**Creating a UD treebank for a low-resourced Modern Greek dialect: the case of Lesbian**

Stavros Bompolas, *ARCHIMEDES Unit – Athena Research Center, Greece*

Keywords: *Modern Greek dialects, Lesbian, universal dependencies, treebank*

## 1. Description

**Aim and scope:** This project proposes the creation of a specialized *Universal Dependencies* (UD) treebank for Lesbian, a dialect of Modern Greek that is still used on the island of Lesbos. Lesbian belongs to the Northern Greek dialect group [1], in contrast to the Southern Greek dialects upon which *Standard Modern Greek* (SMG) is based [2]. Notably, Northern Greek dialects demonstrate the raising of unstressed mid vowels /e/ and /o/ to [i] and [u] respectively, alongside the loss of unstressed high vowels /i/ and /u/ ([3]; Map 1). What is more, Lesbian has been influenced by Italo-Romance and Turkish. As a result, the dialect is characterized by a set of phonological, morphological, lexical and syntactic features (for an overview, see [4], [5], [6], [7]) that pose significant challenges regarding the tokenization, lemmatisation and morphological annotation of its written form. Despite a rich history in the study of this specific dialect and growing efforts to incorporate various Greek dialects into UD treebanks (e.g., Cappadocian/Asia Minor Greek, Messenian, Cretan), there currently exists no UD treebank for Lesbian Greek, nor is any other Northern Greek dialect represented within UD. My aim is to fill this gap and discuss the problems that are particular to Northern Greek dialects. This project is inspired by research carried out at the "Archimedes" Center for Research in Artificial Intelligence (AthenaRC). Its main objective is to blend existing linguistic knowledge with the creation of new algorithms capable of drawing conclusions that involve multiple inference steps. These algorithms will leverage premises and conclusions articulated in natural language and multilingual models, with a specific emphasis on their effectiveness in low-resourced settings, such as those encountered in the exploration of Greek dialects.

**Map 1.** The geographic position of Lesbian and the isogloss delineating Northern and Southern Greek dialects.

**The status of the dialect:** Due to its non-standardized nature, Lesbian Greek primarily functions as a spoken language for the inhabitants of Lesbos (Map 1). Unlike SMG, it lacks a substantial written legacy and is not formally taught in educational institutions. Consequently, the dialect faces growing pressure from the dominant SMG [7].

**Oral/textual sources and script**: The documentation of Lesbian vocabulary is primarily found in specialized glossaries and dictionaries (e.g., [8], [9], [10], [11]). Additionally, contemporary Lesbian literature serves as valuable source of linguistic data, including humorous tales [12] and plays [13], as well as local periodicals and newspapers [14]. These texts employ a modified Greek alphabet to accurately represent the unique phonetic characteristics of the dialect. Given that these texts are written by the community itself, it is crucial to take them seriously into account. Additionally, a significant oral resource comprises a roughly 15-hour corpus of recorded speech from native Lesbian speakers across various island locations. This corpus was compiled by Em. Prof. A. Ralli (2023-2024). Selected portions of this material will be incorporated into the treebank corpus, facilitating the comprehensive representation of both written and spoken language.

