Crossing Dialectal Boundaries: Building a Treebank for the Dialect of Lesbos through Knowledge Transfer from Standard Modern Greek

Stavros Bompolas[1], **Stella Markantonatou**[1,2], **Angela Ralli**[1,3], **Antonios Anastasopoulos**[1,4]
[1]Archimedes, Athena Research Center, Greece
[2]ILSP, Athena Research Center, Greece [3]University of Patras, Greece
[4]George Mason University, USA
Correspondence: s.bompolas@athenarc.gr

Abstract

This paper presents the first treebank for the dialect of Lesbos, a low-resource living northern variety of Modern Greek (MG), annotated according to the Universal Dependencies (UD) framework. So far, the only dialectal treebank available for Greek developed with cross-dialectal knowledge transfer is an East Cretan one, which belongs to the same southern branch as Standard Modern Greek (SMG). Our study investigates the effectiveness of cross-dialectal knowledge transfer between dialectologically less similar varieties of the same language by leveraging knowledge from SMG to annotate the northern dialect of Lesbos. We describe the annotation process, present the resulting treebank, inject additional linguistic knowledge to enhance the results, and evaluate the effectiveness of cross-dialectal knowledge transfer for active annotation. Our findings contribute to a better understanding of how dialectal variation within language families affects knowledge transfer in the UD framework, with implications for other low-resource varieties.[1]

1 Introduction

The Universal Dependencies (UD) project (de Marneffe et al., 2021) has established consistent syntactic representations across 168 languages, but non-standard varieties are underrepresented. This gap arises from challenges in text collection, the scarcity of qualified annotators, and the expertise needed to adapt existing guidelines (Blaschke et al., 2024). Documenting less-used dialects not only preserves linguistic diversity but also provides valuable insights for contrastive linguistic analysis while serving as a testbed for computational approaches in data-scarce scenarios.

[1]The treebank is available at https://github.com/UniversalDependencies/UD_Greek-Lesbian, released as part of UD v2.16 (May 15, 2025).

Two UD treebanks exist for Standard Modern Greek (SMG), and recent efforts aim to develop dialectal ones, including the Lesbian treebank. However, Modern Greek (MG) dialects remain largely unexplored. Moreover, research on cross-dialectal knowledge transfer in NLP for MG is limited. To date, the only effort in this direction has focused on the East Cretan dialect—an intuitively favorable case, given its relative proximity to SMG as a fellow southern MG variety.

Figure 1: The geographic position of Lesbian (and Cretan) and the isogloss delineating northern and southern MG dialects.

This paper addresses this gap with three key contributions. (a) We present the first UD treebank for a northern MG dialect, focusing on Lesbian. (b) We introduce dialect-specific annotation guidelines for Lesbian and a method for integrating orthographic conventions into the annotation scheme, which is adaptable to other (northern MG) dialects without standardized orthography. (c) We extend previous research on knowledge transfer between SMG and southern MG di-

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 39–51

alects (i.e., Cretan) to investigate cross-dialectal transfer between linguistically less similar varieties, specifically from SMG, a southern variety, to a northern one (i.e., Lesbian). In particular, we explore how the injection of linguistic knowledge can improve cross-dialectal transfer.

2 Related Work

Recent work has increasingly focused on developing dialectal treebanks and exploring knowledge transfer for low-resource languages. Treebanks have been created for varieties such as Egyptian Arabic (Maamouri et al., 2014), Norwegian (Øvrelid et al., 2018; Kåsen et al., 2022), Occitan (Miletic et al., 2020), and Bavarian (Blaschke et al., 2024), among others. For surveys of cross-lingual transfer methods in dependency parsing, see Mageuresse et al. (2020), Das and Sarkar (2020), Hedderich et al. (2021), and Pakray et al. (2025).

2.1 SMG Treebanks

The UD framework has been applied to SMG through two treebanks. The first, UD_Greek-GDT (Prokopidis and Papageorgiou, 2017), is based on the Greek Dependency Treebank (http://gdt.ilsp.gr). The more recent and comprehensive treebank, UD_Greek-GUD (GUD), follows UD.v2 morphological guidelines, ensuring improved consistency and coverage (Markantonatou et al., 2025).

These treebanks provide a solid foundation for dialect-oriented NLP by allowing cross-dialectal knowledge transfer from SMG to under-resourced dialects. They also serve as important benchmarks, as new resources for MG should align with their guidelines and validation standards for compatibility and interoperability within the UD ecosystem.

2.2 MG Dialectal Treebanks

Three UD treebanks have been developed for MG dialects recently: two for Cappadocian—UD_Cappadocian-AMGiC (Sampanis and Prokopidis, 2021) and UD_Cappadocian-TueCL (Vligouridou et al., 2024) and one for Eastern Cretan, UD_Greek-Cretan (Vakirtzian et al., 2025). Among these, only the work focusing on Cretan has investigated cross-dialectal knowledge transfer from SMG, leveraging the linguistic proximity between these two southern MG varieties.

In contrast, our work addresses cross-dialectal transfer across less similar dialect groups, specifically from SMG, a southern MG variety, to Lesbian, a northern dialect. To our knowledge, this is the first study within the UD framework to examine knowledge transfer between more distantly related MG dialects.

3 The Lesbos Dialect

As shown in Figure 1, the Lesbos dialect belongs to the northern MG dialect group, which is characterized by the so-called "northern vocalism"—specifically, the raising of unstressed mid vowels /e/ and /o/ into [i] and [u], respectively (e.g., *πιδί* [piˈði] instead of SMG *παιδί* [peˈði] 'child', *κάτου* [ˈkatu] instead of SMG *κάτω* [ˈkato] 'down'), and the deletion of unstressed high vowels /i/, /u/ (e.g., *φίδ* [ˈfið] instead of SMG *φίδι* [ˈfiði] 'snake', *βνό* [ˈvno] instead of SMG *βουνό* [vuˈno] 'mountain'). These features distinguish the dialect of Lesbos from southern dialects, including SMG (Chatzidakis, 1905).

