Towards Universal Dependencies for L2 Learners of Modern Greek: Annotation and Challenges

Christina Klironomou[1], Thelka Pasparaki[2], Arianna Masciolini[1], Alexandros Tantos[3], Despoina Ourania Touriki[3], Konstantinos Tsiotskas[3], Eleni Tsourilla[3]

[1]University of Gothenburg, [2]University of Crete, [3]University of Thessaloniki

gusklirch@student.gu.se, philp1011@philology.uoc.gr, arianna.masciolini@gu.se, alextantos@lit.auth.gr, dtouriki@lit.auth.gr, ktsiotskas@gmail.com, etsourib@lit.auth.gr

Abstract

This paper focuses on annotating the Greek Learner Corpus in Universal Dependencies (UD). It presents the annotation process, development of guidelines and evaluation of the attempted annotation of two annotators. This work is part of a larger annotation project which aims to compile a sizeable learner treebank that can be used to promote research on second language acquisition and its automatic processing.

Keywords: Universal Dependencies, Modern Greek, L2

1. Introduction

Universal Dependencies (UD) (de Marneffe et al., 2021) is a well-established framework for the cross-linguistically consistent annotation of morphosyntax, aiming to promote multilingual grammatical parsing and typological research. Since its first release in 2015, when only a handful of European languages were represented in the corpus, the UD annotation scheme has been used to create hundreds of resources for over 180 languages from all over the world, including multiple minority languages, dialects, and nonstandard varieties.

In a 2017 paper, Lee et al. proposed parallel UD treebanks as a format for L2 corpora, arguing that morphosyntactically annotated learner productions contrasted with teacher corrections are functionally equivalent to – and at the same time richer than – explicitly error-annotated corpora. In the following years, this idea has grown in popularity, leading to the development of nine such UD treebanks (Lee et al., 2017a; Berzak et al., 2016; Pulido et al., 2025; Di Nuovo et al., 2022; Sung and Shin, 2024; Hana and Hladká, 2018; Kyle et al., 2022; Rozovskaya, 2024; Volodina et al., 2025).

The aim of this work is twofold. From a computational perspective, gold-annotated learner productions allow for the inclusion of nonstandard text in parser development, thereby improving the robustness and overall performance of the models. From the perspective of second language acquisition (SLA) research, the treebank can constitute a helpful resource for cross-lingual interlanguage studies, shifting the focus from individual errors to L2 grammar as a whole.

Annotating learner corpora is a challenging undertaking, as standard annotation guidelines are typically designed for well-formed, native-like language data and do not adequately account for the systematic ungrammaticality and non-standard structures found in learner texts. Furthermore, existing learner treebanks do not follow a unified approach to the treatment of learner errors (Masciolini et al., 2025). This lack of convergence underscores the need for universal annotation guidelines tailored for learner corpora, as well as for at least some language-specific extensions that capture the idiosyncrasies of the learner variety of the language under investigation.

In the present paper, we present ongoing work on a new treebank based on version 2 of the Greek Learner Corpus (GLCII) (Tantos et al., 2023), consisting of written data produced by learners of Modern Greek.[1] For the purposes of this project, we annotated 50 texts (671 sentences) produced by L2 learners of Modern Greek across different CEFR levels along with their reference counterparts (658 sentences), extracted from the development set of the GCLII corpus. In addition, we developed treebank-internal annotation guidelines to ensure consistency across annotations and to systematically address language- and treebank-specific phenomena, particularly those arising from learner errors.

2. The Greek Learner Corpus

GLCII, the largest and freely available error-annotated Learner Corpus of Greek as a second

[1]The treebank is and its first version is planned to be released as part of UD 2.18.

Level	Texts
A1	6
A2	7
B1	13
B2	15
C1	7
C2	2

Table 1: Texts per CEFR level in the subset of GLCII currently annotated in UD.

L1	Texts
Russian	15
Serbian, Turkish	5
Albanian, Chinese, English	4
Arabic, Italian	3
Polish	2
Abkhazian, Armenian, French, Georgian, Slovak	1

Table 2: Texts per first language in the subset of GLCII currently annotated in UD.

language (L2), was developed within the Latent Aspects in L2 Acquisition (LAL2A) research program, funded by the Hellenic Foundation for Research and Innovation and designed to address limitations identified in earlier Greek learner corpora (Tantos et al., 2023). GLCII is a growing corpus, comprising 1,101 authentic raw written texts (422,360 tokens) and 318 spoken interviews produced by adult L2 Greek learners. The corpus represents a wide range of sociolinguistic and cultural backgrounds, which are extensively documented through rich metadata. The majority of the learner data was collected in classroom instruction settings. Furthermore, GLCII includes a control subcorpus of Greek L1 data consisting of 42 written texts (66,645 tokens), which is particularly suitable for systematic comparisons of L2-L1.

