Theoretical and Practical Issues in the Construction of a Greek Dependency Corpus

Prokopis Prokopidis[1,2], Elina Desipri[1], Maria Koutsombogera[1],
Harris Papageorgiou[1], Stelios Piperidis[1,2]

[1] Institute for Language and Speech Processing
Artemidos 6 & Epidavrou, 15125 Maroussi, Athens
[2]National Technical University of Athens,
Iroon Polytexneiou 9, 15780 Zografou, Athens
E–mail: {prokopis,elina,mkouts,xaris,stelios}@ilsp.gr

1 Introduction

Development and testing of a large range of NLP applications presuppose corpora annotated at levels more advanced than those of part–of–speech and shallow syntax. Therefore, multi–layered annotation schemes have been designed in order to provide deeper representations of intra– and inter–sentential structure and meaning. Linguistic insights and semi–automatic processing are being combined for the generation of corpora that integrate various types of information like predicate–argument structure, coreference, pragmatic information, etc. As the connection between syntax and semantics is of particular importance in theoretical and applied NLP research, there are many attempts to describe this relation ([8]), investigate the possibility of automatic transition from one level to another, experiment on mapping between syntactic and semantic features, or even develop sets of rules that connect syntactic structures to their corresponding event types. In this paper, we present work in progress for the construction of a resource that we provisionally call *Greek Dependency Treebank*. GDT currently encompasses annotation at the level of syntax and semantics.

The rest of this paper is structured as follows: in the next section we give an overview of the data comprising our corpus and their preprocessing. We describe syntactic representation in section 3, while in section 4 we discuss the semantic layer of our corpus. Section 5 puts our goal into perspective by focusing on further work and exploitation of the resulting resource.

2 Corpus description and data preparation

The GDT corpus comprises texts that were collected in the framework of national and EU–funded research projects aiming at multilingual, multimedia information extraction. While building the annotation collection we tried to address project requirements (by selecting texts from particular domains of interest) and, at the same time, to create resources that would form the initial part of a reference corpus for Modern Greek, annotated at multiple levels. The main categories covered at this stage are manual transcripts of European parliamentary sessions, and web documents pertaining the politics, health, and travel domains. Each annotation file corresponds either to the full text of a web document or to a randomly extracted segment (30–60 sentences long in most cases) from parliamentary sessions. The total size of the currently annotated resource amounts to 70K words.

Annotators working on this collection were presented with data that had been preprocessed via an existing pipeline (Figure 1) of shallow processing tools for Greek. This infrastructure is based on both machine learning algorithms and rule–based approaches, together with language resources adapted to the needs of specific processing stages. Development and performance information for these is given in [13], while their use in the preparation of the particular annotated resource is detailed in the next section.

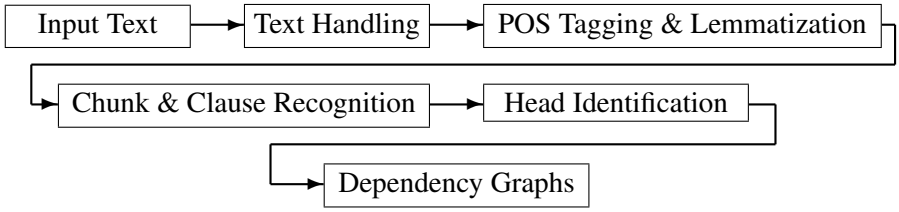


Figure 1: Preprocessing Pipeline

3 Syntactic representation in GDT

The first level of manual annotation in GDT focuses on surface syntax. At this level, we have opted for a dependency–based representation instead of one based on constituency. Dependency analyses represent sentences as graphs where each word corresponds to a node in the graph. Sentences are prototypically headed by the verb of the main clause, which can have zero or more dependents. Words are direct dependents of their heads without any intermediate phrasal nodes. Arcs be-

tween heads and dependents are labeled according to the kind of relation between respective words, although it is common practice to assign the label to the dependent node, together with any other word–specific information like POS tags and lemmas.

