This limitation hinders researchers from employing these features cartographically. A feature of many Aegean dialects, for example, is that the velar fricative consonant underwent lenition (i.e. it acquired a less "strong" pronunciation) and became

[5] We refer the interested reader to Map 4 in Trudgill (2003).

	TEXT	TRANSLITERATION	TRANSLATION	ED	ND
(a)	Η γλώσσα κόκκαλα δεν έχει και κόκκαλα τσακίζει	i glossa kokkala den ekhei kai kokkala tsakizei	Tongue doesn't have bones but bones it crushes	23	28
(b)	Στου κουφού την πόρτα όσο ϑέλεις βρόντα	stou koufou tin porta oso theleis vronta	You can knock on a deaf man's door forever (Kazantzakis, 1996)	20	53
(c)	Ανεμομαζώματα, διαβολοσκορπίσματα	anemomazomata, diavoloskorpismata	Easy come, easy go	18	33

Table 1: The three most common proverbs across locations, along with the number of those locations (Exact Duplicates, ED) and the number of locations we found near-duplicates (ND; examples in Appendix D).

semivocalic (i.e., it became an approximant) before disappearing altogether at a following stage, e.g., μεγάλο (meγalo) (> probably /mewalo/) > μεάλο (mealo). This feature, however, was disregarded in (Trudgill, 2003), due to limited quantitative information.

To address this limitation, we represented each word in a proverb of a specific location with its frequency in proverbs of that location, normalised with the inverse location frequency of that word (i.e., how rarely it appears in a proverb in any location). Inspired by TFIDF, we dub this representation as TFILF, because the document in our case is a concatenation of localised proverbs.[6] Preprocessing comprised lower casing and filtering out features existing in more than half of the places (not informative). Further preprocessing was avoided, in order to avoid harming features revealing the oral, local speech. High TFILF indicates that a word is frequent in the proverbs of a specific location but not in the proverbs of the other locations.

In Cyprus, we observe that the palato-alveolar pronunciation (i.e. the voiced affricate /dz/) is prevalent, verifying prior studies (Trudgill, 2003). Extreme palatalization, in general, is found in various locations, taking for example the alveolo-palatal pronunciation (i.e., the voiceless affricate [tɕ]), which is present in Crete, Evia, Skyros, Kefalinia, Karpathos, Lesvos, etc. The word tokens with the highest TFILF are shown in Appendix F while a thorough investigation of more frequent terms is left out for future work.

3 Benchmark: Geographical Attribution

More than three thousand proverbs remain without any geographical attribution today (§2.1). Each of these proverbs may be attributed geographically by the experts, if provided with some assistance. Being able to filter, for example those that are more likely to come from Cyprus, we could possibly assist experts focusing on that location. Therefore, we surmised that, to some extent, *each location has its own linguistic identity, which can be modelled.* To investigate this hypothesis, we opted for text classification (i.e., detecting the place of origin), and text regression (i.e., estimating the geographical coordinates of the place of origin). We note that, although our hypothesis may be true, proverbs are often edited, "regularised" linguistically by those who collect and record them,[7] increasing the already challenging nature of our investigation. For our experimental purposes, we created a balanced corpus, using 500 proverbs per location and 11,500 overall,[8] splitting randomly into train (90%), dev (5%), and test (5%) subsets.[9]

3.1 Text classification

3.1.1 Text classification methods

Authorship analysis Although there may be different variants within a certain location, we assumed that each location has by and large a distinctive linguistic identity, as if all the proverbs registered within that location were authored by a single author or group of authors. Hence, the task of attributing the geographical location of a proverb can be approached as an authorship task. Given that the oral, collective composition of proverbs resembles, to a significant extent, that of ancient literature, where parts are later interpolations inserted during the process of oral transmission,[10] we experimented with the authorship analysis (AA) method of Pavlopoulos and Konstantinidou (2022) by training one 3-gram character-based language model

[6] We used the scikit-learn implementation of TFIDF.

[7] Edits may occur even unintentionally, e.g., if an odd word was considered as wrongly recorded.

[8] We used the 23 locations with 1k+ proverbs, excl. duplicates, to allow the development of a balanced dataset. We used a high threshold (500) to allow for Monte Carlo sampling.

[9] Used for all our experiments, unless otherwise stated. For Monte Carlo validation, we simply changed the seed (2023,2024,2025).

[10] Pavlopoulos and Konstantinidou (2022) defined authorship analysis as the study of any association between a corpus and a group of authors. We employ their definition.

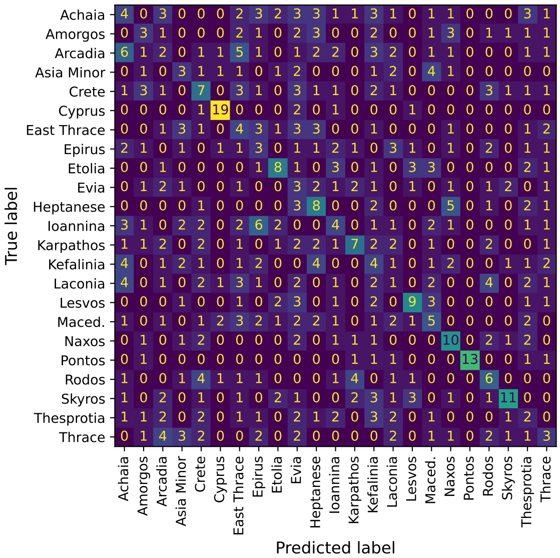
on the proverbs of each location, merged as if they were verses of a single poem from a single author. This process yielded twenty-three language models. Then, for an unseen proverb, we computed the perplexity (Goodfellow et al., 2016; Graves, 2013), or PPL for short, per model and used the lowest value to attribute the location of the proverb.

