Adding multi-layer semantics to the Greek Dependency Treebank

Harris Papageorgiou*, Elina Desipri*, Maria Koutsombogera*, Kanella Pouli*, Prokopis Prokopidis*†

*Institute for Language and Speech Processing / IRIS
Artemidos 6 & Epidavrou, Athens, Greece
†National Technical University of Athens
{xaris, elina, mkouts, kanella, prokopis}@ilsp.gr

Abstract

In this paper we give an overview of the approach adopted to add a layer of semantic information to the Greek Dependency Treebank [GDT]. Our ultimate goal is to come up with a large corpus, reliably annotated with rich semantic structures. To this end, a corpus has been compiled encompassing various data sources and domains. This collection has been preprocessed, annotated and validated on the basis of dependency representation. Taking into account multi-layered annotation schemes designed to provide deeper representations of structure and meaning, we describe the methodology followed as regards the semantic layer, we report on the annotation process and the problems faced and we conclude with comments on future work and exploitation of the resulting resource.

1. Introduction

Treebanks encode information and relations that are considered vital for linguistic knowledge and NLP technologies. In this sense, their design follows the latest developments in linguistic theories and frameworks, according to certain task definitions. In recent years, the 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. In this view, multi-layered annotation schemes have been designed in order to provide deeper representations of intra- and inter-sentential structure and meaning (Böhmová et al., 2003; Meyers, 2005). Linguistic insights and semi-automatic processing are being combined for the generation of corpora that integrate information on various types of linguistic relations. In order to be semantically complete, such merged linguistic representations take into account deep syntactic annotation, basic semantic propositions information, connections to eventuality variables, discourse elements, argument structure for instances of common nouns, coreference, temporal features representation, formal description of valency frames etc. Following these advances our approach involves annotation at the level of semantics and is envisioned to provide deeper representations of structure and meaning for the Greek language.

The rest of the paper is structured as follows: in the next section we give an overview of the data comprising our corpus and we describe their preprocessing in section 3. In section 4 we present the corpus enrichment via multi-layered semantic representations and finally in section 5 we put our goal into perspective by focusing on further work and exploitation of the resulting resource.

2. Corpus description

The corpus comprises texts that were collected in the framework of national and EU-funded research projects aiming at multilingual, multimedia information extraction. The text selection process was guided by the specific needs of the above mentioned projects. The main domains covered at this stage are politics (manual transcripts of European parliamentary sessions, and web documents) and travel (web documents).

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 corpus amounts to 70K words.

3. Data preparation and preprocessing

Human annotators working on semantic annotation were presented with data that had been already annotated and validated at the surface syntax level (Prokopidis et al., 2005).

The dependency–based model chosen for the syntactic representation 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. Dependency representations seem to be more theory–neutral and they are quite similar to constructions of traditional grammars with which annotators are usually quite familiar. The set of labels used in the annotation schema is a derivative of the Prague Dependency Treebank, adapted to cater for Greek language structures. The annotation process was based on guidelines compiled for the main syntactic structures of Greek.

The assignment of syntactic labels was performed on data that had been preprocessed via an existing pipeline of shallow processing tools for Greek. This processing infrastructure (Figure 1) is based on both machine learning algorithms and rule-based approaches, together with language resources adapted to the needs of specific processing stages (Papageorgiou et al., 2002). Specifically, the processing tools include tokenization and sentence boundary recognition, part-of-speech tagging, lemmatization, chunk and clause recognition, and head identification modules.

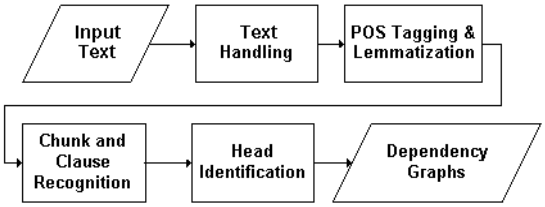

Figure 1: Preprocessing pipeline

For the initial generation of the dependency graphs, the following procedure is undertaken; after tokenization, 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 word node, we record its lemma, morphosyntactic information according to a Parole–compatible tagset for Greek, a label describing the type of dependency between the wordform and its head, and a slot for annotators' comments. The annotators' task was 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 used TrEd, an open source tool (Pajas, 2005) for the annotation of dependency trees. A sample syntactic dependency graph is shown in Figure 2.

