Parsing Modern Greek verb MWEs with LFG/XLE grammars

Niki Samaridi
National and Kapodistrian University of Athens
nsamaridi@gmail.com

Stella Markantonatou
Institute for Language and Speech Processing/ 'Athena' RIC
marks@ilsp.athena-innovation.gr

Abstract

We report on the first, still on-going effort to integrate verb MWEs in an LFG grammar of Modern Greek (MG). Text is lemmatized and tagged with the ILSP FBT Tagger and is fed to a MWE filter that marks Words_With_Spaces in MWEs. The output is then formatted to feed an LFG/XLE grammar that has been developed independently. So far we have identified and classified about 2500 MWEs, and have processed 40% of them by manipulating only the lexicon and not the rules of the grammar.

Research on MG MWEs (indicatively, Anastasiadi-Simeonidi, 1986; Fotopoulou, 1993; Mini et al., 2011) has developed collections of MWEs and discussed classification, syntax and semantics issues. To the best of our knowledge, this is the first attempt to obtain deep parses of a wide range of types of MG verb MWEs with rich syntactic structure.

1 The parsing system

We take advantage of the mature ILSP FBT Tagger (Papageorgiou et al., 2000) that is an adaptation of the Brill tagger trained on MG text. It uses a PAROLE compatible tagset of 584 tags. The tagger works on the output of a sentence detection and tokenisation tool and assigns a lemma and a set of tags corresponding to an exhaustive morphological analysis of tokens (Fig. 1). The tagger is a black box for our system and allows for no preprocessing of MWEs, as it would be possible if the XFST/XLE component was used (Attia, 2006). We have been working on a system that aims to move as much as possible of the parsing burden from the LFG/XLE component to a MWE recognizer (the 'filter') at the same time allowing for 'natural' LFG analyses. Oflazer et al. (2004) discuss a similar preprocessing step but they focus on the retrieval and not on the deep parsing of verb MWEs. Our filter, implemented in Perl, scans the output of the tagger for strings containing verb MWEs and feeds a script ('formatter') that yields a format readable by an LFG/XLE grammar.

1.1 The filter lexicon

The filter consults the 'filter lexicon' where each verb MWE entry is specified for the following:

1. Compositionality. Certain verb MWEs can take a compositional interpretation. For instance, the free subject, flexible (Sag et al, 2001) verbal

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 33–37,
Gothenburg, Sweden, 26-27 April 2014.

MWE κάνω μαύρα μάτια να σε δω (9) has no compositional interpretation while the semi-fixed MWE τις_αρπάζω (2) "to be beaten up", can take the compositional interpretation "grab/steal them-FEM". The filter lexicon specifies which MWEs will be eventually assigned both MWE and compositional XLE parses.

2. The lemmatized form of Words_With_Spaces (WWS) whether they are independent fixed MWEs or substrings of a MWE. For instance, the lemmatised WWS μαύρος_μάτι would be stored for WWS μαύρα μάτια of the MWE (9).

3. PoS of the WWS. For instance, we have classified the WWS ταπί-και-**ψύχραιμος** 'penniless and calm'(6) as adjective; however, only the second conjunct (ψύχραιμος 'calm') is an adjective while the first conjunct ταπί is an indeclinable non-Greek word that occurs with this type of MWE only. Regarding distribution, the conjunction behaves as an adjective. In general, we have relied on distribution criteria in order to assign PoS to WWSs.

4. Morphological constraints on the lemmatized constituents of a WWS that uniquely identify fixed or semi-fixed MWE substrings. For instance, for the adjective μαύρα in the WWS μαύρα μάτια (9) the lemma of the adjective μαύρος is stored together with the tags adjective-plural-accusative-neutral-basic.

5. Multiple WWSs if different word orders of the same WWS occur, for instance πίνει [το αίμα του κοσμάκη]$_{WWS}$ [gloss: drink the blood of people] and πίνει [του κοσμάκη το αίμα]$_{WWS}$ 'takes a lot of money from people by applying force'.

1.2 The filter

The filter, implemented in Perl, reads the tagged sentence from an xml file (the output of the tagger), checks it for MWEs and feeds it to the formatter if no MWE or a MWE that can take a compositional interpretation is found. Strings containing MWEs are preprocessed by the filter: their fixed parts are replaced with the corresponding WWS and morphological constraints and the resulting new string is sent to the formatter. The filter can identify all word permutations available to a listed MWE.

