Chapter 4

Issues in parsing MWEs in an LFG/XLE framework

Stella Markantonatou
Institute for Language and Speech Processing/Athena RIC

Niki Samaridi
Institute for Language and Speech Processing/Athena RIC

Panagiotis Minos
Institute for Language and Speech Processing/Athena RIC

We present an LFG/XLE system coupled with an independent lexicographic environment for encoding and parsing Modern Greek MWEs. The system assigns a flat structure to the fixed sequences of words within MWEs, the so-called "words with spaces" (WWSs) with the help of a preprocessing module that receives the morphologically analysed string from a tagger external to XLE. We describe the overall system and discuss certain implications of the designing choices.

1 Introduction

This paper presents the system for parsing Modern Greek (MG) Multiword Expressions (MWEs) with LFG/XLE grammars that is schematically depicted in Figure 1 and discusses the issues encountered with the LFG/XLE representations. The main idea of the adopted parsing strategy is that the parser treats the sequential fixed parts of the MWEs as a type of "words with spaces" (WWS) (Sag et al. 2002). Our WWSs are fixed sequences of fixed words that may contain one word that declines (for instance, see example 7 in Table 1). The rigid word order is an important criterion of fixedness in the case of MG that has a relatively free word order. Morphological fixedness is also important in a language with rich

Stella Markantonatou, Niki Samaridi & Panagiotis Minos. 2019. Issues in parsing MWEs in an LFG/XLE framework. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 109–126. Berlin: Language Science Press. DOI:10.5281/zenodo.2579039

morphology but, exactly for the same reason, the existence of an inflected word within an otherwise rigid structure is not a surprise. The usage of (this type of) WWSs has practical and theoretical implications.

WWSs have been used by Copestake et al. (2002), by Attia (2006) for parsing Arabic MWEs with LFG grammars, by Korkontzelos & Manandhar (2010) for shallow parsing and was recently shown to be beneficial for a transition-based dependency grammar parser of Modern Greek (Apidianaki et al. 2018). We have adopted the WWS approach in an effort to move as much as possible of the parsing burden from the LFG/XLE component to an external MWE recognizer (the "filter" from now on). At the same time, we have tried to allow for natural LFG analyses. The system depicted in Figure 1 consists of:

1. The ILSP FBT Tagger
2. IDION: A lexicographic tool that allows for formal descriptions of the MWEs
3. The filter
4. The XLE/LFG grammars

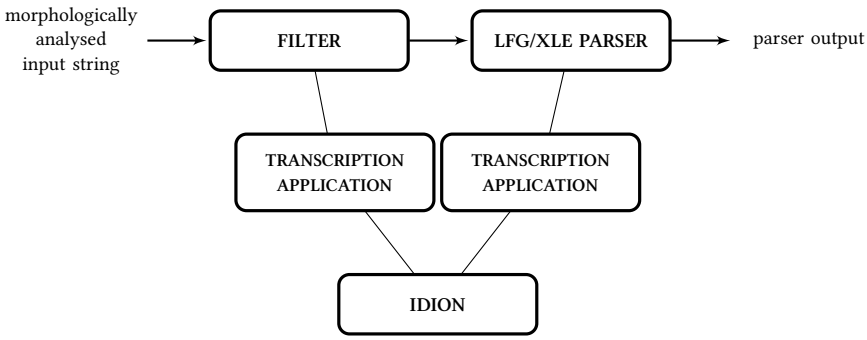

Figure 1: The overall structure of the parsing system

The ILSP FBT Tagger and IDION are independent pieces of NLP software; they are compatible with the "core" parsing system that consists of the filter and the grammars (Samaridi & Markantonatou 2014). In what follows, we describe the parts 1–4 in separate sections in this order. We will use (1) as a working example. (1) is a verb MWE that contains a fixed NP *mavra matia* 'black eyes' and an obligatory sentential complement that is controlled by the MWE subject. The subject is free and fully agrees with the verb of the MWE (MG is a pro-drop language therefore in (1) no explicit subject is present):

(1) kano mavra matia na do kapion / kati
make.1SG black.ACC.PL eye.ACC.PL to see.1SG.PERF someone something
'I have not seen someone/something for a long time.'

2 The LFG analysis adopted: Challenging options

It has already been stated that the main idea of the adopted parsing strategy is that the fixed parts of the MWEs are treated as "words with spaces" (WWSs) (Sag et al. 2002). WWSs are used only if an MWE contains fixed sequences of words; the WWS stands only for the fixed sequence and not for the whole MWE – if the remaining MWE is flexible. The fixed sequences are identified with diagnostics involving word order permutations, the ability to introduce an XP between words and diathesis alternations (if applicable). As an example, in (1) there is the WWS *mavra_matia* 'black eyes'. The sequence *mavra_matia* is morphologically and syntactically fixed, it can be moved to the beginning of a sentence in emphatic structures and it accepts neither a determiner nor modification. The remaining parts of the MWE in (1), with the exception of certain morphological constraints on the subordinated verb, behave like the parts of a compositional structure and are treated as such.

