STSM Report - PARSEME COST Action
Combining cross-lingual and syntactic evidence for Greek MWE identification

Marianna Apidianaki
LIMSI, CNRS, Université Paris-Saclay, Orsay, France
marianna.apidianaki@limsi.fr

General

Host institution:	Institute for Language and Speech Processing (ILSP / “Athena” R.C.), Athens, Greece
Dates:	March 1, 2016 - April 29, 2016

1 Work carried out during the STSM

The goal of this STSM has been the automatic identification of Greek MultiWord Expressions (MWEs). MWEs are word combinations that present idiosyncrasies in their syntax and semantics, making their processing by Natural Language Processing (NLP) systems a real challenge (Sag et al., 2002). We applied a translation-based methodology to Greek MWE identification (Melamed, 1997; de Medeiros Caseli et al., 2010; Moirón and Tiedemann, 2006) and complemented the cross-lingual evidence with shallow syntactic information. Contrary to previous work, we used two foreign language bridges (English and French) rather than just one, based on the assumption that word sequences having non-compositional meaning would tend to be translated consistently in different languages. Using two language pivots (English and French) we built MWE resources of higher quality than when one language was used. To evaluate the quality of the obtained resources, we fed them in two Greek dependency parsers (Prokopidis and Papageorgiou, 2014) and measured the impact of the proposed MWEs on parser performance. The work was carried out in collaboration with Prokopis Prokopidis, Haris Papageorgiou and Stella Markantonatou, the inviting person in the host institution (ILSP / “Athena” R.C.).

2 Results

We detected MWEs through one-to-many alignments (i.e. expressions translated with only one word in the other language) and through many-to-many translation correspondences (i.e. expressions consistently translated with specific word sequences in the other language). MWEs detected through one-to-many alignments were generally of higher quality, so we used this resource in the parsing experiments. Table 1 shows the size of the resources (in number of MWEs) built using one (English) or two language pivots (English and French), and retained after syntactic filtering. In the second case the number of MWEs is of course lower, but the resource is much cleaner as it contains expressions detected through both languages.

Syntactic Role	# of MWEs	
	One pivot	Two pivots
Adverb phrases	810	298
Prepositional phrases	1332	499
Adjective phrases	3768	1084
Noun phrases	1746	667
Total	7656	2548

Table 1: Number of cross-lingually extracted and syntactically filtered MWEs.

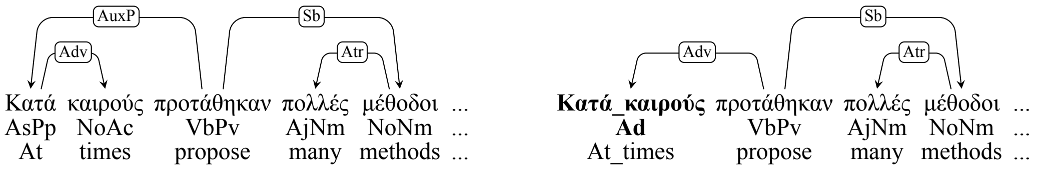
AuxP
Adv
Sb
Atr
Κατά καιρούς προτάθηκαν πολλές μέθοδοι ...
AsPp NoAc VbPv AjNm NoNm ...
At times propose many methods ...
Adv
Sb
Atr
Κατά_καιρούς προτάθηκαν πολλές μέθοδοι ...
Ad VbPv AjNm NoNm ...
At_times propose many methods ...

Figure 1: A tree segment before and after converting a MWE to a word_with_underscores.

3 Evaluation

We evaluated the quality of the extracted MWEs in experiments involving a recently extended version of the Greek Dependency Treebank (GDT) (Prokopidis et al., 2005). We used the GDT to train two well known representatives of the transition and graph-based families of parsers, the Maltparser (Nivre et al., 2007) and the Mateparser (Bohnet, 2010). We exploited only MWEs extracted as adverbial and prepositional phrases. If the components of a MWE were found as a sequence of tokens in a sentence of the train or test partitions of the GDT, we examined whether the sequence constituted a sub-tree. If yes, we joined the words of the sequence using underscores (e.g. ούτως_ή_άλλως (one way or another), κατά_κόρον (extensively), κατά_καιρούς (occasionally)), assigned an adverb part-of-speech tag to the newly created token and attached it as an adverbial modifier of the governor of the original subtree. See Figure 1 for an example.

The results presented in Tables 2 and 3 show the performance of each parser when it had no access to MWE information ("No conversion" setting); when it exploited information on frequent MWEs occurring more than 5, or more than 2, times in the corpus (MWEs >= 5 and >= 2); and when it used all extracted adverbial and prepositional MWEs regardless of their frequency. We report the Labelled and Unlabelled Attachment Scores (LAS and UAS) and the Label Accuracy (LACC) obtained by the parsers.[1] Table 2 presents the results when using Greek MWEs detected through English while Table 3 shows parsers' performance when using MWEs extracted through two language pivots. The results show that the use of two pivots (English and French) provides cleaner MWE resources compared to the use of one foreign language (English). In both cases, we observe consistent improvements of the transition-based Maltparser which achieved best performance when the entire resource (all MWEs) was used. In this case, 41 MWEs were found in the test data, compared to 38 when MWEs with a frequency of >= 2 were used and 35 when a stricter frequency filtering applied (MWEs occurring >= 5 times in the parallel corpus). This shows the good quality of the resource, as using a higher number of MWEs helps the parser without introducing errors. For the graph-based parser, it seemed harder to take benefit of these resources. In future work, we intend to analyse this parser's behaviour in dealing with MWEs and explore ways for taking advantage of this external source of knowledge.