**Methods and challenges:** Developing a treebank for Lesbian presents unique challenges due to its oral nature, lack of standardization, and limited textual and lexical resources. Cross-lingual technology transfer from related languages emerges as a promising approach for low-resourced ones (e.g., [15], [16], [17]). SMG serves as a natural source language for this transfer, utilizing the UD_Greek-GUD treebank as a foundational resource. However, the effectiveness of this transfer depends on the similarity between the source and target languages (e.g., [15], [16], [18], [19], [20]). In our case, several notable challenges in technology transfer from SMG to Lesbian can be foreseen: (a) **Tokenization issues** due to (i) script conventions that have been so far used to represent the loss of certain vowels in Northern Greek phonology, such as the use of apostrophes both within a word and at its boundaries, in contrast to SMG where apostrophes mark the loss of vowels only at word boundaries. E.g., *Τ'ν πρόσβαλι τσι πυρουκουτσίν'σι απ' τ' ντρουπή τ'ς* (Lesbian) *vs. Την πρόσβαλα και πυροκοκκίνισε απ' την ντροπή της* (SMG) 'She felt attacked and blushed from embarrassment'; (ii) phonological peculiarities, such as (semi-)vowel epenthesis between functional words and nouns. E.g., <*τν-ι-μκρή*> [tn-i-mkrí] (Lesbian) *vs.* <τη μικρή> [ti mikrí] (SMG) 'the little (girl)'. (b) **Lexical and morphological differences** due to (i) phonological peculiarities of Northern dialects: <κιρί> [cirí] (Lesbian) *vs.* <κερί> [cerí] (SMG) 'wax', <κυρί> [cirí] (Lesbian) *vs.* <τυρί> [tirí] (SMG) 'cheese'; (ii) the use of vocabulary that differs from SMG, including words from older stages of Greek:

*x(u)ʎár(i)* (Lesbian < Hellenistic Greek *κοχλιάριον*) *vs. kutáli* (SMG) 'spoon'; loanword integration: *parasóʎ(i)* (Lesbian < Venetian *parasol*) *vs. ombréla* (SMG) 'umbrella', *burdízu* (Lesbian < Turkish *burmak*) *vs. liɟízo* (SMG) 'bend'; use of very particular derivational suffixes for diminutives *ðindr-*$_{\text{STEM}}$*éʎ(i)*$_{\text{DIM}}$ (Lesbian) *vs. ðendr-*$_{\text{STEM}}$*áci*$_{\text{DIM}}$ (SMG) 'little tree'. (c) **(Morpho)syntactic parsing issues** due to deviations in tense formations from SMG, resembling forms used in Medieval Greek of the type ['have' + passive participle] in active voice and ['be' + passive participle] in passive voice. E.g., Active voice: *éxo xaménu* (Lesbian) *vs. éxo xási* (SMG) 'I have lost'; Passive voice: *ími xaménus* (Lesbian) *vs. éxo xaθí* (SMG) 'I have been lost'.

**Future prospects:** The treebank for the Lesbian dialect holds significant potential. Many of the linguistic peculiarities observed in Lesbian are shared by other, yet undocumented, Northern Greek dialects. This suggests that the resulting resource can serve as a valuable asset for linguists, NLP researchers, and developers working on language technologies for this broader dialectal group. Ultimately, this project contributes to the preservation and exploration of linguistic diversity in Greek dialects, which present phenomena encountered in the dialects of other languages as well (e.g., for vowel reduction and deletion in other languages and dialects, see [21], [22], [23], [24], [25]), enabling cross-lingual transfer of already-established methods based on Greek data.

## 2. Benefits of participation

The UD annotation of Lesbian data is part of a broader project that aims to develop treebanks for various dialects of Greek, each with its own unique features that pose new annotation challenges. Collaborating with experts and fellow researchers from diverse linguistic backgrounds will offer fresh insights and refine methodologies for creating dialectal treebanks, which can be considered as forming "linguistic families" related to a "standard" language variety. This collaboration is crucial because the standard variety typically provides a sufficient amount of data that facilitates language technology transfer to often under-resourced dialects. The challenge lies in ensuring that this transfer introduces minimal or no bias to the dialectal treebanks.

## 3. Open questions

(a) **Balancing specificity and generalizability:** How can we effectively annotate a low-resource dialect (like Lesbian), capturing its unique features (specificity), while maintaining compatibility with broader linguistic frameworks (generalizability)? What strategies can optimize annotation workflows and tools for streamlined treebank creation in such scenarios?
(b) **Incorporating dialectal nuances:** How can we best integrate dialectal variations and unique linguistic features into the overall annotation scheme?
(c) **Dialect-specific challenges:** How can we develop effective strategies to address the challenges encountered when annotating Lesbian?
(d) **Adapting existing guidelines:** How can we best adapt existing annotation guidelines (e.g., those found on GitHub for Greek) to capture the nuanced linguistic features of a closely related dialect like Lesbian?
(e) **Cross-dialectal compatibility:** How can we ensure the Lesbian UD treebank is compatible and interoperable with existing treebanks of other (Greek) dialects? What potential challenges might arise, and how can we overcome them?

