Dialectal treebanks and their relation with the standard variety: The case of East Cretan and Standard Modern Greek

Socrates Vakirtzian[1], Vivian Stamou[2], Yannis Kazos[2,3], Stella Markantonatou[2,4]
[1]Department of Informatics and Telecommunications, NKUA
[2]Archimedes, Athena R.C.
[3]National Technical University of Athens, NTUA
[4]Institute of Language and Speech Processing, Athena R.C.
socratesvak@hotmail.com, [vivianstamou, kazosj, stiliani.markantonatou]@gmail.com

Abstract

We report on the development of the first treebank and parser for Eastern Cretan in the framework of Universal Dependencies (UD). Eastern Cretan is a living but under-resourced dialect of Modern Greek. We have worked on the transcription of oral material and relied on active annotation and knowledge transfer from GUD, a treebank of Standard Modern Greek. Along with its other phonological and morphosyntactic differences from Standard Modern Greek, Eastern Cretan (and other varieties of Modern Greek) makes heavy use of euphonics and voicing that have not been included in the UD annotation guidelines so far. We have provided annotation guidelines for East Cretan euphonics and voicing and included them in the models. Knowledge transfer from the treebank of Standard Modern Greek to the dialectal models helped to initiate annotation via an active annotation procedure.

1 Introduction

The leaps in NLP in recent years have brought considerable efficiency to language analysis tools. This rapid progress has reduced the cost of the oral material-to-linguistically annotated text pipeline and facilitated knowledge transfer from well resourced languages to less resourced ones. At the same time it is challenging because the resulting representation of the less resourced languages may be biased by the massive evidence collected for the richly resourced ones (Bird, 2020). In the face of the increasingly rapid digitization characterising our era, it is a matter of survival for under-resourced languages to gain an independent digital presence that respects their individual nature so that they can be integrated into modern technologies and methods of study.

Considering dialects, the available linguistic data are not only scarce but are also often characterized by a significant lack of consistency in their orthographic representation. This is due to the primarily oral nature of these language varieties. Since our goal was to create language models capable of understanding the current linguistic reality, it was essential to rely on contemporary speech data.

The Eastern Cretan treebank[1] is the first morphosyntactically annotated treebank of a living Modern Greek dialect. Annotation complies to the Universal Dependencies - Version 2 (UD.V2) guidelines (de Marneffe et al., 2021). For Standard Modern Greek (SMG) there are two UD.V2 treebanks, GDT and GUD, with GUD being the most recent one[2]. GUD contains 1,807 sentences (25,493 tokens) randomly selected from fiction texts. We trained models on the Eastern Cretan treebank only, and on the Eastern Cretan treebank plus the GUD (henceforth Eastern Cretan+GUD), to see whether and to what extent SMG can contribute to the development of Eastern Cretan language models.

In Section 2, the basic linguistic differences of the Eastern Cretan dialect from SMG are briefly presented. In Section 3, we present the linguistic resources we used and in Section 4, we provide details about the compilation of the treebank and the handling of specific morphological and syntactic phenomena. In Section 5, we discuss the annotation method, and in Section 6 and 7, we present and comment on the models we developed. In the last three sections we present the limitations of our approach and the conclusions we reached.

[1] `https://github.com/UniversalDependencies/UD_Greek-Cretan`

[2] `https://github.com/UniversalDependencies/UD_Greek-GUD`

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
March 3-4, 2025 ©2025 University of Tartu Library

2 The Eastern Cretan dialect and its relation with SMG

Cretan is a language variety of Modern Greek (MG) primarily spoken on the island of Crete and by the Cretan diaspora. This includes communities of Cretan descent who relocated to Hamidieh in Syria and Western Asia Minor after the 1923 population exchange between Greece and Turkey. The preservation and development of the dialect have been influenced by Crete's long-term isolation from the mainland and the island's domination by non-Greek-speaking powers such as the Arabs, Venetians, and Turks for more than nine centuries. Cretan is divided into two main dialect groups —western and eastern— based on phonological, morphological and lexical characteristics. The two groups share a lot of features that characterise the Cretan dialects. The division aligns with the island's administrative boundaries between the prefectures of Rethymno and Heraklion.