Stylistic features and supervised learning By approaching the task as a multi-class classification problem, AA can be seen as one out of many available text classification algorithms. Therefore, we experimented also with (multinomial) Logistic Regression (LR), Random Forests (RF), K-Nearest Neighbours (KNN), and linear Support Vector Machines (SVM) (Cortes and Vapnik, 1995). We opted for representations based on character-n-gram frequency and inverse-proverb frequency,[11] which can reveal linguistic information without introducing assumptions (as in feature engineering). Transfer learning was assessed with GrBERT (Koutsikakis et al., 2020), which is a BERT model (Devlin et al., 2018) based on subwords and pre-trained on contemporary Greek text. We fine-tuned this pre-trained Transformer for 100 epochs, using patience of 5 epochs, early-stopping based on the macro-averaged F1, learning rate of 2e-05, batch size of 64, and max length of 32.[12]

3.1.2 Classification experimental results

Table 2 shows that KNN is the least successful, followed by Random Forest. GrBERT achieved the best performance overall together with LR (F1 of 0.30), achieving the best results for different locations. The highest F1 across locations was achieved for Cyprus (GrBERT), followed by Pontos (AA), Skyros (LR), Karpathos (SVM), and Lesvos (LR). The reason why conventional machine learning algorithms (i.e., SVM, LR) outperformed GrBERT for most of the locations (15 out of 23) lies mainly in the linguistic peculiarities of proverbs (Trudgill, 2003), indicating that transferring learning from contemporary Greek is not very valuable.

Besides GrBERT, discouraging results were obtained also when we used embeddings from large language models. Specifically, we experimented with sentence (i.e., proverb) representations by taking embeddings for each proverb from three different large language models: Gemini (`text-multilingual-embedding-002`), OpenAI (`text-embedding-3-large`), and Mistral (`mistral-embed`). We fitted a Logistic Regression model for each one of them. The results were poor (average F1 was 0.21 for OpenAI, 0.20 for Gemini, 0.21 for Mistral). This may come as no surprise if we take into account that the distinctiveness of proverbs lies on their linguistic features, and not on the semantics.



	Achaia	Amorgos	Arcadia	Asia Minor	Crete	Cyprus	East Thrace	Epirus	Etolia	Evia	Heptanese	Ioannina	Karpathos	Kefalinia	Laconia	Lesvos	Maced.	Naxos	Pontos	Rodos	Skyros	Thesprotia	Thrace
Achaia	4	0	3	0	0	0	2	3	2	3	3	1	1	3	1	0	1	1	0	0	0	3	1
Amorgos	0	3	1	0	0	0	2	1	0	2	3	0	0	2	0	0	1	3	0	1	1	1	1
Arcadia	6	1	2	0	1	1	5	1	0	1	2	2	0	3	2	0	1	1	0	0	0	1	1
Asia Minor	0	1	0	3	1	1	1	0	1	2	0	0	0	1	2	0	4	1	0	0	0	0	0
Crete	1	3	1	0	7	0	3	1	0	3	1	1	0	2	1	0	0	0	0	3	1	1	1
Cyprus	0	0	0	0	1	19	0	0	0	2	0	1	0	0	0	1	0	0	0	0	0	0	0
East Thrace	0	0	1	3	1	0	4	3	1	3	3	0	0	1	0	0	0	1	0	0	0	1	2
Epirus	2	1	0	1	0	1	1	3	0	1	1	2	1	0	3	1	0	1	0	2	0	1	1
Etolia	0	0	1	0	0	0	0	1	8	1	0	3	0	1	0	3	3	0	0	0	0	2	1
Evia	0	1	2	1	0	0	1	0	0	3	2	1	2	1	0	1	0	1	0	1	2	0	1
Heptanese	0	0	0	0	1	0	0	0	0	3	8	0	0	2	0	0	0	5	0	1	0	2	1
Ioannina	3	1	0	2	2	0	2	6	2	0	0	4	0	1	1	0	2	1	0	0	0	1	1
Karpathos	1	1	2	0	2	0	1	0	1	2	2	1	7	2	2	0	1	0	0	2	0	0	1
Kefalinia	4	0	1	2	1	0	1	2	0	0	4	0	0	4	1	0	1	2	0	0	1	1	2
Laconia	4	0	1	0	2	1	3	1	0	2	0	1	0	2	1	0	2	0	0	4	0	2	1
Lesvos	0	0	0	1	0	0	1	0	2	3	0	1	0	2	0	9	3	0	0	0	0	1	1
Maced.	1	0	1	0	1	2	3	2	1	2	2	1	0	1	2	1	5	0	0	0	0	2	0
Naxos	0	1	0	1	2	0	0	0	0	2	0	1	1	1	0	0	0	10	0	2	1	2	0
Pontos	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	13	0	0	1	1
Rodos	1	0	0	1	4	1	1	1	0	0	0	1	4	0	1	1	0	0	0	6	0	0	0
Skyros	1	0	2	0	1	0	1	0	2	1	0	0	2	3	1	3	0	1	0	1	11	0	0
Thesprotia	1	1	2	0	2	0	1	1	0	2	1	2	0	3	2	0	1	0	0	0	1	2	0
Thrace	0	1	4	3	2	0	0	2	0	2	0	0	0	0	2	0	1	1	0	2	1	1	3

Figure 3: Confusion matrix on the test data of the twenty-three AA language models, one per location.