The dialect has been shaped by extensive historical contact with Italo-Romance and Turkish (Ralli, 2015, 2019a,c; Alexelli, 2021). During Italo-Romance rule (1355-1462), numerous loanwords and morphological elements of Venetian origin were introduced, such as the diminutive suffix *-ελ(ι)* [-el(i)] (Melissaropoulou and Ralli, 2010). The subsequent Ottoman period (1462-1912) further enriched the dialect with Turkish borrowings. Around the 16th century, speakers from Lesbos settled in the nearby Asia Minor areas of Kydonies and Moschonisia, where the dialects share many similarities with Lesbian. After the Asia Minor Catastrophe (1922) and the Treaty of Lausanne (1923), refugees from these regions permanently resettled in dialectal enclaves in Lesbos, resulting in a complex linguistic system on the island, characterized by intra-dialectal variation and features absent from SMG.

Today, unlike most MG dialects, Lesbian remains vital, serving as the primary means of communication on the entire island.

4 The Treebank

This section outlines the procedure for creating the Lesbian Treebank. As a UD_Greek treebank, it broadly follows the existing annotation guidelines for SMG and adopts the same set of UFeats, UPOS tags, and dependency relations. Accordingly, we focus on dialect-specific deviations and annotation decisions driven by dialectal features not addressed in the SMG guidelines. This overview

is not intended to be exhaustive (for a recent overview on the dialectal variation in Lesbos, see Alexelli, 2021 and the linguistic atlas of Lesbos, Ralli, 2019b).

4.1 Source Materials

The corpus draws from six main sources representing different text types and dialectal variants from the island of Lesbos: (a) example sentences extracted from three comprehensive dialectal dictionaries (Papanis and Papanis, 2004; Ralli, 2017; Anagnostopoulou, 2021), and (b) sentences of oral nature taken from three additional texts of contemporary Lesbian literature (Tsokarou-Mitsioni, 1998; Anagnostou, 2014; Tsokarou-Mitsioni, 2019), including humorous tales, plays, narratives, and personal accounts written in the dialect.

These sources capture internal variation within the Lesbian dialect, including sub-dialectal and stylistic differences across narrative and conversational contexts. They span a broad stylistic range—from standardized dictionary entries to informal, orally styled literary texts—exhibiting significant orthographic and grammatical variation.

One of the major challenges in working with these materials is the lack of a standardized orthography for Lesbian, which, like many dialects, is primarily an oral variety (for an overview of the issues that arise when representing spoken varieties in UD treebanks, see Dobrovoljc, 2022). In the aforementioned sources, the general trend in orthographic representation leans toward conformity with SMG, but several issues arise:

1. Despite the adherence of the authors to SMG spelling, the texts contain a considerable number of orthographic errors.
2. Significant inconsistencies also emerge in the representation of the northern vocalism, particularly regarding the raising of unstressed /e/ to [i] and the deletion of unstressed high vowels /u, i/. Within SMG orthography, vowel /i/ corresponds to multiple graphemic representations (<η, ι, υ, ει, οι>), leading authors to reflect vowel raising with notable inconsistency, influenced by both stylistic considerations and etymological factors. Similarly, the deletion of unstressed high vowels is occasionally marked with an apostrophe (’), yet this orthographic strategy, when employed, exhibits irregular application, resulting in orthographic variation for identical lexical items. For example, the Lesbian counter-form for the SMG word *απόμεινε* [aˈpomine] ‘remained’ is attested as *απόμνι*, *απόμνει*, *απόμ’νι*, and *απόμ’νει* [aˈpomni].
3. These orthographic challenges are further complicated by distinctive phonological features of the Lesbian dialect that lack standardized representation in SMG orthography, such as euphonic sound insertion and consonant voicing phenomena (e.g., SMG *κοντά του* [konˈda tu] ‘close to him/it’ > Lesbian *κουντά τ* [kunˈda t] *vs*. *κουντά ντ* [kunˈda d]), thus introducing additional complexity to orthographic standardization efforts.

4.2 Annotation

The annotation of the Lesbian dialect has primarily followed the UD annotation guidelines established for GUD, complemented by grammatical descriptions (Anagnostou, 1903) and dialect dictionaries (Papanis and Papanis, 2004; Ralli, 2017; Anagnostopoulou, 2021). We use the same set of dependency (sub)relations as defined for SMG. Only deviations, new constructs, and forms have been documented in supplementary guidelines specific to the Lesbian treebank, which are listed as comments on the GUD guidelines. Our main annotations remain compatible with existing (S)MG treebanks in UD, facilitating comparative research. It should be noted, however, that our lemmatization respects Lesbian phonology and morphology rather than conflating with SMG lemmas. As a result, our treebank passes SMG validation rules with minimal exceptions for certain auxiliaries that differ only due to the phonological application of northern vocalism (i.e., the lemmas *έχου* [ˈexu] ‘have’ instead of SMG *έχω* [ˈexo]; *είμι* [ˈimi] ‘be’ instead of SMG *είμαι* [ˈime]).

4.3 Tokenization

Following earlier (S)MG treebanks, we segment adposition-determiner contractions; for example, *στο* [sto] ‘in/to the’ is tokenized as two syntactic words, *σ* [s] ‘in/to’ and *το* [to] ‘the’. Unlike SMG treebanks, we needed to split not only contracted forms but also clitics that are frequently attached to verbs in written dialectal texts (e.g., *τάμπλιξις* [ˈtabliksis] ‘(you) mixed them up’ > *τά* + *μπλιξις*). The same approach applies to possessive pronouns frequently attached to nouns (e.g., *πατέρασιτς* [paˈterasits] ‘her father’ > *πατέρας* + *ιτς*).