The LFG/XLE lexicon has to recognize the WWSs as words that are assigned some part of speech (PoS) value. However, the selection of the PoS value is not always straightforward with MWEs, all the more when no WWS occurs in the MWE. Examples (2–4) illustrate the issue (the identified WWSs are in square brackets "[]"). We often find nouns functioning as adverbs; in (2) the NP headed by *zachari* 'sugar' is normally questioned with *how much*. Furthermore, the WWS in (2) could be analysed as a syntactic complex, consisting of an "object" clitic and a verb; clitics are used widely in MG. We treat this complex as a fully inflected verb. The WWS in (3) could have been generated with the rule NP → Det N; given that the head is a common noun (*dromous* 'roads') probably the PoS tag "N" is a natural choice for the WWS *tous dromous* 'the roads'. In (4), the WWS is a fixed sequence of fixed words that behaves exactly as the WWS in (3) with respect to word order phenomena (4a,b) and unlike the corresponding compositional copula structures of MG (4c,d). However, there is no phrase structure rule that would generate the WWS *to_psomi_psomaki* 'the bread little-bread' and of course, there is no likely head.

(2) [tin pernao] zachari
her.ACC.FEM pass.1ST sugar.ACC
'I have an easy time.'

(3) perno [tous dromous]
take the roads.ACC
'I wander'

(4) a. leo [to psomi psomaki]
call the bread.ACC little-bread.ACC
'to starve'
b. to psomi psomaki leo (emphatic)
c. * to psomi leo psomaki (emphatic)
d. * psomaki leo to psomi (emphatic)

In addition, the identification of the syntactic function of the fixed parts of verb MWEs is not straightforward in LFG. This is so because the governable grammatical functions (GFs) of LFG[1] are defined on the basis of particular semantic and syntactic properties (Dalrymple 2001). Alas it is very often the case that the fixed parts of MG MWEs are not characterized by these particular properties. And still, one cannot avoid using a large choice of grammatical functions to model MG MWE phenomena because the language allows for some word order flexibility within verbal MWEs (4a,b) and often there are control (1) and binding phenomena (5) that have to be accounted for. LFG models these phenomena on the f-structure with the use of syntactic functions. (In (5) the WWS *to ksilo tis chronias tis* 'the beating the.GEN year.GEN hers' can be thought to have a noun head *ksilo* 'beating'; the structure contains a possessive pronoun that is bound by the free subject of the MWE.)

(5) I Maria efage [to ksilo tis chronias tis / *tou]
the Maria.FEM ate the beating the year.GEN hers / *his
'Maria has been beaten up.'

The OBJ function makes a good example of a GF that does not fit well to the MWE data. The WWS *tous_dromous* 'the roads' in (3) is a fixed simple NP; one would be tempted to assign the OBJ function to it but, on the other hand, the fixed NP never turns up as the subject of a passive form although the verb *perno* 'take' passivises. Furthermore, the WWS in (3) presents an idiosyncratic behavior with clitics; normally it cannot be replaced by a clitic, while this is absolutely possible in a compositional structure; the fixed NP can be replaced only in a very

[1]The governable GFs of LFG are: SUBJ, OBJ, OBJ2, POSS, COMP, XCOMP.

restricted context, namely when the same MWE precedes the structure with the clitic (Markantonatou & Samaridi 2018) producing an ironic or emphatic effect. Passivisation is a defining property of the OBJ GF in LFG (Dalrymple 2001) and free replacement by a clitic is definitely a defining property of objects in MG. On the other hand, the WWS in (4) behaves just as the WWS in (3) with respect to passivisation and cliticisation and all the other flexibility diagnostics; evidence mandates that the two WWSs are assigned the same GF and the question is whether they should be assigned the OBJ GF or some other GF. It is possible that the idea that MWEs use exactly the syntax employed in the analysis of compositional structures (Gross 1988a,b; Kay & Sag 2012; Bargmann & Sailer 2018) could be imported in LFG and the classical GFs could be assigned to fixed constituents along with a tree-like structure and constraints on inflection, passivisation, modifiability, cliticisation and linear precedence that do the job (Waszczuk & Savary 2015). The problem with the "compositional structure" approach is that it questions the notion of syntactic functions and the generalizations expressed with them: for instance, the OBJs of MG MWEs will be peculiar in that they hardly passivise and they are not replaced by clitics freely unless they occur in highly constrained contexts.

The system we present here uses the classical LFG GFs. This means that *zachari* 'sugar' in (2) is treated as a noun and the phrasal projection is assigned the OBL(ique) GF; on the same par, the bracketed strings in (3), (4) and (5) are assigned the PoS "No"(un) and project NPs that are assigned the OBJ GF. So far we have not used a set of GFs different from the one established in the literature because linear precedence phenomena in the fixed parts are captured with the use of WWSs and modifiability and cliticisation seem to require a more careful modeling than simply allowing or prohibiting them: cliticisation heavily depends on the context and modifiability seems to be rather restricted in MG. A concrete, corpus-based, analysis of both the phenomena has not been made available yet, to the best of our knowledge. This set-up demands that passivisation is blocked with a feature (and not with the absence of an OBJ GF as it would be the case if some other GF was used in the place of the OBJ GF). Of course, a similar blocking feature would be used in the grammar anyway for several non-passivisable transitive verbs of MG MWEs; this fact definitely emphasizes the problematic situation with the OBJ GF and passivisation. In a nutshell, we have used the OBJ GF not because it served our purposes well but because the in-depth exploration of the alternatives is considered a future challenge.

