Universal Dependencies for Greek

Prokopis Prokopidis
Institute for Language and
Speech Processing
Athena Research Center
Athens, Greece
prokopis@ilsp.gr

Haris Papageorgiou
Institute for Language and
Speech Processing
Athena Research Center
Athens, Greece
xaris@ilsp.gr

Abstract

This paper describes work towards the harmonization of the Greek Dependency Treebank with the Universal Dependencies v2 standard, and the extension of the treebank with enhanced dependencies. Experiments with the latest version of the UD_Greek resource have led to 88.94/87.66 LAS on gold/automatic POS, morphological features and lemmas.

1 Introduction

The Universal Dependencies (Nivre et al., 2016) community effort has led to the development and collection of a large number of treebanks adhering to common and extendible annotation guidelines. These guidelines aim to ease the annotation process and improve the accuracy of parsers and downstream NLP applications in generating useful and linguistically sound representations.

Greek is represented in the UD effort with UD_Greek[1]. In this paper, we provide more details on the annotated resource in section 2 and its conversion to the UD standard. In section 3 we discuss ongoing work for extending GDT with a subset of the enhanced dependencies proposed by Schuster and Manning (2016). Section 4 presents experiments with parsers trained on the different-sized versions of the resource and on manually/automatically annotated morphology and lemmas.

2 The Greek Dependency Treebank and its conversion to UD

UD_Greek is derived from the Greek Dependency Treebank (GDT, Prokopidis et al. (2005)), a resource developed and maintained by researchers at the Institute for Language and Speech Processing[2]. Although the conversion and harmonization to UD is work in progress since UD v1.1, the Greek dataset in the v2.0 release was the first one that involved extensive manual validation and correction of labeled dependencies generated from the originial annotations.

The original annotation scheme used for the annotation of the resource was based on an adaptation of the guidelines for the Prague Dependency Treebank (Böhmová et al., 2003). Trees in the original data were headed by words bearing, in most cases, the `Pred` relation. Coordinating conjunctions and apposition markers headed participating tokens in relevant constructions. Prepositions and subordinating conjunctions acted as mediators between verbs/nouns and their phrasal and clausal dependents. The tagset used for the morphology layer in the original resource contained 584 combinations of basic POS tags and features that capture the rich morphology of the Greek language. As an example, the full tag `AjBaMaSgNm` for a word like *ταραχώδης/turbulent* denotes an adjective of basic degree, masculine gender, singular number and nominative case. The three last features are also used for nouns, articles, pronouns, and passive participles. Verb tags include features for tense and aspect, while articles are distinguished for definiteness. The top tree in Figure 1 presents an example of a dependency tree with basic POS tags.

Annotated documents in GDT are stored in XML files that integrate annotations for semantic roles and events. A procedure based on software described in Zeman et al. (2014) was used for rehanging nodes and changing labels in these files, so that annotations beyond the syntactic level were kept intact. The original heads and labels of the original annotation effort were stored as attributes of the XML elements corre-

[1] https://github.com/UniversalDependencies/UD_Greek

[2] http://www.ilsp.gr

Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 102–106,
Gothenburg, Sweden, 22 May 2017.

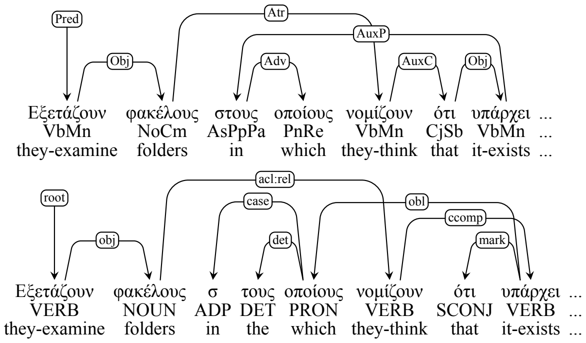
Pred
Obj
Atr
AuxP
Adv
AuxC
Obj
Εξετάζουν φακέλους στους οποίους νομίζουν ότι υπάρχει ...
VbMn NoCm AsPpPa PnRe VbMn CjSb VbMn ...
they-examine folders in which they-think that it-exists ...
root
obj
acl:rel
case
det
obl
ccomp
mark
Εξετάζουν φακέλους σ τους οποίους νομίζουν ότι υπάρχει ...
VERB NOUN ADP DET PRON VERB SCONJ VERB ...
they-examine folders in the which they-think that it-exists ...