GLCII is error-annotated for five basic grammatical categories: Agreement, Voice, Gender, Case and Aspect. Each category is associated with a dedicated tag system designed to capture specific error types.[2] Overall, GLCII is grounded in contemporary learner corpus practices and constitutes a valuable resource for both L2 research and teaching purposes.

3. Annotation process

As mentioned above, we annotated 671 sentences and their reference counterparts (658) from 50 learner texts, extracted from the development set of the GCLII corpus.[3] The distribution of CEFR proficiency levels and L1s for this annotated subset is displayed in Table 1 and 2 respectively.

This effort is the work of two annotators, both native speakers of Modern Greek and currently master-level students of Linguistics, with previous annotation experience and basic knowledge of UD. However, this work is part of a larger project that will lead to the expansion of both the language specific guidelines, discussed in Section 4, and the treebank itself. The complete annotation group consists of six native speakers of Modern Greek.

Syntactic relations offer insights about the learners' processing of L2 structures and were therefore identified as a crucial annotation layer. For that reason, the dependency analysis of the original sentences was performed entirely manually, thus preventing any automatic annotation bias. Conversely, tagging and lemmatization were performed automatically with the *greek-gdt-ud-2.17* UDPipe 2 (Straka, 2018) model and manually validated. The same model was used to pre-annotate corrections, in this case without excluding any CoNLL-U fields.

Given the particular nature of the learner data, the quality of the automatic pre-annotation was overall satisfactory, and undoubtedly saved us a large amount of time that would be spent, e.g., annotating morphological features from scratch. With that said, the parser had serious problems segmenting sentences when the punctuation was nonstandard or when full texts were typed in capital letters (in the latter case, tokenization was also problematic, especially when punctuation was involved), thus a lot of post-processing and decisions had to be made. In addition, misspelled or unstressed words were often lemmatized incorrectly. The feature and MISC columns' values attributed by the parser were for the majority of the times correct, but often enough had to be modified and/or complemented with extra values in correspondence to grammatical errors. It must be noted that learner data pose additional challenges as features cannot always be assigned with certainty when the words are malformed or the sentence is ambiguous. For example, when a learner sentence contains structures such as *ήρθα αύριο* [='I came tomorrow'], we do not know – and cannot always infer from the context – if the learner meant 'I came yesterday' or 'I will come tomorrow'. Having a reference text can often suggest a way out for the annotator, since a lot of choices for resolving such ambiguities are taken while constructing

[2] Further information about the annotation scheme is available through the official GLC Gateway interaction platform, which offers documentation and access to selected portions of the corpus (glc.lit.auth.gr/app/GLC_Gateway)

[3] The total number of sentences varies slightly depending on the annotators' segmentation criteria.

a reference text, but this is not always the case.

Furthermore, Modern Greek presents significant challenges due to widespread syncretism, particularly within the nominal domain where identical morphological forms can map to multiple grammatical values. This leads to two primary types of error ambiguity according to (Tantos and Amvrazis, 2022): classification-level ambiguity, where it is unclear whether an error stems from a failure in agreement or incorrect gender assignment to a noun, and identification-level ambiguity, where a single malformed token could represent multiple non-target cases or genders. During this initial stage, each annotator had to keep the automatic feature annotation or modify it according to their intuition, with discrepancies settled during a conflict resolution process, and ambiguities were resolved by selecting one of the possibly suitable feature values.

To address these challenges without imposing biased or deterministic interpretations, we are considering the possibility of incorporating the descriptive, theory-neutral approach proposed by (Tantos and Amvrazis, 2022) in future versions of the treebank. Specifically, this approach utilizes disjunctive tags to handle identification-level issues and represents disjunctive relations between multiple plausible error domains. Such a strategy would allow the treebank to capture the inherent complexity of the learner's interlanguage while significantly improving inter-annotator agreement (IAA) by reducing subjectivity.

