IDION: A database for Modern Greek multiword expressions

Stella Markantonatou
ILSP/Athena RIC,
Athens, Greece
stiliani.markantonatou@
gmail.com

Panagiotis Minos
ILSP/Athena RIC,
Athens, Greece
pminos@gmail.com

George Zakis
ILSP/Athena RIC,
Athens, Greece
georgizak@gmail.com

Vassiliki Moutzouri
National and Kapodistrian University
of Athens, Greece
vasiliki.moutzouri96@gmail.com

Maria Chantou
University of Patras, Greece
mariachant96@gmail.com

Abstract

We report on the ongoing development of IDION, a web resource of richly documented multiword expressions (MWEs) of Modern Greek addressed to the human user and to NLP. IDION contains about 2000 verb MWEs (VMWEs) of which about 850 have been documented as regards their syntactic flexibility, their semantics and the semantic relations with other VMWEs. Sets of synonymous MWEs are defined in a bottom-up manner revealing the conceptual organisation of the MG VMWE domain.

1 Introduction

We report on the ongoing development of IDION, a web resource of multiword expressions (MWEs) of Modern Greek (MG). [1] IDION is addressed to the human user and to NLP systems. By now, it contains 2000 Greek verb MWEs (VMWEs) that mostly fall in the idioms and light verb constructions categories of the PARSEME annotation guidelines (Savary et al., 2018), of which 850 are fully documented and available under a CC-BY-NC license. It has been developed by a small team of editors who did documentation work and edited material collected with crowdsourcing (about 35 University students of literature participated). The editors compiled a list of VMWEs drawing on published collections, e.g., Sarantakos (2013), dictionaries, e.g., Lexigram, and their intuitions as native speakers of MG; the encoders received a short VMWE list and a documentation manual.

In Section 2 we discuss challenging documentation issues. We pay some attention to VMWE emphasis (Section 3) and synonymy (Section 4). Section 5 is about the developed web editor. Section 6 concludes the presentation.

2 The documentation

Like other MWE databases, which are mentioned as our discussion proceeds, IDION serves both the human user and the NLP (Smørdal Losnegaard et al., 2016). Gantar et al. (2018) list the MWE properties documented in seven dictionaries and NLP databases: phrase structure, variants, morphology of MWE elements, contingency of MWE parts, usage example and definition. IDION documents a superset of the listed properties (Table 1).

We have defined a ‘template’ (Section 5) consisting of fields that we fill according to a set of specifications for the encoding of the MWE properties (Fellbaum and Geyken, 2005). Table 1 approximates the design of the IDION template.

2.1 Entry and lemma definition

A new entry is defined with the unique coupling of a lemma and a definition because lemmas may be coupled with more than one definitions (polysemy), e.g., *vgazo ta sothika mu*, Lit. I take out my guts, is the lemma of five entries meaning “I throw up”, “I express my deeper feelings”, “I

[1] This research has been partly financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE: (T1EΔK-999723442).

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 130–134
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

cough violently", "I sing loudly", "I bust a gut". We use the IDION definition(s) of the VMWE in the contexts where the VMWE was found in order to decide whether multiple entries should be defined. On the other hand, the VMWEs *troo/katapino/cha(ft/v)o/masao to paramithi*, Lit. I eat/swallow/swallow/chew the story, "I swallow something hook, line and sinker" define four entries encoded both as lexical variants, as they have different fixed verb heads, and as synonyms, as they are assigned the same definition.[2]

Lemma form	
Translations	English, French
Codification for NLP	lemma:
	-cranberry words (if any)
	-free XPs (NPs, VPs)
	-optional lemmas
	-morphological constraints
	-contingency
	-control and binding
	-case, animacy (free NPs)
Corpus	web, introspection
	literal usages
Syntactic flexibility and Verb alternations	-word order permutations
	-fixed NPs cliticisation
	-XP interpolation
	-passivisation
	-causatives-inchoatives
	-dative genitives, other
Lexical variation	-multiple entries
	-optionality, disjunction
Semantics	-definition
	-polysemy
	-opposites
	-semantic pairs
	-MWEs in the Possessive and Stative relations
	-polarity, style, emphasis
	-sets of synonymous MWEs

Table 1: VMWE properties encoded in IDION.

