Encoding MWEs in a conceptual lexicon

Aggeliki Fotopoulou, Stella Markantonatou, Voula Giouli

Institute for Language and Speech Processing, R.C. "Athena'

{afotop;marks;voula}@ilsp.athena-innovation.gr

Abstract

The proposed paper reports on work in progress aimed at the development of a conceptual lexicon of Modern Greek (MG) and the encoding of MWEs in it. Morphosyntactic and semantic properties of these expressions were specified formally and encoded in the lexicon. The resulting resource will be applicable for a number of NLP applications.

1 Introduction

Substantial research in linguistics has been devoted to the analysis and classification of MWEs from different perspectives (Fraser, 1970; Chomsky, 1980; M. Gross 1982, 1988; Ruwet, 1983; der Linden, 1992; Nunberg et al., 1994; Howarth, 1996; Jackendoff, 1997; Moon, 1998; Fellbaum, 2007). Moreover, cognitive and psycholinguistic approaches to MWEs (Lakoff, 1993; Gibbs, 1998; Glucksberg, 1993; Diakogiorgi&Fotopoulou, 2012) have accounted for their interpretation. Within the NLP community, there is a growing interest in the identification of MWEs and their robust treatment, as this seems to improve parsing accuracy (Nivre and Nilsson, 2004; Arun and Keller, 2005). In this respect, the development of large-scale, robust language resources that may be integrated in parsing systems is of paramount importance. Representation, however, of MWEs in lexica poses a number of challenges.

2 Basic Notions

Typically, fixed MWEs are identified and classified on the basis of semantic, lexical and morphosyntactic criteria. (M. Gross, 1982, 1987; Lamiroy, 2003), namely:

- non-compositionality: i.e., the meaning of the expression cannot be computed from the meanings of its constituents and the rules used to combine them. Nevertheless, according to (Nunberg et *al,* 1994), compositionality refers to the fact that the constituents of some idioms "carry identifiable parts of the idiomatic meaning". *Variability* has been further emphasised in (Hamblin and Gibbs 1999) and (Nunberg et al. 1994): fixed expressions appear in a continuum of compositionality, which ranges from expressions that are very analysable to others that are partially analysable or ultimately non-analysable.
- non-substitutability: at least one of the expression constituents does not enter in alternations at the paradigmatic axis
- non-modifiability: MWEs are syntactically rigid structures, in that there are constraints concerning modification, transformations, etc.

These criteria, however, do not apply in all cases in a uniform way. The *variability* attested brings about the notion 'degree of fixedness' (G. Gross 1996). The kind and degree of fixedness result in the classification of these expressions as *fixed*, *semi-fixed*, *syntactically flexible* or *collocations* (Sag et al, 2002). It is crucial for a satisfactory MWEs representation in a computational lexicon to provide an accurate and functional formal modelling of *fixedness*, *variability* and *compositionality.*

In this paper, we will discuss the classification and encoding of compounds and fixed MWEs in a conceptually organised lexicon of MG.

3 The conceptual lexicon

The conceptually organised lexicon that is under development (Markantonatou & Fotopoulou, 2007) capitalises on two basic notions: (a) the notion of lexical fields, along with (b) the Saussurian notion of sign and its two inseparable facets, namely, the *SIGNIFIER* and the *SIGNIFIED* as the building blocks (main classes) of the underlying ontology.

Proceedings of the 10th Workshop on Multiword Expressions (MWE 2014), pages 43–47,
Gothenburg, Sweden, 26-27 April 2014.

In this sense, the intended language resource is a linguistic ontology in which words are instances in the *SIGNIFIER* class. At this level, morphological, syntactic and functional information about lemmas is encoded. Similarly, word meanings are instances in the *SIGNIFIED* class. Each instance in the *SIGNIFIER* class is mapped onto a concept, the latter represented as an instance in the *SIGNIFIED* class.

The Instances of the class *SIGNIFIER* are specified for (a) features pertaining to lexical semantic relations (i.e, synonymy, antonymy); (b) lexical relations such as word families, allomorphs, syntactic variants etc.; and (c) morphosyntactic properties (PoS, gender, declension, argument structure, word specific information etc.). Values for these features are assigned to both single- and multi-word entries in the lexicon. MWEs are further coupled with rich linguistic information pertaining to the lexical, syntactic and semantic levels.

4 Encoding MWEs in the lexicon

MWEs are encoded as instances in the SIGNIFIER class of our ontology and are also mapped onto the corresponding concepts or word meanings (instances in the SIGNIFIED class).

In the remaining, we focus on the encoding of MWEs as instances in the SIGNIFIER class. We cater for sub-classes corresponding to grammatical categories (verb, noun, adjective, adverb, preposition, etc) under the class SIGNIFIER in our schema. The class MWEs (as opposed to the class Simple Lexical Units) has been defined further under the verb, noun, adjective and adverb sub-classes.