Parsers and annotation efforts for identification of dependencies between words (or dependencies between constituents) are known to exist for a number of languages including Danish [5] and Turkish [10], while a large dependency annotation project is the Prague Dependency Treebank for the Czech language (PDT), developed by the Institute of Formal and Applied Linguistics [1].

We chose a dependency–based representation because we believe that it allows for more intuitive descriptions of a number of phenomena, including long–distance dependencies, as well as structures specific to languages like Greek that exhibit a flexible word order. At the same time, dependency representations seem to be more theory–neutral since they are based on relations from traditional grammars with which annotators are usually quite familiar. Moreover, while available constituency–based approaches tend to focus on specific constructions, traditional textbook grammars describing the full range of Greek language phenomena are more compatible with dependency–based descriptions. The set of labels in our annotation schema is a derivative of the PDT, adapted to cater for Greek language structures. We compiled guidelines for the main syntactic structures of Greek after an initial study of randomly extracted selections from our corpus and exemplary sentences from textbook grammars.

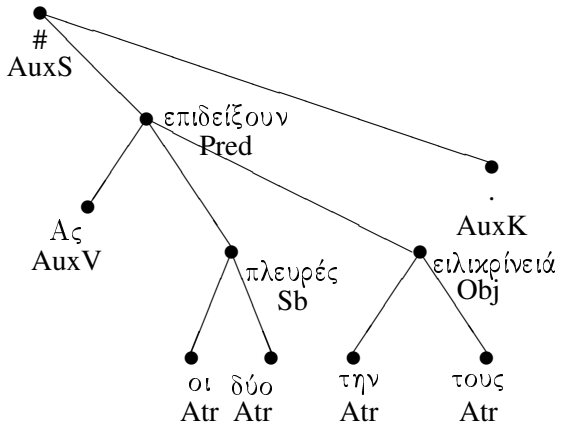
For the initial generation of the dependency graphs that the annotators have to correct the following procedure is undertaken. After POS–tagging and lemmatization, a pattern grammar compiled into finite state transducers recognizes chunk and clause boundaries, while a head identification module based on simple heuristics takes care of spotting the heads of these structures, and assigning labeled dependency links between head words of chunks and clauses, and the rest of the words inside their limits. The head identification module also assigns dependency links between heads of different chunks or clauses inside the limits of the sentence. The output is a dependency graph where, for each wordform, the following information is recorded:

- Lemma
- Morphosyntactic information according to a Parole–compatible tagset for Greek
- A label describing the type of dependency between the wordform and its head
- A slot for annotators' comments

The annotators have to further enrich the sentence graph by providing missing dependencies for unattached words, and/or by correcting automatically generated labeled edges. Thirty students of a postgraduate NLP course were each given an equal size portion of the 70K words corpus to correct. All annotators have used TrEd, an open source tool [11] for the annotation of dependency trees.

3.1 Specific constructions

In configurational languages like English, fixed word order provides strong evidence concerning grammatical relations between elements and their heads. Nevertheless, word order variation is not uncommon in many languages. Greek is an inflected language that uses case to encode grammatical relations, while exhibiting a flexible word order: certain S/V/O combinations are preferred, while others are associated with focal readings. Since we directly encode grammatical relations without presupposing any default constituent structure from which all others are derived, representation for the main relations in a sentence is quite straightforward. In Figure 2, the verb 'επιδείξουν' heads the sentence and is assigned the label *Pred*icate, while two words, 'πλευρές' and 'ειλικρίνειά', are annotated as dependents of the Pred and are assigned labels *Sub*ject and *Obj*ect respectively.