Despite its simplicity, AA was the best for Pontos (also for Rhodes and Macedonia). AA's confusion matrix (see Figure 3) reveals that Cypriot proverbs are easily recognised by the respective model while Pontos, Skyros, Naxos, and Lesvos follow closely. Proverbs from Laconia, on the other hand, are the most difficult to recognise, probably due to their high dispersion (§2.2). We also observe that confusion often concerns neighbouring or partly overlapping locations, i.e., Epirus and Ioannina,[13] and Achaia and Arcadia.[14] Based on these results, our findings indicate that specific places, such as Cyprus,[15] have their own distinct linguistic identity, and thus respective proverbs (or proverb variants) can be attributed correctly.

[11] We used grid search to tune the representation, reaching optimal results with character n-grams (bigrams-fourgrams), max doc. frequency of 50%, and discarding hapax legomena. Algorithms were tuned with Optuna (Akiba et al., 2019) and we used the implementations of the scikit-learn library.

[12] GrBERT has 110m parameters; experimented on T4 GPU.

[13] Ioannina is the largest regional unit in Epirus, but these two sets were not merged due to their high number of proverbs.

[14] Confusion could be due to linguistic factors or due to history, geography, dialect and language contact (e.g. the number of Arvanites in parts of the Peloponnese).

[15] We refer to Cypriot Greek, a variety of Modern Greek.

	LR	SVM	KNN	RF	AA	GrBERT
Cyprus	0.74±0.01	0.58±0.05	0.62±0.03	0.58±0.03	0.66±0.04	**0.80±0.02**
Pontos	0.59±0.04	0.54±0.02	0.57±0.02	0.43±0.04	**0.69±0.03**	0.68±0.01
Skyros	**0.63±0.02**	0.57±0.01	0.51±0.03	0.57±0.04	0.40±0.03	0.54±0.05
Karpathos	0.46±0.04	**0.52±0.02**	0.32±0.03	0.49±0.04	0.40±0.09	0.34±0.03
Lesvos	**0.49±0.05**	0.44±0.02	0.25±0.05	0.30±0.06	0.34±0.10	0.43±0.03
Achaia	0.23±0.08	0.28±0.10	0.15±0.03	0.29±0.10	0.13±0.05	**0.46±0.02**
Aetolia	**0.40±0.03**	0.38±0.05	0.32±0.02	0.31±0.05	0.29±0.06	0.39±0.06
Heptanese	**0.40±0.04**	0.38±0.03	0.32±0.01	0.31±0.04	0.29±0.05	0.32±0.04
Naxos	0.32±0.04	**0.37±0.08**	0.20±0.05	0.24±0.05	0.28±0.02	0.32±0.08
Rhodes	0.31±0.05	0.33±0.02	0.27±0.07	0.27±0.02	**0.36±0.04**	0.27±0.05
Crete	0.26±0.04	**0.34±0.09**	0.22±0.06	0.15±0.02	0.27±0.05	0.15±0.09
Amorgos	0.25±0.04	**0.32±0.03**	0.23±0.02	0.24±0.04	0.32±0.06	0.32±0.01
Ioannina	0.22±0.07	0.18±0.03	0.07±0.03	0.17±0.05	0.16±0.04	**0.28±0.04**
East Thrace	0.18±0.05	0.17±0.04	0.16±0.05	0.19±0.05	0.18±0.00	**0.27±0.02**
Kefalinia	0.19±0.05	0.17±0.06	0.15±0.01	0.16±0.05	0.18±0.03	**0.25±0.01**
Macedonia	0.18±0.04	0.23±0.10	0.09±0.03	0.11±0.06	**0.25±0.02**	0.16±0.06
Thesprotia	0.23±0.03	**0.25±0.03**	0.18±0.04	0.17±0.04	0.17±0.04	0.15±0.01
Evia	0.19±0.02	0.21±0.04	0.15±0.05	0.08±0.00	0.14±0.02	**0.21±0.03**
Thrace	0.10±0.03	**0.16±0.05**	0.09±0.02	0.12±0.07	0.12±0.05	0.12±0.10
Laconia	0.06±0.04	0.06±0.01	0.02±0.02	0.12±0.05	0.11±0.04	**0.13±0.01**
Arcadia	0.12±0.05	0.08±0.01	0.07±0.02	**0.13±0.01**	0.09±0.01	0.11±0.03
Epirus	**0.13±0.04**	0.04±0.02	0.09±0.02	0.06±0.06	0.06±0.03	0.11±0.03
Asia Minor	0.11±0.03	0.11±0.05	0.04±0.02	0.05±0.02	0.09±0.05	**0.12±0.03**
AVERAGE	0.30	0.29	0.22	0.24	0.26	0.30

Table 2: F1 per location per classification algorithm. The ranking is based on the best performance achieved and the mean and the standard error of the mean are shown across three random splits. The best per line is shown in bold. A random classifier achieves an average F1 of 0.04±0.02 (the maximum F1 for a single location is 0.07).

