Greek within the Global FrameNet Initiative: Conclusions and Issues so far

Voula Giouli[1,2]; Vera Pilitsidou[2]; Hephaestion Christopoulos[2]
[1]Institute for Language & Speech Processing, ATHENA RC, [2]National & Kapodistrian University of Athens
voula@athenarc.gr, verapilitsidou@gmail.com, hchristo@turkmas.uoa.gr

Abstract

Large coverage lexical resources that bear deep linguistic information have always been considered useful for many natural language processing (NLP) applications including Machine Translation (MT). The Global FrameNet initiative has been conceived of as a joint effort to bring together FrameNets in different languages. The proposed paper is aimed at describing ongoing work towards developing the Greek (EL) counterpart of the Global FrameNet, based on the database of the Berkeley FrameNet (BFN) project. In the paper, we will elaborate on the annotation methodology employed, the current status and progress made so far, as well as the problems raised during annotation.

Introduction

FrameNet (FN), the lexical database for the English (EN) language for general purposes (Baker et al., 1998), was developed at the University of Berkeley in California based on the theory of Frame Semantics (Fillmore, 1977, 1982, 1985). Over the years, a number of frame-based language resources have been developed for various languages, and this paper is our contribution to this effort.

Our work consists of LU creation and corpus annotation and it is aimed at the development of a frame-based lexical resource for the EL language and its alignment initially with the BFN, but also to FNs developed for other languages. From another perspective, one of our objectives is to examine whether the aligned lexica can be utilized for the translation process. Effort has also been made to detect and categorize the differences spotted between the EL and EN languages.

We report on the progress made and the results obtained so far, as well as the various issues and challenges we faced while working on the EL component of the Global FrameNet project.

Annotation Methodology

Annotation was performed on the transcribed TED Talk "Do schools kill creativity?" (Robinson 2006) and the subtitles provided for the EL counterpart of the TED talk. The EL corpus comprises 251 sentences and 3,012 tokens and the raw text was pre-processed at various levels of linguistic analysis using the UDPipe annotation platform (Straka & Starková, 2017).

Annotation was a two-stage procedure performed by two annotators via the dedicated MLFN WebTool (Matos & Torrent, 2016). In stage (a), creation of the LUs (or lexical annotation), we adopted a purely lexicographic approach; we first extracted all the lemmas from the EL text and then assigned them a frame based on their semantics. For the LUs that we created we also provided a short lexicographic gloss in English. Following the global guidelines, we adopted frames defined in the 1.7 release of the BFN 1.7.

Stage (b) was the annotation of the corpus using the LUs already created and extending or modifying them where needed. Each sentence was annotated at the following layers: (a) Frame and Frame Element (FE) layer, (b) Grammatical Function (GF) layer, and (c) Phrase Type (PT) annotation.

An example of an annotated sentence from the corpus is presented in Figure 1.

Είχαμε **ΓΕΜΙΣΕΙ** [GOAL το μέρος] [THEME με ατζέντηδες (`Filling`) (Implied AGENT: We)
Had1.pl filled the.acc place.acc with agents.acc
"We had filled the place with agents"

Figure 1: Example of annotation from within the corpus

Results

POS	Total unique
verbs	167
nouns	92
adjectives	23
adverbs	4
numerals	7

Table 1: Distributions of unique LUs per POS

Number of LUs created	**626**
Number of LUs annotated	603
Perfect fits	549
Non-perfect fits	54
No frame assigned	23
New frames proposed	2

Table 2 Quantitative results of frame assignment

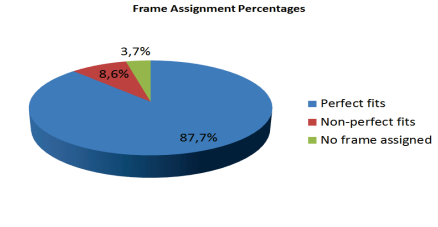

Figure 2: Frame assignment percentages

`Experiencer_focus`		
	Realization	
	like.v	αρέσω.v
EXPERIENCER	Ext.NP	Obj.NP
CONTENT	Obj.NP	Nsubj

Table 3: Realization of the LUs *to like.v* and *αρέσω.v*

Discussion

Our discussion will be focused on the annotation process and the results obtained.