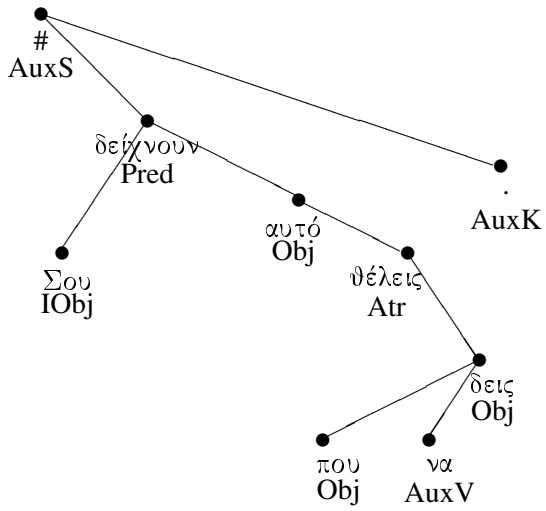


Ας επιδείξουν οι δύο πλευρές την ειλικρίνειά τους .
Let demonstrate$_{-3pl}$ the two sides the sincerity their .
The two sides should demonstrate their sincerity.

Figure 2: A tree for a VSO sentence

Since the representation schema allows non–projective trees, long–distance de-

pendencies and discontinuous constructions can be intuitively annotated. In cases of unbounded dependencies in relative clauses and wh–questions, a labelled arc connecting pronouns or other wh–elements to their governor allows us to represent their relation without the use of coindexation with a trace. This is illustrated in the non–projective tree of Figure 3, where the object pronoun 'που' is attached to its deep governor 'δεις'.



Σου	δείχνουν	αυτό	που	θέλεις	να	δεις	.
you$_{-gen}$	show$_{-3pl}$	this	that	want$_{-2sg}$	to	see$_{-2sg}$	.

They show you what you want to see.

Figure 3: A tree for a sentence with a long–distance dependency

4 Semantic representation in GDT

In this section we give an overview of the approach adopted to add a layer of semantic information to GDT. We consider enrichment via Semantic Role Labeling (SRL), which can be defined as the recognition and labeling of the arguments of a target predicate. Given a sentence, the task consists of identifying, extracting and labeling the arguments that fill a semantic role of the predicates identified in that sentence. Our approach is envisioned to provide consistent argument labeling that would facilitate automatic extraction of relational data, without attempting to justify any theory. However, we incorporate and combine insights from recent

work in the field, especially from PropBank [12] and the Tectogrammatical Level of PDT [3].

4.1 Lexical resource

Parallel to the SRL annotation of the GDT, we compiled a lexicon with semantic information for verb predicates. The information encoded relies on predicate–argument structure, while lexicon building is conducted in 5 steps: a) selection of verbal predicates b) sense discrimination based on corpus evidence c) semantic role labeling, d) verb class assignment and e) event type assignment (for a subset of the selected predicates).

Selection of the verbal predicates – lemmas of the lexicon – is determined by a) the frequency of the verbs in the whole corpus from which annotation material was extracted, and b) the analysis of the data with respect to our end goal, i.e. fact extraction for end–users. The selection process has so far yielded a list of approximately 800 verbs.

The next step concerns sense discrimination based on examination of target verbs in sentences extracted from the corpus. Sense discrimination is based on both syntax and semantics. Each sense is thus turned into a single frameset, that is, a corresponding set of semantic roles (Table 1).

All possible syntactic realizations of a sense are grouped under the same frameset. Therefore, possible differences in the syntactic realizations of the arguments are not considered as criteria for distinguishing between framesets. For instance, the predicate 'αναγγέλλω' allows for both a clause and an NP object as shown in Table 2.

In general, we distinguish framesets in terms of a) the number of the semantic roles b) the semantic role labels and c) the verb class. We intend for synonymous predicates to share similar number of arguments and role labels. The argument list consists of labels that, in general, follow naming conventions for thematic roles. Table 3 presents the list of argument labels.

