UniDive

3nd General Meeting

HUN-REN Hungarian Research Centre for Linguistics,
Budapest, Hungary, 29-30 January 2025
https://unidive.lisn.upsaclay.fr/

WG1, WG2, WG3

GUD: a new Standard Modern Greek treebank enriched with VMWE annotations

Stella Markantonatou[1,2],Vivian Stamou[2], Stavros Bompolas[2], Katerina Anastasopoulou [3],
Iriana Vasileiadi-Linardaki [3], Konstantinos Diamantopoulos [3], Yannis Kazos [4,2]
[1][Institute for Language and Speech Processing, [2]Archimedes]/Athena RC, Greece
[3]National and Kapodistrian University of Athens (NKUA), Greece
[4]National Technical University of Athens (NTUA), Greece

Introduction

- **UD_Greek-GUD: Universal Dependencies (UD).v2 treebank of Standard Modern Greek (SMG)**
- The **first UD treebank** for SMG with annotations for **Verbal Multiword Expressions (VMWEs)**

Key Features

As compared to previous UD treebanks of SMG, GUD:

- Adheres closer to the UD.v2 morphological and syntactic guidelines
- Annotates more phenomena, including diminutives and augmentatives, adjective degrees and **Verbal MWEs**

Experiments

- **Data**:
  - ✓ GUD: 1.807 sentences from fiction texts.
  - ✓ 723 sentences with 100 VMWEs from IDION VMWE database (https://search.idion.athenarc.gr)
- **Stanza models** trained with GUD + 723/500/300 VMWE sentences and 723 VMWE sentences

Annotation and Processing of Verbal MWEs

Annotation of VMWEs on column 10 (C10) of CONLLU*

```
# text = « Δεν θα βάζω για πάντα πλάτη » (I won't put forever my.back/"I won't keep backing this up forever.")
# sent_id = output9_91
1	«	«	PUNCT	OPUNCT	_	4	punct	_	PunctType=Quot
2	Δεν	δεν	PART	OPUNCT	Polarity=Neg	4	advmod	_	_
3	θα	θα	AUX	PtFu	Tense=Fut	4	aux	_	_
4	βάζω	βάζω	VERB	VbMn	Aspect=Imp|Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act	0	root	_	mwe=1:VID
5	για	για	ADP	AsPpSp	_	6	case	_	None=Yes
6	πάντα	πάντα	DET	AdBa	_	4	obl	_	None=Yes
7	πλάτη	πλάτη	NOUN	NoCm	Case=Acc|Gender=Fem|Number=Sing	4	obj	_	mwe=1
8	»	»	PUNCT	NoCm	_	4	punct	_	PunctType=Quot
```

Transferring VMWE annotation to the Deprel CONLLU column (C8)

```
# text = « Δεν θα βάζω για πάντα πλάτη » (I won't put forever my.back/"I won't keep backing this up forever.")
# sent_id = output9_91
1	«	«	PUNCT	OPUNCT	_	4	punct	_	PunctType=Quot
2	Δεν	δεν	PART	OPUNCT	Polarity=Neg	4	advmod	_	_
3	θα	θα	AUX	PtFu	Tense=Fut	4	aux	_	_
4	βάζω	βάζω	VERB	VbMn	Aspect=Imp|Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin|Voice=Act	0	root:vid	_	mwe=1:VID
5	για	για	ADP	AsPpSp	_	6	case	_	None=Yes
6	πάντα	πάντα	DET	AdBa	_	4	obl	_	None=Yes
7	πλάτη	πλάτη	NOUN	NoCm	Case=Acc|Gender=Fem|Number=Sing	4	obj:vid	_	mwe=1
8	»	»	PUNCT	NoCm	_	4	punct	_	PunctType=Quot
```