2 An outline of the LFG analysis

The output of the formatter is parsed with an LFG grammar of MG. The grammar includes sublexical rules that parse the output of the tagger and ensure information flow from the tagger to XLE. The sub-lexical trees can be seen in the c-structure of Fig. 1. MG MWEs are rich in syntactic structure despite any simplifications that might result from the usage of WWSs. In agreement with Gross (1998a; 1998b) and Mini et al. (2011) who argue that MWEs and compositional structures can be treated with more or less the same grammar, we have so far manipulated only the lexicon but not the grammar rules. Identification of phrasal constituents within the MWEs relies on possible permutations and the ability of XPs to intervene between two words, thus indicating the border between two constituents. Grammatical functions are identified with diagnostics that apply to compositional expressions such as morphological marking and WH questions. The types of syntactic structure we have treated thus far are:

1. **Fixed verb WWS** (Table 1:1): no inflection or word permutation.

(1) πάρε πέντε
take-2-sg-IMP five-numeral
'You are silly.'

2. **Free subject-verb** (Table 1:2): inflecting, SV/VS word order.

(2) Ο Πέτρος τις άρπαξε
the Peter-nom CL-pl-fem-acc grab-3-sg-past
'Petros was beaten up.'

3&4. **Impersonal verb-complement**: inflecting, fixed object (Table 1:3) or saturated sentential subject (Table 1:4), intervening XPs between the verb and its object or subject, VO/OV word order (but not VS/SV).

(3) Έριξε καρεκλοπόδαρα χθες.
pour-3-sg-past chair-legs yesterday
'It rained cats and dogs yesterday.'

(4) Έχει γούστο να βρέξει.
have-3-sg-pres gusto-noun to rain
'Don't tell me that it might rain.'

	LFG representation	Sub-WWS	C
1	V: PRED παίρνω_πέντε		Y
2	V: PRED εγώ_**αρπάζω** <SUBJ >		Y
3	V: PRED **ρίχνω** <SUBJ,OBJ>, OBJ PRED= καρεκλοπόδαρο		N
4	V: PRED **έχω**_γούστο<SUBJ>, SUBJ COMPL=να	**έχω**_γούστο	N
5	V: PRED **μένω** <SUBJ,XCOMP>, XCOMP PRED=στήλη_άλας, XCOMP SUBJ=SUBJ	στήλη_άλας	N
6	V: PRED **μένω**< SUBJ,XCOMP>, XCOMP PRED=ταπί-και-**ψύχραιμος**, XCOMP SUBJ=SUBJ	ταπί_και **_ψύχραιμος**	N
7	V: PRED **τρώω/αρπάζω**<SUBJ,OBJ>, OBJ PRED=ο _ξύλο_ο _χρονιά, OBJ POSS PRED= **εγώ**, OBJ POSS TYPE= weak pronoun, OBJ POSS PERSON/NUMBER/GENDER =SUBJ PERSON/NUMBER/GENDER	ο_ξύλο_ο _χρονιά	N
8	V: PRED **ρίχνω** <SUBJ, OBJ, XCOMP>, XCOMP COMPL= να, OBJ PRED=άδειος, XCOMP PRED= **πιάνω**_γεμάτος, XCOMP SUBJ=SUBJ, XCOMP PERF=+, ¬(XCOMP TENSE)	**πιάνω** **_**γεμάτος	N
9	V: PRED **κάνω** <SUBJ, OBJ, XCOMP>, XCOMP COMPL=να, OBJ PRED= μαύρος_μάτι, XCOMP PRED=**βλέπω** <SUBJ, OBJ>, OBJ PRED=**εγώ,** XCOMP SUBJ=SUBJ, XCOMP PERF=+, ¬(XCOMP TENSE)	μαύρος_μάτι	N
10	V: PRED **τραβώ**<SUBJ, OBJ>, OBJ PRED= ο_λινάρι_ο_πάθος	ο_λινάρι ο_πάθος	N

Table 1. LFG analysis of MG verb MWEs used in this text. Boldfaced words inflect within the MWE. C: compositional. Only lemmatised forms are given.

5&6. **Free subject-copula-complement**: inflecting copula, complement fixed (Table 1:5), intervening XPs between the subject and the verb or between the copula and the complement, constituent permutations.

(5) Μένει η Ρέα στήλη άλατος
be-left-3-sg-pres the Rea-nom stele-of-salt
'Rea was left speechless.'

Alternatively, the complement may inflect (Table 1:6) and agree with the free subject.

(6) Και μένει η Ρέα
and be-left3-sg-pres the Rea-sg-fem-nom
ταπί και ψύχραιμη
penniless and calm-sg-fem-nom
'Rea lost all her money.'

7. **Free subject-verb-fixed object with subject bound possessive** (Table 1:7): inflecting verb, object modified with a subject bound possessive weak pronoun, intervening XPs between the object and the verb, constituent permutations.