Semantic arguments of each predicate are sequentially numbered starting from Arg0 up to Arg5. The use of numbered arguments is strongly inspired from PropBank and serves the "easy mapping to any theory of argument structure" [12]. Each frameset is complemented by a set of indicative examples extracted from our corpus, that denote the respective predicate–argument structure described in the frameset. The frameset resource is being produced by 3 computational linguists and it is currently being enriched. As regards the framing rates, framing of each verbal predicate requires approximately 10–15 minutes. However, longer framing times are needed for highly polysemous verbs. The frameset descriptions in this resource are meant to serve as guidelines for the actual labeling procedure by the

‘απαντώ’ sense 1: answer			
Example: ο επίτροπος απάντησε στο Κοινοβούλιο ότι δεν υπάρχουν πλέον κονδύλια (the commissioner replied to the Parliament that there are no more funds)			
Argument	Arg. Label		
0	ACT	επίτροπος	(commissioner)
1	THE	δεν υπάρχουν πλέον κονδύλια	(there are no more funds)
2	ADDR	Κοινοβούλιο	(Parliament)
‘απαντώ’ sense 2: exist			
Example: στην Ισπανία απαντά μεγάλος αριθμός γυναικών που... (there exists a great number of women in Spain that...)			
Argument	Arg. Label		
1	THE	μεγάλος αριθμός	(great number)
2	LOC	Ισπανία	(Spain)

Table 1: Sense discrimination for the verb ‘απαντώ’ (answer)

Frameset for ‘αναγγέλλω’ (announce)	
Arg0: ACT, Arg1: THE, Arg2: ADDR	
Ex.1:	Ο διευθυντής ανάγγειλε στους υπαλλήλους ότι θα [συνταξιοδοτηθεί Arg1–THE] (The director announced to the employees that he will [retire Arg1–THE])
Ex.2:	Ο διευθυντής ανάγγειλε στους υπαλλήλους του τη [συνταξιοδότησή Arg1–THE] του (The director announced his [retirement Arg1–THE] to his employees)

Table 2: Syntactic realizations of the Arg1–THE in the ‘αναγγέλλω’ frameset

Role	Label	Role	Label	Role	Label
Actor	ACT	Attribute	ATTR	End Point	ENP
Theme	THE	Location	LOC	Cause	CAU
Patient	PAT	Time	TMP	Purpose	PNC
Benefactive	BNF	Manner	MNR	Source	SRC
Experiencer	EXP	Instrument	INSTR	Destination	DST
Addressee	ADDR	Extent	EXT		
Recipient	RCP	Start Point	STP		

Table 3: Argument labels

annotators involved.

Apart from the frameset descriptions each verbal predicate is accompanied by two more attributes: verb class and event type. Verb class is referred to as a category defined in terms of both syntactic and semantic properties. Categorization is based on the dependency labels encoded at the syntactic and semantic layers, and can be considered as an adaptation of Levin's verb classes [7] to Greek. As an example let us examine the behavior of the verb 'απαντώ' (Table 1). This particular verb has two distinct senses corresponding to English *reply* and *exist*. Syntactic patterns corresponding to the two meanings are different. In the *reply* sense the verb is linked to 3 arguments, while in the *exist* sense there are two arguments. This distinction leads to the categorization of the two senses into two different verb classes. Each verb class contains semantically related verbs having the same argument syntactic properties and the same behavior with respect to diathesis alternations. Different senses are likely to belong to different verb classes. It should be noted that this information does not concern all the verbs of Greek language but only the sample of the 800 verbs extracted in the previous phase. Moreover, it is not exhaustive as regards the senses of these verbs but it pertains only to the senses identified in our corpus. However, since Greek language lacks resources like Levin's verb classes, we view this attempt as a preliminary step for the development of a comprehensive classification scheme.

As regards the event type attribute, the verbs of the lexicon are assigned an event type that corresponds to a node of a shallow domain specific ontology. We are mainly interested in events that correspond to the domains of our initial data collection, that is politics, health and travel. It should be noted that there are cases where event types and verb classes are identical but the perspective of the categorization is different. The event type assignment task involves spotting the verbal predicates that indicate significant events and give evidence of the target domains.

