Towards a frame semantics lexical resource for Greek

University of Essex
Department of Language and Linguistics

Elina Desipri, Maria Koutsombogera, Prokopis Prokopidis, Harris Papageorgiou
Institute for Language and Speech Processing
Department of Language Technology Applications

George Markopoulos
National & Kapodistrian University of Athens
Department of Linguistics

Abstract

We report on preliminary methodological issues related to the development of a Greek lexical resource based on the theory of frame semantics and supported by corpus evidence. Although our approach is primarily lexicographic, we also address a treebank annotation goal. We are aiming to produce an initial network of Greek words and frame-semantic descriptions that will reliably contribute to the multilingual dimension of the frame semantics framework.

1 Introduction

This paper presents a collaborative initiative aiming to develop a Greek lexical resource based on the theory of frame semantics (Fillmore, 1985). Relying on the English FrameNet project (Baker et al., 1998), our goal is the creation of a database containing frame-semantic descriptions of Greek words. We intend to document the range of semantic and syntactic combinatorial properties (*valences*) of each word in each of its senses in terms of annotated corpus attestations. For the development of the resource we use a corpus collection that amounts to 280M words. Our collection incorporates a variety of textual genres and domains; it comprises texts drawn from the Hellenic National Corpus (HNC)[1], transcripts of European parliamentary sessions (Koehn, 2002), and web documents pertaining to the financial, health, and travel domains. On a parallel track, we address a small-scale full-text annotation goal planning to add frame-semantic information to the Greek Dependency Treebank (GDT) (Prokopidis et al., 2005), a resource that is manually annotated at the level of syntax[2] and amounts to 70K words and 2,9K sentences. We report on preliminary

[1] http://hnc.ilsp.gr/
[2] Currently the GDT incorporates a PropBank-style semantic annotation.

methodological issues related to the first phase of our work. In this phase, our main focus is the production of an initial, balanced network of Greek lexical units and frames that will reliably contribute to the multilingual dimension of frame semantics.

2 Methodological Issues

Frame semantics describes word meaning in terms of underlying conceptual structures. These are encoded in the form of *frames*, i.e. schematic representations of stereotyped situations capturing certain amount of background (real-world) knowledge. Each frame is associated with a set of words (verbs, nouns, or adjectives) or expressions that *evoke* it and a set of semantic roles (*frame elements*) corresponding to the participants and props in the designated prototypical situation.

Our approach is primarily lexicographic. We aim to document the entire sense space of each lexical unit and represent it in terms of the frame semantics paradigm. As explained below, we apply a 'hybrid' methodology working on two levels: (i) word level and (ii) frame level. Our ultimate goal is to cover a variety of semantic domains (not restricted to the domains currently covered by the English FrameNet[3]) in a balanced fashion, so that reliable conclusions on the multilingual applicability of the FrameNet model can be drawn.

Vanguarding process: In the terminology of FrameNet, vanguarding refers to the theoretical, lexical semantic analysis of words which is required for the creation and population of frames. It includes organizing and prioritizing frames and lexical units, selecting the correct sense of polysemous words, sorting and selecting samples that display the variety of syntactic patterns of a given word, choosing the most relevant collocations, etc. (Fillmore, 2006).

Building on an inventory of already existing frames (the English frames), we organize this process as follows. On a first level, we work one lexical unit at a time concentrating (for the time being) on verbal predicates. Our initial set of predicates is a subset of the ones that appear in the Greek Dependency Treebank. For each predicate, we record the entire set of senses as described by Greek dictionaries. We perform certain 'smoothing' of the dictionary-based semantic distinctions, revising extremely fine-grained or vague distinctions and excluding terminological senses as well as colloquial senses. Metaphorical senses are recorded, unless they are exclusively colloquial. For each word sense we additionally report a set of synonymous and antonymous predicates. No frame-semantic criteria are considered in this stage.

[3] FrameNet is an ongoing lexicographic work. Currently, it contains more than 625 frames covering more than 8,900 lexical items.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

On the basis of this report, we perform a first analysis of each predicate extracting sufficient corpus attestations and grouping the recorded senses into a corresponding set of 'host' frames. Note that there is no a priori requirement that the relation between the dictionary-based senses and the 'host' frames be one to one. In some cases we decide to group two senses into one frame, while in others we have to split a single sense in two frames. However, it is noteworthy that although the dictionary-based distinction is used to speed up the process of representing the complete lexical semantic space of each word, a significant overlapping with corresponding frames has been observed so far, which keeps complication to minimum.