In the remainder of this document we will present and discuss the parts of the system as they are depicted in Figure 1.

3 ILSP FBT Tagger

The mature ILSP FBT Tagger (Papageorgiou et al. 2000) is an adaptation of the Brill tagger trained on MG text. It uses a PAROLE compatible tagset (Bilgram & Keson 1998) of 584 different tags that capture the morphological particularities of MG. The tagger works on the output of a sentence detection and tokenisation tool and assigns both a lemma and a set of tags corresponding to an exhaustive morphological analysis of each token. Figure 2 shows the output of the ILSP FBT Tagger for (1). We decided to use the ILSP FBT Tagger because the effort to develop an XFST morphological component is a project on its own. In the set-up of Figure 1, the tagger is a black box that allows for no identification of the fixed parts of MWEs at the level of morphological analysis, as it would be possible if, for instance, the XFST/XLE component was used as in Attia (2006). For this reason, the morphologically analysed output of the ILSP tagger that offers information only about tokens, is processed with a filter (Samaridi & Markantonatou 2014) that scans 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; the filter informs the XLE parser whether an MWE exists, whether it contains any WWSs – if so, the WWSs are marked on the output string that feeds the parser – and whether the input string can receive both a compositional and a MWE interpretation.

4 IDION

The XLE parser receives lexical knowledge on MWEs from IDION[2], an open source lexicographic environment for MWEs that is addressed both to the human user and to NLP applications and encodes, among others, morphosyntactic properties of MWEs in a, as much as possible, theory-neutral formalism. IDION is connected to the parsing system with an application that transcribes the IDION formalism to the XLE formalism (Minos et al. 2016). As opposed to other MWE DBs, such as DUELME (Grégoire 2010), that use a simplified formal language for encoding morphological features, IDION exhaustively describes morphological features with the ILSP-PAROLE compatible tagset that is also used by the ILSP FBT Tagger.

It is important to note that syntactic functions are assigned to phrasal constituents in Modern Greek (and not to parts of a word); therefore, diagnostics for constituent identification are also required along with diagnostics for the

[2]http://idion.ilsp.gr/

```
<cesDoc version="0.4">
  <cesHeader version="0.4"/>
    <text>
      <body>
        <p id="p1">
          <s id="s1" casing="lowercase">
          <t id="t1" word="έκανα" tag="VbMnIdPa01SgXxIpAvXx" lemma="κάνω"/>
          <t id="t " word="μαύρα" tag="AjBaNePlAc" lemma="μαύρος"/>
          <t id="t3" word="μάτια" tag="NoCmNePlAc" lemma="μάτι"/>
          <t id="t4" word="να" tag="PtSj" lemma="να"/>
          <t id="t5" word="τον" tag="PnPeMa03SgAcWe" lemma="εγώ"/>
          <t id="t6" word="δω" tag="VbMnIdXx01SgXxPeAvXx"lemma="βλέπω"/>
        </s>
      </p>
    </body>
  </text>
</cesDoc>
```

Figure 2: The output of the ILSP FBT tagger for the verb MWE in (1)

identification of WWSs. In IDION the following diagnostics are used for these purposes (Markantonatou & Samaridi 2018): possible word order permutations, the ability of XPs (modifiers included) to intervene between two words thus possibly indicating the border between two constituents, passivisability, clitic replacement, wh-questioning and causative-inchoative alternations. Grammatical functions are identified with diagnostics that apply to compositional expressions such as morphological marking and wh-questions (in MG subjects are always in the nominative case and objects almost always in the accusative case, verbs agree with their subjects and objects can be replaced by clitics).

The IDION encoding of the MWE structure corresponds to a rather flat tree and does not make use of powerful expressive means, such as inheritance, that in the literature have been combined with tree-based formalisms (Pollard & Sag 1987; Crabbé et al. 2013). The reason for choosing a perhaps redundant but rather simple encoding is that we aim at ensuring IDION's reusability. For this purpose, we try to make sure that we use expressive means that are shared by or can be easily transcribed to many formalisms and that the encoding does not rely on implicit assumptions concerning the overall grammar of the language.[3] To this end, the IDION representation of verbal MWEs defines the following nodes: (i)

[3]For instance, in MG possession is expressed with the sequence "DET noun Possessive". In IDION the whole sequence is encoded as fixed rather than encoding only the noun as fixed.

the root category (default) (ii) the phrasal categories shown in (6) that are used to denote free nominal constituents of the MWE (iii) leaf nodes (words). Phrasal categories and words are directly linked to the root category. IDION only indexes the fixed contiguous parts of an MWE (the WWSs of our implementation) and does not assign them a phrasal structure.