While this decision aligns with guidelines from earlier MG treebanks regarding tokenization han-

dling, it differs from GUD's approach of pretokenizing contracted forms. In our treebank, we maintain merged word sequences as written in the original sources and treat such cases as multiword tokens. Additionally, we have respected the original written sources by not merging tokens that were erroneously split, using instead the "goeswith" relation for these instances.[2]

4.4 Lemmatization

For lemmatization, we relied primarily on dialectal dictionaries. As already stated, the words of Lesbian that diverge from their SMG counterparts were assigned a lemma form that bears the dialectal characteristics. Nevertheless, this aspect of annotation required significant effort due to several challenges:

1. As previously mentioned, phonological phenomena are not spelled uniformly across or even within sources due to orthographic inconsistencies. Consequently, the same words are often spelled differently in the legacy texts.
2. Although northern vocalism is a defining feature of northern dialects, it is not applied uniformly even within the same source, creating inconsistencies for the same lexical items (e.g., SMG *φοβάμαι* [foˈvame] 'I am afraid' > Lesbian *φοβάμι* [foˈvami] *vs.* *φουβάμι* [fuˈvami]).
3. Sources frequently contain orthographic errors unrelated to dialectal features.

To address these issues:

For 1.: We eliminated apostrophes from lemmas, since these are not used consistently in the texts. This decision allowed the lemma to serve as a unifying element across all texts, regardless of whether they systematically used apostrophes, didn't use them at all, or used them inconsistently. This standardization approach represents a significant contribution to MG dialectal text processing, as apostrophe usage has been a persistent challenge across Greek dialect documentation efforts. Our systematic treatment of this orthographic feature, combined with the annotation approach developed for this treebank, offers a replicable methodology that can benefit future computational work on MG dialects with similar orthographic variation.

For 2.: We consistently used the dialectologically expected form (with vowel raising and deletion) as the lemma, even when texts did not systematically apply these features; i.e., for both *φοβάμι* and *φουβάμι* we assigned the lemma *φουβάμι*.

For 3.: Regardless of orthographic errors in the source materials, we applied standardized (MG) orthography to lemmas.

4.5 Morphology

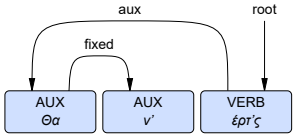
The Lesbian dialect's morphology broadly aligns with SMG, though surface forms differ due to northern vocalism. Notable morphological differences include:

1. Use of the definite article *η* [i] 'the' for masculine nouns in the nominative case of the singular number, phonetically matching the feminine definite article. The SMG definite article /o/ (raised to [u]) occasionally appears in free variation with the latter.
2. 3rd-person plural present and future active verb forms take the inflectional suffix *-in* instead of SMG *-un* (e.g., Lesbian *χάν-ιν* [ˈxan-in] *vs.* SMG *χάν-ουν* [ˈxan-un] 'they lose').
3. Distinctive diminutive suffixes, particularly the highly productive *-ελ(ι)* [-el(i)], attach to bases of all genders, loanwords, and proper names (instead of SMG *-άκι* [-aci]). For these forms, we lemmatize to the base word without the diminutive suffix.

4.6 (Morpho-)Syntax

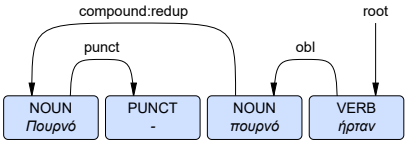
Based on available written sources and oral material, few syntactic divergences have been identified between the Lesbian dialect and SMG (Ralli, 2019a), documented in our sources and annotated in the treebank:

1. Alternation between genitive and accusative case in examples such as *τ*.GEN *έδουσι ένα δικάρ* [t ˈeðusi ˈena ðiˈkar] *vs.* *τουν*.ACC *έδουσι ένα δικάρ* [tun ˈeðusi ˈena ðiˈkar] 'he/she gave him a dim'.
2. In addition to the SMG future particle *θα* [θa] 'will', the form *θα ν(α)* [θa n(a)] 'will' is frequently used in the dialect, which is not attested in SMG. We annotated these structures as follows:
   - *Θα ν' έρτ' ς* [θa n ˈerts] 'you will come'



[2] https://universaldependencies.org/u/overview/typos.html#wrongly-split-word

3. Although reduplication is also attested in SMG, it occurs with notably higher frequency in Lesbian, likely due to the dialect's prolonged contact with Turkish (0 occurrences in 1807 sentences of GUD and 12 occurrences among 270 sentences in the Lesbian treebank). These reduplicated forms typically involve two identical lexical elements with occasional morphophonological alternations. Following the UD guidelines,[3] reduplication is annotated as follows in the Lesbian treebank:
   - *Πουρνό-πουρνό ήρταν* [pur'no pur'no 'irtan] 'They came very early in the morning'



   Additionally, we annotate the head of the compound construction with the UFeat Degree=Aug, as these constructions primarily function to express intensity or augmentation of the core meaning.
4. Given the oral nature of some sources, we frequently employ the parataxis relation, which is much less common in GUD (26 occurrences in GUD and 82 occurrences in the Lesbian treebank). This relation connects pairs of potentially standalone sentences treated as a single sentence. In spoken corpora, this occurs naturally as sentence boundaries often align with utterance turns. When more than two sentences join in this manner, we make all subsequent sentences dependents of the first one, reflecting the structural parallel between parataxis and conjunction relations.

4.7 Voicing and Euphonics

Following Vakirtzian et al. (2025, 780-781), we integrated voicing and euphonic annotations in the MISC (10th) column, explicitly documenting these phenomena to support comparative analysis, e.g., dialectometry.

Euphonics are vowels or consonants that appear within, between, or at the end of words. In Lesbian (and MG more broadly), they create open syllables and avoid hiatuses. We tag these elements as "euphonic" using the MSeg|MGloss format, allowing us to treat them as separate tokens and distinguish them from the rest of the word.