The phenomenon of the gradual decline of MG dialects in the face of SMG is observed. Beyond the social and economic reasons for the depopulation of rural areas, which are the natural speaking environments for these language varieties, efforts to preserve and reproduce them have not yet taken on a systematic character. Specifically, the dialects have not been systematized regarding their orthographic representation and are not taught.

Unlike most other MG dialects, Cretan is not endangered and remains widely used as the primary mode of communication in many parts of the island. However, as all MG dialects, it is under-resourced, in particular with regard to resources that would support its presence in the contemporary technological landscape.

Below we will mention some of the distinctive features that the Cretan dialect retains, according to the studies by Kontosopoulos (1969, 2008).

Phonological level

1. Palatalization and affrication of /k/, /g/, /x/, /ɣ/ before the phonemes /e/, /i/. The corresponding cretan allophones in the aforementioned phonetic environment are respectively: [ʧ], [dʑ], [ɕ], [ʑ]

2. Fricativation of /t/ to /θ/ and /d/ to /ð/ before semivocalic phonemes:
   - [ta 'ma. **t**ça] → [ta 'ma. **θ**ça]
   - [ku.ve.'**d**ja.zo] → [ku.ve.'**ð**ja.zo]

3. Realization of the clusters <μπ>, <ντ>, <γκ> as voiced plosive phonemes [b], [d], [g] without the nasal element in any position.

4. Development of the euphonic sounds [e], [n], and [j] to avoid hiatus in cases of word coarticulation (see also Section 4.3):
   - <τον βάνω>, [ton 'va. no] → <τονε βάνω>, [ton**e** 'va. no], 'I put him'
   - <ούτε όμπιασε>, ['u.te 'o.bja.se] → <ούτε νόμπιασε>, ['u.te '**n**o.bja.se], 'nor did it swell'
   - <η αφορμή>, [i a.for.'mi] → <η γιαφορμή> [i **j**a.for.'mi], 'the occasion'

5. Elision of the final /n/ in the genitive plural:
   - [to**n** spit.'ço**n**] → [t**o** spiθ.'ç**o**]

6. Stress on the fourth syllable from the end as opposed to SMG where the so-called 'law of three syllables' demands that no word carries a stress beyond the third syllable from the end.
   - [ef. **'ta**.ksa.me.ne]
   - [e.**'fi**.ɣa.me.ne]

7. Development of prothetic /a/ or /o/.
   - [**a**.mo.na.'xos]
   - [**o**ɣ.'λi.ɣo.ra]

Morphological level

1. Use of different article forms than SMG.
   - <τση>, *the*.GEN.FEM.SG, [tsi] instead of SMG <της>, [tis]
   - <τσι>, *the*.ACC.F.PL, [tsi] instead of SMG <τις>,[tis]
   - <τσοι>, *the*.ACC.MASC.PL, [tsi] instead of SMG <τους>, [tus]

2. Inflection suffix <-ομε >instead of SMG <-ουμε> in the first person plural of verbs in active voice:
   - <έχομε>, ['e.xo.me], instead of SMG <έχουμε>, ['e.xu.me], 'we have'
   - <κάνομε>, ['ka.no.me], instead of SMG <κάνουμε>, ['ka.nu.me], 'we do'

3. Several masculine nouns in <-ος> are used as neuter nouns:
    - <το λαός>, [to la.'os].neuter, instead of SMG <ο λαός>, [o la.'os].masc, 'the people'
    - <το πλούτος>, [to 'plu.tos].neuter, instead of SMG <ο πλούτος>, [o 'plu.tos].masc, 'the wealth'
4. Extension of forms of demonstrative pronouns:
    - <τουτοσές>, [tu.to.'ses], instead of SMG <τούτος>, ['tu.tos], 'this.NOM.MASC.SING'
    - <εκειοσές>, [e.cio.'ses], instead of SMG <εκείνος>, [e.'ci.nos], 'that.NOM.MASC.SING'
5. Verbs ending in <-εύγω> instead of <-εύω>:
    - <χορεύγω>, [xo.'re.v**ɣ**o] instead of SMG <χορεύω>, [xo.'re.vo], 'I dance'
6. In both SMG and Cretan, the future tense is expressed periphrastically. In contrast to SMG, which employs one auxiliary element, Cretan uses two: the subordinating conjunction <να> and the verb <θέλει>. The verb <θέλει> can appear in two forms: either in its indeclinable form, which is considered the infinitive form of <θέλω> ('I want'), or in finite form, but only for the singular (Chairetakis, 2020), e.g.,
    - infinitive form: <να πας θέλει>, [na 'pas 'θe.ʎi], 'You will go'
    - finite form: <να πας θες> [na 'pas 'θes], 'You will go'
    - instead of SMG <θα πας>, [θa 'pas], 'You will go'
7. The use of <ξ>, [ks] as a perfective aspect marker:
    - <τραγούδηξα>, [tra.'ɣu.ði.**ks**a] instead of SMG <τραγούδησα>, [tra.'ɣu.ði.sa], 'I sang'

