Parsing Modern Greek verb MWEs in LFG/XLE

Niki Samaridi[1] and Stella Markantonatou[2]

[1]National and Kapodistrian University of Athens, nsamaridi@gmail.com

[2]Institute for Language and Speech Processing/ "Athena" RIC, marks@ilsp.gr

WP2

1 Introduction

Drawing on previous research on Modern Greek (MG) MWEs (Fotopoulou, 1993), we are working on the integration of MWE parsing into an LFG grammar of MG. So far we have identified and classified 2500 MWEs, while our system is able to process 40% of them.

2 The architecture of the parsing system

Lemmatization and morphological analysis is performed with the ILSP FBT Tagger that uses a PAROLE compatible tagset of 584 different tags. The tagger is a black box that returns lemmata and their tags and does not recognise MWEs. We developed a MWE recogniser (the 'filter') that filters the output of the tagger for strings containing MWEs and feeds a script ('formatter') that transforms the output to a format readable by an LFG/XLE grammar.

2.1 The filter lexicon

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

1. Compositionality. Which MWEs can take a compositional interpretation.

2. The 'signifier': the lemma of the substring of a flexible MWE that instructs the filter to look at the appropriate filter lexicon entries. For the flexible MWE πίνω το αμίλητο νερό "to drink the non-speaking water"= to be silent, the signifier is the lemma πίνω. If the expression is fixed, the symbol '~' is listed instead.

3. The lemmatised form of Words_With_Spaces (WWS) (Sag et al., 2001) whether they are independent fixed MWEs or substrings of a MWE. For instance, the lemmatised WWS ο_αμίλητος_νερό would be stored for the underlined substring of the MWE πίνω <u>το αμίλητο νερό</u>.

4. PoS and morphological constraints for the headword of a WWS. The lemmatised headword of the WWS ο_αμίλητος_νερό is νερό and is tagged as common_noun-neutral-singular-accusative.

5. Constraints on the lemmatised forms of the remaining constituents of a WWS (apart from its headword) that uniquely identify fixed or semifixed MWE substrings. For instance, the form αμίλητο is the accusative-singular-neutral-basic form of the adjective αμίλητος.

2.2 The filter

The filter, implemented in Perl, reads the tagged sentence from an xml file (the output of the tagger) and stores it. Then, it:

A. Checks whether a signifier exists and,

A1. If no signifier is found, the string is copied as it is to the formatter.

A2. If a signifier is found, the filter lexicon is scanned for some WWS entry. The filter checks whether the morphological constraints on the filter lexicon entries (headword and remaining words) match the lemma and the tags on the input string and:

B1. If they do not match, the input string is copied as it is to the formatter.

B2. Else, the filter lexicon is consulted if the MWE can take a compositional reading and

C1. If it can, it sends the input string to the formatter and moves to step C2

C2. Else, the part of the string that has been recognised is replaced with the corresponding WWS and morphological constraints and the resulting new string is sent to the formatter.

3 An outline of the LFG analysis

Fixed MWEs allow for no morphological modification, no intervening words, no word order variations, no alternations (eg. passivisation). For these MWEs the filter returns a Verb WWS.

Semi fixed and flexible verb MWEs inflect and allow for the identification of constituents and grammatical functions on the basis of word order permutations, whether an XP can occur within the MWE, morphology and wh-questions. The structures treated so far are: **Free subject-verb**: inflected verb, **Impersonal verb:** inflected impersonal verb with a fixed object or a saturated sentential complement, **Free subject-copula-complement**: the complement may be fixed or inflected in which case it fully agrees with the subject, **Free subject-verb-fixed object with subject-bound possessive**: the object is modified by a possessive weak pronoun bound by the subject, **Free subject (controller)-verb-object subordinated sentence with controlled subject**: the object may be fixed and/or the subordinated sentence may be semi-fixed and **Free subject-verb-object** (Fig. 1): fixed or non-fixed object.

4 Conclusions and future research

We have obtained natural analyses of the MG MWEs with the existing machinery of LFG. Using the same basic method, we will parse the remaining MWEs in our collection and will consider both a more sophisticated engineering of the filter and the issue of semantics.

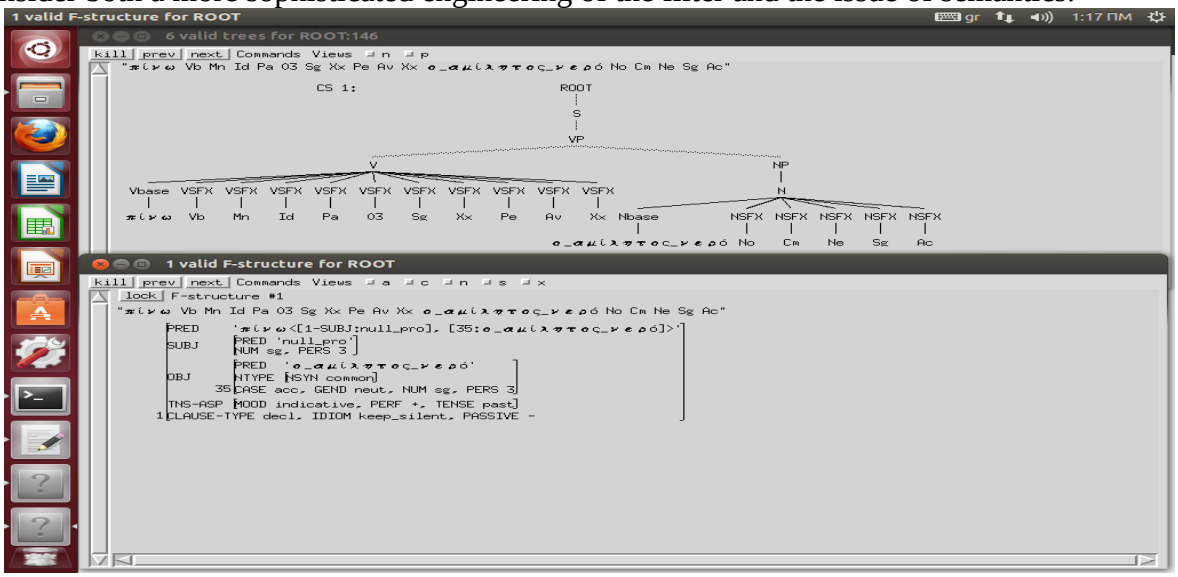

Fig. 1. The XLE output for the free subject-verb-fixed object flexible MWE πίνω το αμίλητο νερό.

References

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.

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

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2001. 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