3.2 Text regression

3.2.1 Text regression methods

By geocoding all the locations in our dataset (§2.1), we are able to approach the task of geographical attribution as a multi-output regression problem, i.e., learning to predict the latitude and the longitude from the text. Using the balanced version of the corpus and TFIDF features,[16] we experimented with Linear Regression, Elastic Net, K-Nearest Neighbours Regression, Random Forests, and, the best for small-difficult datasets, Extremely Randomised Trees (Xtrees) (Fernández-Delgado et al., 2019).

3.2.2 Regression experimental results

Table 3 shows that Elastic Net and GrBERT are the best regressors in latitude and longitude respectively. Even though Elastic Net provides slightly better results than Linear Regression, the parameter $\alpha \approx 0.0006$; when $\alpha = 0$, Elastic Net is equivalent to ordinary least squares. We note that all systems exhibit a higher error for longitude compared to latitude, with the explanation likely being rooted in the historical processes of their spread, because dispersion is higher in the former (2.79 vs 1.74).

4 Applications

Our work showed that specific locations, wherein Greek proverbs have been registered, have a distinctive linguistic identity, as was evidenced by our supervised learning benchmark (tables 2, 3). Such locations are Cyprus and Skyros, which are easier to classify, most probably due to their linguistic distinctiveness (e.g., extreme palatalization; §2.4). Building beyond this observation, we discuss two applications. First, we have extracted and discuss words that move the regression outcome toward the North, South, East, or West. Second, we use conformal prediction to attribute unregistered proverbs.

4.1 Cardinal direction driving words

A best-performing regression model in our benchmark was Elastic Net (Table 3). Given that important features can "drive" the model's predicted geographical coordinates to specific directions, we used feature importance as a means to identify words[17] that push the prediction towards a higher longitude (East), a lower longitude (West), a higher latitude (North), or a lower latitude (South). In Table 4, which shows the most important features per cardinal direction, we can see that South and

[16] We used scikit-learn, employing the text representation of our classification benchmark, and we tuned with Optuna.

[17] We changed character n-grams to word unigrams in this experiment to maximise interpretability.

	LAT		LON		AVG	
	MAE ↓	MSE ↓	MAE ↓	MSE ↓	MAE ↓	MSE ↓
Elastic Net	**1.31±0.01**	**2.69±0.04**	1.88±0.12	5.99±0.81	1.59	4.34
Linear	1.32±0.01	2.73±0.02	1.90±0.12	6.09±0.79	1.61	4.41
Random Forest	1.33±0.02	2.81±0.05	1.88±0.09	6.17±0.74	1.61	4.49
Extra Trees	1.34±0.02	2.82±0.05	1.89±0.10	6.21±0.76	1.61	4.52
K-NN	1.38±0.01	2.95±0.03	2.00±0.10	6.96±0.87	1.69	4.96
GrBERT	1.33±0.04	2.70±0.08	**1.81±0.04**	**5.34±0.37**	**1.57**	**4.02**
Baseline	1.40±0.02	3.08±0.06	2.05±0.11	7.84±0.93	1.73	5.46

Table 3: Mean absolute and squared error per regressor, averaged and st. error of mean across four runs (best in bold, ranked based on average performance). The average lat/lon is used as a baseline predictor.

East share the phenomenon of tsitakism (/dz/).

North		South		East		West	
τρώγει	(trogei)	τζ	(tz)	τζ	(tz)	ναν	(nan)
πε	(pe)	τζαι	(tzai)	τζαί	(tzai)	τσου	(tsou)
ατ	(at)	τζαί	(tzai)	τζαι	(tzai)	αφέντη	(afenti)
έφαγε	(efage)	τζι	(tzi)	τζι	(tzi)	μύλο	(mylo)

Table 4: Top-features (i.e., tokens, not necessarily words) of ElasticNet ranked based on latitude: descending (North), ascending (South); and longitude: descending (East) and, ascending (West).

4.2 Conformal prediction

Conformal prediction is a general methodology for constructing prediction intervals (Vovk et al., 2005, 1999). In our multi-class classification setting, a set of all the possible classes (locations) are returned for a single proverb, instead of a single class label (location), which also come with a mathematically guaranteed coverage (Sadinle et al., 2019; Romano et al., 2020).[18] Following the work of Sadinle et al. (2019), we calibrated the conformity scores strategy on a development set.[19] We used LR and we opted for an alpha value of 0.05.[20] The probability that the true class is in the prediction set, measured using a separate test set, was found to be 97%. Figure 4 shows the number of proverbs per prediction set size. Although right-skewed (i.e., most proverbs have wide prediction sets), the histogram also comprises proverbs with relatively narrow prediction sets (e.g., half-sized). Hence, although it cannot narrow down the possible locations for any proverb, it can suggest narrow prediction sets for specific proverbs, providing hints to facilitate the experts.