Lexicological level In the Cretan dialect, a wealth of words is attested that are not found in SMG. Most of these words are loanwords from Turkish and Venetian. The influence of each of these languages on the Cretan dialect spans four centuries, with Turkish linguistic influences being comparatively stronger due to the fact that the Turkish conquest was more recent. These loanwords are frequently used to name objects and processes related to the material culture of the people.

- <θιαμπόλι>, [θça.'bo.ʎi], 'cretan flute' < italic <fiabuolo>

- <ντελικανής> [de.ʎi.ka.'ɲis], 'the young man' < turkic <delikanli>

Some Cretan word forms are used in SMG with a different meaning.

- <κουράδι>, [ku.'ra.ði] 'flock of sheep' instead of SMG 'faeces'

- <ξανοίγω>, [ksa.'ni.ɣo], 'to see' instead of SMG 'fade out (for a colour)'

Finally, the Cretan dialect also attests to stereotypical expressions not found in SMG.

- <μια ολιά>, [mɲa o.'ʎa], Lit. one sip, 'a little'

- <δίδω των αμμαθιώ μου>, ['ði.ðo ton a.ma.'θço mu], Lit. I give to my eyes, 'I flee upset'

Syntactic level The weak pronouns that are functioning as objects are placed after the verb in contrast with the SMG that places them before the verb:

- <ρωτώ σε>, [ro.'to se] instead of SMG <σε ρωτώ>, [se ro.'to], 'I ask you'

Many verbs take objects in genitive case; in SMG the same verbs take objects in the accusative case:

- <ζηλεύγω σου>, [zi.'le.vɣo su] instead of SMG <σε ζηλεύω>, [se zi.'le.vo], 'I envy you'

3 Resources

For the compilation of this corpus, we collected 32 tapes containing material from radio broadcasts in digital format, with permission from the Audiovisual Department of the Vikelaia Municipal Library of Heraklion, Crete. The broadcasts were recorded and aired by Radio Mires, in the Messara region of Heraklion, during the period 1998-2001, totaling 958 minutes and 47 seconds. The recordings primarily consist of narratives by one speaker, Ioannis Anagnostakis, who is responsible for their composition. The material belongs to the Eastern Cretan dialect group. In terms of textual genre, the linguistic content of the broadcasts consists of folklore narratives. Out of the total volume of material collected, we utilized nine tapes. Criteria for material selection were digital clarity of speech and the representative sampling among the entire three-year period of radio recordings.

For the transcription of the recorded speech to text, the Whisper large−v2 model was utilized. At the time this process was carried out (April 2023), Whisper large−v2 returned the best results to small trials with the Cretan data. The transcriptions were edited by a linguist who is native speaker of the Eastern Cretan dialect. Given that the Cretan dialect is primarily an oral language variety, there is no standardized orthography. The general trend in the orthographic representation of Cretan is conformity with that of Standard Modern Greek. We followed that trend, in an effort to strike a balance among facilitating knowledge transfer from GUD, representing the linguistic characteristics of the dialect in the orthography and aligning with the dominant orthographic trends adopted by the dialect's native speakers. The handling of the distinctive phonological phenomena of the Cretan language variety, such as the frequent insertion of euphonic sounds and the occurrences of voicing, will be discussed below.

