Combining cross-lingual and syntactic evidence for Greek multi-word expression identification

Marianna Apidianaki[1], Prokopis Prokopidis[2], Haris Papageorgiou[2]

[1] LIMSI, CNRS, Université Paris-Saclay, Rue John von Neumann, Orsay
[2] Institute for Language and Speech Processing/Athena R.C., Athens

Abstract

We propose a methodology for Greek multiword expression (MWE) identification which combines alignment information from parallel corpora with syntactic knowledge. The main assumption behind the use of alignments for MWE extraction is that a sequence of words in one language that is often translated by a specific word or word sequence in another language might constitute a possible MWE. We apply this translation-based methodology to Greek MWE identification using alignments with two foreign languages (English and French) and filter the proposed candidates using shallow syntactic information. The quality of the extracted MWEs is evaluated by measuring their impact on parsing performance. The observed increase in the performance of a Greek dependency parser, compared to a setting where it has no access to MWEs, indicates the good quality of the extracted resource which we expect to be useful in both lexicographic work and in end-to-end NLP applications.

Keywords: multi-word expression, parallel corpora, alignment, syntactic chunking, dependency parsing

1. Introduction

Multiword expressions (MWEs) are word combinations that present idiosyncrasies in their syntax and semantics. The term describes a large number of distinct but related phenomena (e.g. non-decomposable idioms, phrasal verbs, nominal compounds, decomposable idioms, institutionalised phrases) which defy conventions of grammar and compositional interpretation to varying degrees, making their processing by Natural Language Processing (NLP) systems a real challenge (Sag et al., 2002). Some of these expressions exclude while others allow different degrees of internal variability and modification. They might allow some of their constituents to freely inflect while restricting the inflection of others, or they may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation (Tsvetkov & Wintner, 2011). Syntactically, MWEs might behave like words or phrases, occur in one rigid pattern or permit various syntactic transformations. From a semantics point of view, the compositionality of MWEs is gradual, ranging from fully compositional to idiomatic (Bannard et al., 2003). Some MWEs are more opaque in their meaning while others have more transparent meanings that can be inferred from their parts. Applying compositional methods of linguistic analysis to MWE processing might however (i) result in overgeneration and production of invalid word combinations, and (ii) pose idiomaticity problems, reflecting the difficulty to predict that the meaning of a grammatically correct expression might be unrelated to the meaning of its parts. Failing to handle MWEs may cause serious problems for NLP tasks that involve syntactic or semantic processing, such as parsing and Machine Translation.

We present a methodology for Greek MWE identification that combines shallow syntactic information with evidence obtained through cross-linguistic analysis. Translations have been shown to be a useful source of information for different semantic tasks: they can successfully guide disambiguation (Apidianaki & Gong, 2015) and can serve for sense induction (Apidianaki, 2008), sense annotation (Diab & Resnik, 2002) and paraphrasing (Bannard & Callison-Burch, 2005). Alignment information has also been shown to be particularly useful for MWE extraction (Melamed, 1997; De Medeiros Caseli et al., 2010) and for distinguishing between literal expressions and more idiomatic and opaque ones with non-compositional meaning (Moirón & Tiedemann, 2006; Tsvetkov & Wintner, 2010). In this work, we apply for the first time a translation-based methodology to Greek MWE identification.

Contrary to previous work which has mainly exploited alignments with one foreign language, we show that the use of two language pivots generates better MWE candidates. Alignments involving a morphologically-rich language, like Greek, can be noisy. The use of a second language pivot permits to discard erroneously identified multiword units. Moreover, rather than retaining only candidates that correspond to specific grammatical patterns (De Medeiros Caseli et al., 2010), we discard noisy sequences by complementing the cross-lingual information with knowledge provided by a shallow syntactic parser for Greek (Boutsis et al., 2000). To evaluate the quality of the retained MWEs, we feed them in a graph-based and a transition-based Greek dependency parser (Prokopidis & Papageorgiou, 2014). The observed improvement in the transition-based parser's performance, compared to a setting where it has no access to MWE information, highlights the good quality of the resource which we expect to be useful in both lexicographic work and in end-to-end NLP applications.

This chapter is organized in five sections. In the next section, we present related work on translation-based MWE identification. Section 3 presents our Greek MWE identification mechanism which combines alignment and shallow syntactic information. Section 4 reports the experiments that served to evaluate the quality of the extracted MWEs in a parsing framework and the obtained results. Finally, before concluding, we provide insights on the nature of the MWE candidates gained from a manual inspection of the automatically generated resources.

2. Related work

Alignment-based approaches to MWE identification use word alignments in parallel corpora for distinguishing between idiomatic expressions and more transparent ones (Tsvetkov & Wintner, 2010). The main assumption behind the use of alignments for MWE extraction is that if a sequence of words S ($S = s_1 \dots s_n$ with $n \geq 2$) in a text is aligned to a word sequence T ($T = t_1 \dots t_m$ with $m \geq 1$) in a translation (i.e. $S \leftrightarrow T$), then: (a) S and T share some semantic features, and (b) S may constitute a MWE. In other words, if a group of words in a source language is translated as a single word or as a fixed expression in one or more target languages, this can be considered as an indication of a fixed source expression with non-compositional meaning (Moirón & Tiedemann, 2006; De Medeiros Caseli et al., 2010; Zarrieß & Kuhn, 2009; Salehi et al., 2014).