The relatively free word order of MG allows us to use two (default) 'canonical' (maximal) orders: Free(NP_Subject) + Fixed(+Verb+NP_Direct Object + PPs)+Free(XPs) and Fixed(NP Subject +Verb+ NP Direct Object+PPs)+Free(XPs) for the lemma definition provided that no other more frequent order exists, e.g., (1) is used in the word order PP + Direct Object(Clitic) + Verb. Additionally to the lemmatisation conditions used in MG grammar we postulate that: (i) tenses are divided into past, present and future ones and (ii) the 'order' of grammatical persons is 1st>2nd>3rd, e.g., (1) appears with 2nd/3rd person subjects only and 1st person singular possessives, therefore the verb's 'lemma' is in the 2nd person singular.

The maximum length of the fixed string may vary (Fellbaum and Geyken, 2005). We model this phenomenon as optionality denoted with brackets (2). Variation on fixed functional parts such as prepositions is indicated on the lemma form with disjunction (2).

(1) *apo to stoma mu to pernis*
Lit. from the mouth mine it.CL take.2nd.SG
"you say exactly what I was about to say before I utter it"
(2) *afino* (*gia / os*) *kavatza kati*
Lit. leave (for / as) buffer something
"I put something aside"

2.2 Morphosyntactic information

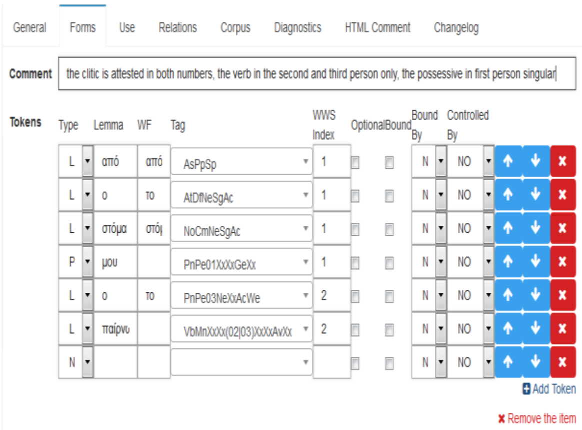
We use a template that facilitates encoding (Figure 1) to structure an NLP oriented representation of the VMWE in an as much as possible theory independent way mainly aiming at reusability and less at representing linguistic generalisations (Villavicencio et al., 2004): the theoretical constructs used are part-of-speech (PoS) and simple phrasal categories (NP, VP). Information about contingency, subject control, anaphor binding and optionality is provided. We use regular expressions on the MG PAROLE (Labropoulou et al., 1996) to exhaustively document the morphological constraints on the VMWE parts. Figure 1 shows the encoding of the morphological constraints on (1): verb person is constrained to 2nd and 3rd, verb/clitic number is not constrained and the possessive is specified as 1st singular (Xx=unspecified value in a closed set of values). The free parts of the MWE are characterised for phrasal category and, in the case of NPs, for case and animacy. More than one NLP-oriented representations can be defined for the same VMWE enabling us to treat certain types of lexical variation without creating a new lemma, namely lexical variation on functional categories (2) and morphological variation/diminutives on

[2] Synonymous MWEs with identical verb heads and different fixed NP parts define distinct entries unless the fixed NP parts are morphological variants such as gender variables, for instance *nerofida*.FEM-*nerofido*.NEUT ``grass snake", or diminutives.

fixed content parts. An experiment to convert the NLP oriented representation to an XLE/LFG lexicon was successful (Minos et al., 2016).

