Vasiliki Foufi, Luka Nerima and Eric Wehrli
WG1-WG2-WG3

Identification of Modern Greek collocations with a syntactic parser

This presentation will show some preliminary results obtained with the Greek version of our multilingual parser with respect to collocation identification. Modern Greek is a morphology-rich language, with case marking and a relatively free word order. In our grammar-based processing model, collocations, if present in the lexicon, are identified in the input sentence during the analysis of that sentence, as soon as the second term of the collocation is added to the syntactic structure. The grammar used for the computational modeling comprises rules and procedures. Attachment rules describe the conditions under which constituents can combine, while procedures compute properties such as long-distance dependencies, agreement, control properties, argument-structure building, and so on. In our approach, priority is given to parsing alternatives involving collocations, and hence collocational information helps the parser through the maze of alternatives. The integration of the collocation identification procedure in the parsing process leads to significant improvements of both parsing and collocation identification. Our system identifies a wide range of Greek collocations both nominal (ψυχρός πόλεμος 'cold war') and verbal (βγάζω συμπέρασμα 'to draw a conclusion'). The first step of this research concerned the enhancement of the grammar of Modern Greek used by the parser, and of its associated lexical database. In its current state, the lexicon used by the parser comprises 16,648 lexemes (corresponding to almost 155,000 inflected forms), and 20,982 collocations. Figure 1 shows the analysis of a sentence containing a verbal collocation (παίρνω μέρος 'to take part') and a nominal collocation (εκπαιδευτική άσκηση 'instructional exercise'):

Έπαιρνε	VERB-IND-IMP-3-SIN	1	παίρνω	**παίρνω μέρος**
μέρος	NOUN-SIN-NEU	9	μέρος	
χτες	NOUN-SIN-NEU	15	χτες	
το	DET-SIN-NEU	20	ο	
πρωί	NOUN-SIN-NEU	23	πρωί	
σε	PREP	28	σε	
εκπαιδευτική	ADJ-SIN-FEM	31	εκπαιδευτικός	
άσκηση	NOUN-SIN-FEM	44	άσκηση	**εκπαιδευτική άσκηση**
.	PUNC-point	50		

Figure 1. Identification of a verbal and a nominal collocation.

In our examples, column 1 displays the words of the sentence, column 2 the POS label, column 3 the position of the first letter of each word in the sentence (starting at the beginning of the input document), column 4 the citation form, and column 5 the collocation.

Collocations, and in particular verbal collocations, may present a fair amount of syntactic flexibility. Several researchers (for instance, Seretan, 2011) have pointed out that the constituents of a collocation can be far away from each other and not necessarily in their canonical order. The main advantage provided by a syntactic parser is its ability to identify collocations even when complex grammatical processes disturb the canonical order of constituents. In the example Έκκληση στους διοικητικούς υπαλλήλους να σταματήσουν την απεργία έκανε ο υπουργός 'An appeal to the administrative staff to stop the strike made the minister' (literal translation), the constituents of the Greek verb-object collocation κάνω έκκληση 'to make an appeal' do not follow the canonical verb-object order. Instead, the object έκκληση 'appeal' is placed in the beginning of the sentence long before the verb έκανε 'made':

Έκκληση	NOUN-SIN-FEM	1	έκκληση	
στους	PREP	9		
διοικητικούς	ADJ-PLU-MAS	15	διοικητικός	
υπαλλήλους	NOUN-PLU-MAS	28	υπάλληλος	διοικητικός υπάλληλος
να	CONJ-SUB	39	να	
σταματήσουν	VERB-SUB-AOR-3-PLU	42	σταματώ	
την	DET-SIN-FEM	54	ο	
απεργία	NOUN-SIN-FEM	58	απεργία	
έκανε	VERB-IND-IMP-3-SIN	66	κάνω	**κάνω έκκληση**
ο		72	ο	
υπουργός	NOUN-SIN-MAS	74	υπουργός	
.	PUNC-point	82		

Figure 2. Identification of the verbal collocation κάνω έκκληση 'to make an appeal'.

Our parser can also cope with long-distance dependencies, such as the ones found in *wh*-questions or relative clauses. As we can see below, the parser has correctly identified the collocation σπάζω το ρεκόρ 'to break the record' in the *wh*-question Ποιο ρεκόρ θέλει να σπάσει ο Μελισσανίδης; 'Which record does Melissanidis want to break?':

Ποιο	DET-SIN-NEU	1	ποιος	
ρεκόρ	NOUN-SIN-NEU	6	ρεκόρ	
θέλει	VERB-IND-PRE-3-SIN	12	θέλω	
να	CONJ-SUB	18	να	
σπάσει	VERB-SUB-AOR-3-SIN	21	σπάζω	**σπάζω το ρεκόρ**
ο	DET-SIN-MAS	28	ο	
Μελισσανίδης	NOUN	30	Μελισσανίδης	
;	PUNC-point-virgule	42		

Figure 3. Identification of the verbal collocation σπάζω το ρεκόρ 'to break the record'.