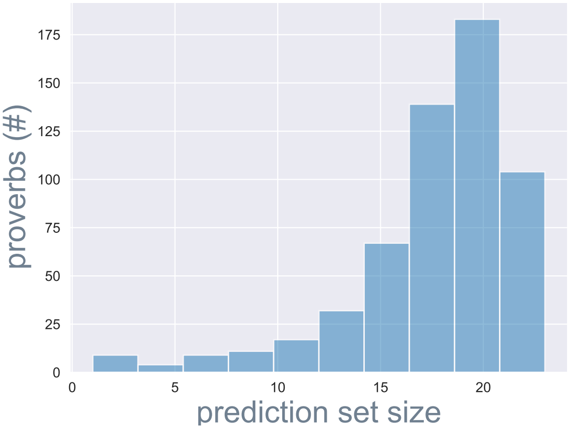


Figure 4: Histogram of the number of test proverbs (vertically) according to the number of locations predicted per proverb (prediction set size).

Locating unregistered proverbs We applied the conformal predictor to the 3,676 proverbs whose location is unregistered (§2.1).[21] In Figure 5, we present the locations which have been excluded at most from the unregistered proverbs. Aetolia, Skyros, Cyprus, Pontos, and Lesvos were the ones most frequently omitted. Excluding Karpathos, these are the locations LR achieved the highest F1 (Table 2). Each location, however, was excluded from the prediction set of at least ten proverbs. The proverb with the most narrow prediction set is Αυτός είναι Παύλος (/aftos einai pavlos/; 'This is Paul'), which could be registered with Achaia, Naxos, Aetolia, Heptanese, Thesprotia, or Thrace. We release our predictions publicly, to allow their further investigation and facilitate future research.

[18] We used the MAPIE library. We also experimented with conformal regression, but the produced prediction intervals were too wide to be functional.

[19] We followed the Least Ambiguous set-valued Classifier (lac) strategy. We also experimented with other strategies (Romano et al., 2020; Angelopoulos et al., 2020), but we did not observe any significant benefits.

[20] We consider LR and SVM as equally good options.

[21] We excluded twenty-two proverbs which were found to be noisy (e.g., written in Italian).

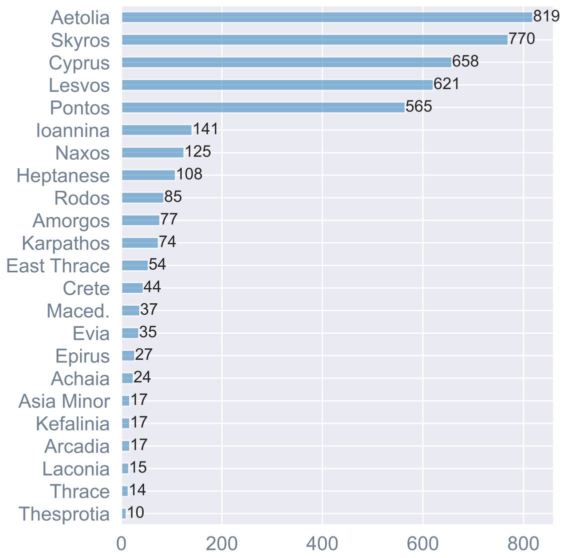


Figure 5: The number of unregistered proverbs (horizontally) for which a location was excluded from the respective prediction set.

5 Discussion

5.1 Error analysis

By focusing on classification (§3.1), the predictions of GrBERT and the best-performing n-gram model (LR) differ substantially (Jaccard similarity of 0.21). This low figure is derived by all predictions. If we restrict our attention to when the models make the right predictions, then if GrBERT is correct there is a 59.39% probability that LR will also be correct; conversely, if LR is correct, there is a 66.48% probability that GrBERT will also be correct. That recalls Tolstoy's remark in Anna Karenina: *when they err, they err in their own ways*, but in our case ways are regions.

By noting the loci of agreement and disagreement, we observe that the two agree more frequently for texts registered in locations with strong linguistic characteristics and for which they both perform best (Cyprus, Skyros, Pontos). When they disagree, this happens for texts of different locations, where either one or the other is the best performing model. GrBERT is correct more often for texts registered in Achaia, Kefallinia, Ioannina (where GrBERT achieved the best F1, as can be seen in Table 2) while LR is correct more often for texts from Etolia, Heptanese, Lesvos (where LR is the best). In 25% of these texts, all the other ML models agree with LR. When GrBERT is correct and LR is not, this percentage drops to 5%. Hence, we find that GrBERT may be disregarding features exploited by a character-level n-gram ML model.

5.2 Impact

Often deeply rooted in local tradition, proverbs reflect the linguistic nuances of a community. Also, they can serve as valuable data points for understanding how language varies across regions, dialects, or time periods, since they encapsulate both linguistic and cultural aspects in a concise form (Michel et al., 2011; Karsdorp and Fonteyn, 2019). Although this study focused on the former, we will release also normalised English translations to allow future exploration on cultural transmission (Bortolini et al., 2017) and variation (Ross et al., 2013). Next, we discuss the potential impact of our study across different fields, in order to suggest more detailed routes.