(6) NP-NOM/NP-NOM-anim/NP-NOM-nonanim;
NP-GEN/NP-GEN-anim/NP-GEN-nonanim;
NP-ACC/NP-ACC-anim/NP-ACC-nonanim;

The Java-based transcription application provides for the remaining phrasal categories needed for an LFG representation that requires the definition of constituents and typically involves trees deeper than the ones defined in IDION. All in all, IDION only specifies the phrasal categories shown in (6) and it is on the transcription applications to specify the categories that are necessary for any given formalism.

The IDION encoding of the MWE in (1) is given in Figure 3. On the first column it is specified whether the annotated part of the MWE is a phrasal category (phrasal categories are shown in 6) or a word and whether it is optional or not (for instance, the MWE of example (1) that is depicted in Figure 3 has only obligatory parts). Words are encoded as lemmas and only complementisers are encoded as such (in Figure 3, the depicted MWE contains a complementiser). On the second column, the lemmas of the parts of the expression are listed, namely the verb head *kano* 'make', the lemmatized parts of the WWS *mavros mati* 'black eye', the complementizer *na* 'to' that always introduces a sentential complement and the lemma form of the irregular verb head *vlepo* 'see' of the sentential complement. On the third column are encoded the actual form of the WWS and the control facts; in the case depicted in Figure 3, the sentential complement is controlled by the NP-NOM-anim. The fourth column provides the full morphological analysis of the fixed or semi-fixed parts of the MWE, for instance it is specified that the head verb of the controlled sentential complement is always in the active voice and in a form denoting perfect aspect; person and number of the controlled verb are not specified as they are determined by the free subject of the MWE. On the last column the parts of the WWS are indexed.

We developed a Java transcription application that generates XLE entries from the IDION specifications.

The LFG/XLE entries listed below are developed out of the IDION representation of (1) shown in Figure 3. As a first step, the transcription application generates lexical entries for the WWSs that are indexed in the IDION representation of

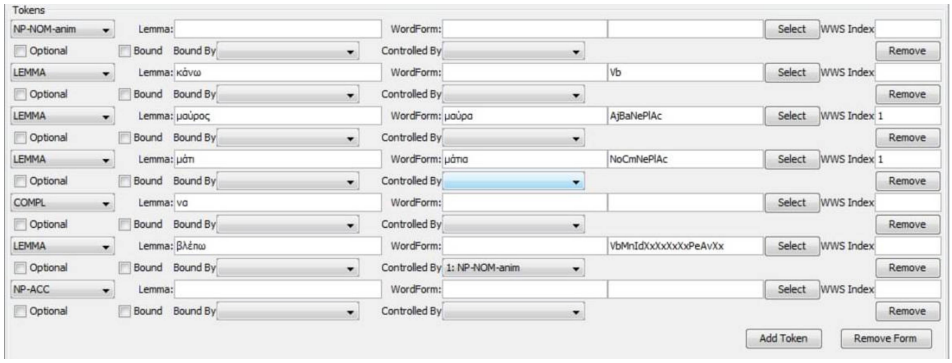

Figure 3: The IDION encoding of the MWE in (1)

the MWE; if one or more WWSs have been indexed in the IDION representation of the MWE, a corresponding number of XLE entries are produced and stored in the XLE lexicon. Morphological information about the entries, here the WWS and the verb head of the controlled sentential complement, is received from the annotation encoded on the fourth column. Next, the application generates the entry for the head verb of the MWE as follows: the NP-NOM-anim slot in the first column shows that the verb selects a free subject NP, the WWS that contains a noun and an adjective both in the accusative case shows that the head verb selects a fixed object and finally, the existence of a COMPL(ementiser) slot in the first column coupled with the control information on the third column shows that the head verb subcategorises for an XCOMP controlled by the subject of the main verb. This information generates the entry of the head verb *kano*. Finally, the head verb of the sentential complement is retrieved from the second column as it immediately follows COMPL. The application knows that the verb *vlepo* is transitive because it has a controlled subject and it is followed by an NP-ACC.

The WWS in MWE (1): mavra_matia, NoCmPlAc

The verb head of MWE (1): kano<SUBJ,OBJ,XCOMP>
↑ OBJ PRED = mavra matia
↑ XCOMP PRED = vlepo<SUBJ,OBJ>
↑ XCOMP PRED FINITE = +
↑ XCOMP SUBJ= ↑SUBJ

5 The filter

The filter consists of two parts: the filter lexicon and the filtering part proper.

5.1 The filter lexicon

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

1. Compositionality: Certain MWEs can take a compositional interpretation. For instance, the free subject verbal MWE in (1) has no compositional interpretation while the semi-fixed MWE in (7) can also take the compositional interpretation 'I grab them.FEM'.

   (7) tis arpazo
   them.FEM grab.1SG
   'I am beaten up.'