To this end, verbs of general language or verb senses that do not adhere to the domains of interest are not assigned a specific event type. Our approach is based on guidelines released by LDC[1] in the framework of the ACE project.

4.2 Corpus Annotation

At this annotation stage, a new label is attached to dependents of verbal predicates, depicting their semantic relation to their head. Preprocessing for this phase consists of assigning default semantic relations to nodes annotated as *Sb*, *Obj* or *IObj* at the syntactic level. The annotation process is a two–pass procedure. The annotation team was asked to correct the automatically generated labels and to assign labels to all arguments attached to the verbal predicates of each sentence. This first pass was then checked and corrected according to modifications that resulted from the problems encountered during the annotation process. The annotation team worked on the data for a period of 4 weeks.

The annotators were provided with the frameset descriptions and a set of guidelines defining the dependency relations that correspond to semantic roles and their labels. The guidelines were accompanied with indicative examples, concerning problematic cases like null subjects, passive or ergative constructions, alternations and the disambiguation between similar roles, as in the case of the recipient/addressee pair. One of the major issues encountered is handling of null elements that were not annotated in the previous phase. We introduced new nodes to restore only null subjects, in order to fill important semantic roles like Actor and Theme (Figure 4).

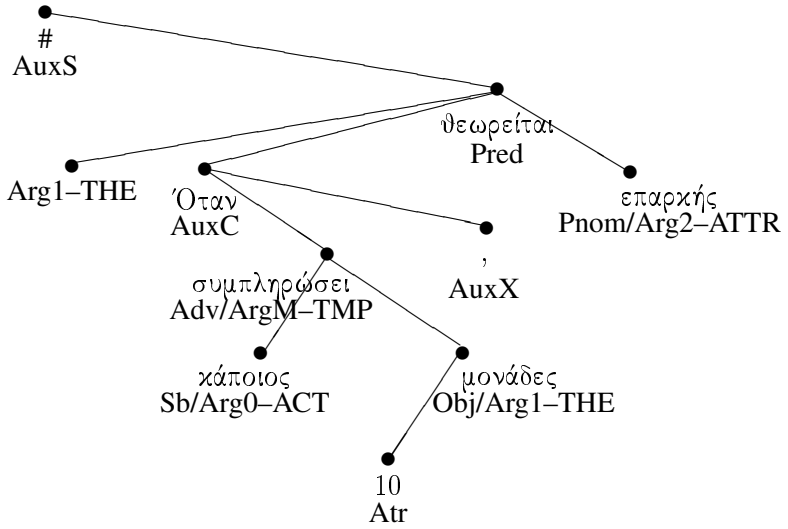
Apart from arguments filling semantic roles, adverbial modifiers were also annotated during this phase. The list of the semantic labels assigned to these modifiers is provided in Table 4. Although annotated throughout the corpus, these adjuncts were not included in the frame files.

As regards interannotator agreement, discrepancies mainly concerned a) the distinction between highly numbered arguments (from Arg2 up to Arg5) and adverbial modifiers and b) the type of adjunct labels like TMP, ENP and STP. Inconsistencies concerning inanimate agents of passive constructions were also frequent, as annotators assigned either ArgM–CAU or Arg0–ACT labels to these arguments.

5 Future Work

We plan to use the syntactically annotated resource for the development of a dependency parser for Greek, following recent advances in parsing literature ([14];

[1] http://www.ldc.upenn.edu/Projects/ACE/Annotation/2005Tasks.html



Όταν	κάποιος	συμπληρώσει	10	μονάδες	,	θεωρείται	επαρκής
When	someone	reach $_{-3sg}$	10	points,	,	consider$_{-pass}$	adequate

When someone reaches 10 points, (s)he is considered adequate

Figure 4: Semantic role label annotation introducing the role Arg1–THE in the null subject position of the predicate 'θεωρείται' (passive construction)