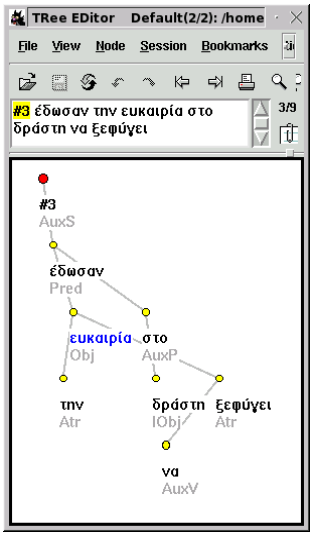

Figure 2: A dependency graph with syntactic annotation for the sentence "έδωσαν/gave την/the ευκαιρία/chance στο/to the δράστη/culprit να/to ξεφύγει/escape" (they gave the culprit the chance to escape).

4. Semantic annotation

In this section we consider enrichment of the GDT via two seemingly distinct levels of representation a) semantic role labeling and b) event detection and annotation and we explain the underlying concept.

Semantic Role Labeling (SRL), is defined as *the recognition and labeling of the arguments of a target predicate*. Given a sentence, the task consists of detecting, extracting and labeling the arguments that fill a semantic role of the predicates identified in that sentence. This 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 (Palmer et al., 2005) and the Tectogrammatical Level of PDT (Hajičová et al., 2000).

On the other hand, the event type and subtype attributes reflect an addition to the semantic annotation scheme. The task involves defining these verbal predicates that indicate significant events of specific predefined domains and assigning an event type and subtype based on a shallow, domain specific ontology. In this context, 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 tracks the latest developments and guidelines released by LDC in the framework of the ACE project (http://www.ldc.upenn.edu/Projects/ACE/docs/English-Events-Guidelines_v5.4.3.pdf).

With respect to the above definitions, our approach combines the two tasks in order to provide a multi-layered semantic representation. In our framework, event annotation strongly relies on semantic role labeling information. However, while semantic role labeling is applied to the whole corpus, event annotation pertains so far only to predicates of the political domain.

In the following sections we present in detail the semantic role labeling and the event type assignment tasks, as well as the infrastructure used by the annotators involved.

4.1. Lexical Resource

Prior to the SRL and event annotation of the GDT we compiled a lexical resource with semantic information for verb predicates in order to ensure a) consistent annotation of argument labels across different realizations of the same verb and b) consistent annotation of argument labels across predicates of the same event type.

The information encoded relies on predicate–argument structure while lexicon building is conducted in 4 steps: a) selection of verbal predicates b) sense discrimination based on corpus evidence, c) selection and labeling of arguments for each sense and d) event type assignment for a subset of senses.

Selection of the verbal predicates – lemmas of the lexicon – is determined by a) the frequency of the verbs in the whole corpus, and b) the analysis of the data with respect to further goals, i.e. fact extraction for end–users. The selection process has so far yielded a list of approximately 800 verbs that we expect to denote events of interest.

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. These instances were grouped into one or more major senses. Each sense is thus turned into a single frameset, that is, a corresponding set of semantic roles. More specifically, 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. An example of sense discrimination is shown in table 1.

In general, we distinguish framesets in terms of a) the number of the semantic roles and b) the semantic role labels. The argument list consists of labels that in general,

follow naming conventions for thematic roles. We intend for synonymous predicates to share similar number of arguments and role labels. When distinguishing framesets we also attempt to verify the hypothesis of a particular event type corresponding to a set of semantic roles that differentiate it from all other event types. Table 2 presents the list of argument labels.

«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)

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 adheres to its stipulation about the "easy mapping to any theory of argument structure" (Palmer et al, 2005). Each frameset is complemented by a set of indicative examples extracted from our corpus and denoting the respective predicate–argument structure described in the frameset.