When applied to a large parallel corpus, this alignment-based approach proposes a high number of word sequences that are coherent with word alignments. This long list of MWE candidates needs to be filtered to discard noisy word sequences. De Medeiros Caseli et al. (2010) clean the list of candidates using a part-of-speech filter that matches specific grammatical patterns (i.e. sequences of part-of-speech tags or words) and discard sequences whose frequency is below a certain threshold. Tsvetkov and Wintner (2010) rank and filter the extracted MWE candidates using statistics gathered from a large monolingual corpus. Candidate MWEs are decomposed into bi-grams and each bi-gram is associated with its Pointwise Mutual Information (PMI)-based score computed from the monolingual corpus. A word sequence of any length is then considered a MWE if all the adjacent bi-grams it contains score above a threshold.

Zarrieß and Kuhn (2009) use word alignment together with syntactic information for MWE identification. They detect translation correspondences on dependency parsed aligned sentence pairs, and identify single lexical items in one language that are aligned to a group of lexical items in the other language. In more recent work, Tsvetkov and Wintner (2011, 2014) use translation equivalents as one among other linguistically-motivated features (e.g. orthographic variation, partial morphological inflection) aimed at capturing the properties with respect to which MWEs may differ from non-MWEs. Using a dictionary they generate word-by-word translations of candidate MWEs to English and check the number of occurrences of the English literal translation in a large English corpus. The intuition behind this is that non-MWEs are expected to have some literal translational equivalents whereas for MWEs no or few literal translations are expected. Salehi et al. (2014a) propose an approach to detecting non-compositional components in MWEs which makes use of definitions, synonyms and translations in Wiktionary. Attia et al. (2010) use translations of Wikipedia titles in several languages and machine translation for Arabic MWE detection. Salehi et al. (2014b) predict the compositionality of MWEs using translations into multiple languages to estimate the distributional similarity between each component word and the overall MWE expression. They show that the estimation of compositionality is improved when using translations into multiple languages as compared to using distributional similarity in the source language, and that string similarity complements distributional similarity.

Our approach to Greek MWE identification relies on the same intuition as these alignment-based techniques. If a group of words in one language is translated as a single word in one or more languages, this can be considered as an indication of the presence of a fixed expression with non-compositional meaning.

3. Greek MultiWord Expression identification

3.1. Translation-based candidate MWE detection

We apply an alignment-based MWE detection method to two parallel corpora consisting of Greek texts, and their English and French translations. Aligning with only one language extracts a high number of candidate MWEs, a large part of which are not valid and need to be discarded. Our hypothesis for aligning with two languages, rather than just one, is that word sequences having non-compositional meaning will tend to be translated consistently in different languages. The use of a second language reinforces the confidence of the predictions and can rule out noisy sequences retained due to misalignments. As will be shown in the next section, the Greek MWE databases extracted by keeping word sequences identified through both English and French are smaller and “cleaner” compared to the ones obtained by pivoting through one language.

We carry out experiments on the Greek-English and Greek-French part of the multilingual Europarl corpus (version 7) (Koehn, 2005). The two corpora are sentence aligned, tokenized and lowercased, and have been cleaned from empty lines and lines with a great difference in length to ease the alignment procedure.[1] We apply GIZA++ (Och & Ney, 2003) to each parallel corpus (Greek-English and Greek-French) and obtain word alignments in both translation directions (source to target and target to source). Each word in one language is aligned to at most one word in the other language (one-to-many alignments).[2] To combine the directional alignments into a many-to-many alignment, i.e. an alignment linking sequences of >1 words on both sides, we apply the phrase-extract algorithm (Och & Ney, 2004).[3] In a phrase-based SMT system, all phrase pairs extracted by the algorithm are generally preserved. Since here the extracted phrases will be used for MWE identification, we only keep phrase pairs that contain up to five words on either side.

We carry out experiments using the Greek MWE candidates proposed by aligning with one (English) and two foreign languages (English and French). In the latter case, we keep MWEs identified through both languages, i.e. found in the intersection of the two candidate sets constructed through alignments.

[1] For English, we use the tokenization and lowercasing scripts available in the Moses Statistical Machine Translation (SMT) toolkit (Koehn et al., 2007). The Greek corpus has been pre-processed using a Greek tokenizer (Prokopidis et al., 2011). Lines on either side that are too long (i.e. more than 50 tokens) or which violate the 9-1 sentence ratio limit of the GIZA++ word aligner are discarded.

[2] Note that the parallel corpora are not syntactically analyzed prior to alignment and that we only apply syntactic constraints at a post-processing stage.

[3] The *phrase-extract* algorithm is implemented in the symmetrization heuristic ‘grow-diag-final’ of the Moses toolkit. To extract bilingual contiguous phrases, the algorithm starts by intersecting two asymmetrical alignments and adds additional alignment points that lie in the union of the two alignments and are diagonally adjacent. Links for unaligned words are added in a final step.