Application of evaluation metrics

```
# text = Ο πήχυς για φέτος έχει ανέβει πολύ ψηλά (The bar for this.year has risen very high/"The bar has risen very high this year.")
# sent_id = output6_106
1	Ο	ο	DET	AtDf	Case=Nom|Definite=Def|Gender=Masc|Number=Sing|PronType=Art	2	det:vid	_	mwe=1
2	πήχυς	πήχης	NOUN	NoCm	Case=Nom|Gender=Masc|Number=Sing	6	obj:vid	_	mwe=1
3	για	για	ADP	AsPpSp	_	6	mark	_	None=Yes
4	φέτος	φέτος	ADV	AdBa	_	3	fixed	_	_
5	έχει	έχω	AUX	VbMn	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act	6	aux	_	_
6	ανέβει	ανέβω	VERB	VbMn	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|VerbForm=Fin|Voice=Act	0	root	_	mwe=1:VID
7	πολύ	πολύ	ADV	AdBa	_	8	advmod	_	None=Yes
8	ψηλά	ψηλά	ADV	AdBa	_	6	advmod	_	_
```

Per-token recall: 2(TP)/(2(TP)+1(FN) = 2/3 = 0.66

```
# text = Δεν βγήκε ποτέ από το μυαλό μου και ούτε πρόκειται . (It never got.3.SING out of my mind and it will not/"It never left my mind and it never will.")
# sent_id = output6_1
1	Δεν	δεν	PART	PtNg	Polarity=Neg	2	advmod	_	_
2	βγήκε	βγαίνω	VERB	VbMn	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act	0	root:vid	_	mwe=1:VID
3	ποτέ	ποτέ	ADV	AdBa	PronType=Neg	2	advmod	_	None=Yes
4	από	από	ADP	AsPpSp	_	6	case:vid	_	mwe=1|None=Yes
5	το	ο	DET	AtDf	Case=Acc|Definite=Def|Gender=Neut|Number=Sing|PronType=Art	6	det:vid	_	mwe=1
6	μυαλό	μυαλό	NOUN	NoCm	Case=Acc|Gender=Neut|Number=Sing	2	obl:vid	_	mwe=1
7	μου	εγώ	PRON	PnPo	Case=Gen|Number=Sing|Person=1|Poss=Yes|PronType=Prs	6	nmod	_	_
8	και	και	CCONJ	CjCo	_	10	cc	_	None=Yes
9	ούτε	ούτε	CCONJ	CjCo	_	10	cc	_	None=Yes
10	πρόκειται	πρόκειται	VERB	VbMn	Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Pass	2	conj	_	_
11	.	.	PUNCT	PTERMP	_	2	punct	_	PunctType=Peri
```

Per-token/Per-VMWE Recall = 1, Per-token Precision = 1

Results

Setting	Lemma	UPOS	UFEATS	UAS	LAS
GUD+723	90.99	94.78	87.18	88.03	81.62
GUD+500	90.99	94.97	87.80	87.94	82.27
GUD+300	90.23	94.69	86.93	88.03	81.25
723	90.12	94.01	86.39	86.67	78.94

Evaluation → Definitions:

- ➢ TP when a correct annotation occurs on C8 and on C10
- ➢ FP when a correct annotation occurs on C8 but not C10
- ➢ FN when a correct annotation does not occur on C8 but occurs on C10

- Recall (R=TP/(TP+FN) and Precision (P=TP/TP+FP) are measured:
- 1. Per-token
- 2. Per-VMWE
- *Per-VMWE precision: undefinable

	Per-token	Per-VMWE
Precision	0.84	Undefinable
Recall	0.94	0.53

Applications

- **Cross-dialectal knowledge transfer** for Modern Greek dialects
- **Enhancement of downstream NLP tasks**,
- e.g., **offensive language detection**

Future Work

- Further investigate VMWE evaluation issues
- Develop the full suite of tools to support VMWE active annotation
- Experiment with larger datasets/more MWE types

**Annotation done in the spirit of Savary et al. (2023)*

{stiliani.markantonatou, vivianstamou, stavros.bompolas, katerinaanastasopoulou8, hrvaslin, kostasdiama, kazosj} @gmail.com

cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union