Syntactic configurations pertaining to each class are also represented as distinct sub-classes hierarchically organised under the verb, noun, adjective and adverb classes. Morphosyntactic properties, selectional preferences, and semantic interpretation patterns are provided for each MWE depending on the grammatical category it pertains to; encoding is based on a set of parameters represented as feature-value pairs.

More precisely, a typology of Greek verbal MWEs has been defined in (Fotopoulou, 1993, Mini, 2009) (NP V NP1 NP2…) and of nominal MWEs in (Anastasiadis, 1986) (Adj N, NN…) on the basis of the lexical and syntactic configurations involved. This typology has been mapped onto a hierarchy under classes *verb* and *noun*).

In our approach, the main distinction between *collocations* and *fixed MWEs* is made explicit. The degree and type of fixedness are then encoded as features. Further morphosyntactic information is also encoded depending on the grammatical category of the MWE (i.e., declension of one or more constituents, *only_singular* or *only_plural* for nouns, etc.). In this way, information that may be useful for the automatic identification and interpretation of the MWEs may be retained. Moreover, the standard set of features inherited from the class SIGNIFIER is also retained (PoS, Gender, Number, Tense, synonyms, antonyms, etc.).

4.1. The encoding schema

We have so far implemented an encoding schema for nominal and verbal MWEs. We aimed at encoding rich linguistic knowledge in a formal way that would be exploitable in computer applications. The two types of fixedness (collocations and fixed) are encoded as features: (a) *Lexical_variance*, and (b) *Is_actually*.

The feature *Lexical_variance*[1] has as possible values (yes or no). Collocations (assigned a yes value) are further specified with respect to alternative lemmas; these lemmas are encoded in the appropriate feature *Variants.* For instance, in example (1) the two alternative lemmas are *καταστάσεις* and *περιστάσεις*:

(1) *έκτακτες (καταστάσεις / περιστάσεις)*
(=emergency (situations / circumstances))

The feature *Is_actually* (with possible values yes or no) encodes information about the interpretation pattern: a value *yes* signifies a compositional or partially compositional meaning; on the contrary, a value *no* denotes a non-compositional interpretation (fixed meaning).

Collocations are by default assigned feature values corresponding to a compositional meaning. In these cases, the feature *maintains_meaning* further specifies the constituent(s) that contribute to the non-fixed interpretation of the expression. For example, the meaning of the compound in (2) is retained from the meaning of the first noun *ταξίδι* (=trip), which, in turn, is the value assigned to the *maintains_meaning* feature:

[1]In our MWE classification scheme, a lexical unit is considered 'fixed' at the lemma level. This is because MG is a heavily inflected language.

(2) *ταξίδι αστραπή* (trip - lightning (=very sudden and short trip)

<*maintains_meaning = ταξίδι* />

Finally, the feature *has_meta_meaning* signifies further the constituent(s) – if any – bearing a figurative meaning. For example, the compound *ταξίδι αστραπή* in (2) assumes the figurative meaning of the second noun *αστραπή* (=very sudden and short-term).

On the contrary, verbal and nominal expressions with a non-compositional meaning are assigned a negative value *(no)* for the *Is_actually* feature since their constituents do not contribute to a compositional meaning; therefore, the features *maintains_meaning* and *has_meta_meaning* are left empty as non-applicable. This is exemplified in (3) below; the constituents *παιδική* (=kids') and χαρά (=joy) of the expression *παιδική χαρά* (=playground) do not contribute to the overall interpretation:

(3) *παιδική χαρά* (=playground*)*

<*maintains_meaning/*>

<*has_meta_meaning/*>

This schema that applies to both nominal and verbal MWES, is presented in Table 1 below.

Slot	**Values**
mwe_type	*Fixed; collocation*
Lexical_variance	*Boolean (yes, no)*
Variants	*string*
Is_actually	*Boolean (yes, no)*
maintains_meaning	*String*
has_meta_meaning	*String*

Table 1 The encoding schema for nouns & verbs

4.2. Nominal MWEs

Furthermore, nominal MWEs are also assigned values for features that are specific to the nominal MWEs. Information on inflected constituents - if any – is provided in the declension feature; values for *only_singular* and *only_plural* provide further morphological/usage information; when used in combination with other features (i.e, *is_actually*) this type of information is evidence of fixedness. Frequent co-occurrence patterns with verbs are provided in the *verb_combined* feature; finally, alternative nominalised forms are listed as values of the feature *nominalization*. The schema is presented in the table below:

only singular	*Boolean (yes, no)*
only plural:	*Boolean (yes, no)*
N_declension	*N1, N2, N1_N2, Adj_N*
verb_combined	*string*
Nominalization	*string*