Passive is also a syntactic process which impacts the identification of verb-object collocations, turning the "deep" direct object –which takes accusative case– into the "surface" subject with nominative case. Figure 4 presents the analysis of the verb-object collocation χάνω χρόνο 'to waste time' in its passive form:

με	PREP	1	με	
αποτέλεσμα	NOUN-SIN-NEU	4	αποτέλεσμα	
να	CONJ-SUB	15	να	
χαθεί	VERB-SUB-AOR-PASSIVE-3-SIN	18	χάνω	**χάνω χρόνο**
πολύτιμος	ADJ-SIN-MAS	24	πολύτιμος	
χρόνος	NOUN-SIN-MAS	34	χρόνος	

Figure 4. Identification of the verb-object collocation χάνω χρόνο 'to waste time' in its passive form.

Our system recognizes even more complex structures. The phrase Είχαν υποβάλει ερωτήσεις στον επίτροπο, αλλά κανένας δεν βρισκόταν εδώ για να τις θέσει 'They had posed questions to the commissioner, but there was nobody here to make them' comprises two verb-object collocations similar in meaning but with a different verb υποβάλλω ερώτηση 'to pose a question' and θέτω ερώτηση 'to make a question'. The parser has successfully identified both collocations: in the first collocation, the verb υποβάλλω 'to pose' is followed by the direct object (DO) ερωτήσεις 'questions' while in the second collocation, the DO of the verb θέτω 'to make' is the pronoun τις 'them' that refers to the noun ερωτήσεις 'questions':

Είχαν	AUX-VERB-IND-IMP-3-PLU	1	έχω		
υποβάλει	VERB-SUB-AOR-3-SIN	7	υποβάλλω	SU:pro**DO:ερωτήσεις** PO:επίτροπο	**υποβάλλω ερώτηση**
ερωτήσεις	NOUN-PLU-FEM	16	ερώτηση		
στον	PREP	26		PO	
επίτροπο	NOUN-SIN-MAS	31	επίτροπος		
,	PUNC-virgule	39			
αλλά	CONJ-COO	41	αλλά		
κανένας	PRO-IND-SIN-MAS	46	κανείς	SU	
δεν	adverbe-NEG	54	δεν		
βρισκόταν	VERB-IND-IMP-PASSIVE-3-SIN	58	βρίσκω	SU:κανένας	
εδώ	ADV-LOC	68	εδώ		
για να	CONJ-SUB	72	για να		
τις	PRO-CLI-PLU-FEM	1	του		
θέσει	VERB-SUB-AOR-3-SIN	83	θέτω	**DO:τις**	**θέτω ερώτηση**
.	PUNC-point	88			

Figure 5. Identification of the collocations υποβάλλω ερώτηση 'pose a question' and θέτω ερώτηση 'make a question'.

We have measured the performance of our parser to identify collocations that are present in the lexical database and the impact of the collocation knowledge on the performance of the parser (in percentage of complete analyses). To achieve the evaluation, we took a small newspaper corpus of about 20,000 words and we manually identified 638 collocations (both nominal and verbal). We ran the parser twice on the corpus: the first time before and the second time after enrichment of the collocation database. On the first run, the parser achieved 43.26% of complete analyses and identified 124 collocations. On the second run, after enrichment of the lexicon, the percentage of complete analyses increased to 44.33% and three quarters of the corpus collocations were identified (482/638). Thus, over this small corpus, the parser achieved a 100% precision in the collocation identification task, with a recall of 75.54% and F-measure of 0.86. The collocations that were not identified (156 out of 638), were part of sentences for which the parser did not achieve a complete analysis.

References:

Heid Ulrich. 1994. On ways words work together – research topics in lexical combinatorics. *Proceedings of the 6th Euralex International Congress on Lexicography (EURALEX '94)*: 226-257.

Sag A. Ivan., Baldwin Timothy, Bond Francis, Copestake Ann and Dan Flickinger. 2002. Multi-word Expressions: A Pain in the Neck for NLP. *LinGO Working Paper 2001-01*, Stanford University, CA.

Seretan Violeta. 2011. *Syntax-Based Collocation Extraction*. Springer Verlag.