Figure 1: Annotation of a sentence fragment with a non-projective arc, according to the the original (top) and the current representation.

sponding to tree nodes. In view of the UD v2.0 release, the results of the automatic conversion were manually examined and corrected, in an effort focusing on errors related to core arguments of content words; heads of the copula; nodes participating in coordinating conjunctions; non-projective dependencies; and multi-word expressions acting as clause-introductory markers. Another difference to previous versions of the resource concerned preposition-article combinations (e.g. *στις/in-the/case/Prep3rdPersFemPlurAcc*). These multi-word tokens were split into words that were assigned morphological information and syntactic heads. The second tree in Figure 1 is an example involving several of the conversions mentioned above. The `acl:rel` relation in the example is a language specific extension used for the annotation of relative clauses. Another extension is `obl:arg`, which in the current version of the resource is used for prepositional arguments that cliticize and are described by many Greek grammars (e.g. Holton et al. (1997)) as indirect objects.

GDT is regularly updated with new material from different genres, and its current version comprises 178207/7417 tokens/sentences. The data in UD_Greek have also increased since v1.1 and currently[3] consist of 63441/2521 tokens/sentences. UD_Greek data are derived from annotated texts that are in the public domain, including Wikinews articles and European Parliament sessions. For the UD v2.* versions sentences are not shuffled and documents are not split across train/dev/test partitions. There are 10927/5894/6375 types/lemmas/hapax legomena in the resource, while the average sentence length is 25.17 tokens. Non-projective trees (12.38% of all sentences) allow for the intuitive representations of long-distance dependencies and non-configurational structures common in languages with flexible word order. The relatively free word order of Greek can also be inferred when examining typical head-dependent structures in the resource. Although determiners and adjectives almost always precede their nominal heads, the situation is different for arguments of verbs. Of the 2776 explicit subjects in UD_Greek, 32.89% occur to the right of their parent, while the percentage rises to 46.12% for subjects of verbs heading dependent clauses. The situation is more straightforward for non-pronominal objects, of which only 2.66% occur to the left of their head. Of those subjects and objects appearing in “non-canonical” positions, 21.58% and 29.63%, respectively, are of neuter gender. This fact can pose problems to parsing, since the case of nominative and accusative neuter homographs is particularly difficult to disambiguate, especially due to the fact that articles and adjectives often preceding them (e.g. *το/the κόκκινο/red βιβλίο/book*) are also invariant for these two case values.

3 Enhanced dependencies

A recent addition to the resource is semi-automatic annotation for the enhanced dependencies proposed by Schuster and Manning (2016). We have initially focused on a subset of these dependencies involving coordination and control structures.

For coordination structures, we have exploited the fact that conjunctions headed these constructions in the previous representation and that the ids of the heads of the conjuncts are still available in the current annotation files. We were thus able to convert trees like the one in Figure 2 to the enhanced dependency graph shown in the same example.

In the latest GDT version, no `ccomp/xcomp` distinction was included for Greek finite clauses that depend on verbs of obligatory subject or object control. We are currently using Lexis (Anagnostopoulou et al., 2000), a computational lexicon with syntactic and semantic information for Greek verbs, to annotate instances of these verbs with two extensions of the `xcomp` relation, `xcomp:sc` and `xcomp:oc`. These annotations allow us to gener-

[3] The dataset is available from `https://github.com/UniversalDependencies/UD_Greek/tree/dev`. The experiments described in Section 4 correspond to commit: `https://goo.gl/fhPmbN`.

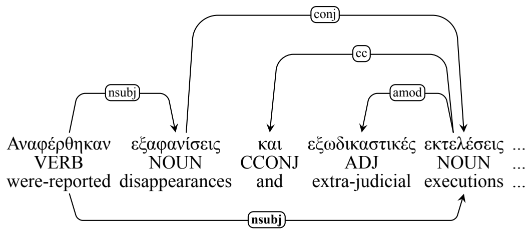
conj
cc
nsubj
amod
Αναφέρθηκαν εξαφανίσεις και εξωδικαστικές εκτελέσεις ...
VERB NOUN CCONJ ADJ NOUN ...
were-reported disappearances and extra-judicial executions ...
nsubj

Figure 2: Enhanced dependency graph for a coordination structure.