Syntactic role	# of MWEs	
	One Language	Two Languages
AdvP	810	298
PP	1332	499
AdjP	3768	1084
NP	1746	667
Total	**7656**	**2548**

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

3.2. Syntactic constraints on cross-lingually defined MWEs

Phrase pairs that are consistent with word alignments include many non-intuitive phrases that have no specified linguistic status (e.g. υπερβαίνουν κατά (*ypervenoyn kata*) 'exceed by', όμως απόλυτη (*omos apolyti*) 'but absolute') (Koehn et al., 2003). To eliminate the noise present in the extracted phrase pairs and restrict our MWE candidates to linguistically motivated phrases, we identify shallow phrases (chunks) in the Greek part of the parallel corpus using the parser developed by Boutsis et al. (2000).

The syntactic information does not serve to identify the phrases to be aligned but is used on top of the alignments, in order to drop MWE candidates proposed by the translation-based method that do not form syntactic constituents. More precisely, we retain phrases[4] falling into the following categories:

adverb phrase (advp); e.g. ξανά και ξανά (*xana ke xana*) 'over and over again', σχετικά με το θέμα (*schetika me to thema*) 'on this topic', ευθύς εξαρχής (*efthis exarchis*) 'from the very beginning'

prepositional phrase (pp); e.g. κατά συνέπεια (*kata synepia*) 'as a consequence', προς το παρόν (*pros to paron*) 'for the time being', *κατά νου* (*kata noy*) '(bear) in mind', κατά κάποιο τρόπο (*kata kapio tropo*) 'somehow'

adjective phrase (adjp); e.g. συντομότερο δυνατόν (*syntomotero dynaton*) 'the earliest possible', κοινοτικό κεκτημένο (*kinotiko kektimeno*) 'community acquis', πολιτικούς υπεύθυνους (*politikoys ypefthinoys*) 'politically responsible'

noun phrase (np); e.g. κενό γράμμα (*keno gramma*) 'empty letter', βήμα σημειωτόν (*vima simeioton*) 'mark time', Βοσνία Ερζεγοβίνη (*Vosnia Erzegovini*) 'Bosnia and Herzegovina', τοις μετρητοίς 'cash'

In general, MWEs detected through one-to-many alignments (i.e. expressions translated with only one word in the other language) are of higher quality than the MWEs spotted through many-to-many translation correspondences. Consequently, in the parsing experiments described in the next section we use the one-to-many alignments. Note that these are alignments where a sequence of *n* Greek words is translated by one English or French word in the corpus and not the inverse. Going from a morphologically-rich language (like Greek) to a relatively simpler one in this respect (like English or French) could trigger one-to-many alignments that would not be MWEs as we would expect different grammatical cases to be translated into specific word combinations.

[4] The phrase category (*advp, adjp, pp, np*) is determined by the part of speech of the head of the phrase. The *adjp* category is dominated by instances of the combination of the negation particle *μη* 'not'+ adjective. We intend to check the contents of this resource more carefully before using it in future parsing experiments.

Table 1 shows the size of the resources (in number of MWEs) built from one-to-many alignments correspondences using one (English) or two languages (English-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.[5]

4. Evaluation

We evaluate the quality of the extracted MWEs by experimenting with the Greek Dependency Treebank (GDT) (Prokopidis et al., 2005), which we split in a train (5668/173849 sentences/tokens) and a test partition (570/13442 sentences/tokens). Apart from labeled dependencies, GDT sentences have been manually validated for part-of-speech (PoS), morphosyntactic features and lemmas. The tagset used contains 584 combinations of PoS tags (Table 2) and features that capture the rich morphology of the Greek language. As an example, the full tag *AjBaMaSgNm* for a word like ταραχώδης (*tarachodis*) 'turbulent' denotes an adjective of basic degree, masculine gender, singular number and nominative case. The three last features are also used for nouns, articles, pronouns, and passive participles. Verb tags include features for voice, tense and aspect, while articles are marked for definiteness.

PoS	**Description**	**PoS**	**Description**
Ad	Adverb	AsPpPa	Prep. + Article combination
AjBa	Adjective (basic degree)	CjCo	Coordinating conjunction
AsPpSp	Preposition	CjSb	Subordinating conjunction
AtDf	Definite article	NoCm	Common noun
AtId	Indefinite article	PnPo	Possessive pronoun
VbMn	Finite verb	PnRe	Relative pronoun

Table 2: Common fine-grained PoS tags in GDT.

The dependency-based annotation scheme used for the syntactic layer of the GDT is based on an adaptation of the guidelines for the Prague Dependency Treebank (Böhmová et al., 2003), and allows for intuitive representations of long-distance dependencies and non-configurational structures common in languages with flexible word order. Most trees are headed by a word that bears the Pred relation to an artificial root node. Other tokens depending on this root node include sentence-final punctuation marks and coordinating conjunctions. Coordinating conjunctions and apposition markers head tokens participating in relevant constructions. Table 3 contains some of the most common dependency relations used in the treebank, while Figure 1 presents a sentence fragment that contains a non-projective arc connecting the verb of a complement clause and its extraposed argument.