The frameset resource was initialized by 3 computational linguists and 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 annotators involved.

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 2: Argument labels

4.1.1. Event annotation

The event annotation procedure strongly relies on semantic role labeling information as described in the previous section and is conducted in 3 steps a) the building of a taxonomy concerning the political domain b) the development of the annotation scheme and c) the actual annotation.

Events have been widely discussed in the framework of both theoretical linguistics and NLP. There have been attested significant variances in what exactly an event is and through which structure is instantiated. (Chung & Timberlake, 1985; Pustejovsky, 2000; Siegel & McKeown, 2000; Ingria & Pustejovsky, 2004). At the same time, various methodologies have been reported concerning the event detection, extraction and tracking (Allan et al., 1998; Filatova & Hovy, 2001; Filatova & Hatzivassiloglou, 2003).

In our framework, event annotation is considered as the key issue in the extraction of the core information concerning the domains of interest. In this context we are restricted to predicates denoting events that give evidence to the domain of politics while we plan to move onto other domains represented in our corpus i.e. travel.

The procedure starts with the identification of the most *significant* events in order to come up with a shallow ontology representing the data. The notion of *significance* is defined in terms of rich semantic representation as regards the domain under exploration. To this end, events correspond to schematic representations of specific situations together with the participants involved. Our approach is inspired by the efforts of the ACE project (http://www.ldc.upenn.edu/Projects/ACE/). The objective of the ACE project is to "develop extraction technology to support automatic processing of source language data (in the form of natural text, and as text derived from ASR and OCR)". Thus with respect to this objective the research efforts pertain to the detection and characterization of entities, relations and events. In this framework, event annotation is limited to a specific set of types and subtypes.

Our study shows that texts from the political domain usually convey information about societal and political issues and show reactions of parties against them. In addition, they include positions and evaluation of solutions to a variety of issues. Therefore, political events can be characterized as *neutral*, *for* or *against* something. Drawing on these remarks we developed the political events taxonomy through the following steps:

The grouping of all predicates included resulted into 3 major type categories, according to their literal or metaphorical meaning:

- verbs that do not signify events
- verbs that signify attitude towards events in life
- verbs that signify events in life

A further statistical analysis showed that 47% of the verbs presented in the corpus express states of consciousness[1] (perception, positive and negative emotions, knowledge etc.) compared to 46% that express

[1] This category was mapped to the "attitude towards events" group.

action[2] (activity, transfer of possession and movement). 7% of the verbs could not be classified.
In respect with our goals, the verbs of interest are those that signify attitude towards events in life.

Examples of the first category, which were excluded from further analysis, are:
a) impersonal verbs (*it seems that The United States has established with its European allies "a common approach"…*),
b) copular verbs and those that have a copular function (*nuclear technology is not negotiable, the passengers stood still on the train platform*),
c) existential verbs (*there is a great number of women in Spain*),
d) aspectual verbs (*a legal investigation has begun, this war lasts for decades*), and
e) the majority of the verbs that express logical relations (*…the solvents industry continues to contribute to the improved air quality in Europe).*

As it is discussed above the kernel of our initial hypothesis is that political events can be neutral, for or against something. This distinction concerns the relation between the main event and its arguments which might be an event as well. This practically means that if an event is a "for event" it contributes to the realization of its argument-event (*We {**intend, desire, hope,…**} to help you*, *the council {**agrees with, confirms, accepts, supports,…**} the government's decision*). On the contrary if an event is an "against event", it blocks the realization of its argument-event (*The government {**objects, is opposed,...**} to these requests*, *The President **contradicted** Democrats' charges that the tax cuts have damaged the economy*).

Taking this distinction into account we came up with a predicate taxonomy based on two factors: first, the verbs' meaning, literal or metaphorical and, second, the meaning of the verbs role sets. Predicates of the same category are expected to share a common role set. Thus, we concluded to the taxonomy of event types and subtypes, a sample of which is depicted in the table below:

The type and subtype attributes were then encoded for each verbal predicate of the lexical resource updating the frameset descriptions.