4. Internal guidelines

4.1. Segmentation

Original texts are segmented into sentences based on their punctuation unless there are very strong signs that the learner forgot or misplaced the period (e.g. another sentence beginning with a capitalized word is following, but the period is missing). The annotators tried to keep the text segmentation as faithful to the learner's judgment as possible, sometimes diverging from the choice made in the reference text. To address the potential sentence misaligment between learner and reference text, a three-level naming system for sentence IDs was developed: If, for instance, the second sentence of the sixth text in the reference file (6.2) corresponds to two sentences in the original text, in the learner file, the sentence id will be 6.2.1 for the first sentence and 6.2.2 for the second. This allows obtaining a fully sentence-parallel treebank, similar to the ones available for Chinese (Lee et al., 2017a), English (Berzak et al., 2016), Italian (Di Nuovo et al., 2022), Russian (Rozovskaya, 2024) and Swedish (Volodina et al., 2025).

When dealing with under-segmented words where the merged tokens are assumed to be a result of mistyping, tokenization and annotation is done as if they were correctly separated, indicating that a space is missing in the MISC field of the first token. In *το ιδιο ισχυει* ***καιγια*** *την νυχτερίδα* [='the same applies alsofor the bat'], for instance, we assume that the learner knows that *και* ('and, also´) and *για* ('for') are separate words and annotate them as separate tokens. When the under-segmentation is not clearly a result of mistyping, the under-segmented tokens are treated as one and analyzed based on what is considered as the head of the merged tokens.

Following the KSL treebank (Sung and Shin, 2024), when dealing with over-segmented tokens, if the resulting tokens are not existing words in the language and no reasonable morphosyntactic annotation can be drawn, the first token is given a dependency relation, the following tokens depend on the first with the `goeswith` relation, and POS tag X and no features are assigned to the latter. In addition, `CorrectSpaceAfter=No` is placed in the MISC column of the following token.

Ellipsis points ('...') are annotated as one token, but multiple punctuation marks that do not form a unit (i.e. their meaning does not change when they are combined, e.g. '!!', ';;', '..') are treated as individual tokens.

4.2. Derivational morphology

Following the VALICO treebank (Di Nuovo et al., 2022), lemmas in the annotation of the learner texts keep the spelling and stress of the tokens as they were typed by the learner. The features are also descriptive of the observed form, independent of the learner's intended meaning, unless it causes problems in producing a syntactic annotation. POS tags and syntactic relations are also assigned based on the form of the given lemma. Thus, in the sentence *Αν και ήθελα να πάω όχι στην* ***Αλβανός*** *αλλά στην Ελλάδα* [='Even though I wanted to go not to Albania(n) but to Greece'], the word *Αλβανός* ('Albanian', lit. person of Albanian origin) is annotated as a noun:

PART	ADP	DET	NOUN	CCONJ	ADP	DET	PROPN
όχι	σ	την	**Αλβανός**	αλλά	σ	την	Ελλάδα
not	in	the	Albanian	but	in	the	Greece
			Αλβανός				
			Case=Nom				
			Gender=Masc				
			Number=Sing				

A fully interpretative annotation would lemmatize according to the hypothesized intended form *Αλβανία* ('Albania'), which would be tagged as a proper noun with the features `Case=Acc|Gender=Fem|Number=Sing`.

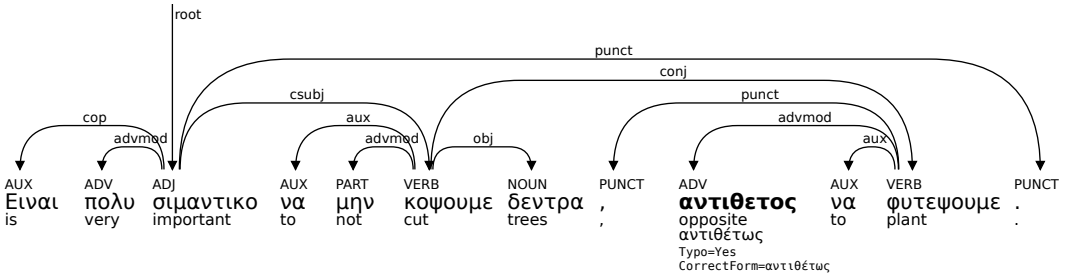

Figure 1: Application for the guidelines for misspellings. The adjectival form *αντιθετος* is used instead of the corresponding adverb, but since this is the result of a phonetic misspelling, the lemma, UPOS tag and DEPREL are assigned based on the target form *αντιθέτως*. The presence of a misspelling and the target form itself are indicated in the FORM and MISC fields. Another spelling error is present in the above example sentence (*σιμαντικο* instead of *σημαντικό*), but the wrong choice of */i/* does not change the morphosyntactic analysis of the token or the sentence.