4 The treebank

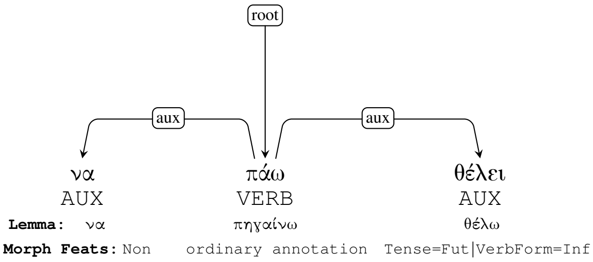
The annotation of East Cretan has relied on the UD annotation guidelines for GUD.[3] Only deviations and new constructs and forms have been documented in the guidelines for the East Cretan treebank that are listed as comments of the GUD guidelines.

4.1 Morphology

1. For the case of nouns and adjectives, which form diminutives and augmentatives, it was decided to list the basic word as the lemma, meaning the word without the diminutive or augmentative suffix, e.g., <μπεγιρ-άκι>, ‘the little horse’ has been assigned the lemma ‘μπεγίρι’ that does not contain the diminutive suffix <-άκι>.

2. As mentioned in Section 2, Morphological level 6, the (Eastern) Cretan dialect uses a distinctive periphrastic structure for the future tense, which is not found in SMG. We annotated these structures as follows:

- <να πάω θέλει>, [na 'pa.o 'θe.ʎi], ‘I will go’



3. The perfect tense is expressed, in addition to the usual SMG way, with the following structure: auxiliary verb έχω ‘have’ + passive participle (Chairetakis, 2020):

το 'χει λεομένο
(1) *it.*ACC *has said.*PARTICIPLE.ACC
‘He/She has said it’

4. All words of the Cretan dialect that appear slightly different from their SMG counterparts were assigned a lemma form that bears the dialectal linguistic characteristics:

- <βρίχνω>, ['vri.xno] instead of SMG <βρίσκω>, ['vri.sko], ‘I find’

4.2 Syntax

Because of the oral nature of the collected linguistic material, we encountered many elliptical sentences in the corpus. Copulas were often omitted as well as verb heads: in the example below the subject φτωχός is promoted as the head of the sentence.

ο φτωχός μια φουρνιά κουτσούβελα
(2) *the poor.man a bunch kids*
‘the poor man had a bunch of kids’

According to the UD guidelines, the non-promoted dependents (here: <φουρνιά>) are con-

[3] https://github.com/UniversalDependencies/docs/tree/pages-source/_el

nected with the promoted one using the special relation "orphan".

4.3 Voicing and Euphonics

Both voicing and euphonics are phenomena due to the phonetic environment but with no effect on the syntax and meaning of an utterance. In the Cretan treebank they are annotated separately.

Voicing in MG is a phonological phenomenon where, given the sequence of two words, the initial unvoiced consonant (/ts/, /t/, /p/, /k/) of the second word is voiced, e.g., /tsi/→/dzi/, /t/→/d/, /p/→/b/, /k/→/g/.

In contrast, euphonics are sounds that are added with the phonological procedure of epenthesis, in order to avoid the hiatus produced by vowel sequences, e.g., /'u.te 'om.bja.se/ → /'u.te 'n om.bja.se/ or sequences of consonants, e.g., /'an 'θe.ʎi/ > /'an e 'θe.ʎi/. In all cases, the result of the epenthesis are two open syllables of the type consonant+vowel.

Below, we first discuss the phenomena briefly and then we make a proposal for their representation in the Eastern Cretan UD treebank.

4.3.1 Euphonics in the East Cretan UD treebank

Euphonics are vowels or consonants that occur within a word or between words (3, 4) or at the end of a word (5). In Cretan (and Modern Greek in general) their function is to create open syllables and eliminate hiatuses. For instance, in Eastern Cretan, the 'γι' euphonic is used within phonological words as a hiatus breaker, so the condition for its occurrence is the particular hiatus and the existence of a (phonological) word (Kappa, 2014).