Dependency Relation	**Description**	**Dependency Relation**	**Description**
Pred	Main sentence predicate	Adv	Adverbial dependent
Sb	Subject	Atr	Attribute
Obj	Direct object	Coord	A node governing coordination
AuxC	Subordinating conjunction node	AuxP	Prepositional node

Table 3: Common dependency relations in the Greek Dependency Treebank.

[5] Examples of MWEs wrongly identified through English and filtered out by aligning with the second language: προς βήμα (*pros vima*) 'by step' (pp), όπως εκείνα (*opos ekina*) 'like those' (advp), από εκείνα (*apo ekina*) 'from those' (pp), περί εκπομπών (*peri ekpompon*) 'on emissions' (pp), υπόλοιπα δασωμένα (*ypolipa dasomena*) 'remaining forested' (adjp), οποιοδήποτε σημείο (*opiodipote simio*) 'any point' (np).

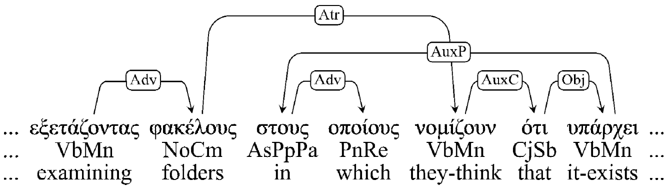
Atr
AuxP
Adv
Adv
AuxC
Obj
... εξετάζοντας φακέλους στους οποίους νομίζουν ότι υπάρχει ...
... VbMn NoCm AsPpPa PnRe VbMn CjSb VbMn ...
... examining folders in which they-think that it-exists ...

Figure 1: An analysis for a sentence fragment with a non-projective arc.

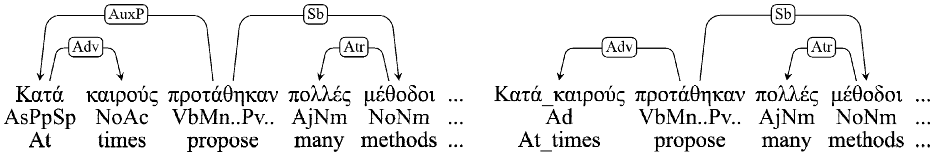
AuxP
Adv
Sb
Atr
Κατά καιρούς προτάθηκαν πολλές μέθοδοι ...
AsPpSp NoAc VbMn..Pv.. AjNm NoNm ...
At times propose many methods ...
Adv
Sb
Atr
Κατά_καιρούς προτάθηκαν πολλές μέθοδοι ...
Ad VbMn..Pv.. AjNm NoNm ...
At_times propose many methods ...

Figure 2: A tree segment before and after conversion of a MWE to a *word_with_underscores*.

We use the GDT to train Maltparser (Nivre et al., 2007) and Mateparser (Bohnet, 2010), two well known representatives of the transition and graph-based families of dependency parsers. In our initial experiments, we exploit only MWEs extracted as adverb and prepositional phrases in the following way: if the constituent tokens of a MWE are found as a sequence in a sentence of the train partition of the GDT, we examine whether the sequence constitutes a sub-tree; if yes, we join them into a single token using underscores (e.g. ούτως_ή_άλλως (*oytos_i_allos*) 'one way or another', *κατά_κόρον* (*kata_koron*) 'extensively', κατά_καιρούς (*kata_keroys*) 'occasionally') following Nivre and Nilsson (2004) and Eryiğit et al. (2011). We repeat the same procedure on the test set but, in order to avoid using manual syntactic information in the gold test set, we first parse the test set sentences with a parser trained on the unconverted train set. Then we join sequences into a single token in the gold test set, only if they constitute a sub-tree in the automatically parsed test set. We assign an adverb (*Ad*) part-of-speech tag to the newly created token and attach it as an adverbial modifier of the governor of the original subtree. See Figure 2 for an example.

The results presented in Tables 4 and 5 show the performance of each parser when it has no access to MWE information ('No conversion' setting); when it exploits information on frequent MWEs occurring more than 5, or more than 2, times in the corpus (MWEs >= 5 and >=2); and when it uses all extracted adverb 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.[6]

Table 4 presents the results when using Greek MWEs detected through English, while Table 5 shows parsers' performance when using MWEs extracted with the alignment with two languages. The results show that the alignment with two languages (English and French) provides cleaner MWE resources compared to the use of one foreign language (English). In both cases, we observe small but consistent improvements of the transition-based Maltparser which achieves best performance when the entire resource (all MWEs) is used. In this case, 41 MWEs are found in the test data, compared to 38 when MWEs with a frequency of >=2 are used and 35 when a stricter frequency filtering applies (MWEs occurring >=5 times in the parallel corpus). This highlights 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 seems harder to take benefit of these resources. In future work, we intend to analyse this parser's

[6] 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.

behaviour in dealing with MWEs and explore ways for taking advantage of this external source of knowledge.

	Mateparser (graph-based)			Maltparser (transition-based)		
Setting	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 4: Parser performance when using MWEs obtained through English.

	Mateparser			Maltparser		
Setting	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 5: Parser performance when using MWEs obtained through English and French.