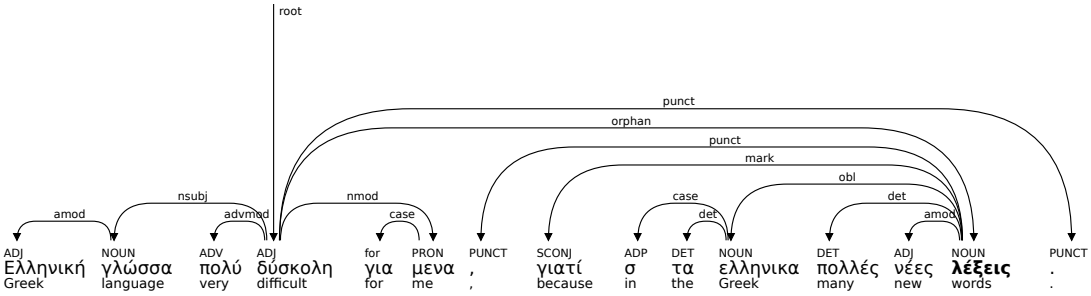

Figure 2: Application of the guidelines for ellipsis. The sentence, which translates approximately to '(the) Greek language (is) very difficult for me, because in Greek (there are) many new words', lacks a determiner and an auxiliary in the main clause, but these do not pose a problem as they would be leaves in the UD tree. In the subordinate, however, *λέξεις* appears to be the subject of an (omitted) existential verb (e.g. *υπάρχουν*, 'there exist'). It is therefore attached to the root and labeled `orphan`.

4.3. Orthography

When the spelling or stress point to a UPOS tag that does not permit any kind of reasonable syntactic annotation but it is obvious that the learner misspelled, *'Typo=Yes'* is added to the feature column and the lemma and features correspond to the token that is assumed that the learner meant. The intended token is also specified to the MISC column using `CorrectForm=[...]`. For example, in the sentence *Ειναι πολυ σιμαντικο να μην κοψουμε δεντρα,* ***αντιθετος*** *να φυτεψουμε* ['It is very important not to cut trees, converse(ly) to plant] we suggest that the learner used the adjective *αντιθετος* ('opposite') instead of the adverb *αντιθέτως* ('on the contrary') not because they were unaware of the target form, but rather as a result of a phonetic misspelling that is common even among L1 speakers. We therefore annotate on the basis of the target form (cf. Figure 1).

Spelling and stress errors are not marked with the current guidelines in the L1 CoNLL-U file, but can be found by comparing the UD annotation of the learner file with the corrected (L1) counterpart. Note also that a script that automatically extracts such errors from comparing UD annotated files is currently developed by authors of this paper.

For words transliterated from other languages, we apply the guidelines for borrowed analysis of code-switched material,[4] i.e. assign no features and specify the source language in the MISC column. For instance, the word *Ηαουαι* – a transliteration of the name 'Hawaii', used instead of its proper Greek translation *Χαβάη* – is annotated with `OrigLang=en`.

4.4. Syntax

In our annotated subset of texts, written mostly by B2 learners within the CEFR framework, most of the non-trivial cases of syntactic annotation concern missing or redundant tokens. When it comes to missing tokens, we follow the guidelines for promotion by head elision and ellipsis of the CLF and Russian learner treebanks, as described in Mascioloni et al. (2025) (cf. Figure 2).

Repeated tokens (not for emphatic reasons), in case they do not serve any semantic purpose or cannot be assigned any syntactic function, are assigned the *reparandum* relation to their following counterparts that can be analyzed syntactically

[4] universaldependencies.org/foreign.html#option-2-borrowed-analysis

	LEMMA	UPOS	UAS	LAS
Originals	97.5%	97%	94.1%	91.32%
References	98.95%	98%	95.26%	92.99%

Table 3: IAA on original learner texts and corrected references.

(the presumed 'repair'). This comes as an extension to the typical use of *reparandum*, where it is used to indicate disfluencies overridden in a speech repair.

5. Inter-Annotator Agreement

The manual syntactic annotation of the original learner sentences proved less challenging than expected. This can also be seen from the extent of the inter-annotator agreement (IAA) (cf. Table 3) and is perhaps also related to the relatively high level of proficiency of the learners (cf. Section 1). Having a reference text, to which we had chosen to remain faithful, was also helpful sometimes, giving us a way out of structures that posed ambiguity (semantic or syntactic) to their interpretation. However, the shortage of learner-specific/error-specific guidelines posed additional requirements of effort and attention to the annotation process.

We conducted a three-level quantitative evaluation of the annotation:

- Raw agreement on lemma assignment (counts of identical lemmas / total number of tokens).
- IAA on POS tags, measured with Cohen's kappa coefficient, a metric that also accounts for potential cases where the two annotators selected the same label by chance.
- IAA on dependencies, using UAS (Unlabeled Attachment Score) and LAS (Labeled Attachment Score).