Seven syntactic flexibility tests (free or fixed subject, word order permutations, whether an XP (=NP, AdjP, AdvP, …) can be inserted among the lexicalised parts of the VMWE, passivisation, dative genitive alternation (3), fixed object NP cliticisation and causative-inchoative alternation) are exemplified with corpus examples.

(3) *mu spas ta nevra - spas ta nevra mu*
Lit. me.DATGEN break.2nd the nerves - break.2nd the nerves mine.POSS
"you grate on me"



Comment the clitic is attested in both numbers, the verb in the second and third person only, the possessive in first person singular

Tokens

Type	Lemma	WF	Tag	WWS Index	Optional	Bound	Bound By	Controlled By
L	από	από	AsPpSp	1	□	□	N	NO
L	ο	το	AtDfNeSgAc	1	□	□	N	NO
L	στόμα	στόμ	NoCmNeSgAc	1	□	□	N	NO
P	μου		PnPe01XxXxGeXx	1	□	□	N	NO
L	ο	το	PnPe03NeXxAcWe	2	□	□	N	NO
L	παίρνω		VbMnXxXx(02|03)XxXxAvXx	2	□	□	N	NO
N					□	□	N	NO

Figure 1: Representation of (1) for NLP purposes.

2.3 Collection of annotated examples

Because there are no sizeable corpora of MG, examples are retrieved from the web (about 8 examples/VMWE). Introspective examples (less than 10% in total) mainly demonstrate the unacceptability of certain structures. Examples are selected to illustrate the syntactic flexibility of the VMWE and whether it accepts emphasis; in short, the corpus contains examples annotated for acceptability and the phenomena they exemplify. Literal examples are included if the MWE accepts both a literal and a fixed interpretation. The corpus currently offers only evidence about the participation (or not) of a VMWE to a certain linguistic phenomenon; crucially, it provides no frequency information. Other databases drawing on large corpora include frequency information [DuELME, Gregoire (2010); The Berlin Idiom Project, Fellbaum and Geyken (2005)]. We plan to enhance IDION with the ability of encoding the frequence of occurrence of the VMWE alternants.

2.4 Semantics

IDION documents a set of semantic relations among VMWEs (Table 1). Online dictionaries and lexicographic databases, such as Algemeen Nederlands Woordenboek, WordNet, provide synonyms and opposites. We devised the term 'semantic pair' to denote pairs of morphologically unrelated predicates that stand in a causative/non-causative relation (4). The 'Opposite' relation is encoded for VMWE pairs with opposite meanings; the 'Stative' relation for VMWE pairs that denote an event and a situation resulting from it, e.g., *meno misos*, Lit. I remain half, "I lose a lot of weight" - *ime petsi ke kokalo*, "I am skin and bones"; the 'Possessive' relation for VMWE pairs that denote an event and a result situation in which an entity has something in his/her 'possession/control', e.g., *vazo stin akri kati*, Lit. I put at the edge something, "I lay up something" - *echo stin akri kati*, Lit. I have something at the edge, "I have something in store".

(4) *afino* anavdo kapion - *meno* anavdos
Lit. leave.1st speechless somebody.ACC - stay.1st speechless.NOM
"I leave somebody speechless - I become speechless"

The 'Verb alternation' relation, e.g., *erchete keramida se kapion*, Lit. comes tile.SUBJ to somebody, "someone is floored"- *kati erchete keramida se kapion*, Lit. something comes as a tile to somebody, "something floors somebody" is an intransitive verb/verb-copula pair with the same verb head. This set of relations, along with 'Synonymy', will be exploited to define a network of VMWEs expressing a concept. Such concepts 'emerge' from the synonyms sets in a bottom-up way (Section 4), e.g., the concept in (1) or of being let down exactly the moment when a desire is about to be satisfied, etc.