2. The "signifier": the lemma of the substring of an MWE that instructs the filter to look at the appropriate filter lexicon entries. For the MWE in (1), the signifier is the lemma *kano* 'make, do'. If the expression is fixed as in (8) the symbol "~" is used as a signifier. (8) has no translation, it is a kind of swearing (often accompanied with an offensive gesture) meaning that someone has made a serious mistake or is totally idiot:

   (8) pare pente
   take.2SG.IMP five

3. The lemmatised form of "words with spaces" (WWSs) whether they are independent fixed MWEs as in (8) or substrings of an MWE as in (1). In the case of (8) the lemmatised WWS would be *perno pente* 'take five'. In the case of (1) the fixed part is *mavra matia* 'black eyes' and the corresponding lemmatised form is *mavros mati* 'black eye'.

4. PoS and morphological constraints on the parts of the WWS. For the fixed part of (1) *mavra matia* the constraints would be: *mavros*: adjective, plural, accusative, basic; *mati*: noun, common, plural, accusative.

5.2 The filtering part

The filter proper, implemented in Perl, reads the tagged sentence from an XML file (the output of the tagger) and stores it. Then, it checks whether a signifier exists and,

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

B2. If they match, the filter consults the filter lexicon whether the MWE can take a compositional reading and,

C1. if it can, it sends to the formatter the input string and goes to step C2

C2. if it cannot, the part of the string that has been recognized is replaced with the corresponding WWS and morphological constraints and the resulting new string is sent to the formatter.

6 The LFG analysis (implemented with XLE grammars)

The output of the formatter is processed with an LFG grammar of Modern Greek with sub-lexical rules that can parse the output of the tagger. The grammar runs on XLE, a parsing environment dedicated to writing, running and debugging LFG grammars.[4] The trees generated by the sub-lexical rules can be seen in the c-structure of Figure 5.

Modern Greek verbal MWEs are rich in syntactic structure despite any simplifications that might result from the usage of WWSs. In Section 2 we discussed why we have adopted an LFG analysis that applies the classical LFG Grammatical Functions on MWEs despite the obvious problems. Thus, so far we have manipulated the lexicon by introducing the idiomatic lexical entries but we have not manipulated the grammar rules.

[4]XLE is the basis for the Parallel Grammar Project, which is developing industrial-strength grammars for English, French, German, Norwegian, Japanese, and Urdu. XLE is written in C and uses Tcl/Tk for the user interface. It currently runs on Solaris Unix, Linux, and Mac OS X.

With the reservations discussed in Section 2 in mind, we proceed to present Table 1 where the various types of parsed MWE structures are listed. In all, simple sentences containing 850 verb MWEs have been parsed. In Table 1 we give the basic form of the MWEs: the reader should keep in mind that MG is a pro-drop language with no infinitives, therefore the 1st person singular present indicative (or the 3rd person present indicative if the verb is an impersonal one) are used as the verb's lemma. Our system parses strings in the Greek alphabet but in Table 1 we have used Latin characters for reasons of readability. We represent WWSs as sequences of words joined with underscores, e.g. *pare_pente* (**1** in Table 1 and example 8). The column headed with "C" indicates whether the MWE receives a compositional interpretation (Y) or not (N). Lastly, the column headed with "FX" shows whether the MWE is flexible (FL), semi-flexible (SF) or fixed (F). We have marked as SF the MWEs that allow for no word order permutations but their head verb declines fully. MWEs that allow for word order permutations and their head verb declines fully are marked as FL.

With the approach described here, the lexicon has to be enriched with verb-like predicates such as *ego_arpazo* (**2** in Table 1) and *piano_gematos* (**9** in Table 1), noun-like predicates such as *mavros_mati* (**10** in Table 1) and adjective-like predicates such as *tapi_ke_psichremos* (7 in Table 1) and their morphological paradigms. Therefore, the morphological paradigm of the verb *arpazo* has to be duplicated in order to develop the paradigm of *tis_arpazo*. Similarly, (7 in Table 1) *meno tapi_ke_psichremos* contains a WWS that consists of the cranberry word *tapi*, the conjunction *ke* 'and' and a fully declinable adjective *psichremos* 'cool' that occurs freely in compositional structures. However, the overall amount of new lexical entries is not more than the entries required when MWEs are parsed like compositional structures (that is, without assuming WWSs) because in a "compositional approach" the same number of entries (or more) would be listed as "idiomatic". We have already pointed out that if the presented system is provided with the appropriate lexical entries and their morphological paradigms, it uses the grammar developed for compositional structures to parse sentences containing verb MWEs.