Historical Linguistics The proposed resource of geolocated proverbs can help us trace language change and the spread of varieties over time. By identifying the origins and migration patterns of specific phrases or idioms, historical linguists can gain insights into language change, dialect formation, and the influence of historical events (and other extra-linguistic factors) on language. The proposed dataset, then, can help us explore the paths through which language and culture travel, showing how specific expressions adapt and evolve as they move from one region to another.

Anthropology Understanding where certain proverbs originate and how they are used in different locales can offer insights into cultural values, norms, and practices. Proverbs often encapsulate cultural wisdom, ethical guidelines, and communal values. By mapping these expressions geographically, anthropologists can explore how cultural identities are formed and maintained, and how they differ across geographic regions. This can also enhance our understanding of cultural exchanges and interactions among communities.

Sociology Sociologists can use our dataset to examine social structures and group identities through the lens of colloquial language. Proverbs reflect societal attitudes and collective experiences, and their usage can reveal much about social norms, class distinctions, gender roles, and community relationships within and across regions. By analysing the distribution and variations of proverbs, sociologists can study the dynamics of social influence and the role of language in social cohesion and identity.

Linguistics (Theoretical) Linguists can leverage this dataset to study semantic, syntactic, and phonetic variation in language use across different regions. The geolocation of proverbs allows for an analysis of linguistic features that are geographically bounded. This can contribute to dialectological studies, sociolinguistics, and the development of language models that are sensitive to geographical variation in language use.

6 Related Work

Resources Davis et al. (2021) measured the frequency of English proverbs in Twitter, the Google Books n-gram Corpus, articles of the New York Times, and the Gutenberg Corpus. In a recent study, Ghosh and Srivastava (2022) used 250 English proverbs in context as a benchmark for abstract language understanding, showing that large language models struggled for such tasks. Özbal et al. (2016) presented a bilingual resource of 1,054 English proverbs and their equivalents in Italian. Noah and Ismail (2008) used 500 Malay proverbs to train a Naive Bayes model to learn to classify a proverb between family, life, destiny, social, and knowledge. The same task was addressed by Baptista and Reis (2022), who used 32,000 Portuguese proverbs in order to benchmark supervised machine learning algorithms. In this work, we presented the first publicly-available, machine-actionable dataset of proverbs in Greek. Our benchmark showed that geographical attribution from the proverb's text is feasible and the accuracy greatly depends on the location. Our experimental results also showed that the state-of-the-art in NLP (BERT) is struggling with the linguistic peculiarities of proverbs (outperformed in most locations), indicating the challenging nature of the specific attribution task.

Information extraction Information extraction from a textual input is a known problem. To provide an example, researchers have focused on toponym extraction (Radford, 2021; Kulkarni et al., 2020). Our work, however, is different. We also predict a location given a text, but that location is not explicitly mentioned in the text, and it can only be inferred by location-specific linguistic characteristics (Trudgill, 2003). In this sense, our task is rather related to text geocoding, concerning the prediction of the geographic coordinates of a text.[22]

Text geocoding In their survey of text geocoding, Melo and Martins (2017) classified the respective approaches to those using language models to represent different geospatial locations (Wing and Baldridge, 2011; Roller et al., 2012), and to those using supervised machine learning classifiers, on top of text representation features (Wing and Baldridge, 2014). All the surveyed approaches were applied to data from Wikipedia (long texts, using toponyms) and Twitter (short texts, using abbreviated and slang language).[23] The latter was considered as more difficult, which is probably why researchers used geo-indicative words to extract features, e.g., dialectal words (Han et al., 2014). We experimented with both of these approaches, language modelling and supervised learning, also investigating geocoding as a classification and regression task. Besides the benchmark, however, we suggested two applications, one to extract cardinal-direction-driving works and another to attribute un-located proverbs. Especially for the latter, we leveraged conformal prediction to let the attribution model facilitate (and not substitute) the experts.

7 Conclusion

This work focused on Greek proverbs. We have developed and publicly release the first large-scale machine-actionable dataset of Greek proverbs, quantifying their spatial distribution across different locations. We used this dataset to set a benchmark in geographical attribution, approaching the task with classification and regression. A BERT model did not clearly outperform its conventional counterparts in both tasks, sharing best performance with LR (classification) and Elastic Net (regression). To assess the benefits and explore the potential of our benchmark, we introduced two applications. First, we used Elastic Net to extract terms that push the prediction towards the four cardinal directions, showing that tsitakism pushes toward the South and East. Second, we leveraged conformal prediction, narrowing the possible locations for 3,676 unregistered proverbs, which we release publicly along with our data and code. Directions for future work comprise a thorough investigation of more distinctive local terms; the study of the propagation of more proverbs; and mining linguistic paths in more possible ways.

[22] Text geocoding is completely different from address geocoding, where an address is converted into geographic coordinates (Goldberg et al., 2007).

[23] Twitter is now renamed to X.

Acknowledgements

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Furthermore, we acknowledge the help of Maria Gavriilidou, whose advice was vital toward establishing the release of a properly licensed version of our dataset. We also acknowledge the help of Dr. Aikaterini Polymerou-Kamilaki, who shared important information regarding the proverb collectors.