As regards the annotation scheme, apart from the detailed description of verb-members of each type and subtype of the ontology, it further clarifies the event trigger and the event extent as well. Event trigger is defined as the word that clearly denotes the occurrence of an event. In our framework this practically corresponds to single word verbal predicates as, for the time being, we don't deal with nominalizations and multi-word predicates. On the other hand, event extent relies on the dependency structure and corresponds to the text region within which the event is spanned. More specifically, it corresponds to all the arguments and modifiers that depend on the node of the verbal predicate in the dependency structure.

The annotation process described in the next section relies on the information and the lexical resources analyzed above.

[2] This category was mapped to the "events in life" group.

***neutral* events**	
Presentation event Discussion event …	a) Announcement (official presentation) b) Report (unofficial presentation)
***for* events**	***against* events**
Desire event Intention event Suggestion event Approval event Support event …	Rejection event Opposition event Invalidation event Hindrance event …

Table 3: Sample of the event taxonomy

4.2. Annotation

The annotation process encompasses two distinct phases corresponding to the two different levels of semantic information, that is, SRL and event annotation.

Based on the manually annotated syntactical dependencies assigned by annotators, a new version of the resource was automatically generated. In this version, a new label has been attached to dependents of verbal elements, depicting their semantic relation to their head. Only single-word verbal predicates are being targeted at this stage, while nominalizations and multi-word predicates will be annotated at a later stage.

Preprocessing for this phase is implemented as a running procedure on the dependency annotated data. It mainly involves assigning semantic roles to nodes annotated as *Sb*, *Obj* or *IObj* at the syntactic level, according to rules resulting from the analysis of the annotated data. Functional and auxiliary words are still attached to the head but are not assigned any role. The output of this phase is then manually corrected, using the TrEd tool, as in the previous phase.

The SRL annotation process is a two-pass procedure. The annotators' team worked on the data for a period of 4 weeks. Apart from the frameset descriptions, the annotators were provided with guidelines describing the annotation schema adopted, in order to further ensure consistency of the annotation process. These guidelines were enriched with indicative examples, concerning handling of several problematic cases such as null subjects, passive and ergative constructions, alternations, disambiguation between similar roles and labels i.e. Recipient and Addressee etc.

Specifically, the annotators were 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.

One of the major issues encountered during the SRL annotation is the handling of null elements that were not annotated in the previous phase. We decided to introduce new nodes to restore only null subjects, in order to fill important semantic roles such as actor and theme (in passive and ergative constructions).

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

Regarding the 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.

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

As regards the event type annotation, annotators were asked to detect events of interest and to assign type and subtype attributes. The annotators were given guidelines describing the proposed ontology, along with examples and remarks on what should be marked as an event of interest. Specifically, we annotate the event trigger and the event extent in terms of the verbal predicate and the participants and modifiers involved. During this phase, the initial hypothesis of event participants being mapped on verbal arguments was verified justifying the interrelation of the two levels of representation (table 5)

Presentation Events – Announcements	
Event Participants	Argument Labels
Sb that makes an announcement	0-ACT
The announcement itself	1-THE
Sb that the announcement is made to	2-ADDR

Table 5: event participants and argument labels mapping

A sample of the SRL and event annotation is presented in figure 3.

5. Future work

Drawing on a synthesis of the work described above, our goal is to develop event extraction technologies on the basis of predicate-argument lists, and to automatically recognize spatiotemporal relations between the most significant events of predefined domains. Intermediate plans include the development of an automatic SRL system by examining machine learning techniques for the training of classifiers that disambiguate between role labels and with the exploitation of the dependency relations for this particular task (Hacioglu, 2004).

Since our current collection is limited for training purposes we are currently enriching it with more annotated data. Furthermore we plan the enrichment of the GDT with coreference and spatiotemporal links concerning events and their participants as well.