Table 2 The encoding schema for nouns

4.3. Verbal MWEs

In the typology adopted for the verbal idiomatic expressions, fixedness can be limited to only certain constituents of the sentence; a combination of fixed and non-fixed constituents in *Subject* or *Object* position is permitted. For example, in sentences (4) and (5) below, fixedness relies on the relation among the verbs and the nouns that function as *Objects* (direct and indirect) and as *Subject* respectively:

(4) *δίνω* <u>*τόπο*</u>$_{\underline{NP}\text{-acc, Obj}}$ <u>*στην οργή*</u>$_{PP}$

to give way to anger (=to swallow one's pride/anger)

(5) *ανάβουν* <u>*τα λαμπάκια μου*</u>$_{NP\text{-nom, Subj}}$

my lights are switched on (=to become very angry)

Moreover, the typology allows for a restricted alternation of fixed elements of the expression. For example, in the MWE in (6), the two alternative lemmas are *τάζω* and *υπόσχομαι*:

(6) <u>*τάζω / υπόσχομαι*</u> *τον ουρανό με τ' άστρα*

to <u>undertake to offer / promise</u> the sky with the stars

This information is encoded in verbal MWEs, namely: (a) the syntactic properties of the verb that occurs in the expression (*valency*); and (b)

fixed and non-fixed arguments either in *Subject* or *Object* position. Moreover, selectional restrictions applied to the arguments (such as +/- human) are also added.

The encoding schema that applies to verbal MWEs specifically is presented in Table 3. In this schema, *N* signifies a non-fixed noun, whereas C denotes a fixed one; number *0* (in *N0* and *C0*) is used to represent a noun (either fixed or non-fixed in Subject position), and *1, 2, 3,* etc. denote complements in *Object* position (or complements of prepositional phrases). Other features provide rich linguistic information regarding facets of the expression in terms of: (a) selectional restrictions (i.e., the features *N0_type, N1_type,* etc., accept as values the semantic category in which a noun in *Subject* or *Object* position respectively, belongs to), (b) syntactic alternations (i.e., *Poss_Ppv* encodes the alternation among possessive and personal pronoun), grammatical information (i.e., *Ppv_case* encodes the case of the personal pronoun), etc.

Slot	**Value**
N0_type	*hum, -hum, npc*
C0_variants	*string*
Poss=Ppv	*Boolean (yes or no)*
Ppv_case	*gen, acc*
N1_type	*hum, -hum, npc (Nom de partie du corps/noun of the part of body)*
N2_type	*hum, -hum, npc*
N3_type	*hum, -hum, npc*
C1_variants	*string*
C2_variants	*string*
C3_variants	*string*

Table 3. The encoding schema for verbs

Alternative nouns (in Subject or Object position) that oftern co-occur with the verbal expression are also provided for (C0_variant, C1_variant, etc).

5. Discussion

As it has been shown above, in our lexicon we have opted for an approach to MWE representation that builds on rich linguistic knowledge. The linguistic classifications adopted deal with morphology, syntax, and semantics interface aspects. Thus, a lexicon – grammar representation of MWEs has been constructed by encoding key morphosyntactic and semantic information.The typology of verbal MWEs shares common characteristics with similar efforts for other languages (i.e, DuELME, Gregoire, 2010 Morphosyntactic properties and selectional preferences account better for a number of phenomena, inherent in the Greek language, as for example word order and gaps attested in running text.

More specifically, Greek is a language with a relatively free word order, and idiomatic expressions often occur in texts in various configurations. The encoding of fixed and non-fixed constituents provides, therefore, extra information for the identification of expressions in texts. Moreover, the identification of MWEs as collocations entails a relatively loose fixedness, allowing, thus, for gaps and discontinuities as shown in (7):

(7) *Το κόμμα έχει αριθμό υποψηφίων-ρεκόρ*

The political party has a number of candidates record (=many candidates)

6. Conclusions and Future work

We have given an overview of the conceptual lexicon currently under development and the treatment of MWEs in it. We have so far treated nominal and verbal MWEs (~1000 entries). Future work involves the population of the lexicon with new expressions also pertaining to the grammatical categories adjective and adverb and the definition of a fine-grained typology for the latter. Moreover, a more granular representation of fixedness will be attempted. Compatibility of the resource with diverse syntactic approaches will also be investigated. The evaluation of the final resource will be performed by integrating it in a tool that automatically recognizes MWEs in texts.

References

Αναστασιάδη-Συμεωνίδη Ά. (1986). *Η νεολογία στην Κοινή Νεοελληνική*. Θεσσαλονίκη: Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης (Επιστημονική Επετηρίδα Φιλοσοφικής Σχολής).