	οι	γι-άλλοι
(3)	*the*.NOM.M.PL	EUPH-*other*.NOM.M.PL
	'the others'	
	ούτε	ν-όμπιασε
(4)	*nor*.CCONJ	EUPH-*swell*.PERF.3SG.PAST
	'not did it swell'	
	κάν'	τον-ε
(5)	*do*.3SG.IMP	*I*.PRON.ACC.M.3SG-EUPH
	'do it'	

The textual encoding of euphonics is an issue. In SMG orthography, the euphonic 'e' is attached to the preceding word (5). We had to define additional guidelines for Cretan. We did not encode them as orthographic words because they are single sounds and have no morphosyntactic impact on the utterance. In all cases, we have attached euphonics to the word that precedes or follows them, on the condition that open syllables are created:

- παιδιών-ε, *child*.PL.GEN-EUPH
- τον-ε, I.PRON.ACC.M.3SG-EUPH
- αν-ε, *if*-EUPH
- γι-άλλοι, EUPH-*other*.NOM.M.PL.

The euphonic 'γι' (3) is encoded with two characters because the Greek alphabet does not have a dedicated character for the sound [ʝ]. We could probably use non-Greek characters for them, for instance in (3) we could use 'j'. As explained in Section 3, we retain the Greek alphabet, which is also used by the speakers of the dialect.

4.3.2 Voicing in the orthography of SMG

SMG orthography uses the following conventions for encoding voicing; these conventions are adopted by most authors who write in other Greek varieties:

1. A '-ν' /n/ is added to the end of an article with the features CASE=ACC|GENDER=MASC|FEM when it is followed by another word whose first consonant is voiced in this context but unvoiced in other contexts.

(6) την πατρίδα */'ti ba.'tri.ða/*
the-ACC.FEM.SG *homeland*-ACC.FEM.SG
'the homeland'

2. In all other [word1 word2] sequences where word2 appears with a voiced first consonant (while occurrences of the word with a non-voiced first consonant are attested in other contexts of the same dialect) and word1 is not independently found with a final '-ν', voicing is represented on word2. In the example below the word 'μύτη' is in the nominative case that does not accept a final -ν with this type of nouns.

(7) η μύτη τζη */i 'mi.ti dzi/*
the-NOM.FEM.SG *nose*-NOM.FEM.SG
her-GEN.FEM.SG
'her nose'

The Greek alphabet has no single letter corresponding to the sounds /dz/, /d/ and /b/ so Modern Greek orthography represents them with two characters (τζ, ντ and μπ respectively).

4.3.3 Annotation of euphonics and voicing in the Eastern Cretan treebank

We use the `MSeg|MGloss` representation and the label `euphonic` for annotating euphonics in the Cretan treebank. With the MSeg annotation schema we are able to isolate the euphonic segments from the rest of the word and handle each part as a separate token.

```
γιάλλοι DET
MSeg=γι-άλλοι|MGloss=euphonic-others
```

We cannot resort to the MSeg|MGloss representation in order to annotate voicing because the results of voicing cannot be separated from the rest of the word, e.g., in the form of an affix. For instance, 'τζη' (/dzi/) cannot be divided as 'τζ-η' because '-η'(/i/) is not a word with the same morphosyntactic features as 'τζη' (recall that voicing has no syntactic or semantic effect). Instead, we define a feature that differentiates the unvoiced form from the voiced one. This is a new MISC feature of the Cretan treebank called `Voicing` with values `Voiced` and `Unvoiced`.

```
τζη PRON ... Case=Gen ...
Voicing=Voiced
```

Voicing characterises all MG dialects, including SMG, in the environment of the Accusative case and contributes to the distinction between Accusative and Nominative case. We do not annotate this type of expected voicing. However, sometimes the voiced version of a word is also used in environments where no voicing is expected, suggesting that the voiced version is lexicalised and co-exists and competes with the unvoiced one, e.g., (dialect of the island of Lemnos) 'η μπατρίδα' (/i ba.'tri.ða/) coexists with 'η πατρίδα' (/i pa.tri.ða/), both in the nominative case, singular number. The question is which lemma should be assigned to each of the two versions. We assign the unvoiced version of the lemma to both versions; in addition, the voiced form is assigned the feature-value pair `Voicing=Voiced`. Our choice of the unvoiced version contributes to the consistency of the annotation and to knowledge transfer from SMG to the dialects because SMG usually has the unvoiced version of the lemma, if it has this lemma at all.