Role	Label	Role	Label
Location	LOC	End Point	ENP
Time	TMP	Cause	CAU
Manner	MNR	Purpose	PNC
Instrument	INSTR	Source	SRC
Extent	EXT	Destination	DST
Start Point	STP		

Table 4: Adjunct labels

[6]). Since our current collection is limited for training purposes[2] we are currently enriching it with more annotated data. The semantic layer of the resulting resource will serve as training material for the development of a system for automatic SRL, as in the shared task of CoNLL–2004[2]. We will examine machine learning techniques for the training of classifiers that disambiguate between role labels, and examine approaches that exploit dependency relations [4] for this particular task. We plan to experiment on automatic verb class and event type assignment exploiting the relevant information encoded in the lexical resource, together with syntactic and semantic labels from the dependency trees.

Acknowledgements

We would like to thank three anonymous reviewers for useful suggestions and comments. Work described in this paper was fully supported by the research project "Multimedia Content Management Systems" (MUSE), funded in the framework of Axis 3, Measure 3.3 of the Concerted Programme for Electronic Business of the General Secreteriat for Research and Technology of the Greek Ministry of Development.

References

[1] Alena Böhmová, Jan Hajič, Eva Hajičová and Barbora Hladká (2003) The Prague Dependency Treebank: A Three–Level Annotation Scenario. In A. Abeillè (ed.) *Treebanks: Building and Using Parsed Corpora.* Kluwer.

[2] Xavier Carreras and Lluís Màrquez (2004) Introduction to the CoNLL–2004 Shared Task: Semantic Role Labeling. In *Proceedings of the Eighth Conference on Natural Language Learning.* Boston, MA.

[3] Eva Hajičová, Jarmila Panevová and Petr Sgall (2000) A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank. UFAL/CKL Technical Report TR–2000–09. Prague, Czech Republic.

[4] Kadri Hacioglu (2004) Semantic Role Labeling using Dependency Trees. In *Proceedings of COLING 2004.* Geneva, Switzerland.

[5] Matthias Trautner Kromann (2003) The Danish Dependency Treebank and the DTAG Treebank Tool. In *Proceedings of TLT 2003.* Växjö, Sweden.

[2] Sparse data problems are reported in an experiment with a dependency treebank of about 100K words in [9].

[6] Ryan McDonald, Fernando Pereira, Kiril Ribarov and Jan Hajič (2005) Non–projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT–EMNLP 2005.* Vancouver, Canada.

[7] Beth Levin (1993) *English Verb Classes and Alternations: A Preliminary Investigation.* University of Chicago Press.

[8] Adam Meyers (2005) Introduction to Frontiers in Corpus Annotation II: Pie in the Sky. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky (ACL 2005).* Ann Arbor, Michigan.

[9] Joakim Nivre and Jens Nilsson (2005) Pseudo–Projective Dependency Parsing. In *Proceedings of the 43nd Annual Meeting of the ACL.* Ann Arbor, Michigan.

[10] Kemal Oflazer, Bilge Say, Dilek Zeynep Hakkani–Tür and Gökhan Tür (2003) Building a Turkish Treebank. In Anne Abeillè (ed.) *Treebanks: Building and Using Parsed Corpora.* Kluwer.

[11] Petr Pajas (2005) Tree Editor TrEd. http://ckl.mff.cuni.cz/pajas/tred/

[12] Martha Palmer, Daniel Gildea and Paul Kingsbury (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics,* 31(1):71–106.

[13] Harris Papageorgiou, Prokopis Prokopidis, Iason Demiros, Voula Giouli, Alexis Konstantinidis and Stelios Piperidis (2002) Multi–level XML–based Corpus Annotation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002).* Las Palmas, Spain.

[14] Hiroyasu Yamada and Yuji Matsumoto (2003) Statistical Dependency Analysis with Support Vector Machines. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 2003).* Nancy, France.