[1] The LAS corresponds to the percentage of tokens that are assigned a correct head and a correct dependency type. The UAS corresponds to the percentage of tokens that are assigned a correct head, and the LACC corresponds to the percentage of tokens with the correct dependency.

	Mateparser (graph-based)			Maltparser (transition-based)		
	LAS	UAS	LACC	LAS	UAS	LACC
No conversion	82.65	88.45	89.69	79.76	85.27	88.42
MWEs >= 5	82.41	88.11	89.58	79.88	85.33	88.50
MWEs >= 2	82.61	88.59	89.65	79.88	85.27	88.50
all MWEs	82.46	88.18	89.65	**80.00**	**85.33**	**88.64**

Table 2: Parser performance when using MWEs obtained through English.

	Mateparser (graph-based)			Maltparser (transition-based)		
	LAS	UAS	LACC	LAS	UAS	LACC
No conversion	82.65	88.45	89.69	79.76	85.27	88.42
MWEs >= 5	**82.75**	88.44	**89.90**	79.96	85.38	88.58
MWEs >= 2	82.48	88.25	89.63	80.00	85.37	88.59
all MWEs	82.48	88.40	89.62	**80.20**	**85.50**	**88.70**

Table 3: Parser performance when using MWEs obtained through English and French.

4 Future collaboration with the host institution

This STSM has paved the way for future collaboration with members of the host institution on several topics. Future extensions will involve the exploitation of the extracted MWEs for cross-lingual knowledge transfer and Greek Semantic Role Labelling (van der Plas et al., 2014). Moreover, we intend to extract paraphrases of the identified MWEs from the Paraphrase Database (PPDB).[2] Apart from their exploitation for improving Greek syntactic and semantic processing, the extracted MWEs and their paraphrases will also serve to enrich the lexicographic resource IDION, dedicated to the documentation of Greek idioms.[3] Given the high number of extracted MWEs, we plan to proceed to a manual evaluation and selection of MWEs to be included in IDION in collaboration with students from the Department of Mediterranean Studies/University of the Aegean, Lab of Linguistics. Last but not least, we plan to explore alternative MWE representations that could be beneficial for the graph-based parser which turned out to be more difficult to improve using this knowledge.

5 Foreseen publications

The MWE identification methodology, the parsing experiments and the obtained results have been submitted for publication to the "Special issue on MWEs in Greek and other languages: from theory to implementation" of the Bulletin of Scientific Terminology and Neologisms of the Academy of Athens. Submission consisted in a extended (5-page) abstract. Notification of acceptance is expected by the end of June 2016. The long version of accepted abstracts is due for the end of September 2016. The order of the authors is as follows: Marianna Apidianaki (LIMSI, CNRS, Université Paris-Saclay), Prokopis Prokopidis (ILSP / "Athena" RC) and Haris Papageorgiou (ILSP / "Athena" RC).

We are also planning a joint poster submission for the 7th PARSEME General Meeting which will take place on 26-27 September 2016 in Dubrovnik, Croatia.

[2] The PPDB is a database that contains millions paraphrases in 16 languages. The resource is freely available for research purposes and can be found at `http://paraphrase.org/`

[3] IDION can be downloaded from `http://idion.ilsp.gr`

6 Confirmation by the host institution of the successful execution of the STSM

Dr. Markantonatou reports:

Dr. Apidianaki worked at ILSP from March, 1 to April, 29 2016 on methods for retrieving good quality Greek MWEs from parallel corpora. The retrieved materials were used to improve the performance of the ILSP dependency parser of Modern Greek and the results were promising. This work has been submitted to the volume "Special issue on MWEs in Greek and other languages: from theory to implementation", Bulletin of Scientific Terminology and Neologisms of the Academy of Athens that I am co-editing with Dr. Anastasia Christofidou. Apart from the interesting tangible results, Dr. Apidianaki's research has opened new possibilities for cooperation regarding (i) the usage of the retrieved material for enriching the IDION resource of Modern Greek that has been developed in the framework of PARSEME, a task that is planned to take place in July with the help of students from the Department of Mediterranean Studies/University of the Aegean, Lab of Linguistics (ii) the very interesting research possibility of using paraphrases of MWEs to enrich IDION – this is a promising way for a systematic usage of corpus knowledge to encode the "meaning" of MWEs.

References

Bohnet, B. (2010). Top Accuracy and Fast Dependency Parsing is not a Contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 89–97, Beijing, China.

de Medeiros Caseli, H., Ramisch, C., das Graças Volpe Nunes, M., and Villavicencio, A. (2010). Alignment-based extraction of multiword expressions. *LRE Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2):59–77.

Melamed, D. (1997). Automatic Discovery of Non-Compositional Compounds in Parallel Data. In *Proceedings of EMNLP*, pages 97–108, Providence, RI.

Moirón, B. V. and Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. In *Proceedings of the EACL 2006 Workshop on Multiword Expressions in a multilingual context*, pages 33–40, Trento, Italy.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

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)*, pages 149–160, Barcelona, Spain.

Prokopidis, P. and Papageorgiou, H. (2014). Experiments for Dependency Parsing of Greek. In *Proceedings of the SPMRL-SANCL Workshop*, pages 90–96, Dublin, Ireland.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing*, Berlin/Heidelberg. Springer.

van der Plas, L., Apidianaki, M., and Chen, C. (2014). Global Methods for Cross-lingual Semantic Role and Predicate Labelling. In *Proceedings of COLING 2014*, pages 1279–1290, Dublin, Ireland.