In the example Ψυχοπόνεσέ ντον**ε** το παπαδάκι, Lit. felt.sorry him the altar.boy, 'The altar boy felt sorry for him' both unexpected voicing and euphonics are used because the verb form 'Ψυχοπόνεσε' never appears with a final -ν:

```
ντονε    Voicing=Voiced|MSeg=ντον-ε
|MGloss=him-euphonic
```

5 Active annotation

To annotate the Cretan treebank we used active annotation (Vlachos, 2006) implemented in 6 iterative cycles. The first set of 40 unlabelled Cretan samples was annotated with a model trained on GUD, which represents SMG. In each cycle, the annotator edited 40 samples from the output, split in 30 for the training set and 10 for the development set, added them to the existing training and development sets and the model was retrained on the revised data. For the test set, 30 manually annotated samples were used. All samples were randomly selected, with the only criterion being that each sample contained more than five tokens to avoid sentences with minimal linguistic information.

		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**	448	903	1395	1880	2398	2976
	Dev	175	348	504	728	939	1129
	Test	523	523	523	523	523	523

Table 1: East Cretan sentences and tokens per round.

During this first attempt to develop a UD treebank of Cretan, the annotation guidelines were developed as research progressed. Any revisions to the annotation guidelines were implemented across the entire training, development and test sets.

6 Including euphonics and voicing in the models

To introduce euphonics in the model, we process the input CoNNLU representations of sentences by transferring information from the MSeg annotation (column 10 of the CoNNLU format) on the LEMMA, UPOS and XPOS columns and training the model on the modified treebank[4]. The original word's UPOS and DEPREL tags are inherited by the piece of the token that remains after the euphonic is removed (the 'original word') and the

[4] The script for the transformation of the input CONLLU files can be found at `https://anonymous.4open.science/r/euphonics-7F98`

euphonic is represented as a separate token with the new XPOS/UPOS tag EUPH that depends on the original word with the new dependency relation 'euph'. The XPOS tag EUPH and the dependency 'euph' have been defined for the purposes of the Eastern Cretan treebank and are used in ongoing work on other varieties of Modern Greek, including SMG. The tag EUPH was introduced to the UPOS column to satisfy a requirement of the processing tool.

We did not use the UPOS X because euphonics can hardly be called words, at least in the sense of self-standing linguistic entities that combine a form with some type of semantic contribution. But even if euphonics were considered a type of word, again the X UPOS would not be a choice because euphonics are clearly parts of the language varieties we study and play a well defined role. These two facts contrast with the UD annotation guidelines about UPOS X: "(UPOS) X is discouraged for words that clearly belong to the language, even if they are idiosyncratic in form or distribution and thus do not neatly fit into other syntactic categories." Neither did we use the UPOS PART(icle). UD define particles as "function words that must be associated with another word or phrase to impart meaning and that do not satisfy definitions of other universal parts of speech". Euphonics do not impart any meaning at all. Finally, we did not consider them clitics as suggested by one of our reviewers, because clitics do not define a POS of their own and we have argued that euphonics cannot be assigned any of the POS available in UD.

The output of the model that knows about euphonics cannot be used in the active annotation cycle because its form differs from the form of the UD treebank. This output contains a modified XPOS column (which may not be a problem), no information on the MISC column about voicing and euphonics while the UPOS column is modified with an extra tag. We have not applied active annotation on voicing and euphonics but for future needs, since the phenomena occur in many MG dialects, we will have to post-edit the model's output and make it comply with the form of the annotation of the input.