The scores acquired are comparable to the ones of other learner treebanks (cf. for example Berzak et al. 2016; Lee et al. 2017a; Di Nuovo et al. 2022; Sung and Shin 2024). For the case of the dependency relations on the reference texts, as the normalized data are less ambiguous and the parser annotation bias is present, agreement scores were higher. There was no dependency that distinguished particularly for causing annotator disagreement.

6. Limitations

The present work is part of an ongoing annotation process, thus not all choices followed are definitive and the size of the treebank is still relatively small, especially compared to that of the source corpus (the annotation covers 4,5% of the whole corpus). We are currently working on extending the treebank to include more pairs of learner and reference texts, and enhancing our annotation technique in a way to better handle the ambiguity issues discussed in Section 3. This will allow to derive error labels compatible with the original GLCII guidelines from the UD annotations themselves.

7. Conclusions

This study presented the annotation and evaluation of a Greek learner treebank within the UD framework. The annotation process required the adaptation of existing UD guidelines and led to the development of new ones for non-standard linguistic data, addressing phenomena characteristic of learner language. Inter-annotator agreement scores indicate a satisfactory level of consistency, suggesting that the proposed guidelines can be applied reliably despite the variability inherent in learner production. Nevertheless, certain constructions remain challenging, highlighting the need for further refinement of annotation criteria and additional data. Overall, the resulting dataset, scheduled for a first release as part of 2.18, provides systematically categorized and linguistically analyzed material, suitable for research on SLA and pedagogical applications. The present work is part of a broader annotation effort focusing on GLCII annotation for UD. Future work will focus on assessing its utility in downstream tasks, particularly dependency parser training and fine-tuning.

8. Acknowledgments

This work has been supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; project id 2023-00161). In addition, we would like to thank Onassis Foundation, who supports the main author of this work under the Scholarship Program for Greek students 2025-2026.

9. Bibliographical References

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal Dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746, Berlin, Germany. Association for Computational Linguistics.

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

Elisa Di Nuovo, Manuela Sanguinetti, Alessandro Mazzei, Elisa Corino, and Cristina Bosco. 2022. VALICO-UD: Treebanking an Italian learner corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 8(8-1).

Jirka Hana and Barbora Hladká. 2018. Universal dependencies and non-native Czech. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, Oslo University, Norway. LiU Electronic Press.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A dependency treebank of spoken second language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

John Lee, Herman Leung, and Keying Li. 2017a. Towards Universal Dependencies for learner Chinese. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 67–71, Gothenburg, Sweden. Association for Computational Linguistics.

John Lee, Keying Li, and Herman Leung. 2017b. L1-L2 parallel dependency treebank as learner corpus. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 44–49, Pisa, Italy. Association for Computational Linguistics.

Arianna Masciolini, Aleksandrs Berdicevskis, Maria Irena Szawerna, and Elena Volodina. 2025. Annotating second language in Universal Dependencies: a review of current practices and directions for harmonized guidelines. In *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 153–163, Ljubljana, Slovenia. Association for Computational Linguistics.

Emiliana Pulido, Robert Pugh, and Zoey Liu. 2025. I speak for the árboles: Developing a dependency treebank for Spanish L2 and heritage speakers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 814–822, Vienna, Austria. Association for Computational Linguistics.

Alla Rozovskaya. 2024. Universal Dependencies for learner Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17112–17119, Torino, Italia. ELRA and ICCL.

Hakyung Sung and Gyu-Ho Shin. 2024. Constructing a dependency treebank for second language learners of Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3747–3758, Torino, Italia. ELRA and ICCL.

Alexandros Tantos and Nikolaos Amvrazis. 2022. Classification and identification level ambiguity in error annotation. *Applied Corpus Linguistics*, 2(3):100035.

Alexandros Tantos, Nikolaos Amvrazis, and Eleni Drakonaki. 2023. Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics*, 36.

Elena Volodina, Arianna Masciolini, Beáta Megyesi, Julia Prentice, Lisa Rudebeck, Gunlögand Sundberg, and Mats Wirén. 2025. SweLL with pride: How to put a learner corpus to good use. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis, and Elena Volodina, editors, *Huminfra Handbook: Empowering digital and experimental humanities*, number 59 in NEALT Proceedings Series, pages 251–306. University of Tartu Library.

10. Language Resource References

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Alexandros Tantos, Nikolaos Amvrazis, and Eleni Drakonaki. 2023. Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics*, 36.