Building a Contact Language Treebank with UniDive Support: The Asia Minor Greek in Contact (AMGiC) UD Resource

Anonymous ACL submission

Relevant UniDive working groups: WG1, WG4

1 Introduction

The UniDive COST Action (CA21167) has provided a framework for advancing Universal Dependencies (UD) resources for typologically diverse and under-documented languages. In this abstract, we report on how UniDive activities—Short-Term Scientific Missions (STSMs), working group discussions, and training schools—have directly contributed to the development of the *Asia Minor Greek in Contact* (AMGiC) treebank, a UD resource for annotating contact-induced morphosyntactic phenomena (CIMSP) in Asia Minor Greek (AMG) varieties.

AMG designates a group of Greek dialects spoken in Central Anatolia until 1923, when the Treaty of Lausanne mandated a population exchange between Greece and Turkey. These varieties—Silliot, Cappadocian, and Pharasiot Greek—underwent intensive contact with Turkish, resulting in significant typological restructuring: borrowing of free grammatical morphemes, agglutinative-like noun inflection, copula encliticization, and adoption of head-final syntactic configurations (Janse, 2009; Thomason and Kaufman, 1988). The AMGiC treebank was initiated by Sampanis and Prokopidis (2021) and has since expanded through UniDive-supported collaboration.

2 UniDive Contributions to AMGiC

STSMs. Two UniDive STSMs were central to advancing the treebank. In June 2024, an STSM at the University of Vienna (titled "Extending the AMGiC UD treebank") focused on developing annotation infrastructure: the BoAT (Boğaziçi Annotation Tool) web application was extended with metadata editing and sentence management functionalities tailored to AMGiC's needs, and the tool was introduced to the treebank's annotator, who had previously relied on a text editor. The tool and the treebank were also presented in the "Historical Grammar of Modern Greek" course at the University of Vienna. In September 2024, a reciprocal STSM at Boğaziçi University enabled intensive collaborative annotation work, resulting in the completion of the first 36 Silliot sentences and their preparation for publication.

Working groups. AMGiC is relevant to WG1 (Corpus Annotation), as it develops novel annotation conventions for contact phenomena within the UD/CoNLL-U framework, and to WG4 (Quantifying and Promoting Diversity), as it contributes a UD resource for a severely under-documented and endangered contact variety, expanding the typological and genealogical coverage of the UD project.

Training schools. One of the authors participated in both UniDive Training Schools: the 1st (Chişinău, July 2024), which covered dependency syntax, UD annotation, treebank validation and quality control; and the 2nd (Yerevan, January 2026), which focused on low-resource and typologically diverse languages in NLP.

3 The AMGiC Treebank

The initial 36 Silliot sentences were completed as a direct outcome of the two STSMs. The treebank has since been expanded to **72 annotated sentences** (851 tokens) covering two dialect groups: Silliot (36 sentences; Kostakis, 1968; R. M. Dawkins, 1916) and Cappadocian (36 sentences, Delmesó subdialect; R. M. Dawkins, 1916). The treebank follows the UD annotation framework (Nivre et al., 2020; De Marneffe et al., 2021) and introduces two distinctive features:

CIMSP annotation. A systematic taxonomy of contact-induced phenomena is encoded in the CoNLL-U MISC field using three features: `LC=Yes`, `MorphSynC` (broad category), and `MorphSynSC` (specific subcategory). Table 1 shows the distribution across categories. The most

Tag	Category	*n*
FrGrEl	Free Grammatical Elements	37
SynIn	Syntactic Interference	7
MorphIn	Morphological Interference	5
BGrEl	Bound Grammatical Elements	3
L	Lexical	2
ELIs	Embedded Language Islands	1
	Total	**55**

Table 1: CIMSP categories in AMGiC (72 sentences).

frequent category is Free Grammatical Elements (FrGrEl), covering borrowed question particles, subordinating conjunctions, and quantifiers from Turkish.

Sociodemographic metadata. Each sentence carries metadata encoding extralinguistic parameters relevant to contact intensity—population coexistence patterns, proximity to urban centres, availability of Greek education, enclave status, and diaspora connections—informed by the framework of Karantzola et al. (2021). This enables correlations between sociolinguistic conditions and CIMSP frequency as the treebank grows.

4 Outcomes and Future Work

The UniDive-supported development of AMGiC has produced the following outcomes: (a) the completion of the initial 36-sentence Silliot portion of the treebank through two STSMs, with subsequent expansion to 72 sentences covering a second dialect group; (b) a replicable CIMSP taxonomy for annotating contact phenomena within the UD framework, applicable to other contact situations; (c) the BoAT collaborative annotation tool, extended with metadata editing capabilities specifically for AMGiC during the Vienna STSM; and (d) sociodemographic metadata enabling quantitative sociolinguistic analysis of contact intensity.

Future directions include expanding the treebank to cover additional Cappadocian subdialects and Pharasiot Greek, and conducting statistical analyses of CIMSP distribution across dialects and sociodemographic conditions. The treebank will be made publicly available through UD (UD_Cappadocian-AMGiC) under a Creative Commons licence.

References

Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 1–54.

Mark Janse. 2009. Greek-Turkish Language Contact in Asia Minor. *Études Helléniques/Hellenic Studies*, 17:37–54.

Eleni Karantzola, Anatoli Theodoridi, and Konstantinos Sampanis. 2021. The Interplay of External and Sociolinguistic Factors in Contact-Induced Language Change: Cappadocian Greek as a Case Study. *Mediterranean Language Review*, 28:21.

Athanasios P. Kostakis. 1968. *To glossiko idioma tis Sillis*. Centre for Asia Minor Studies, Athens. The dialect of Silli.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

R. M. Dawkins. 1916. *Modern Greek in Asia Minor*. Cambridge 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.

Sarah Grey Thomason and Terrence Kaufman. 1988. *Language Contact, Creolization, and Genetic Linguistics*, first paperback printing 1991, reprint 2020 edition. University of California Press, Berkeley, CA.