A complete example is included below featuring the word <ντονε> that contains the voiced masculine, singular, accusative form of the personal pronoun <εγώ> 'I' with the euphonic 'ε' /e/ attached to it. Similarly, the feature-value pair "Voicing=Voiced" is added to the list of morphological features.

```
2  ντονε  εγώ  PRON  Case=Acc...|Gender=Masc|Number=Sing|Person=3|PronType=Prs

1  obj  _  Voicing=Voiced|MSeg=ντον-ε|MGloss=him-euphonic

2-3  ντονε  _    _  _  _  _    Voicing=Voiced|MSeg=ντον-ε|MGloss=him-euphonic

2  τον  εγώ  PRON  _

Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs|Voicing=Voiced    _  1  obj    _

3    ε    ε    _    EUPH    _    2    euph    _    _
```

7 Models

For the experiments we used the open source Stanza package (Qi et al., 2020). The embeddings for all experiments were generated by combining the GUD treebank with the Cretan corpus. We used two different settings for the treebanks: GUD plus the Eastern Cretan data (henceforth GUD+Cretan treebank) that increased at each round by 40 samples (30 in the training set and 10 in the development set) and, the Eastern Cretan samples only that increased exactly in parallel with the GUD+Cretan treebank. In both settings, we finetuned the Greek BERT model (Koutsikakis et al., 2020) for the tasks of PoS tagging and dependency parsing.

Metric	R1	R2	R3	R4	R5	R6
UPOS	80.12	83.57	85.80	88.64	87.83	89.25
XPOS	79.31	78.09	80.12	82.56	82.35	83.37
UFeats	55.38	63.49	72.82	77.08	76.47	78.70
AllTags	48.68	53.75	59.84	65.92	65.92	68.15
Lemmas	66.53	73.02	77.28	80.12	81.74	81.34
UAS	65.31	73.02	75.25	78.09	75.25	78.50
LAS	45.84	58.22	63.29	65.92	65.52	67.75
CLAS	32.59	46.54	51.47	55.76	55.22	59.33
MLAS	10.37	20.00	30.51	36.06	33.58	40.67
BLEX	14.81	29.23	34.19	40.15	40.67	43.28
ELAS	0.00	0.00	0.00	0.00	0.00	0.00
EULAS	0.00	0.00	0.00	0.00	0.00	0.00

Table 2: Accuracy scores across rounds: East Cretan treebank. R=Round.

8 Discussion

The results are depicted in Table 2 and 3, and were obtained using the pre-tokenized text option provided by Stanza. Figure 1 shows that the model trained on the GUD+Eastern Cretan treebank consistently outperforms the model trained on the Eastern Cretan treebank across all rounds and tasks. Therefore, GUD was an excellent resource for knowledge transfer from SMG to Eastern Cretan models. This result must have been

Metric	R1	R2	R3	R4	R5	R6
UPOS	89.25	92.29	92.49	92.09	92.90	92.90
XPOS	89.25	89.45	89.66	89.45	88.84	89.45
UFeats	83.77	85.40	84.99	87.22	85.60	85.60
AllTags	76.27	78.50	77.28	78.30	77.28	77.48
Lemmas	83.98	87.42	87.83	87.42	89.05	88.44
UAS	84.58	83.98	85.40	87.02	87.02	85.40
LAS	73.83	74.85	77.08	76.88	78.50	78.30
CLAS	66.54	68.56	70.30	70.57	71.64	72.76
MLAS	51.88	55.30	57.14	56.60	55.97	57.09
BLEX	53.76	57.58	60.90	58.87	61.19	61.57
ELAS	0.00	0.00	0.00	0.00	0.00	0.00
EULAS	0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Accuracy scores across rounds: GUD+Cretan treebank. R=Round.

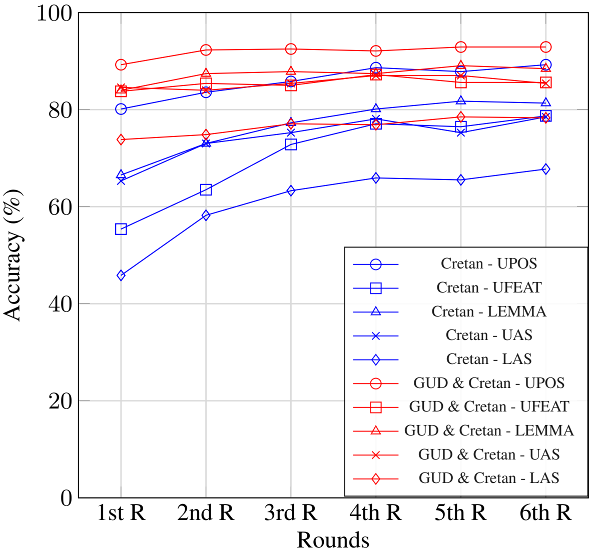


Figure 1: Accuracy scores for the GUD+Cretan & Cretan datasets.

corroborated by the fact that the texts of both language varieties are written with the same orthographic conventions.