Arun, A. and F. Keller. 2005. Lexicalisation in crosslinguistic probablisitic parsing: The case of french. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp 306–313. Ann Arbor, MI

Chomsky, N. 1980. *Rules and Representations*. New York: Columbia University Press.

Diakogiorgi, K. & Fotopoulou, A. 2012. Interpretation of Idiomatic Expressions by Greek Speaking Children: implications for the Linguistic and Psycholinguistic Research. An interdisciplinary approach. *Lingvisticae Investigationes*, Volume 35:1. 1-27, John Benjamins, Paris, France

Fellbaum, C. 2007. Introduction. Fellbaum, C. (ed). *Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies*. London: Continuum, 1-

Fotopoulou, A. 1993. *Une Classification des Phrases à Compléments Figés en Grec Moderne*. Doctoral Thesis, Universite Paris VIII.

Fraser, B. 1970. Idioms within a Transformational Grammar. *Foundations of language*, 1, 22-42.

Fritzinger, F., Weller, M., and Heid. U. 2010. A survey of idiomatic preposition-noun-verb tiples on token level. In *Proceedings of LREC-10.*

Grégoire, N. 2010. DuELME: a Dutch electronic lexicon of multiword expressions*; Lang Resources & Evaluation* (2010) 44:23–39

Gibbs R.W. 1998. The Fight over Metaphor in Thought and Language. In A.N. Katz, C. Cacciari, R.W. Gibbs & M. Turner (eds.), *Figurative Language and Thought*. OUP, 88-118.

Glucksberg, S. 1993. Idiom meanings and allusional context. In *Idioms: Processing, structure, and intepretation*. C. Cacciari and P. Tabossi (eds.). Hillsdale, NJ: Erlbaum, 201-225.

Gross, G. 1996. Les expressions figées en français. Noms composés et autres locutions. Paris/Gap: Ophrys.

Gross, M. 1982. Une classification des phrases figées du français. Revue Québécoise de Linguistique 11 (2), 151-185.

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

Gross, Maurice. 1988b. Sur les phrases figées complexes du français. *Langue française* 77: 4770.

Hamblin, J., and Gibbs, W. R. 1999. Why You Can't Kick the Bucket as You Slowly Die: Verbs in Idiom Comprehension. *Journal of Psycholinguistic Research*. 28 (1): 25-39.

Howarth P.A. 1996. Phraseology in English academic writing. *Lexicographica Series* 75. Tübingen: Max Niemeyer. Jackendoff R. 1997. *The Architecture of the Language Faculty.* MIT Press.

Lakoff G. 1993. The Contemporary Theory of Metaphor. In A. Ortony (ed.), *Metaphor and Thought,* 2nd edition Cambridge University Press, 202-251.

Lamiroy, B. 2003. Les notions linguistiques de figement et de contrainte. *Lingvisticae Investigationes* 26:1, 53-66, Amsterdam/Philadelphia: John Benjamins.

van der Linden E-J. 1992. Incremental processing and the hierarchical lexicon. *Computational Linguistics*, 18, 219-238

Markantonatou, Stella and Fotopoulou, Aggeliki. 2007. The tool "Ekfrasi". In *Proceedings of the 8th International Conference on Greek Linguistics, The Lexicography Workshop*. Ioannina, Greece.

Markantonatou, S., Fotopoulou, A., Mini, M. & Alexopoulou, M. 2010. In search of the right word. In *Proceedings of Cogalex-2: Cognitive Aspects of the Lexicon, 2nd SIGLEX endorsed Workshop*. Beijing.

Mini, M. 2009. *Linguistic and Psycholinguistic Study of Fixed Verbal Expressions with Fixed Subject in Modern Greek: A Morphosyntactic Analysis, Lexicosemantic Gradation and Processing by Elementary School Children*. Unpublished doctoral dissertation. University of Patras.

Moon, R. 1998. *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: OUP.

Nivre, J. and Nilsson, J. 2004. Multiword units in syntactic parsing. *Workshop on Methodologies and Evaluation of Multiword Units in Real-World Applications*.

Nunberg ,G., Sag I., Wasow, T. 1994. Idioms. *Language* 70, 491-538.

Ruwet, N. 1983. Du Bon Usage des Expressions Idiomatiques dans l'Argumentation en Syntaxe Générative. *Revue Québecoise de Linguistique* 13 (1): 9-145.

Sag, I.A., Bond, F., Copestake A., Flickinger, D. 2001. Multiword Expressions. LinGO Working PaperNo.2001-01.

Sag, Ivan A., T.Baldwin, F.Bond, A. 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 COLING-2002.