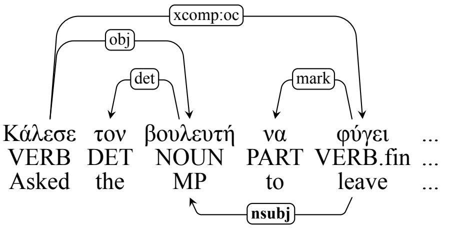
xcomp:oc
obj
det
mark
Κάλεσε τον βουλευτή να φύγει ...
VERB DET NOUN PART VERB.fin ...
Asked the MP to leave ...
nsubj

Figure 3: Enhanced dependency graph for an object control structure.

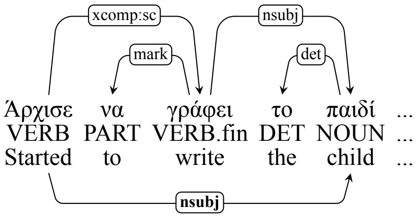
xcomp:sc
nsubj
mark
det
Άρχισε να γράφει το παιδί ...
VERB PART VERB.fin DET NOUN ...
Started to write the child ...
nsubj

Figure 4: Enhanced dependency graph for a backward subject control structure.

ate graphs like the object control one in Figure 3 and also the backward, subject control one in Figure 4.

4 Parsing experiments with the UD representation

In this section, we report on experiments with the current version of UD_Greek and its GDT superset. In all experiments reported below, we remove the annotations related to the enhanced dependencies described in Section 3, since they do not yet cover the whole resource. We examined parsing accuracy in scenarios involving manual and automatic annotations for morphology and lemmas. In the latter setting, POS tagging is conducted with a tagger (Papageorgiou et al., 2000) with an accuracy of 97.49 when only basic POS is considered. When all features (including, for example, gender and case for nouns, and aspect and tense for verbs) are taken into account, the tagger's accuracy drops to 92.54. As an indication of the relatively rich morphology of Greek, the tags/word ratio in the tagger's lexicon is 1.82. Tags for a word typically differ in only one or two features like case and gender for adjectives. However, distinct basic parts of speech (e.g. Vb/No) is also a possibility. Following POS tagging, a lemmatizer retrieves lemmas from a lexicon of 2M different entries. When a token under examination is associated in the lexicon with two or more lemmas, the lemmatizer uses information from the POS tags for disambiguation. For example, the token+POS input *εξετάσεις*/Vb guides the lemmatizer to retrieve the lemma *εξετάζω (examine)*, while the lemma *εξέταση (examination)* is returned for *εξετάσεις*/No.

We use the graph-based Mateparser (Bohnet, 2010) and the transition-based version of Bistparser (Kiperwasser and Goldberg, 2016). For the latter, we projectivise datasets by lifting non-projective arcs (Nivre and Nilsson, 2005), and we use 100-dimensional word-embeddings obtained with the fastText library (Bojanowski et al., 2016) from a 350M token corpus.

Table 1 summarizes the results. Using the whole resource with gold POS, morphological features and lemmas (GDT-MPL), the Mate and Bist LAS are 90.29/89.36, respectively. The difference between the two parsers on input with automatic annotations (GDT-APL) is smaller (88.82/88.36). When comparing the performance of both parsers on the different size datasets, the LAS improvement on the bigger dataset is more evident for Bistparser, with a 1.97% increase from the APL setting with the UD_Greek dataset (UD-APL). For both parsers, best LAS is observed for small sentences of 5-15 tokens long, with the accuracy remaining relatively stable for sentences of 15-25 tokens (cf. Fig. 5).

In related work, Prokopidis and Papageorgiou (2014) trained the Mateparser on a version of GDT of 130K tokens annotated according to the PDT-compatible representation, and reported a LAS of 80.16 on manually validated POS tags and lemmas. The automatically converted UD_Greek v1.* (59156/2411 tokens/sentences) has been used in evaluations for multilingual parsing, including the experiments by Straka et al. (2016), where 79.4/76.7 LAS were reported for manual/automatic POS tags, respectively.