	Adverb MWEs			Prepositional MWEs		
	Responses	%	Agreement	Responses	%	Agreement
0Y	121	40.88	0Y or 4Y: 54.73%	226	45.29	0Y or 4Y: 59.72%
4Y	41	13.85		72	14.43	
3Y - 1N	34	11.49	4Y & 3Y: 25.34%	71	14.2	4Y & 3Y: 28.66%
2Y - 2N	39	13.18		56	11.2	
1Y - 3N	61	20.61		74	14.9	
	296			499		

Table 6: Distribution of annotations for the adverb and prepositional MWEs.

In a subsequent experiment, we asked four students in linguistics to examine the lists of MWEs extracted as adverb and prepositional phrases and filter out those entries that were not actual MWEs. The MWEs were the ones extracted by the translation-based method through both language pivots and which satisfied the syntactic criterion (i.e. formed syntactic constituents). The annotators were asked to judge the quality of the MWEs (assign a “Yes” (Y) or “No” (N) value to each of them) with respect to their stable structure and non-compositional meaning. The validation procedure focused on 296 adverb and 499 prepositional MWEs. Table 6 gives a detailed picture of the agreement between the annotators. The first two rows contain information about MWEs that were rejected or accepted by all annotators (i.e. that were assigned 0 or 4 “Yes” values). The annotators agreed on the goodness of the MWEs in 54.73% of the cases. The other rows correspond to MWEs where there was disagreement between the annotators (for instance, MWEs that were assigned 2 “Yes” and 2 “No” values) which correspond to 45.28% of the cases. We retained the MWEs that were judged as good ones by at least three annotators. We include examples of MWEs with high agreement in the Appendix.

We observed that the accepted MWEs often corresponded to constructs inherited from older or ancient variations of Greek which are not used compositionally in everyday Modern Greek (e.g. εκ προοιμίου (*ek proimioy*) ‘from the very beginning’, (φέρνω) εις πέρας ((*ferno) is peras*) ‘(bring) to en end’).

Table 7 presents the parsers' performance when using the filtered list. The same pattern of improvements for the transition-based parser appears in this set of experiments, with the best results obtained when no threshold is used. In this case, 22 MWEs are found in the test data, compared to the 19 and 17 MWEs found when >=2 and >=5 thresholds apply, respectively.

	Mateparser			Maltparser		
Setting	LAS	UAS	LACC	LAS	UAS	LACC
No conversion	**82.65**	**88.45**	89.69	79.76	85.27	88.42
MWEs >= 5	82.62	88.39	**89.82**	79.94	85.35	88.48
MWEs >= 2	**82.65**	88.47	89.61	79.93	85.35	88.57
all MWEs	82.28	88.03	89.63	**80.14**	**85.53**	**88.63**

Table 7: Parser performance when using manually validated MWEs obtained through English and French.

Original test set, words involved in MWE conversion ignored							
		Mateparser			Maltparser		
#Tokens	Setting	LAS	UAS	LACC	LAS	UAS	LACC
13406	MWEs >= 5	**82.68**	**88.45**	89.73	79.78	85.25	88.44
13402	MWEs >= 2	**82.68**	88.45	**89.73**	79.78	85.25	88.44
13396	all MWEs	**82.70**	**88.45**	**89.75**	79.79	85.24	88.46
Converted test set, words_with_underscores ignored							
		Mateparser			Maltparser		
#Tokens	Setting	LAS	UAS	LACC	LAS	UAS	LACC
13406	MWEs >= 5	82.64	88.42	**89.83**	**80.01**	**85.42**	**88.55**
13402	MWEs >= 2	82.66	**88.49**	89.62	**79.94**	**85.36**	**88.57**
13396	all MWEs	82.29	88.04	89.64	**85.54**	**85.54**	**88.64**

Table 8: Results with words involved in MWE conversion ignored. Scores in bold are the best when comparing settings between the original and the converted test set.

It should be noted that when we merge tokens and replace them with words with underscores in the experiments discussed above, the number of tokens in the test datasets for each setting changes. This makes comparing results problematic. In an effort to remedy this, we applied the following procedure. For each setting involving the manually validated MWEs obtained through English and French (MWEs >=5, MWEs >=2, all MWEs) we identify all words involved in a MWE in the original and the converted test dataset as well as in the output of the parser trained in each setting. We then ignore these words when we calculate the performance of the parser. To better illustrate this procedure, if the sentence of Figure 2 was part of the test dataset, the dependent tokens of the AuxP and Adv relations of the original sentence, as well as the word with underscores dependent token of the Adv relation in the converted sentence, would have been ignored when calculating performance. We report results after applying this procedure in Table 8, where the upper part of the table corresponds to the "no conversion" setting, but with all words involved in MWE conversion (e.g. *Κατά*, *καιρούς* (*kata, keroys*) 'occasionally') ignored during scoring. In the lower part of the table we evaluate with all words_with_underscores (e.g. *Κατά_καιρούς* (*kata_keroys*) 'occasionally') ignored during scoring. This allows us to have the same number of tokens when comparing settings. We see that, although the performance of the graph-based parser is negatively affected when it is trained on a dataset where MWEs are merged to one token, the pattern of improvement for the transition-based parser is corroborated in all settings.