A wide variety of structures is shown in Table 1. **1** is a sentence but functions as an adverb, the MWE in **2** and **3** function as intransitive verbs, **4** and **5** function as transitive verbs with **5** featuring a case of where the subject binds a possessive selected by the fixed object. **6** and **7** are predicative structures that contain a controlled adjectival constituent normally modeled as an XCOMP in LFG. **8**, **9** and **10** are MWEs that contain sentential complements, either free (8) or subject to constraints such as control **9**, **10** and strong selection requirements on the form of the subordinated verb. These structures capture the typology of the 850

Table 1: Types of MG verb MWEs

	MWE	LFG representation	WWS lemma	C	FX
1	perno pente take five :a type of swearing	PRED pare_pente	ADV: perno_pente	Y	F
2	tis arpazo them.CL.ACC grab 'I am beaten up'	PRED ego_arpazo <SUBJ>	V: ego_arpazo	Y	SF
3	tin pernao zachari her.ACC pass sugar.ACC 'I have an easy time'	PRED ego_pernao <SUBJ,OBL> ↑OBL PRED =zachari	V: ego_pernao	N	FL
4	richni touloumia nero pours bags.ACC water.ACC 'It rains cats and dogs'	PRED richno <SUBJ,OBJ> ↑OBJ PRED=touloumia_nero	N: touloumia_nero	N	FL
5	troo to ksilo tis chronias mou eat the beating the year.GEN mine 'I am beaten up'	PRED troo <SUBJ,OBJ> OBJ PRED=o_ksilo_tis_chronias<POSS> ↑OBJ POSS PRED= ego ↑OBJ POSS PERS =↑SUBJ PERS ↑OBJ POSS NUM =↑SUBJ NUM ↑OBJ POSS GEN =↑SUBJ GEN	N: o_ksilo_o_chronia	N	FL
6	meno stili alatos remain stele.ACC salt.GEN 'I am left speechless'	PRED meno <SUBJ,XCOMP> ↑XCOMP PRED=stili_alas<SUBJ> ↑XCOMP SUBJ=↑SUBJ	N: stili_alas	N	FL
7	meno tapi ke psichremos remain tapi and cool 'I lose all my money'	PRED meno <SUBJ,XCOMP> ↑XCOMP PRED=tapi_ke_psichremos <SUBJ> ↑XCOMP SUBJ=↑SUBJ	ADJ: tapi_ke_psichremos	N	FL
8	echi yousto na S has preference to S 'don't tell me that S'	PRED echi_yousto <COMP> ↑COMP COMPL=να (impersonal)	V: echo_yousto	N	SF
9	richno adia na piaso yemata throw empty to catch full 'I fish out of/from'	PRED richno <SUBJ,OBJ,XCOMP> ↑XCOMP COMPL=na ↑OBJ PRED=adios ↑XCOMP PRED=piano_yematos<SUBJ> ↑XCOMP SUBJ=↑SUBJ ↑XCOMP PERF=+, ¬(↑XCOMP TENSE)	V: piano_yematos	N	FL
10	kano mavra matia na do NP make black eyes to see NP 'I have not met NP for a long time'	PRED kano <SUBJ,OBJ,XCOMP> ↑XCOMP COMPL=na ↑OBJ PRED= mavros_mati ↑XCOMP PRED=vlepo <SUBJ, OBJ> ↑OBJ PRED=ego ↑XCOMP SUBJ=↑SUBJ ↑XCOMP PERF=+, ¬(↑XCOMP TENSE)	N: mavros_mati	N	FL

verb MWEs that we parsed. Below we give selected parse-outs of the material in Table 1. Please notice that all f-structures contain a sentential feature IDIOM that is of semantic nature and conveys the meaning of the MWE. Figure 4 shows the f-structure of (9) that features the verb MWE **5** in Table 1. This MWE contains an OBJ GF headed by a WWS and a possessive anaphor that is analysed as a specifier of the projection of the WWS and is bound by the free subject; as a result the free subject and the anaphor are of the same gender and number.

(9) I Maria efage to ksilo tis chronias tis
the Maria.3SG.FEM ate the beating the year hers.3SG.FEM
'Maria was beaten up.'

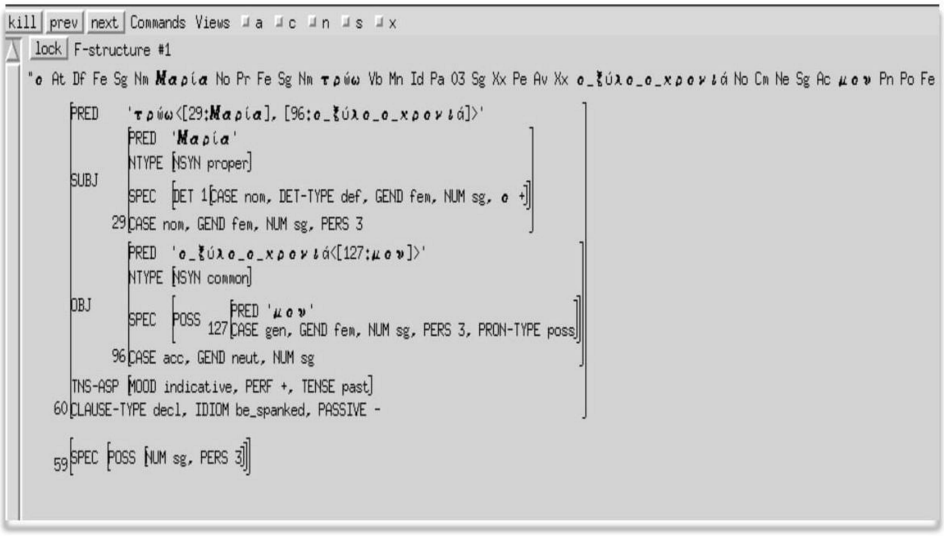

Figure 4: f-structure for *I Maria efage to ksilo tis chronias tis.* 'Maria was beaten up.', example (9), MWE **5** in Table 1