(7) έφαγε/άρπαξε η Ρέα$_j$ το
eat/grab-3-sg-past the Rea-nom the
ξύλο της χρονιάς της$_j$
beating the year-gen weak-pron-fem-gen
'Rea was beaten up.'

8&9. **Free subject (controller)-verb-object-subordinated clause with controlled subject**: inflecting verb, object possibly fixed (Table 1: 9), the subordinated clause possibly semi-fixed (Table 1:8), intervening XPs, VSO/OVS word orders.

(8) Έριξαν άδεια να πιάσουν γεμάτα
throw-3-pl-past empty to catch-3-pl full
'They tried to obtain information.'

(9) έκανε η μάνα του
make-3-sg-past the mother-sg-nom his$_j$
μαύρα μάτια να τον δει
black eyes to him$_j$ see-3-sg
'It took his mother a long time to meet him.'

The transitive verb ρίχνω "throw" (8) is used as a control verb only in (8). An alternative analysis that would insure identity of subjects could treat the exemplified MWE as a coordination structure. We opted for the control approach and defined a special entry of the verb ρίχνω "throw" because the particle να typically introduces

(probably controlled) subordinated clauses and the constraints on verbal forms are those of να-subordination and not of coordination.

10. **Free subject-verb-object** (Table 1:10): inflecting verb, fixed or non-fixed object, intervening XPs and OVS/VOS word order.

(10) Οι άνθρωποι τράβηξαν τότε
the people-pl-nom pull-3-pl-past then
του λιναριού τα πάθη
the linen the sufferings
'People suffered a lot then.'

3 Conclusions and future research

This is ongoing work but up to this point, natural analyses of the verb MWEs are possible with the standing rule component of our LFG grammar of MG. On the other hand, the entries of the two lexica we have developed, namely the filter and the XLE lexicon, provide a rich resource for studying the features of the idiomaticity that verb MWEs bring into 'normal' MG (indicatively, see discussion of (8)). In the immediate future, we will use the same methodology to parse the remaining types of MWE in our collection and will draw on the accumulated evidence to study the linguistic phenomena observed in verb MWEs against more general semasio-syntactic properties of MG, for instance the role of control constructions and of animacy in this language. We will consider a more sophisticated design of the filter. Last, we plan to investigate the issue of semantic representation of MWEs.