5. Conclusion

We have built resources of Greek multiword expressions from parallel corpora by combining translation and shallow syntactic information. The translation information used in our experiments consists of alignment correspondences automatically established between Greek texts and their English and French translations. This translation-based methodology allows the development of large-scale resources which however contain noisy MWEs, mainly due to alignment errors.

By filtering the MWE resources using syntactic evidence provided by a shallow parser, we discarded noisy word sequences that did not correspond to syntactic constituents. Furthermore, by aligning with a second language (French) we managed to further refine the contents of the MWE resource by keeping expressions that are consistently translated in the two languages and this is a piece of evidence that increases the chances of their meaning being non-compositional. For evaluation, we integrate the knowledge contained in the extracted MWE resources in the training and test steps of two Greek dependency parsers. The MWEs improve the performance of the transition-based parser; however the graph-based parser is more difficult to improve using this knowledge. In future work, we intend to explore alternative MWE representations that could be beneficial for the graph-based parser as well. Other future extensions will involve Greek MWE paraphrasing and exploitation of the extracted MWEs for cross-lingual knowledge transfer and Semantic Role Labelling (Titov & Klementiev, 2012; Van der Plas et al., 2014).

6. Περίληψη

Οι πολυλεκτικές εκφράσεις (ΠΛΕ) είναι συνδυασμοί λέξεων που παρουσιάζουν ιδιαιτερότητες όσον αφορά τη σύνταξη και τη σημασιολογία τους. Σε αυτό το άρθρο παρουσιάζουμε μια μεθοδολογία για τον εντοπισμό ελληνικών ΠΛΕ, η οποία συνδυάζει πληροφορία από τα αποτελέσματα αυτόματης συντακτικής ανάλυσης με πληροφορία από στοιχισμένες προτάσεις που περιέχονται σε παράλληλα σώματα κειμένων (ΠΣΚ). Η βασική υπόθεση πίσω από τη συγκεκριμένη μεθοδολογία, που πιστεύουμε ότι χρησιμοποιείται πρώτη φορά για την Ελληνική γλώσσα, είναι ότι, αν μια πολυλεκτική ακολουθία από λέξεις στη μία γλώσσα μεταφράζεται συχνά με μία συγκεκριμένη λέξη ή ακολουθία λέξεων σε μία άλλη γλώσσα, τότε α) οι δύο ακολουθίες μοιράζονται κάποια κοινά σημασιολογικά χαρακτηριστικά και β) η πρώτη πιθανώς συνιστά μια ΠΛΕ.

Χρησιμοποιήσαμε για τα πειράματά μας με βάση αυτή τη μέθοδο δύο ΠΣΚ που αποτελούνται από τις Αγγλικές-Ελληνικές (EN-EL) και τις Γαλλικές-Ελληνικές (FR-EL) προτάσεις του παράλληλου σώματος κειμένων Europarl. Εφαρμόσαμε έναν στοιχιστή λέξεων (word aligner) στις προτάσεις κάθε γλωσσικού ζεύγους και εξαγάγαμε λεκτικές (1-σε-n) και στη συνέχεια φραστικές (n-σε-m) στοιχίσεις, όπου 1 < (n, m) <= 5. Το αποτέλεσμα αυτής της διαδικασίας μπορεί να οδηγήσει σε υψίσυχνες φραστικές στοιχίσεις που δεν έχουν καλά ορισμένη γλωσσολογική υπόσταση (π.χ. υπερβαίνουν κατά ↔ exceed by) και δεν συνιστούν ΠΛΕ. Για την απαλοιφή αυτών των περιπτώσεων, χρησιμοποιήσαμε έναν επιφανειακό συντακτικό αναλυτή με τον οποίο επεξεργαστήκαμε το Ελληνικό τμήμα των ΠΣΚ. Η συντακτική πληροφορία δεν χρησιμοποιήθηκε για την επισήμανση των υποψήφιων ΠΛΕ, αλλά για την απαλοιφή όσων δεν αποτελούν συντακτικά συστατικά και, συγκεκριμένα, υποψηφίων ΠΛΕ που δεν αναγνωρίστηκαν αυτόματα ως επιρρηματικές, προθετικές, επιθετικές, ή ονοματικές φράσεις. Σε μια πρώτη εξέταση των αποτελεσμάτων, παρατηρήσαμε ότι το σύνολο των υποψηφίων ΠΛΕ που εξήχθησαν από *1*-σε-*n* στοιχίσεις (*1* Αγγλική/Γαλλική λέξη μεταφρασμένη με *n* Ελληνικές λέξεις) είναι υψηλότερης ποιότητας από τις υποψήφιες που προέκυψαν από φραστικές στοιχίσεις. Παρατηρήσαμε επίσης ότι, αν και οι υποψήφιες ΠΛΕ που εξήχθησαν τόσο από τα EN-EL όσο και από τα FR-EL ΠΣΚ είναι λιγότερες από αυτές που εξήχθησαν μόνο από την εφαρμογή της διαδικασίας στο EN-EL ΠΣΚ, το πρώτο σύνολο ήταν πολύ καθαρότερο καθώς περιείχε υποψήφιες ΠΛΕ που βασίζονταν σε πληροφορίες στοίχισης από δύο ΠΣΚ.