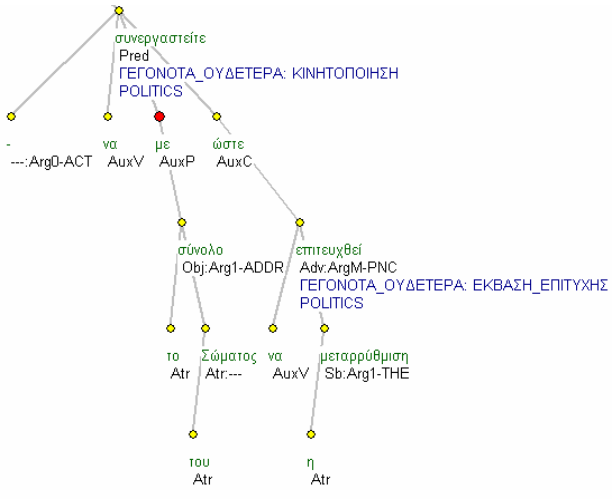

Figure 3: Sample of the SRL and event annotation for the sentence "να/to συνεργαστείτε/cooperate [*neutral event – mobilization*] με/with το/the σύνολο/entire του/of Σώματος/body ώστε/so να/to επιτευχθεί/is accomplished [*neutral event – achievement*] η/the μεταρρύθμιση/reform (cooperate with the entire body so that the reform is accomplished).

6. 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 "Retrieval of Video and Language for The Home user in an Information Society" (REVEAL THIS), FP6-IST-511689, funded in the framework of the specific research and technological development programme "Integrating and strengthening the ERA".

We are indebted to the 30 students who have collaborated in corpus compilation and annotation tasks. Their comments have been surprisingly helpful.

7. References

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. (1998). Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003). The Prague Dependency Treebank: A Three–Level Annotation Scenario. In A. Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer, pp. 103–127.

Carreras, X. and Màrquez, L. (2004). Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling: Data, Systems, and Results. In *Proceedings of CoNLL-2004*. Boston, MA.

Chung, S. and Timberlake, A. (1985). Tense, aspect, and mood. In T. Shopen (ed.), *Language typology and syntactic description*, Vol III. Cambridge University Press, pp. 202–258.

Filatova, E. and Hatzivassiloglou, V. (2003). Domain-Independent Detection, Extraction and Labeling of Atomic Events. In *Proceedings of the Recent Advances in Natural Language Processing conference. (RANLP 2003).*

Hacioglu, K. (2004). Semantic Role Labeling Using Dependency Trees. In *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland.

Hajičová, E., Panevová, J., and Sgall, P. (2000). A Manual for Tectogrammatic Tagging of the Prague Dependency Treebank. Technical Report TR–2000–09, UFAL/CKL, Prague, Czech Republic.

Ingria, B. and Pustejovsky, J. (2004). TimeML: A Formal Specification Language for Events and Temporal Expressions. Version 1.2, http://www.cs.brandeis.edu / jamesp/arda/time/timeMLdocs/TimeML12.htm.

LDC (2005). Automatic Content Extraction, http://www.ldc.upenn.edu/Projects/ACE/Annotation/2005Tasks.html.

Meyers, A. (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: Association for Computational Linguistics.

Pajas, P. (2005). Tree Editor TrEd, http://ckl.mff.cuni.cz/pajas/tred/.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):pp. 71–106.

Papageorgiou, H., Prokopidis, P., Demiros, I., Giouli, V., Konstantinidis, A., and Piperidis, S. (2002). Multi–level XML–based Corpus Annotation. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas, Spain.

Prokopidis, P., Desypri, E., Koutsombogera, M., Papageorgiou, H., and Piperidis, S. (2005). Theoretical and practical issues in the construction of a Greek Dependency Treebank. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Barcelona, Spain, pp. 149–160.

Pustejovsky, J. (2000). Events and the Semantics of Opposition. In Tenny and Pusteypvsky (eds.), *Events as Grammatical Objects*. Stanford, CA: CSLI, pp. 445–482.

Siegel, E. V. and McKeown, K. R. (2000). Learning methods to combine linguistic indicators: improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, 26(4):pp. 595–628.