Unlike SMG, where voicing phenomena are not represented orthographically, dialectal texts often explicitly spell voiced consonants that reflect dialect-specific voicing patterns. These dialectal voicing patterns, which may differ from those in SMG, create additional orthographic complexity when represented in writing. To annotate this phenomenon, the corresponding unvoiced form is used as the lemma, and we add the feature-value pair Voicing=Voiced in the MISC column. This approach contributes to annotation consistency and facilitates knowledge transfer from SMG, which typically uses the unvoiced version of these lemmas.

4.8 Standardization and Translation

Sometimes the text underlying a UD treebank does not conform to canonical spelling or other grammatical rules of the language. In most situations, it is desirable to preserve the error because taggers and parsers that learn their models from the data should learn how to deal with noisy input too. On the other hand, it is also desirable to mark such places as errors and to show the correct spelling so that an application can hide bad sentences or present their correct version when necessary.

Working with dialectal text sources presents a significant challenge due to the absence of standardized orthography. The literary and folkloric texts in our corpus use inconsistent spelling conventions to represent dialectal features, complicating computational processing. While we decided against normalizing the original data sources to an artificial Lesbian standard—as no such written or spoken standard exists and would contradict our goal of curating diverse sources—we still needed to address orthographic inconsistency. Therefore, to facilitate language modeling and enhance cross-dialectal knowledge transfer, we implemented a standardization process integrated with the UD annotation scheme, following the guidelines for handling non-standard forms described in the UD documentation.[4] In our approach (Examples 1-2 in Appendix B):

1. Original dialectal forms were preserved in the FORM (2nd) column of the CoNLL-U.
2. Standardized forms (i.e., correct forms closer to SMG orthography or systematized spelling for

[3] https://universaldependencies.org/u/dep/all.html#compound-redup-reduplicated-compounds

[4] https://universaldependencies.org/u/overview/typos.html

northern vocalism) were annotated in the MISC (10th) column following the annotation guidelines of UDs.

3. We developed a dedicated script that processes the CoNLL-U files, extracting the standardized forms from the MISC column and inserting them into the FORMs[5].

This standardization was crucial for the subsequent active annotation cycles, as it allowed our models to better leverage lexical and morphological knowledge from SMG while preserving the original characteristics of the resources.

Furthermore, we have incorporated translations of each sentence in SMG, produced by the annotators. These translations maintain maximum fidelity to the original sentences while adhering to SMG conventions, thereby establishing a parallel corpus that may facilitate future comparative research and computational applications.

5 Transfer Experiments

This section presents baseline experiments for evaluating dependency parsing performance on the Lesbian treebank. The experimental setup aligns with recent approaches to NLP in low-resource settings, as surveyed by Hedderich et al. (2021).

5.1 Active Annotation

To annotate the Lesbian treebank, we employed active annotation (Vlachos, 2006). In order to facilitate comparative analysis with Vakirtzian et al.'s (2025) prior research and results on knowledge transfer between SMG and a southern MG dialect (Cretan), and to investigate the extent to which SMG can contribute to modeling a northern MG dialect (Lesbian), we replicated their experimental regime.

Initially, 40 unlabeled Lesbian samples were annotated with a model trained on GUD (that represents SMG). In each subsequent cycle, 40 samples from the model's output were corrected, out of which 30 were allocated to the training set and 10 to the development set. The corrected samples were incorporated into the existing datasets, and the model was retrained on the augmented data. For evaluation, we used a test set of 30 manually annotated samples. All samples were randomly selected to ensure unbiased representation. To enhance cross-dialectal transfer, we utilized standardized forms from the MISC column (Section 4.8).

	1st	2nd	3rd	4th	5th	6th
Sentences						
Train	30	60	90	120	150	180
Dev	10	20	30	40	50	60
Test	30	30	30	30	30	30
Tokens						
Train	392	810	1220	1604	1994	2381
Dev	166	308	449	595	695	826
Test	396	396	396	396	396	396

Table 1: Lesbian sentences and tokens per round of active annotation.

During the development of the UD treebank for the Lesbian dialect, annotation guidelines evolved alongside our research, with revisions consistently applied across all datasets to ensure consistency.

5.2 Reducing Dialectal Distance

While Lesbian's morphosyntax is similar to SMG, phonological variation, especially northern vocalism, complicates cross-dialectal knowledge transfer. To address this, we used simple linguistic rules to reduce the dialectal gap by generating synthetic data (Aufrant et al., 2016).

Specifically, we created a Python script to transform the GUD treebank according to key features of northern MG phonology.[6] This resulted in a "northernized" version of GUD (NGUD) that more closely approximates the phonological profile of Lesbian.

The transformation involved two primary modifications (Examples 3-4 in Appendix B):

1. **Applying northern vocalism rules**, including the raising of unstressed /e/ to [i] and /o/ to [u], and the deletion of unstressed high vowels /i, u/, implemented within the constraints of MG orthography.
2. **Altering definite articles**: masculine nominative singular *ο* was systematically replaced with *η*, reflecting patterns attested in northern MG dialects.

Importantly, all syntactic structures and dependency relations were preserved, ensuring compatibility with UD annotation standards.

This synthetic data augmentation strategy enabled us to test whether the reduction in orthographic distance between SMG and Lesbian Greek due to the phonological distance between these two dialects can improve the effectiveness of cross-dialectal transfer. It also allowed us to isolate the

[5] https://github.com/stavros-bompolas/conllu-correct-forms

[6] https://github.com/stavros-bompolas/ngud-transformer

role of phonological divergence as a potential barrier to transfer between dialects belonging to different clusters—southern (SMG) *vs.* northern (Lesbian).

5.3 Models

For the experiments, we used the open-source Stanza package (Qi et al., 2020) with three distinct settings:[7]

1. **Lesbian-only**: A model trained exclusively on the Lesbian data that increased at each round by 40 samples (30 in the training set and 10 in the development set).
2. **GUD+Lesbian**: A model trained on the combination of GUD (1,807 sentences, 25,493 tokens) and Lesbian data, with the Lesbian component increasing exactly as in the Lesbian-only model.
3. **NGUD+Lesbian**: A model trained on the northernized GUD treebank plus the Lesbian data, to test whether reducing dialectal distance through synthetic data enhances knowledge transfer.