Αξιολογήσαμε τις αυτόματα εξαχθείσες ΠΛΕ με μια σειρά πειραμάτων αξιολόγησης με βάση μία ελληνική δενδροτράπεζα (treebank) όπου η κάθε πρόταση έχει αναπαρασταθεί ως ένα δέντρο εξαρτήσεων (dependency tree). Σε κάθε πρόταση αυτού του γλωσσικού πόρου ενσωματώσαμε την πληροφορία από τις ΠΛΕ, μετασχηματίζοντας κάθε ακολουθία λέξεων που ταυτιζόταν με μία από τις ΠΛΕ, σε μία λέξη (π.χ. *Κατά, καιρούς* → *Κατά_καιρούς*). Χρησιμοποιήσαμε δύο διαφορετικούς συντακτικούς αναλυτές εξαρτήσεων (dependency parsers) από τους οποίους ο πρώτος ανήκει στην κατηγορία των αναλυτών που βασίζονται στις μεταβάσεις (transition-based parser), ενώ ο δεύτερος δημιουργεί ένα γράφο με όλες τις πιθανές ακμές μεταξύ των λέξεων μιας πρότασης (graph-based parser) τις οποίες στη συνέχεια βαθμολογεί για να δημιουργήσει ένα συντακτικό δέντρο. Σε ένα από τα πειράματα αξιολόγησης με τους δύο αναλυτές ακολουθήσαμε την παρακάτω διαδικασία. Πρώτα αξιολογήσαμε χειρωνακτικά ορισμένες υποψήφιες ΠΛΕ, ζητώντας από τέσσερις γλωσσολόγους να εξετάσουν τις επιρρηματικές και προθετικές φράσεις που εξήχθησαν βάσει και των δύο ΠΣΚ. Αυτή η διαδικασία επικύρωσης οδήγησε σε έναν κατάλογο με 75 και 143 ΠΛΕ από τις συνολικά 296 και 499 επιρρηματικές και προθετικές φράσεις που εξετάστηκαν. Εκπαιδεύσαμε και αξιολογήσαμε την επίδοση των συντακτικών αναλυτών με και χωρίς το μετασχηματισμό, και παρατηρήσαμε ότι ο μετασχηματισμός οδήγησε στη βελτίωση της ακρίβειας του συντακτικού αναλυτή που βασίζεται στις μεταβάσεις.

Acknowledgements

The work presented in this chapter has been partly supported by the PARSEME (PARSing and Multi-word Expressions) and ENeL (European Network of e-Lexicography) COST (European Cooperation in Science and Technology) Actions.

Abbreviations

Abbreviated form	Full form
ΠΛΕ	Πολυλεκτικές εκφράσεις
ΠΣΚ	Παράλληλα σώματα κειμένων
GDT	Greek Dependency Treebank
LACC	Label Accuracy
LAS	Labelled Attachment Score
MWE	Multiword expression
PoS	Part of Speech
SMT	Statistical Machine Translation
UAS	Unlabelled Attachment Score

E-mail:

Marianna Apidianaki: marianna@limsi.fr
Prokopis Prokopidis: prokopis@ilsp.gr
Haris Papageorgiou: xaris@ilsp.gr

References

Apidianaki, M. (2008). Translation-oriented Word Sense Induction Based on Parallel Corpora. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)* (pp. 3269–3275). Marrakech, Morocco. ELRA.

Apidianaki, M. & Gong, L. (2015). LIMSI: Translations as Source of Indirect Supervision for Multilingual All-Words Sense Disambiguation and Entity Linking. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)* (pp. 298–302). Denver, Colorado. ACL.

Attia, M., Toral, A., Tounsi, L., Pecina, P. & van Genabith, J. (2010). Automatic Extraction of Arabic Multiword Expressions. *Proceedings of the Workshop on Multiword Expressions: from Theory to Applications* (pp. 19–27). Beijing, China. ACL.

Bannard, C., Baldwin, T. & Lascarides, A. (2003). A Statistical Approach to the Semantics of Verb-Particles. *Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment* (pp. 65–72). Sapporo, Japan. ACL.

Bannard, C. & Callison-Burch, C. (2005). Paraphrasing with Bilingual Parallel Corpora. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 597–604). Ann Arbor, Michigan. ACL.

Böhmová, A., Hajič, J., Hajičová, E. & Hladká, B. (2003). *The Prague Dependency Treebank: A Three-Level Annotation Scenario* (pp. 103-127). Dodrecht: Kluwer Academic Publishers.

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

Boutsis, S., Prokopidis, P., Giouli, V. & Piperidis, S. (2000). A Robust Parser for Unrestricted Greek Text. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)* (pp. 467–474). Athens, Greece. ELRA.

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

Diab, M. & Resnik, P. (2002). An Unsupervised Method for Word Sense Tagging using Parallel Corpora. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 255–262). Philadelphia, Pennsylvania. ACL.

Eryiğit, G., Ilbay, T. & Can, O. A. (2011). Multiword Expressions in Statistical Dependency Parsing. *Proceedings of the Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL)* (pp. 45–55). Dublin, Ireland. ACL.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of MT Summit* (pp. 79–86). Phuket, Thailand. AAMT.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (2007).