Deciding on the 'host' frames constitutes the most difficult step of the process. Following common practice, we examine extracted corpus instances of each word sense and check whether some FrameNet frame applies. On the basis of criteria that have been documented in development of FrameNet-like resources for other languages (Ellsworth et al., 2004 and Lönneker-Rodman, 2007)[4], our final decision usually takes one of the following forms: (i) some English frame is used without any changes (ii) it is slightly modified to accommodate the Greek data (iii) a new frame is introduced for Greek. As is the case with other approaches, we are faced with the problem of limited coverage of FrameNet. For word senses not represented in FrameNet we follow the SALSA Project policy of creating *predicate-specific proto-frames* (Burchardt et al., 2006).

Greek predicate	**Sense**	**FrameNet frame**	**Host frame**
χαιρετίζω	greet	no_frame	χαιρετίζω_gr
δικαιολογώ	justify	Justifying	Jystifying_gr

Table 1: Example Greek predicates and frames

Table 1 shows two cases of Greek data that deviate from the existing FrameNet database. In the case of *χαιρετίζω* a proto-frame has been created for Greek. In the second case the FrameNet frame *Justifying* has been modified to meet the meaning of the Greek predicate *δικαιολογώ*. Our provisional version of *Justifying* has an extended frame definition and a slightly different set of frame elements compared to the English frame. While FrameNet *Justifying* involves an Agent[5] giving a Reason for the licitness of an Act that he has done or omitted, or for a State_of_Affairs that a Judge deems to constitute a violation of an

[4] These criteria include questions like: (i) Is word meaning adequately described by a given frame definition? (ii) Do frame elements describe all semantic arguments of the predicate at hand? (iii) Does frame element description correspond to the attested properties of each semantic argument?

[5] Frame elements are marked with capitals.

obligation, in *Justifying_gr* a Justifier gives a Reason for the licitness of a State_of_Affairs for which a Justified_person (that *may or may not* be the Justifier himself) is held responsible. *Justifying_gr* is exemplified in the example below:

[JUSTIFIER Ο πρόεδρος] *δικαιολόγησε* [STATE_OF_AFFAIRS την απουσία] [JUSTIFIED_PERSON της Ομάδας των Πρασίνων] στη χθεσινή συνάντηση.
The chair justified the absence of the Green Party in yesterday's meeting.

Example 1: Annotated sentence for the Greek predicate *δικαιολογώ*

A second methodological level involves frame analysis. Initial frame processing seeks to prioritize a set of new lexical units related (in at least one of their senses) to the already considered frames. This set comprises two (usually) overlapping sets: (i) the translations of all verbal predicates included in the FrameNet frames that have been applied or adapted to Greek, (ii) the set of synonyms and antonyms reported for the processed Greek predicates. As new lexical units are being added and frames are populated, frame analysis includes repeated consistency checks of frame and frame element definitions. Furthermore, proto-frames are grouped together into larger frames.

Lexical unit and frame analysis are two parallel methodological levels that ensure a balanced expansion of both word and frame space. This enables systematic observations regarding cross-lingual frame parallelism.

3 Future work

Frame-semantic annotation of the Greek Dependency Corpus is planned to start at the end of the first phase. We view this as an additional step towards further refinement of the created frames. Exhaustive annotation will follow the previously described analysis, proceeding one predicate at a time. However, it will have to deal with a number of phenomena for which meaning representation is not straightforward, such as metaphoric usages, idioms, etc. We plan to address these issues in the immediate future.

Acknowledgements

Work described in this paper is fully supported by the research project "TV++" (A/V Digital Archive Management), funded in the framework of Measure 3.3 of the Operational Programme "Information Society" of the 3rd CSF.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

References

Baker C. F., Fillmore C. J., and Lowe J. B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*. Montreal, Canada.

Burchardt A., Erk K., Frank A., Kowalski A., Padó S. and Pinkal M. (2006). The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

Ellsworth M., Erk K., Kingsbury P. and Padó S. (2004). PropBank, SALSA, and FrameNet: How Design Determines Product. In *Proceedings of the LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora.* Lisbon.

Fillmore C. J. (1985). Frames and the semantics of understanding. In *Quaderni di Semantica,* Vol. 6.2: 222-254.

Fillmore C. J. (2006). The current state of FrameNet. *Presentation in Multilingual Semantic Annotation Workshop*. Saarbruecken.

Koehn P. (2002). Europarl: A multilingual corpus for evaluation of machine translation. *Unpublished Draft*.

Lönneker-Rodman, B. (2007). Multilinguality and FrameNet. *ICSI Technical Report TR-07-001*. Berkeley, CA

Prokopidis P., Desipri E., Koutsombogera M., Papageorgiou H. and Piperidis S. (2005). Theoretical and practical issues in the Construction of a Greek Dependency Corpus. TLT-2005. Barcelona, Spain.