## 4. Project phases

This project leverages the expertise of the GUD and Pomak treebank teams to create a UD treebank for Lesbian Greek, following this plan:

**Phase 1: Building foundational knowledge (January 2024 - February 2024)**
- Gain a comprehensive understanding of the UDs annotation scheme.
- Master the CoNLL-U format and annotation tools used in UDs.

**Phase 2: Collaboration with GUD Team (March 2024 - April 2024)**
- Actively participate in finalizing the GUD treebank and its guidelines.
- Conduct an in-depth study of the GUD treebank and its annotation guidelines.
- Gain insights and best practices from GUD team members.

**Phase 3: Lesbian corpus creation (May 2024 - August 2024):**
- Utilize OCR and speech-to-text techniques to compile the Lesbian corpus.
- Ensure data quality through cleaning and pre-processing steps.

**Phase 4: Methodology development and treebank construction (July 2024 - December 2024)**
- Explore and address methodological challenges associated with creating a UD treebank.
- Develop and construct the Lesbian UD treebank.

**References**


[1] P. Trudgill, "Modern Greek dialects: A preliminary classification," *J Greek Linguist*, vol. 4, no. 1, pp. 45–63, 2003, doi: 10.1075/jgl.4.04tru.

[2] N. Pantelidis, "Πελοποννησιακός ιδιωματικός λόγος και κοινή νεοελληνική [Peloponnesian dialectal accent and Standard Modern Greek]," in *Proceedings of the Fourth International Conference of Greek Linguistics*, P. Pavlou and A. Roussou, Eds., Thessaloniki: University Studio Press, 2001, pp. 480–486.

[3] G. Hatzidakis, "Περὶ φθογγολογικῶν νόμων καὶ τῆς σημασίας αὐτῶν είς τὴν σπουδὴν τῆς νέας ἑλληνικῆς [Regarding phonological laws and their significance in the study of Modern Greek]," in *Μεσαιωνικὰ καὶ Νέα Ἑλληνικὰ Α' [Medieval and Modern Greek 1]*, P. D. Sakellarios., Athens, 1905, pp. 154–201.

[4] S. Anagnostou, "Λεσβιακά: Ήτοι συλλογή λαογραφικών περί Λέσβου πραγματειών [Lesbian: A Collection of Ethnographic Accounts about Lesbos]," 1903.

[5] G. Sakkaris, "Περί της διαλέκτου των Κυδωνιέων εν συγκρίσει προς τα λεσβιακά [Regarding the Dialect of Kydonies in Comparison to Lesbian]," *Mikrasiatika Chronika*, vol. 3, 1940.

[6] A. Ralli, V. Alexelli, and C. Tsimpouris, "The electronic linguistic atlas of the Aegean island of Lesbos (EDAL)," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:180260743

[7] V. Alexelli, "Χαρτογράφηση της γλωσσικής ποικιλίας της Λέσβου [Mapping the linguistic variety of Lesbos]," PhD thesis, University of Patras, Patras, 2021. doi: 10.12681/eadd/50075.

[8] D. Papanis and G. Papanis, *Λεξικό του Αγιασώτικου Ιδιωματικού Λόγου. Ερμηνευτικό-Ετυμολογικό [Dictionary of the Agiasotic Idiomatic Speech. Interpretive-Etymological]*, 3rd ed. Mytilene, 2004.

[9] A. Ralli, *Λεξικό Διαλεκτικής Ποικιλίας Κυδωνιών, Μοσχονησίων και Βορειοανατολικής Λέσβου [Dictionary of the Dialectal Varieties of Kydonies, Moschonisia, and Northeastern Lesvos]*. Athens: Hellēniko Hidryma Historikōn Meletōn, 2017.