Moses: Open Source Toolkit for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 177–180). Prague, Czech Republic. ACL.

Koehn, P., Och, F. J. & Marcu, D. (2003). Statistical Phrase-based Translation. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* (pp. 48–54). Edmonton, Canada. ACL.

Melamed, D. (1997). Automatic Discovery of Non-Compositional Compounds in Parallel Data. *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 97–108). Providence, Rhode Island. ACL.

Moirón, B. V. & Tiedemann, J. (2006). Identifying idiomatic expressions using automatic word alignment. *Proceedings of the Workshop on Multiword Expressions in a Multilingual Context* (pp. 33–40). Trento, Italy. ACL.

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

Nivre, J. & Nilsson, J. (2004). Multiword Units in Syntactic Parsing. *Proceedings of the Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA) Workshop* (pp. 39–46). Lisbon, Portugal. ELRA.

Och, F. J. & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19–51. MA, USA: MIT Press.

Och, F. J. & Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, *30*(4), 417–449. MA, USA: MIT Press.

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

Prokopidis, P., Geograntopoulos, B. & Papageorgiou, H. (2011). A suite of NLP tools for Greek. *Proceedings of the Tenth International Conference of Greek Linguistics* (pp. 511–519). Komotini, Greece.

Prokopidis, P. & Papageorgiou, H. (2014). Experiments for Dependency Parsing of Greek. *Proceedings of the Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages (SPMRL-SANCL)* (pp. 90–96). Dublin, Ireland. ACL.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science*, v. 2276. Berlin/Heidelberg: Springer.

Salehi, B., Cook, P. & Baldwin, T. (2014a). Detecting Non-compositional MWE Components using Wiktionary. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp.1792–1797). Doha, Qatar: ACL.

Salehi, B., Cook, P. & Baldwin, T. (2014b). Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 472–481). Gothenburg, Sweden. ACL.

Titov, I. & Klementiev, A. (2012). Crosslingual Induction of Semantic Roles. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 647-656). Jeju Island, South Korea. ACL.

Tsvetkov, Y. & Wintner, S. (2010). Extraction of Multi-word Expressions from Small Parallel Corpora. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)* (pp. 1256–1264). Beijing, China. ACL.

Tsvetkov, Y. & Wintner, S. (2011). Identification of Multi-word Expressions by Combining Multiple Linguistic Information Sources. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 836–845). Edinburgh, UK. ACL.

Tsvetkov, Y. & Wintner, S. (2014). Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources. *Computational Linguistics*, *40*(2), 449–468. MA, USA: MIT Press.

Van der Plas, L., Apidianaki, M. & Chen, C. (2014). Global Methods for Cross-lingual Semantic Role and Predicate Labelling. *Proceedings of the 25th International Conference on Computational Linguistics (COLING)* (pp. 1279–1290). Dublin, Ireland. ACL.

Zarrieß, S. & Kuhn, J. (2009). Exploiting Translational Correspondences for Pattern-Independent MWE Identification. *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications* (pp. 23–30). Suntec, Singapore. ACL.

APPENDIX

Examples of multi-word expressions

The following table contains a sample of the adverb and prepositional Greek multi-word expressions we extracted and used in the parsing experiments.

Adverb MWEs		Prepositional MWEs	
ως εκ τούτου	αρκετά πια	κατά συνέπεια	κατά κεφαλήν
ούτω καθεξής	εκτός θέματος	κατά πόσον	κατά προτίμηση
εκτός αυτού	όλως αντιθέτως	εκ τούτου	επί μακρόν
ως αποτέλεσμα	μπροστά μας	μέχρι στιγμής	προ πάντων
ούτως ή άλλως	αμέσως τώρα	από κοινού	άνευ προηγουμένου
όχι μόνο	κάτω κάτω	έως ότου	υπό όρους
ευθύς εξαρχής	πάνω κάτω	εκ νέου	προ πολλού
σιγά σιγά	πόσω μάλλον	κατά δεύτερον	ανά χείρας
εντός ολίγου	ως γνωστόν	κατά κύριο λόγο	άνευ όρων
εκτός τούτου	ξανά και ξανά	παρά μόνο	αντί αυτού
όλως ιδιαιτέρως	ακόμη καλύτερα	μέχρις ότου	εξ αρχής
ευθύς αμέσως	ούτως ειπείν	προ ολίγου	εξ ονόματος
τώρα πια	ως δια μαγείας	εκ βάθρων	μετά βίας
σαν αποτέλεσμα	διόλου τυχαία	κατά μέσο όρο	μετά χαράς
πριν καν	έτσι κι αλλοιώς	κατά βάση	εκ πρώτης όψεως
πρώτα πρώτα	πρόσω ολοταχώς	κατά πρώτον	εξ αποστάσεως
έτσι κι αλλιώς	μαζί σας	εξ ολοκλήρου	συνεπεία τούτου
εκεί έξω	αποκλειστικά και μόνο	εξ ορισμού	επί τόπου
τώρα κιόλας	ακόμη και σήμερα	εις βάθος	άνευ σημασίας