Figure 5 shows the c- and the f-structure of (10) that features an example of use of the verb MWE **10** in Table 1 and of example (1) that contains an OBJ GF headed by a WWS and a controlled sentential complement, an XCOMP in LFG terms. The result of the application of the sub-lexical rules is shown on the c-structure.

(10) Ekana mavra matia na tin do.
made.1SG black eyes to her see.1SG
'I have not seen her for a very long time.'

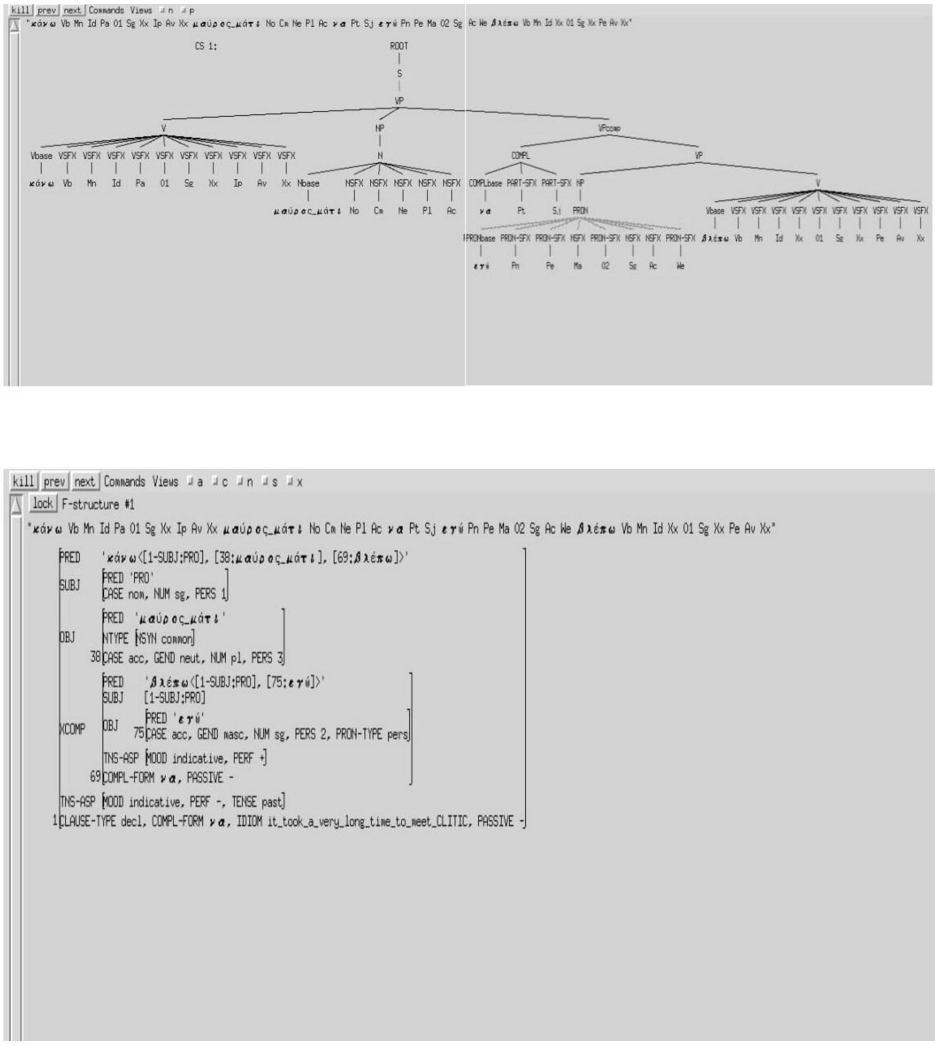

Figure 5: c- and f-structure for *Ekana mavra matia na tin do.* 'I have not seen her for a long time.', example (10), MWE **10** in Table 1

7 Discussion

We have presented a symbolic system for parsing MWEs that uses XLE and LFG grammars as its main components. MWEs are recognized as such before entering the XLE component and their sequential fixed parts are processed to form words with spaces (WWS). WWSs are processed as words by the XLE component. This system definitely reduces ambiguity since fewer parsings are available by definition; furthermore, the system does not require a lexicon more elaborate than the one required by a "compositional" approach. However, we have no way to measure whether the system (with the components that have been implemented so far) performs faster as there is no base system that we can use for a comparison – for instance, it would be interesting to evaluate the effect of the ambiguity that occurs in the filter.