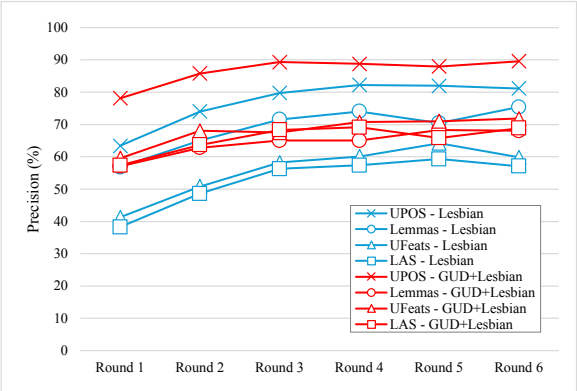
In all settings, we fine-tuned the Greek BERT model (Koutsikakis et al., 2020).

5.4 Results

Figure 2 displays the precision metrics for UPOS, Lemmas, UFeats, and LAS in six evaluation rounds. The remaining metrics can be found in Appendix A.[8] In presenting the results, we compare our findings with Vakirtzian et al.'s (2025) work on knowledge transfer between SMG and Cretan (a southern MG dialect) to examine how SMG contributes to modeling Lesbian (a northern MG dialect). We refer to Vakirtzian et al.'s GUD+Cretan model as "GUD+Cretan" and their Cretan-only model as "Cretan-only" throughout our discussion.

Overall, results show that both knowledge transfer approaches significantly outperform the dialect-only model.

UPOS: NGUD/GUD+Lesbian both reach 89.62%, compared to 82.24% for the Lesbian-only model. However, GUD+Cretan achieves 92.90%, suggesting more effective transfer for the southern dialect.



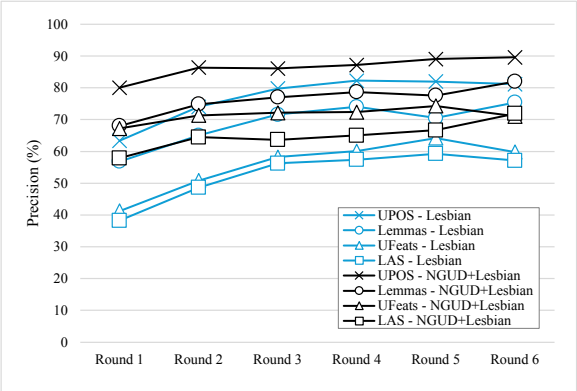


Figure 2: Precision scores between the Lesbian and GUD+Lesbian (top) / NGUD+Lesbian (bottom).

UFeats: NGUD+Lesbian (74.32%) outperforms GUD+Lesbian (71.86%) and Lesbian-only (64.21%), but lags behind both Cretan-only (78.70%) and GUD+Cretan (87.22%), highlighting the greater challenge for northern dialects in knowledge transfer.

Lemmas: NGUD+Lesbian achieves the highest performance (81.97%), dramatically outperforming GUD+Lesbian (68.31%) and matching Cretan-only (81.34%), demonstrating the impact of synthetic data augmentation.

LAS: NGUD+Lesbian (71.86%) outperforms both Lesbian-only (59.29%) and Cretan-only (67.75%), though GUD+Cretan still leads with 78.50%, indicating that dialectal proximity remains a key factor in successful knowledge transfer.

6 Discussion

Comparing knowledge transfer between SMG and a northern (Lesbian) *vs.* a southern (Cretan) dialect reveals several key insights:

[7] https://osf.io/yacxu/?view_only=37105ee21ef64ee297109b99dc875c38

[8] For evaluation, we used the standard CoNLL shared task evaluation script, which computes precision, recall, F1-score, and accuracy metrics. We report the precision values from this script's output. Available at: https://github.com/universaldependencies/tools?tab=readme-ov-file#evalpy

The impact of dialectal distance: A southern dialect benefits more from SMG knowledge transfer than a northern one, as evidenced by the superior performance of GUD+Cretan over GUD+Lesbian across all metrics. This stems from Cretan's closer vocalic system to SMG, while Lesbian's northern vocalism introduces greater surface differences at multiple linguistic levels, demonstrating that phonological distance between northern and southern dialects limits cross-dialectal transfer, regardless of orthographic consistency (see Vakirtzian et al. 2024 for similar results in ASR for MG dialects; see also Faisal and Anastasopoulos 2022).

The effectiveness of adaptation through synthetic data: Our synthetic data approach significantly narrows this gap, enabling the northern dialect model to match or outperform standalone southern models across several metrics, particularly lemmatization. Its effectiveness correlates with the degree of transformation: of the 25,493 tokens processed in the GUD treebank, only 39.82% of surface forms and 31.82% of lemmas remained unaffected by the script, with the rest undergoing northern vocalism adaptations. These results underscore the potential for further improvements through expanded rules and additional resources such as dictionaries (Zhao et al., 2009) and parallel corpora (Yarowsky et al., 2001).

The importance of available resources for the standard variety: The effectiveness of (N)GUD highlights the crucial role of robust standard variety resources in cross-dialectal transfer. Even for less similar dialects, such as Lesbian, high-quality SMG resources enhance performance, especially when combined with appropriate adaptation techniques. High-resource standard varieties cover greater linguistic variability, providing valuable baselines for transfer without the challenges of dialectal resource development (Snæbjarnarson et al., 2023).

The role of source characteristics in knowledge transfer: The Lesbian treebank integrates diverse resources, introducing variation that complicates cross-dialectal transfer but enhances representativeness (Dobrovoljc, 2022). In contrast, Cretan's data are more uniform as they come from a single speaker. Additionally, the orthography used to transcribe Cretan oral material is identical to SMG orthography, which likely facilitated knowledge transfer from the SMG treebank.