[10] M. Anagnostopoulou, *[Λεσβιακοί ιδιωματισμοί: χαρακτηρισμοί – ξομπλιάσματα ανδρών και γυναικών] Lesbian Idioms: Characterizations – Terms used by Men and Women*. Mytilene: Aeolida, 2013.

[11] P. S. Paraskevaidis, *Τουρκικές λέξεις στο μανταμαδιώτικο γλωσσικό ιδίωμα [Turkish Words in the Mantamadiotiko Linguistic Idiom]*. Mytilene: Women's Association of Mantamados, 2020.

[12] S. Karna, *Μικρές ιστορίες της Λέσβου [Short stories of Lesbos]*. Mytilene: iWrite, 2020.

[13] C. Kanimas, *Τι να τα κάνω τα καλά. Θέατρο [What Should I Do with the Good Deeds. Theater]*. Athens, 2013.

[14] S. Dedekis, *Γιατζίδα. Σατιρική εφημεριδούλα (σε Λεσβιακή και Αϊβαλιώτικη διάλεκτο) [Giatzida. Satirical Newspaper (in Lesbian and Aivaliot dialect)]*. Patras: Laboratory of Modern Greek dialects / University of Patras, 2015.

[15] S. Markantonatou, N. Th. Constantinides, V. Stamou, V. Arampatzakis, P. G. Krimpas, and G. Pavlidis, "Methodological issues regarding the semi-automatic UD treebank creation of under-resourced languages: the case of Pomak," in *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, L. Grobol and F. Tyers, Eds., Washington, D.C.: Association for Computational Linguistics, Nov. 2023, pp. 27–35. [Online]. Available: https://aclanthology.org/2023.udw-1.4

[16] C. Tsoukala *et al.*, "ASR pipeline for low-resourced languages: A case study on Pomak," in *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 40–45. doi: 10.18653/v1/2023.fieldmatters-1.5.

[17] S. Ahmadi, Z. Azin, S. Belelli, and A. Anastasopoulos, "Approaches to Corpus Creation for Low-Resource Language Technology: the Case of Southern Kurdish and Laki," 2023, doi: 10.48550/ARXIV.2304.01319.

[18] K. Abe, Y. Matsubayashi, N. Okazaki, and K. Inui, "Multi-dialect Neural Machine Translation and Dialectometry," in *Pacific Asia Conference on Language, Information and Computation*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:198918735

[19] E. S. Tellez, D. Moctezuma, S. Miranda, M. Graff, and G. Ruiz, "Regionalized models for Spanish language variations based on Twitter," *Lang Resources & Evaluation*, vol. 57, no. 4, pp. 1697–1727, Dec. 2023, doi: 10.1007/s10579-023-09640-9.

[20] F. Faisal and A. Anastasopoulos, "Phylogeny-Inspired Adaptation of Multilingual Models to New Languages," 2022, doi: 10.48550/ARXIV.2205.09634.

[21] K. M. Crosswhite, "Vowel reduction in Russian: a unified account of standard, dialectal, and 'dissimilative'patterns," *University of Rochester working papers in the language sciences*, vol. 1, pp. 107–172, 2000.

[22] K. M. Crosswhite, "Vowel reduction," in *Phonetically Based Phonology*, B. Hayes, R. Kirchner, and D. Steriade, Eds., Cambridge University Press, 2004, pp. 191–231.

[23] E. O'Rourke, "Dialect differences and the bilingual vowel space in Peruvian Spanish," in *Selected proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*, Cascadilla Proceedings Project Somerville, MA, 2010, pp. 20–30.

[24] M. Llompart and M. Simonet, "Unstressed Vowel Reduction Across Majorcan Catalan Dialects: Production and Spoken Word Recognition," *Lang Speech*, vol. 61, no. 3, pp. 430–465, Sep. 2018, doi: 10.1177/0023830917736019.

[25] J. J. McCarthy, "How to delete," *Perspectives on Arabic linguistics XXX*, pp. 7–32, 2019.