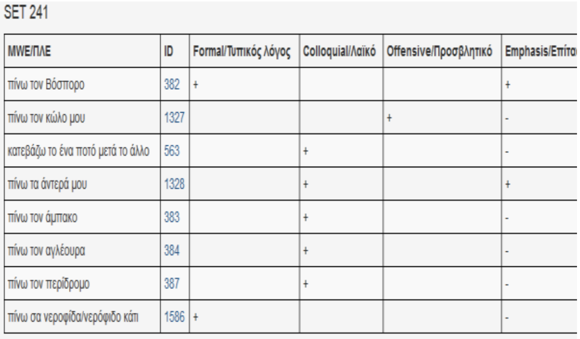
3 Polarity, Style and Emphasis

IDION encodes polarity and style information [DuELME, Gregoire (2010); Polytropon, Fotopoulou et al. (2014)]. For style, the VMWE is assigned one of the values Formal, Colloquial, Offensive (Christopoulou, 2016). To distinguish between a formal and a colloquial VMWE, as a rule of thumb, formal VMWEs should occur in the political articles of established Greek newspapers. For polarity, three values are used, (-) for VMWEs occurring in negative environments only, (+) for VMWEs occurring in positive environments only and ‘unspecified’ otherwise.

To the best of our knowledge, emphasis with VMWEs has received little attention in the international and MG literature (Gavriilidou, 2013). DuELME encodes fixed lexical modifiers of VMWEs and diminutives (Grégoire, 2010) both of which may express emphasis with MG VMWEs. To form an operational view of emphasis (a detailed view would require dedicated research), we have studied 180 VMWEs encoded in IDION, 90 headed by the verb *afino* “leave” and 90 by the verb *vazo* “put”. Drawing on this and on IDION’s material, we assigned the values (+/-) to the feature Emphasis; e.g., the VMWE *pino ton peridromo*, Lit. I drink the catch (fishing) (Sarantakos, 2013), “to hit the bottle” has not been found in an emphatic construction yet and is assigned the value (-). We observed that VMWEs often adopt the general MG emphasis mechanisms while certain VMWEs prefer own fixed emphatic means. We encoded fixed phrasal/ lexical emphasis as an optional part of the VMWE, for instance, *ginome thirio* (*animero*), Lit. I become beast (untamed.ADJ), “I fly off the handle” and diminutives with alternative NLP oriented representations and added a comment on their emphatic function.

4 Sets of synonymous VMWEs

IDION provides sets of synonymous VMWEs and indicates their style and emphasis similarities and differences (Figure 2: VMWEs about drinking a lot). Synonyms sets are defined in a bottom-up manner; IDION applies the transitive property on the pairs of synonymous VMWEs. Therefore, it is not necessary for the encoder to exhaust the list of possible synonyms of a VMWE and the synonyms sets are dynamic; each time the synonyms sets facility is called, the result reflects the current situation of IDION. Since about 80% of the IDION entries were documented with crowdsourcing, it was not advisable to pose strict specifications on the semantic relations because it would complicate the task and reduce the encoders' creativity. Although the editors had to check the validity of the provided synonymous VMWEs against appropriate contexts, the encoders' creativity was proven valuable given the lack of large corpora and lexicographic resources of MG that would provide a variety of synonyms for each VMWE. The bottom-up definition of synonyms sets reveals the concepts which MG expresses with VMWEs---these concepts are not always expressed by existing MG verb predicates, eg. the concept “to let down somebody exactly when his/her desire is about to be satisfied”.



SET 241

MWE/ΠΛΕ	ID	Formal/Τυπικός λόγος	Colloquial/Λαϊκό	Offensive/Προσβλητικό	Emphasis/Επίτα
πίνω τον Βόσπορο	382	+			+
πίνω τον κώλο μου	1327			+	-
κατεβάζω το ένα ποτό μετά το άλλο	563		+		-
πίνω τα άντερά μου	1328		+		+
πίνω τον άμπακο	383		+		-
πίνω τον αγλέουρα	384		+		-
πίνω τον περίδρομο	387		+		-
πίνω σα νεροφίδα/νερόφιδο κάτι	1586	+			-

Figure 2: Synonyms set for drinking a lot.