Limitations

Coverage Our dataset includes the names of the collectors, but we note that collectors may have contributed unequally to the corpus. This means that more proverbs may appear in a specific location due to the systematic research of a specific collector working in that location.

Chronological attribution The presented corpus lacks metadata regarding the chronological attribution of each proverb. This is an inherent weakness in this domain due to the nature of the genesis of proverbs, travelling in space and time before they are established. One way to approximate the genesis of a proverb is to investigate its spatial propagation, connecting its route with historical events, as well as take into account other factors, linguistic (e.g., language contact, borrowing) and extra-linguistic alike.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. 2020. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*.

Jorge Baptista and Sónia Reis. 2022. Automatic classification of portuguese proverbs. In *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Eugenio Bortolini, Luca Pagani, Enrico R Crema, Stefania Sarno, Chiara Barbieri, Alessio Boattini, Marco Sazzini, Sara Graça Da Silva, Gessica Martini, Mait Metspalu, et al. 2017. Inferring patterns of folktale diffusion using genomic data. *Proceedings of the National Academy of Sciences*, 114(34):9140–9145.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Ethan Davis, Christopher M Danforth, Wolfgang Mieder, and Peter Sheridan Dodds. 2021. Computational paremiology: Charting the temporal, ecological dynamics of proverb use in books, news articles, and tweets. *arXiv preprint arXiv:2107.04929*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manuel Fernández-Delgado, Manisha Sanjay Sirsat, Eva Cernadas, Sadi Alawadi, Senén Barro, and Manuel Febrero-Bande. 2019. An extensive experimental survey of regression methods. *Neural Networks*, 111:11–34.

Sayan Ghosh and Shashank Srivastava. 2022. ePiC: Employing proverbs in context as a benchmark for abstract language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3989–4004, Dublin, Ireland. Association for Computational Linguistics.

Daniel W Goldberg, John P Wilson, and Craig A Knoblock. 2007. From text to geographic coordinates: the current state of geocoding. *URISA journal*, 19(1):33–46.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*, volume 1. MIT Press, Massachusetts. http://www.deeplearningbook.org.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500.

Hrisztalina Hrisztova-Gotthardt and Melita Aleksa Varga. 2014. *Introduction to paremiology: A comprehensive guide to proverb studies*. De Gruyter Open.

Folgert Karsdorp and Lauren Fonteyn. 2019. Cultural entrenchment of folktales is encoded in language. *Palgrave Communications*, 5(1).

Nikos Kazantzakis. 1996. *Zorba the Greek*. Simon and Schuster.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. Greek-bert: The Greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, pages 110–117.

Sayali Kulkarni, Shailee Jain, Mohammad Javad Hosseini, Jason Baldridge, Eugene Ie, and Li Zhang. 2020. Spatial language representation with multi-level geocoding. *arXiv preprint arXiv:2008.09236*.

Fernando Melo and Bruno Martins. 2017. Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21(1):3–38.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Google Books Team, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Wolfgang Mieder. 2004. *Proverbs: A handbook*. Greenwood Publishing Group.

Wolfgang Mieder. 2008. *"Proverbs speak louder than words": folk wisdom in art, culture, folklore, history, literature and mass media*. Peter Lang.

Sahrul Azman Noah and Febrin Ismail. 2008. Automatic classifications of malay proverbs using naïve bayesian algorithm. *Information Technology Journal*, 7:1016–1022.

Gözde Özbal, Carlo Strapparava, and Serra Sinem Tekiroğlu. 2016. Prometheus: A corpus of proverbs annotated with metaphors. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (Lrec'16)*, pages 3787–3793.

John Pavlopoulos and Maria Konstantinidou. 2022. Computational authorship analysis of the homeric poems. *International Journal of Digital Humanities*, pages 1–20.

Shreyas Pimpalgaonkar, Dhanashree Lele, Malhar Kulkarni, and Pushpak Bhattacharyya. 2021. Introduction to proverbnet: An online multilingual database of proverbs and comprehensive metadata. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 638–650.

Jelena Prokić and John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing*, 2(Special Issue on Language Variation):153–172. 2009/j.nerbonne/pub014 DOI: 10.13366/E1753854809000366 Special Issue on Language Variation ed. by John Nerbonne, Charlotte Gooskens, Sebastian Kurschner, and Renée van Bezooijen.

Benjamin J. Radford. 2021. Regressing location on text for probabilistic geocoding. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 53–57, Online. Association for Computational Linguistics.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1500–1510.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591.

Robert M Ross, Simon J Greenhill, and Quentin D Atkinson. 2013. Population structure and cultural geography of a folktale in europe. *Proceedings of the Royal Society B: Biological Sciences*, 280(1756):20123065.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

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

Damien Villers. 2022. What makes a good proverb? on the birth and propagation of proverbs. *Lexis. Journal in English Lexicology*, (19).

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. 1999. Machine-learning applications of algorithmic randomness.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 955–964.

Benjamin Wing and Jason Baldridge. 2014. Hierarchical discriminative classification for text-based geolocation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 336–348.