7 Conclusions

In this paper, we presented the first UD treebank for a northern MG dialect (Lesbian), along with tailored annotation guidelines and cross-dialectal transfer experiments.

Our findings suggest that effective cross-dialectal knowledge transfer depends on several factors: (a) greater dialectal distance reduces transfer effectiveness; (b) the nature and diversity of source materials affect performance; (c) simple rule-based transformations of high-resource varieties can substantially improve performance for distant dialects.

This research extends to other MG dialects, particularly northern varieties that share similar phonological features with Lesbian. Hence, the treebank provides a foundation for developing resources for other northern dialects and contributes significantly to advancing dialectal diversity in Greek NLP.

Our work highlights an inherent paradox in cross-dialectal knowledge transfer: while we aim to leverage pre-existing resources from high-resource varieties, doing so effectively often requires developing additional dialect-specific technologies—the very situation we sought to avoid through transfer learning. Similarly, a tension exists between preserving dialectal characteristics and adapting linguistic representations to enhance cross-dialectal transfer. These contradictions underscore the complex relationship between linguistic authenticity and technological pragmatism in developing NLP resources for dialectal varieties.

8 Limitations

Several limitations should be acknowledged:

First, the current treebank is small (270 sentences), limiting model performance and comprehensive linguistic documentation.

Second, our data relies solely on written sources that may reflect authors' idealized forms rather than authentic dialectal usage. Although we have recently collected oral Lesbian dialectal data for future incorporation, the current resource lacks this direct representation.

Third, our rule-based approach to northern vocalism adaptation applies changes deterministically, whereas actual dialectal usage shows considerable variation in the application of these phonological rules. A probabilistic transformation model

might better capture this natural variation.

Fourth, our experimental design, while enabling direct comparison with previous work on Cretan, may not represent the optimal approach for cross-dialectal transfer between distant varieties. Alternative methods such as leveraging larger pre-trained language models specifically fine-tuned for dialectal variation could potentially yield better results.

Finally, we did not explore the extent to which our findings generalize to other dialectal pairs with similar degrees of distance, either within Greek or in other languages with comparable dialectal landscapes. Such comparative analysis would provide a more comprehensive understanding of the relationship between dialectal distance and transfer effectiveness.

Acknowledgments

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. It also received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

References

Vasileia Alexelli. 2021. *Chartografisi tis glossikis poikilias tis Lesvou [Mapping the linguistic variety of Lesbos]*. Ph.D. thesis, University of Patras, School of Humanities and Social Sciences, Department of Philology, Linguistics Section.

Maria Ach. Anagnostopoulou. 2021. *Thematiko Lexiko tis Lesviakis Dialektou [Thematic Dictionary of the Lesbos Dialect]*. Mythos BOOKS, Mytilene.

Spyridon Anagnostou. 1903. *Lesviaka, iti, Sylogi laografikon peri Lesvou pragmateion [Lesbian Studies, or, Collection of Folkloric Treatises on Lesbos]*. From the Press of the "Anestis Konstantinidis" establishments, Athens.

Vasilis Tz. Anagnostou. 2014. *Tsi sta th'ka mas: Komodia sta k'stariot'ka [Tsi sta th'ka mas: Comedy in the K'stariot'ka Dialect]*, first edition. Estia Technon Skoutarou "T'Apono to Scholio".

Lauriane Aufrant, Guillaume Wisniewski, and François Yvon. 2016. Zero-resource Dependency Parsing: Boosting Delexicalized Cross-lingual Transfer with Linguistic Knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 119–130, Osaka, Japan. The COLING 2016 Organizing Committee.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Georgios Chatzidakis. 1905. *Mesaionika kai Nea Ellinika [Medieval and Modern Greek]*, volume 1–2. Sakellarios, Athens.

Ayan Das and Sudeshna Sarkar. 2020. A survey of the model transfer approaches to cross-lingual dependency parsing. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(5):67:1–67:60.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Kaja Dobrovoljc. 2022. Spoken language treebanks in Universal Dependencies: an overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. *Preprint*, arXiv:2205.09634.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Preprint*, arXiv:2010.12309.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The Greeks visiting Sesame Street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, pages 110–117, New York, NY, USA. Association for Computing Machinery. Event-place: Athens, Greece.

Andre Kåsen, Kristin Hagen, Anders Nøklestad, Joel Priestly, Per Erik Solberg, and Dag Trygve Truslew Haug. 2022. The Norwegian Dialect Corpus Treebank. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4827–4832, Marseille, France. European Language Resources Association.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2348–2354, Reykjavik,

Iceland. European Language Resources Association (ELRA).

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Stella Markantonatou, Vivian Stamou, Stavros Bompolas, Katerina Anastasopoulou, Irianna Linardaki Vasileiadi, Konstantinos Diamantopoulos, Yannis Kazos, and Antonios Anastasopoulos. 2025. VMWE identification with models trained on GUD (a UDv.2 treebank of Standard Modern Greek). In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 14–20, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Dimitra Melissaropoulou and Angela Ralli. 2010. Greek derivational structures: restrictions and constraints. *Morphology*, 20(2):343–357.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. A Four-Dialect Treebank for Occitan: Building Process and Parsing Experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Dimitris Papanis and Giannis D. Papanis. 2004. *Lexiko tou Agiasotikou Glosikou Idiomatos (Erminetiko - Etimologiko) [Dictionary of the Agiasos Dialect (Explanatory - Etymological)]*, 3rd improved and expanded edition. Private edition, Mytilene.

Prokopis Prokopidis and Haris Papageorgiou. 2017. Universal Dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 102–106, Gothenburg, Sweden. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Angela Ralli. 2015. Strategies and Patterns of Loan Verb Integration in Modern Greek Varieties. In Angela Ralli, editor, *Contact Morphology in Modern Greek Dialects*, pages 73–88. Cambridge Scholars Publishing.