5 The web editor

The web editor is a PHP based application that takes advantage of the *Symfony PHP framework*, a set of reusable PHP components and a PHP framework for web applications (Shklar and Rosen, 2009). The data are stored in a database (*MySQL*) and a persistence provider (*Doctrine*) is used as a database abstraction layer between the database engine and the rest of the application, allowing for easy migration to any RDBMS. Only a web browser and a computer with an internet connection are required to access the editor that can be used from all major operating systems and browsers. An encoding ‘template’ is provided structured in 7 tabs: General, Forms (MWE morphosyntax), Usage example, Corpus, Diagnostics (flexibility tests), Relations and logistics tab. Editable controlled vocabularies in pull down menus and string matching facilities are used. Special machinery has been developed for defining and editing the semantic relations.

6 Conclusion and future work

IDION is a state-of-the-art resource addressed to humans and the NLP with detailed qualitative information about MG MWEs. Future priorities include: further populating IDION, adding more types of MWEs (nominal, adjectival, adverbial), developing the full network of semantic relations among VMWEs that define a “concept”, using the web to identify usage tendencies.

References

Algemeen Nederlands Woordenboek http://anw.inl.nl/search

Doctrine https://www.doctrine-project.org/

Lexigram https://www.lexigram.gr/lex/enni/

MySQL https://www.mysql.com/

PHP https://php.net/

Symfony https://symfony.com/

WordNet https://wordnet.princeton.edu/

Katerina Christopoulou. 2016. *A lexicological approach to the Modern Greek marginal vocabulary*. PhD Thesis, University of Patras.

Christiane Fellbaum and Alexander Geyken. 2005. Transforming a Corpus into a Lexical Resource The Berlin Idiom Project. *Revue française de linguistique appliqué, X*(2):49—62. https://www.cairn.info/revue-francaise-de-linguistique-appliquee-2005-2-page-49.htm

Aggeliki Fotopoulou, Stella Markantonatou, and Voula Giouli. 2014. Encoding MWEs in a conceptual lexicon. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. Association for Computational Linguistics, pages 43-47. https://doi.org/10.3115/v1/W14-0807.

Polona Gantar, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso. 2018. Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*, ecy012, https://doi.org/10.1093/ijl/ecy012

Zoe Gavriilidou. 2013. *Aspects of Intensity in Modern Greek.* Thessaloniki: Kyriakidis Brothers, Ltd. ISBN 978-960-467-445-9.

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

Penny Labropoulou, Elena Mantzari, Maria Gavrilidou. 1996. Lexicon-Morphosyntactic Specifications: Language Specific Instantiation-PP-PAROLE, MLAP (Report) (1996).

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. PARSEME Survey on MWE Resources. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA, pages 2299-2306.

Anne Osherson and Christiane Fellbaum. 2010. The Representation of Idioms in WordNet. In *Proceedings of the Fifth Global WordNet Conference*. Mumbai, India. http://globalwordnet.org/2010/07/10/proceedings-5th-gwa-conference-online-2/.

Panagiotis Minos, Stella Markantonatou, George Zakis, Elpiniki Margariti. 2016. Generating LFG/XLEMWE entries from IDION (a theory neutral lexical DB) Parseme 6th general meeting in Struga/FYROM http://typo.uni-konstanz.de/parseme/index.php/2-general/156-selected-posters-struga-7-8-april-2016

Nikos Sarantakos. 2013. *"Logia toy aera" and more than 1000 fixed expressions.* Athens: Publications of the 21st Century.

Agata Savary et al. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary and Veronika Vincze (eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*, 87-147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471591

Leon Shklar and Richard Rosen. 2009. Web Application Architecture: Principles, Protocols and Practices. West Sussex, England : John Wiley & Sons, Ltd. ISBN 978-0-470-51860-1.

Aline Villavicencio, Timothy Baldwin, and Benjamin Waldron. 2004. A Multilingual Database of Idioms. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1127-1130. http://www.lrec-conf.org/proceedings/lrec2004/.