A The Hellenic Folklore Research Centre

The Hellenic Folklore Research Centre constitutes the National Documentation Centre for traditional and contemporary Greek culture. It has a specialised library and a rich archive of unpublished material on all aspects of traditional Greek life and culture. Its activities include field trips, research projects, publications, conferences, exhibitions and other events. It provides academic support for museums and for cultural bodies in the provinces of Greece. In the framework of the project "Hellenic National Documentation Network of the Intangible and Tangible Cultural Heritage" (2012-2015), a digitization of the special archives of Proverbs and

Popular Legends took place, and a digital repository was created.

B The collectors

The enormous online proverb dataset of the Academy of Athens is the result of long-term fieldwork and library/scholarly work. The collection of the material took place over a very long period (several decades) and the group of collectors for such a huge and long-term was inevitably diverse, e.g. school teachers and other erudite local people collaborating with the Academy of Athens, researchers / collaborators of the Academy, etc.

C Locations of dispersed proverbs

While Asia Minor was a region of frequent travel and exchange, Epirus remained geographically isolated. Nevertheless, certain segments of its population, such as shepherds engaged in transhumance, craftsmen, and particularly merchants, did travel. We observe the same for Thrace. Particularly during the Byzantine and Ottoman periods, Thrace was inter- and intra-connected enough thanks to Constantinople (modern-day Istanbul), a major urban, political, and economic hub. Constantinople attracted people for trade, employment, pilgrimage, and other opportunities, thus making Thrace a region of significant inter- and intra-regional movement. Although the degree of connectivity across all of Thrace varied depending on local geography and infrastructure (with coastal areas and regions near Constantinople being more accessible), movement to and from Thrace was notably higher compared to more isolated regions, such as Epirus.

D Near duplicates

We present selected near duplicates of the most common proverb (i.e., “i glossa kokkala den ekhei ki kokkala tsakizei”; see Table 1a), to illustrate the scale of the pre-processing problem. In Epirus (Thrace and elsewhere) it is found as “i glossa kokkala den ekh ki kokkala tsakiz”. In Rhodes, it is “i glossa kokkala den ekhei kai kokkala tsakka”. In Crete, it is “i glossa kokkala den ekhei kai kokkala tsakei”. In Cyprus, it is “i glossa kokkala en esei dzai kokkala tsakizei”. More such near duplicates for more proverbs exist in our repository.

E Linguistic alternations clustering

Figure 6 presents the agglomerative clustering of the linguistic alternations for the most widespread Greek proverb (Table 1 a). We used the Euclidean distance and n-gram-based TFIDF embeddings. We note that the same proverb may exist altered (in multiple forms) in the same location. Hence, the same location may appear multiple times.

Location	**1st**	**2nd**	**3rd**
Epirus	τουν	μι	όλ
Aetolia	τουν	είνι	μι
Amorgos	μηδέ	όγοιος	είντα
East Thrace	νε	αμάξι	νάχουν
Arcadia	κάμπους	κάνω	στις
Achaia	γρόσι	τογ	τομ
Heptanese	ναν	εκειός	ειν
Evia	τσαί	τσαι	βγέλλει
Thesprotia	πάρεξ	ζυγό	ντράβαλα
Thrace	πε	τς	διάβολο
Ioannina	μι	σι	τουν
Karpathos	τσαι	εγιώ	τσαί
Kefalinia	τσου	όθεν	τσι
Crete	ντου	καλλιά	τση
Cyprus	τζ	τζαι	τζαί
Lesvos	τσι	τουν	μι
Laconia	ςτο	τες	γλέντι
Macedonia	σι	μι	τουν
Asia Minor	τουν	κη	σι
Naxos	ια	τζη	έρο
Pontos	σο	ατ	σην
Rhodes	κάμνει	λωλλός	γαπά
Skyros	τσαί	τσαι	έναι

Table 5: Tokens with the highest TFILF per location.

F Distinctive local terms

The distinctive local terms found are presented in Table 5. We show the original form of the tokens, instead of their transliteration or translation, which comprises accents.

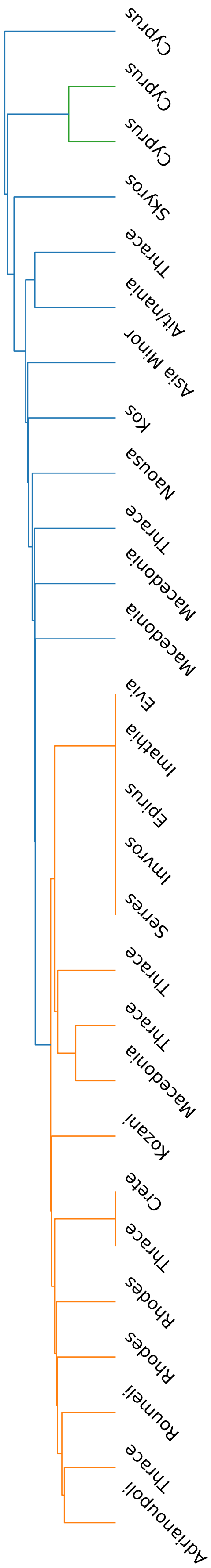


Figure 6: Agglomerative clustering of the near-duplicates of the most widespread Greek proverb (Table 1 a). The different clusters arranged in the tree-like diagram are best viewed in colour.