Angela Ralli. 2017. *Lexiko dialektikis poikilias. Kydonion-Moschonision kai Voreioanatolikis Lesvou. Orthografiko-Proforas-Erminetiko-Christiko-Etymologiko-Synonymon [Dictionary of dialectal variety. Kydonies-Moschonisia and Northeastern Lesbos. Orthographic-Pronunciation-Interpretive-Practical-Etymological-Synonyms]*. Hellenic Foundation for Historical Studies, Athens.

Angela Ralli. 2019a. Affixoids and Verb Borrowing in Aivaliot Morphology. In Angela Ralli, editor, *The Morphology of Asia Minor Greek*, pages 221–254. BRILL.

Angela Ralli. 2019b. Glossiki chartografisi: O ilektronikos dialektikos atlas tis Lesvou [Linguistic mapping: The electronic dialect atlas of Lesbos]. In Grammatiki A. Karla, Io Manolessou, and Nikolaos Pantelidis, editors, *Lexeis: Timitikos tomos gia tin Christina Basea-Bezantakou [Words: Festschrift for Christina Basea-Bezantakou]*, pages 435–456. Kardamitsa, Athens.

Angela Ralli. 2019c. Greek in Contact with Romance. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Konstantinos Sampanis and Prokopis Prokopidis. 2021. Asia Minor Greek in Contact (AMGiC): Towards a dialectal Treebank comprising contact-induced grammatical changes. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 86–95, Sofia, Bulgaria. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on faroese. *Preprint*, arXiv:2304.08823.

Eustratia Tsokarou-Mitsioni. 1998. *Palies Istories ap tn Agiasiou [Old Stories from Agiasio]*, 2nd edition. Private Edition, Mytilene.

Eustratia Tsokarou-Mitsioni. 2019. *Prosfygiá [Refugeehood]*, first edition. D. Doukas & Sia O.V.E.E. / Eustratia Tsokarou-Mitsioni.

Socrates Vakirtzian, Vivian Stamou, Yannis Kazos, and Stella Markantonatou. 2025. Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 776–784, Tallinn, Estonia. University of Tartu Library.

Socrates Vakirtzian, Chara Tsoukala, Stavros Bompolas, Katerina Mouzou, Vivian Stamou, Georgios Paraskevopoulos, Antonios Dimakis, Stella Markantonatou, Angela Ralli, and Antonios Anastasopoulos. 2024. Speech Recognition for Greek Dialects: A Challenging Benchmark. In *Interspeech 2024*, pages 3974–3978. ISCA.

Andreas Vlachos. 2006. Active Annotation. In *Proceedings of the Workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Eleni Vligouridou, Inessa Iliadou, and Çağrı Çöltekin. 2024. A Treebank of Asia Minor Greek. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1715–1721, Torino, Italia. ELRA and ICCL.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*.

Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 55–63, Suntec, Singapore. Association for Computational Linguistics.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA Treebank of Spoken Norwegian Dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Appendix: Precision Metrics

	1st	2nd	3rd	4th	5th	6th
UPOS	63.39	74.04	79.78	82.24	81.97	81.15
UFeats	41.26	50.82	58.20	60.11	64.21	59.84
AllTags	37.16	46.72	56.01	57.38	61.20	58.74
Lemmas	56.83	65.03	71.58	74.04	70.49	75.41
UAS	64.21	69.40	75.41	76.23	75.14	74.86
LAS	38.25	48.63	56.28	57.38	59.29	57.10
CLAS	24.06	33.18	44.44	42.40	46.58	44.80
MLAS	6.60	14.55	23.11	23.96	29.22	26.24
BLEX	9.91	17.27	26.67	28.11	30.59	31.22

Table 2: Precision metrics across rounds (Trained on Lesbian sentences only).

	1st	2nd	3rd	4th	5th	6th
UPOS	78.14	85.79	89.34	88.80	87.98	89.62
UFeats	59.56	68.03	67.49	70.77	71.04	71.86
AllTags	56.28	66.12	65.85	68.03	69.13	70.22
Lemmas	57.10	62.84	65.03	65.03	68.31	68.03
UAS	73.22	77.87	82.79	83.06	80.33	82.51
LAS	57.38	63.66	68.31	69.13	65.85	68.85
CLAS	44.86	52.13	57.08	59.91	51.38	56.81
MLAS	24.77	32.70	35.38	37.74	34.40	39.44
BLEX	22.90	33.18	36.79	38.68	36.24	38.50

Table 3: Precision metrics across rounds (Trained on GUD+Lesbian sentences).

	1st	2nd	3rd	4th	5th	6th
UPOS	80.05	86.34	86.07	87.16	89.07	89.62
UFeats	67.21	71.31	72.13	72.40	74.32	71.04
AllTags	63.11	67.76	68.85	69.40	72.13	68.85
Lemmas	68.03	74.86	77.05	78.69	77.60	81.97
UAS	79.51	80.87	80.60	80.87	82.51	86.34
LAS	57.92	64.48	63.66	65.03	66.67	71.86
CLAS	45.97	52.34	51.18	54.38	54.07	60.37
MLAS	28.91	34.58	32.23	36.87	38.28	39.63
BLEX	30.33	35.98	37.44	41.94	41.63	45.62

Table 4: Precision metrics across rounds (Trained on NGUD+Lesbian sentences).

B Appendix: Examples

text = Πατέρασιτς άμα τν είδει καταφαρμακόστσει .

text_el = Ο πατέρας της, όταν την είδε, καταφαρμακώθηκε.

text_en = When her father saw her, he was devastated.