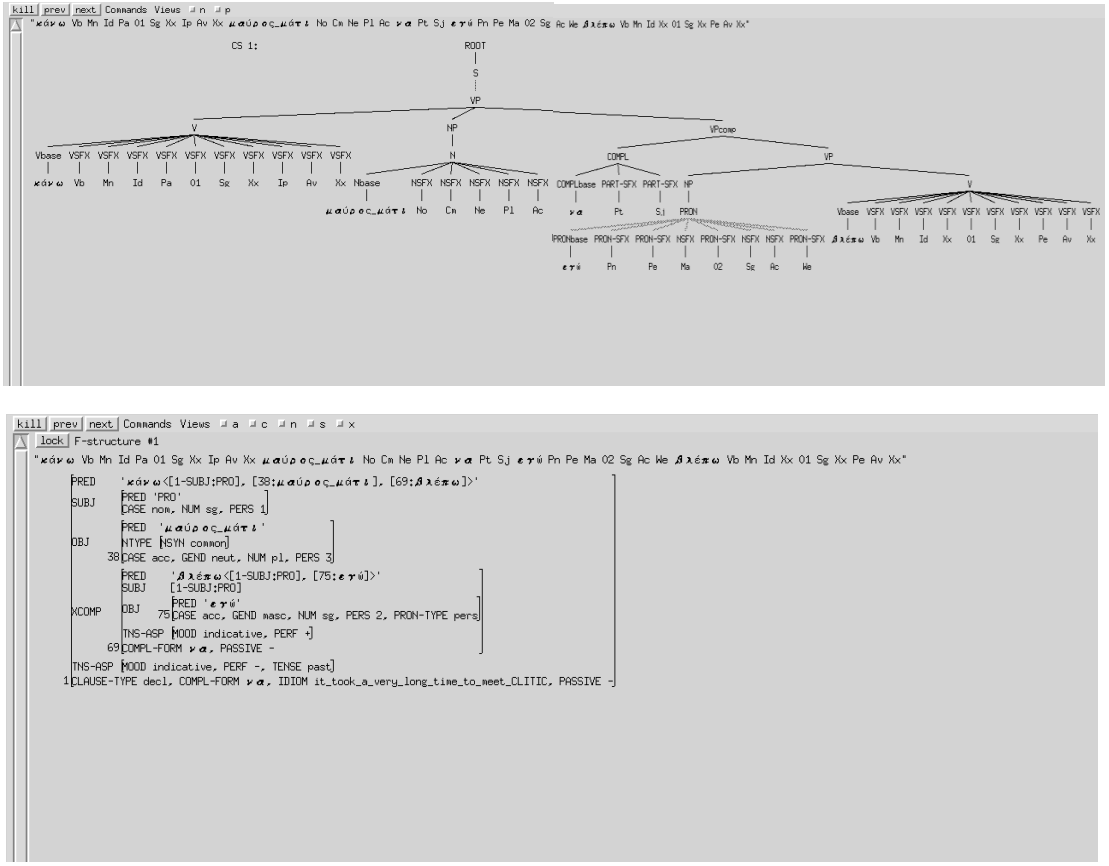

Fig. 1. The XLE output for the flexilbe verb MWE έκανα μαύρα μάτια να σε δω (Table 1: 9).

Acknowledgements

We thank Tracy Holloway-King for her contribution to the development of the sublexical rules.

References

Αναστασιάδη-Συμεωνίδη, Άννα. 1986. *Η Νεολογία στην Κοινή Νεοελληνική*, Θεσσαλονίκη. ΕΕΦΣ του ΑΠΘ, Παράρτημα αρ. 65.

Attia, Mohammed A. 2006. Accommodating Multiword Expressions in an Arabic LFG Grammar. Salakoski, Tapio, Ginter, Filip, Pahikkala, Tapio, Pyysalo, Tampo: *Lecture Notes in Computer Science: Advances in Natural Language Processing, 5th International Conference, FinTAL*. Turku, Finland. Vol. 4139: 87-98. Springer-Verlag Berlin Heidelberg.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Balwin, Ivan A. Sag, and Dan Flickinger. 2002. Multiword expressions: linguistic precision and reusability. *Proceedings of the 3rd International Conference on Language Resources and Evaluation*. Las Palmas, Canary Islands.

Fotopoulou, Aggeliki. 1993. *Une Classification des Phrases a Complements Figes en Grec Moderne*. Doctoral Thesis, Universite Paris VIII.

Gross, Maurice. 1988a. Les limites de la phrase figée. *Langage* 90: 7-23.

Gross, Maurice. 1988b. Sur les phrases figées complexes du français. *Langue française* 77: 47-70.

Mini, Marianna, Kleopatra Diakogiorgi and Aggeliki Fotopoulou. 2011. What can children tell us about idiomatic phrases' fixedness: the psycholinguistic relevance of a linguistic model. *DISCOURS (Revue de linguistique, psycholinguistique et informatique)(9*).

Oflazer, Kemal, Ozlem Cetinoglu and Bilge Say. 2004. Integrating Morphology with Mutli-word Expression Processing in Turkish. *Second ACL Workshop on Multiword Expressions: Integrating Processing*: 64-71.

Papageorgiou, Haris, Prokopis Prokopidis, Voula Giouli and Stelios Piperidis. 2000. A Unified POS Tagging Architecture and its Application to Greek. *Proceedings of the 2nd Language Resources and Evaluation Conference*. Athens.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. LinGO Working Paper No. 2001-03. In Alexander Gelbukh, ed., (2002) *Proceedings of CICLING-2002*. Springer.

ILSP FBT Tagger http://lrt.clarin.eu/tools/ilsp-feature-based-multi-tiered-pos-tagger

XLE documentantion http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html

References used for the development of the filter:

http://interoperating.info/courses/perl4data/node/26
http://stackoverflow.com/questions/2970845/how-to-parse-multi-record-xml-file-ues-xmlsimple-in-perl
http://stackoverflow.com/questions/2039143/how-can-i-access-attributes-and-elements-from-xmllibxml-in-perl
http://stackoverflow.com/questions/7041719/using-perl-xmllibxml-to-parse
http://stackoverflow.com/questions/10404152/perl-script-to-parse-xml-using-xmllibxml
http://www.perlmonks.org/index.pl?node_id=490846
http://lethain.com/xml-simple-for-non-perlers/

Perl:
http://perldoc.perl.org/perlintro.html
http://learn.perl.org/
http://qntm.org/files/perl/perl.html
http://www.perl.org/books/beginning-perl/
http://www.it.uom.gr/project/perl/win32perltut.html

http://www.comp.leeds.ac.uk/Perl/sandtr.html
http://www.troubleshooters.com/codecorn/littperl/perlreg.htm
http://www.cs.tut.fi/~jkorpela/perl/regexp.html
http://www.somacon.com/p127.php
http://perlmaven.com/splice-to-slice-and-dice-arrays-in-perl
http://www.perlmonks.org/?node_id=822947
http://www.perlmonks.org/?node_id=911102