After the 4th cycle the GUD+Eastern Cretan models tend to decrease or stabilize across all accuracy measures, while the Cretan-only models still improve. This suggests that after the 4th cycle information from GUD added noise. Therefore, 4 or 5 cycles with GUD were enough for successful knowledge transfer for this variety of Greek and the set up we used (40 new samples at each cycle).

In a 7th training round, we applied on the Eastern Cretan treebank the transformation that introduces euphonics and voicing in the models (see Section 4.3). The results are shown on Table 4. The East Cretan treebank returns still improving results. The test set contained 10 instances of these phenomena and the training and development sets 67 instances. The model achieved a 100% Recall and Precision, probably because the forms of voicing and euphonics are very distinctive.

Metric	Accuracy
UPOS	89.45
XPOS	85.27
UFeats	78.00
AllTags	68.36
Lemmas	82.36
UAS	78.73
LAS	69.27
CLAS	59.80
MLAS	39.86
BLEX	44.59

Table 4: Accuracy scores for the 7th Round that includes EUPHONICS-VOICING. East Cretan treebank.

9 Conclusion

We have developed the first UD treebank of Eastern Cretan, which is a living, non standardised variety of Modern Greek. We have attempted to model phenomena new to UD guidelines such as voicing and euphonics; these phenomena abide in the dialects of Modern Greek. The successful active annotation procedure and the knowledge transfer from the GUD treebank of SMG to the models of Eastern Cretan suggests that a similar pipeline can facilitate the modelling of other varieties of MG, starting from Western Cretan. We hope that this treebank will support future efforts to provide additional digital material from more native speakers, the textual legacy of East Cretan as well as other, linguistically challenging, dialects of Modern Greek.

10 Limitations

A weak point of our approach is that we have relied on data from one speaker only. However, this was the first time that the full pipeline from oral data to annotated UD treebanks was studied for a Greek dialect (here we report on the work after the Speech-to-Text step). We are currently collecting data from more speakers from the same area (the Heraklion prefecture) and aim to enrich the Cretan UD treebank soon. The orthography we used to transcribe the East Cretan oral material is identical to the orthography used for SMG and has probably facilitated knowledge transfer from the treebank of SMG; however, as it has been mentioned, this is the orthography preferred by many speakers of Cretan (and many other dialects of MG). Future work may try to exploit the existing textual legacy of the Cretan dialect that occasionally adopts an orthography partially different from the orthography of SMG. The exploitation of non-standardised textual legacy, especially for under-resourced language varieties, for model development is a well-known problem (Plank, 2016). These said, we would like to add that we relied a lot on the GUD guidelines in order to develop the Eastern Cretan UD guidelines and, while doing so, we did not have to suppress or alter information particular to this dialect; this may be an indication of the proximity of these two varieties of Modern Greek and of the relatively little bias that SMG exerted on the models of Eastern Cretan.

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.

References

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

George Chairetakis. 2020. *The morphology of the Cretan dialect: Inflection and derivation.* Phd thesis, University of Patras.

Ioanna Kappa. 2014. Epenthetic consonants in the western cretan dialect. In G. Kotzoglou et al., editors, *Selected Papers of the 11th International Conference on Greek Linguistics*, pages 674–688. University of the Aegean, Rhodes.

Nikolaos G. Kontosopoulos. 1969. *Linguistic and geographic investigations of the Cretan Dialect, [Γλωσσογεωγραφικαί διερευνήσεις εις την Κρητικήν διάλεκτον]*, 1st ed. edition. Graphic Arts Efstathios Papoulias, Athens.

Nikolaos G. Kontosopoulos. 2008. *Dialects and idiolects of Modern Greek*, 5th ed. edition. Grigoris, Athens.

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, page 110–117, New York, NY, USA. Association for Computing Machinery.

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

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in nlp.

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.

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