1-2 Πατέρασιτς _ _ _ _ _ _ _ _ _

1 Πατέρασ Πατέρας NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _

2 ιτς μ PRON _ Case=Gen|Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs 1 nmod _ MGloss=euphonic-her|MSeg=ι-τς

3 άμα άμα SCONJ _ _ 5 mark _ _

4 τν ιγώ PRON _ Case=Acc|Gender=Fem|Number=Sing|Person=3|PronType=Prs 5 obj _ _

5 είδει βλέπου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|**Typo=Yes**|VerbForm=Fin|Voice=Act 6 advcl _ **CorrectForm=είδι**

6 καταφαρμακόστσει καταφαρμακώνου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|**Typo=Yes**|VerbForm=Fin|Voice=Pass 0 root _ **CorrectForm=καταφαρμακώστσι**

7 . . PUNCT _ _ 6 punct _ _

Example 1: Integration of orthographic standardization for the dialect in the 10th (MISC) column.

text = Πατέρασιτς άμα τν είδει καταφαρμακόστσει .

text_el = Ο πατέρας της, όταν την είδε, καταφαρμακώθηκε.

text_en = When her father saw her, he was devastated.

1-2 Πατέρασιτς _ _ _ _ _ _ _ _ _

1 Πατέρασ Πατέρας NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _

2 ιτς μ PRON _ Case=Gen|Gender=Fem|Number=Sing|Person=3|Poss=Yes|PronType=Prs 1 nmod _ MGloss=euphonic-her|MSeg=ι-τς

3 άμα άμα SCONJ _ _ 5 mark _ _

4 τν ιγώ PRON _ Case=Acc|Gender=Fem|Number=Sing|Person=3|PronType=Prs 5 obj _ _

5 είδι βλέπου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Typo=Yes|VerbForm=Fin|Voice=Act 6 advcl _ CorrectForm=είδι

6 καταφαρμακώστσι καταφαρμακώνου VERB _ Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|Typo=Yes|VerbForm=Fin|Voice=Pass 0 root _ CorrectForm=καταφαρμακώστσι

7 . . PUNCT _ _ 6 punct _ _

Example 2: Orthographic standardization applied automatically from the MISC column via processing script.

text = Ο υπάλληλος σ την είσοδο κουνάει το κεφάλι του , όταν μαθαίνει το σκοπό της επίσκεψής μας.

text_en = The employee shakes his head at the entrance when he learns the purpose of our visit.

1 Ο ο DET _ Case=Nom|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 2 det _ _

2 υπάλληλος υπάλληλος NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _

3 σ σε ADP _ _ 5 case _ _

4 την ο DET _ Case=Acc|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 5 det _ _

5 είσοδο είσοδος NOUN _ Case=Acc|Gender=Fem|Number=Sing 6 obl _ _

6 κουνάει κουνώ VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 0 root _ _

7 το ο DET _ Case=Acc|Definite=Def|Gender=Neut|Number=Sing|PronType=Art 8 det _ _

8 κεφάλι κεφάλι NOUN _ Case=Acc|Gender=Neut|Number=Sing 6 obj _ _

9 του μου PRON _ Case=Gen|Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=Prs 8 nmod _ _

10 , , PUNCT _ _ 12 punct _ PunctType=Comm

11 όταν όταν SCONJ _ _ 12 mark _ _

12 μαθαίνει μαθαίνω VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 6 advcl _ _

13 το ο DET _ Case=Acc|Definite=Def|Gender=Mas|Number=Sing|PronType=Art 14 det _ _

14 σκοπό σκοπός NOUN _ Case=Acc|Gender=Masc|Number=Sing 12 obj _ _

15 της ο DET _ Case=Gen|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 16 det _ _

16 επίσκεψής επίσκεψη NOUN _ Case=Gen|Gender=Fem|Number=Sing 14 nmod _ _

17 μας μου PRON _ Case=Gen|Number=Plur|Person=1|Poss=Yes|PronType=Prs 16 nmod _ SpaceAfter=No

18 . . PUNCT _ _ 6 punct _ PunctType=Peri

Example 3: Example from the GUD treebank prior to applying the transformation script.

```
# text = Η υπάλληλους σ τν είσουδου κνάει του κιφάλ τ , όταν μαθαίν του σκουπό τς ιπίσκιψής μας.
# text_en = The employee shakes his head at the entrance when he learns the purpose of our visit.
1	Η η DET _ Case=Nom|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 2 det _ _
2	υπάλληλους υπάλληλους NOUN _ Case=Nom|Gender=Masc|Number=Sing 6 nsubj _ _
3	σ σι ADP _ _ 5 case _ _
4	τν η DET _ Case=Acc|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 5 det _ _
5	είσουδου είσουδους NOUN _ Case=Acc|Gender=Fem|Number=Sing 6 obl _ _
6	κνάει κνώ VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 0 root _ _
7	του η DET _ Case=Acc|Definite=Def|Gender=Neut|Number=Sing|PronType=Art 8 det _ _
8	κιφάλ κιφάλ NOUN _ Case=Acc|Gender=Neut|Number=Sing 6 obj _ _
9	τ μ PRON _ Case=Gen|Gender=Masc|Number=Sing|Person=3|Poss=Yes|PronType=Prs 8 nmod _ _
10	, , PUNCT _ _ 12 punct _ PunctType=Comm
11	όταν όταν SCONJ _ _ 12 mark _ _
12	μαθαίν μαθαίνου VERB _ Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act 6 advcl _ _
13	του η DET _ Case=Acc|Definite=Def|Gender=Masc|Number=Sing|PronType=Art 14 det _ _
14	σκουπό σκουπός NOUN _ Case=Acc|Gender=Masc|Number=Sing 12 obj _ _
15	τς η DET _ Case=Gen|Definite=Def|Gender=Fem|Number=Sing|PronType=Art 16 det _ _
16	ιπίσκιψής ιπίσκιψ NOUN _ Case=Gen|Gender=Fem|Number=Sing 14 nmod _ _
17	μας μ PRON _ Case=Gen|Number=Plur|Person=1|Poss=Yes|PronType=Prs 16 nmod _ SpaceAfter=No
18	. . PUNCT _ _ 6 punct _ PunctType=Peri
```

Example 4: Transformation of Example 3 after applying the script with rules to convert GUD to northernized GUD (NGUD).