An interesting feature of the system presented here is that it receives lexical knowledge from a lexicographic resource (IDION) that has been developed independently. The embedding of IDION into the LFG/XLE parsing system is a way of evaluating it. IDION has been designed with reusability issues in mind. However, the development of the transcription software indicated that some additional structural information would be beneficial, such as the marking of the head verbs and the marking of PPs (at the moment PPs are constructed by the transcription application that reads the IDION encoding and generates XLE entries). In the future, we aim to expand and improve the system in several ways, including an enrichment of IDION with other types of MWEs (nominal, adverbial), a more sufficient implementation of the filter and, of course, a grammar capturing a wider range of MG structures.

Abbreviations

GF	grammatical function	PoS	part of speech
LFG	lexical functional grammar	PP	prepositional phrase
MG	Modern Greek	SUBJ	subject
NP	noun phrase	WWSs	words with spaces
OBJ	object		

References

Apidianaki, Marianna, Prokopis Prokopidis & Haris Papageorgiou. 2018. Combining cross-lingual and syntactic evidence for Greek multiword expression

identification. *Bulletin of Scientific Terminology and Neologisms of the Academy of Athens, Special issue on MWEs in Greek and other languages: From theory to implementation.*

Attia, Mohammed A. 2006. Accommodating multiword expressions in an Arabic LFG grammar. In Tapio Salakoski, Filip Ginter, Tapio Pahikkala & Tampo Pyysalo (eds.), *Advances in natural language processing: 5th international conference, FinTAL, proceedings* (Lecture Notes in Computer Science 4139), 87–98. Berlin & Heidelberg: Springer.

Bargmann, Sascha & Manfred Sailer. 2018. The syntactic flexibility of semantically non-decomposable idioms. In Manfred Sailer & Stella Markantonatou (eds.), *Mutliword expressions: Insights from a multi-lingual perspective*, 1–40. Language Science Press. DOI:10.5281/zenodo.1182587

Bilgram, Thomas & Britt Keson. 1998. The construction of a tagged Danish corpus. In *Proceedings of the 11th Nordic conference of Computational Linguistics (NODALIDA 1998)*, 129–139. Copenhagen, Denmark: Center for Sprogteknologi, University of Copenhagen, Denmark. http://www.aclweb.org/anthology/W98-1614.

Copestake, Ann, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan Sag & Dan Flickinger. 2002. Multiword expressions: Linguistic precision and reusability. In *Proceedings of the third international conference on Language Resources and Evaluation (LREC 2002)*, 1941–1947. Las Palmas, Canary Islands, Spain.

Crabbé, Benoît, Denys Duchier, Claire Gardent, Joseph Le Roux & Yannick Parmentier. 2013. XMG: eXtensible MetaGrammar. *Computational Linguistics* 39(3). 591–629. DOI:10.1162/COLI_a_00144

Dalrymple, Mary. 2001. *Lexical functional grammar*. Vol. 34 (Syntax and Semantics). San Diego, CA: Academic Press.

Grégoire, Nicole. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1–2). 23–39.

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.

Kay, Paul & Ivan Sag. 2012. *A lexical theory of phrasal idioms*. Stanford, CA. http://www1.icsi.berkeley.edu/~kay/idiom-pdflatex.11-13-15.pdf.

Korkontzelos, Ioannis & Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Proc. of the 11th annual conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies NAACL/HLT 2010*, 636–644. Los Angeles, California.

Markantonatou, Stella & Niki Samaridi. 2018. Revisiting the grammatical function "object" (OBJ and OBJ_θ). In Manfred Sailer & Stella Markantonatou (eds.), *Mutliword expressions: Insights from a multi-lingual perspective*, 187–213. Language Science Press. DOI:10.5281/zenodo.1182599

Minos, Panagiotis, Stella Markantonatou, Georgios Zakis & Elpiniki Margariti. 2016. *Generating LFG/XLE MWE entries from IDION (a theory neutral lexical DB)*. Struga, fYROM. http://typo.uni-konstanz.de/parseme/index.php/2-general/156-selected-posters-struga-7-8-april-2016. 6th PARSEME general meeting.

Papageorgiou, Haris, Prokopis Prokopidis, Voula Giouli & Stelios Piperidis. 2000. A unified POS tagging architecture and its application to Greek. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhauer (eds.), *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC 2000)*, 1455–1462. Athens, Greece.

Pollard, Carl & Ivan Sag. 1987. *Information-based syntax and semantics* (CSLI Lecture Notes 13). Stanford: CSLI.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd international conference on Computational Linguistics and Intelligent Text Processing* (Lecture Notes in Computer Science 2276), 1–15. Springer.

Samaridi, Niki & Stella Markantonatou. 2014. Parsing Modern Greek verb MWEs with LFG/XLE grammars. In *10th workshop on Multiword Expressions (MWE 2014), European chapter of the Association for Computational Linguistics 2014*, 33–37. Gothenburg, Sweden.

Waszczuk, Jakub & Agata Savary. 2015. *Modeling syntactic properties of MWEs in LFG*. La Valletta. 4th PARSEME general meeting.