In total, 603 LUs were annotated that evoke c. 250 frames; regarding the verbs of the EL corpus, which are the main focus, more than about 220 frames have been assigned to the 167 unique verbs.

We often had to diverge from the frames BFN assigns to certain LUs or make our own choices in cases of LUs that are not indexed.

In general, the major challenges during the creation of LUs and annotation was word sense discrimination for polysemous lemmas and selection of the appropriate frame for closely related frames or frames with no translational equivalent of the LU in EN.

In the case of systematic polysemy, which was a recurring issue, we noticed that BFN often ignores this aspect and, e.g., indexes LUs such as *university.n* or *school.n*, and subsequently their EL counterparts, only under `Locale_by_use`; however, these words do not only denote the building itself, but also the corresponding institution and the activities that take place there.

Some issues arised from gaps in the BFN frames, inconsistencies between languages, or from instances of MWEs or idioms and certain instances of mistranslation. However, in most cases, we were able to find a good fit, either by strictly following the BFN index or by searching through the BFN index and choosing a frame based on context.

Finally, we encountered several instances of non-perfect fits, the main causes of which were different perspective and different entailment, followed by too specific or too general frame, missing FE and different causative alternation.

Below we list some examples where frame assignment was not straightforward and perhaps not optimal:

One example that showcases the differences between EL and EN is depicted in Table 3. The EL verb *αρέσω* (to like) was assigned to the `Experiencer_focus` frame, however this was proven to be a non-perfect fit; the main difference between the verb *to like* and the EL LU *αρέσω.v* is that in EN the EXPERIENCER is always realized as the Subject of the verb; in EL, however, the EXPERIENCER is realized either as the complement of the preposition or as the object complement in genitive case, meaning that the EXPERIENCER is not the main focus. In certain cases, there may be no complement at all.

Another example arising from the idiosyncrasies of EL is the differences between the active and middle/passive morphology of a verb. While the annotation tool treats the active and middle voice of a verb as a single lemma, the voice can be frame-defining. For example, the LU *εμφανίζω.v* (to reveal, to present) in active voice needs to be assigned to a different frame as opposed to its middle voice counterpart *εμφανίζομαι.v* (to appear or arrive).

Conclusion

We presented a methodology for annotating a TED EL corpus within the MLFN shared task and based on BFN 1.7. We described the results obtained and the issues we encountered, a large number of which arise from the differences between EL and EN, as well as shortages in the BFN index. The overall results, however, were quite satisfactory and we were able to assign fitting frames to a large percentage of the LUs.

Future work is already planned towards enriching the EL data with new corpora and annotations and towards using the resource for aiding the translation process. Within the GFN project, comparisons and alignments with FNs in other languages will also be performed.

Acknowledgements

The authors would like to thank the anonymous reviewers and the editors for their valuable comments to the manuscript that contributed to improving the final version of the paper. The research leading to the results presented here was partially supported by the Translation and Interpreting MA programme of the Faculty of Turkish and Modern Asian Studies of the National and Kapodistrian University of Athens.

References

Fillmore, C. J. (1977). Scenes-and-frames Semantics. In A. Zampolli (ed) *Linguistic Structures Processing. Fundamental Studies in Computer Science*, vol. 59. North Holland Publishing: 55-81.

Robinson, K. (2006). *Do Schools Kill Creativity?* TED Talk. Available at https://www.ted.com/talks/ken_robinson_says_schoo ls_kill_creativity.

Straka, M. and Starková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Association for Computational Linguistics: 8899. Vancouver, Canada. http://www.aclweb.org/anthology/K/K17/K17-3009.pdf

Torrent, T., Elsworth, M., Baker, C., da Silva Matos, E. E. (2018). The Multilingual FrameNet Shared Annotation Task: a Preliminary Report. *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructions*.