	UD-MPL		UD-APL		GDT-MPL		GDT-APL	
	Bist	Mate	Bist	Mate	Bist	Mate	Bist	Mate
LAS	86.47	**88.94**	86.39	**87.66**	89.36	**90.29**	88.36	**88.82**
UAS	89.29	**90.78**	89.49	**90.49**	91.49	**92.01**	91.06	**91.38**
LACC	92.73	**93.68**	**92.45**	92.22	94.55	**94.70**	**93.56**	93.33

Table 1: Results from parsing UD_Greek and GDT with the Bist- and Mate parsers. UD_Greek contains 63K tokens, a subset of GDT's 178K tokens. (M/A)PL suffixes refer to training and testing on gold and automatic POS, morphological features and lemmas, respectively. All scores are calculated with punctuation excluded, on a test partition containing circa 10% of the tokens of each dataset.

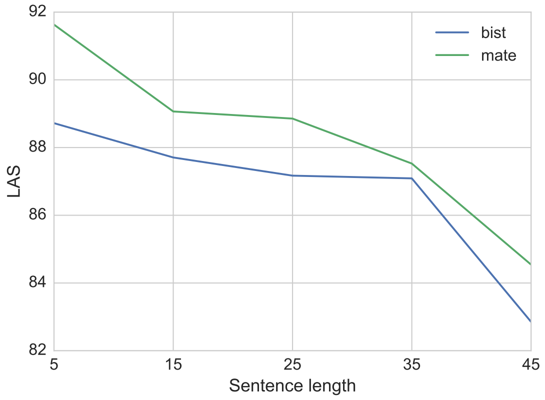

Figure 5: LAS relative to sentence length in the UD-APL setting.

5 Conclusions

We presented work for the harmonization of the syntactic trees in the Greek Dependency Treebank to the UD v.2 standard. We also discussed how we exploited previous annotations and a lexical resource to generate enhanced dependencies for the treebank. Finally, we reported a LAS of 88.94 for UD_Greek, by training the Mateparser on gold POS and lemmas. A 90.29 LAS on a larger version of the resource indicates that there is still room for accuracy improvements with additional data. While training on automatically preprocessed data, we obtain LAS scores (88.82) that are relatively high for morphologically rich languages like Greek. In future work, we plan to improve the enhanced dependencies annotation and augment the UD_Greek resource with sentences involving questions and commands.

References

Anagnostopoulou, D., E. Desipri, P. Labropoulou, E. Mantzari, and M. Gavrilidou (2000). Lexis - Lexicographical Infrastructure: Systematising the Data. In: *Computational Lexicography (COMLEX 2000)*.

Böhmová, A., J. Hajič, E. Hajičová, and B. Hladká (2003). The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: *Treebanks: Building and Using Parsed Corpora*. Kluwer.

Bohnet, B. (2010). Very High Accuracy and Fast Dependency Parsing is Not a Contradiction. In: *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 89–97.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov (2016). Enriching Word Vectors with Subword Information. In: *CoRR* abs/1607.04606.

Holton, D., P. Mackridge, and I. Philippaki-Warburton (1997). Greek: A comprehensive grammar of the modern language. London and New York: Routledge.

Kiperwasser, E. and Y. Goldberg (2016). Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations. In: *Transactions of the Association for Computational Linguistics* 4, pp. 313–327.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajic, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Nivre, J. and J. Nilsson (2005). Pseudo-Projective Dependency Parsing. In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 99–106.

Papageorgiou, H., P. Prokopidis, V. Giouli, and S. Piperidis (2000). A Unified POS Tagging Architecture and its Application to Greek. In: *Proceedings of the 2nd Language Resources* *and Evaluation Conference*. European Language Resources Association, pp. 1455–1462.

Prokopidis, P., E. Desypri, M. Koutsombogera, H. Papageorgiou, and S. Piperidis (2005). Theoretical and practical issues in the construction of a Greek Dependency Treebank. In: *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories*.

Prokopidis, P. and H. Papageorgiou (2014). Experiments for Dependency Parsing of Greek. In: *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pp. 90–96.

Schuster, S. and C. D. Manning (2016). Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Straka, M., J. Hajic, and J. Straková (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Zeman, D., O. Dušek, D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, and J. Hajič (2014). HamleDT: Harmonized multi-language dependency treebank. In: *Language Resources and Evaluation* 48.4, pp. 601–637.