IDION: A database for Modern Greek multiword expressions*

Stella Markantonatou[1], Panagiotis Minos[1], George Zakis[1], Vassiliki Moutzouri[2] & Maria Chantou[3]
[1]Institute for Language and Speech Processing/Athena RIC, [2]National and Kapodistrian University of Athens, [3]University of Edinburgh
stiliani.markantonatou@gmail.com, pminos@gmail.com, georgizak@gmail.com, vasiliki.moutzouri96@gmail.com, mariachant96@gmail.com

Περίληψη

Με το παρόν άρθρο επιχειρούμε να αναδείξουμε την διαρκή εξέλιξη του ΙΔΙΟΝ (IDION), ενός διαδικτυακού πόρου με πληθώρα τεκμηριωμένων πολυλεκτικών εκφράσεων (ΠΛΕ/ MWEs) της νεοελληνικής γλώσσας (MG) που απευθύνεται εξίσου στον άνθρωπο-χρήστη και στην Επεξεργασία Φυσικής γλώσσας (NLP), ενώ προσφέρεται και για εφαρμογή σε ποικίλους τομείς. Το ΙΔΙΟΝ περιέχει 2.500 ρηματικές ΠΛΕ (VMWEs), από τις οποίες 850 έχουν τεκμηριωθεί ως προς τη συντακτική τους ευκαμψία, τη σημασία τους και τις σημασιολογικές τους σχέσεις με άλλες ρηματικές πολυλεκτικές εκφράσεις. Η οργάνωση αυτού του διαδικτυακού λεξικογραφικού περιβάλλοντος, συμβάλλει στην βέλτιστη τεκμηρίωση των ΠΛΕ της νεοελληνικής.

Λέξεις-κλειδιά: ΙΔΙΟΝ, Ρηματικές πολυλεκτικές εκφράσεις, Επεξεργασία Φυσικής Γλώσσας, λεξικογραφικό περιβάλλον, τεκμηρίωση λήμματος

1 Introduction

We report on the ongoing development of IDION, a web resource of multiword expressions (MWEs) of Modern Greek (MG). IDION is addressed to the human user and to NLP systems. By now, it contains 2500 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 the documentation work and edited the material which was 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 about challenging documentation issues. In Section 3 we provide some detailed information about the developed web editor. Finally, in Section 4 we conclude and discuss about our future priorities.

* 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: (T1ΕΔΚ- 999723442).

2 The documentation

Like other MWE databases, which will be mentioned as our discussion proceeds, IDION serves both interests and research purposes for the human user as well as for natural language processing tasks (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, including the lemma form, the translations, the codification for NLP purposes, the corpus, the elements of syntactic flexibility and verb alternations, possible lexical alternations, as well as elements of semantics (Table 1).

We have defined a 'template' (Section 3) consisting of fields that we fill according to a set of specifications for the appropriate encoding of the MWE properties (Fellbaum and Geyken 2005). Table 1 is used to approximate the design of the IDION template in its present form.

1.1 Entry and lemma definition

A new entry is defined with the unique coupling of a lemma and a definition. This unique coupling is important because many lemmas may be coupled with more than one definition (what is described as "polysemy"); for instance, *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 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 resolve whether multiple entries should be defined or not. 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 different entries. These entries, because of the four distinct lemmas, are encoded both as lexical variants, as they have different fixed verb heads, and as synonyms, as they are assigned the same definition.[1]

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. For instance, example (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.

Lemma form	
Translations	English, French
Codification for NLP	lemma:
	-cranberry words (if any)
	-free XPs (NPs, VPs)
	-optional lemmas
	-morphological constraints
	-contingency

[1] 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.

	-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

It is a universal observation that the maximum length of the fixed strings of VMWEs may vary greatly (Fellbaum and Geyken 2005). We model this phenomenon as optionality, which is denoted with brackets to show the free choice between the appearance of the element or its absence, as illustrated with the preposition in example (2). Variation on fixed functional parts such as prepositions is indicated on the lemma form with disjunction, as illustrated with the variation of *gia* and *os* in example (2).

(1) *apo* *to* *stoma* *mu* *to* *pernis*
from the mouth.ACC mine.POSS it.CL take.2nd.SG
'you say exactly what I was about to say before I utter it'

(2) *Afino* (*gia / os*) *kavatza* *kati*
leave.1st.SG (for / as) buffer something
'I put something aside'

1.2 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. In this way we are mainly aiming towards the reusability of the data 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 also provided in the editor. We apply the use of regular expressions on MG PAROLE strings (Labropoulou et al. 1996) to exhaustively document the morphological constraints on the VMWE

parts. Figure 1 shows the encoding of the morphological constraints on example (1): verb person is constrained to 2nd and 3rd, whereas verb/clitic number is not constrained and the possessive is specified as 1st singular; there is also the possibility for "Xx", which represents an unspecified value in a closed set of values. As for the free parts that constitute the MWE, these are characterised for phrasal category; in the case of noun phrases, these are characterised for animacy and case. More than one NLP-oriented representations can be defined for the same VMWE. This characteristic enables us to treat certain types of lexical variation without creating a new form of the lemma, namely lexical variation on some functional categories (2) and morphological variation/diminutives on most of the fixed content parts. An experiment took place in the past which aimed to convert the NLP oriented representation to an XLE/LFG lexicon and it provided successful results (Minos et al. 2016).

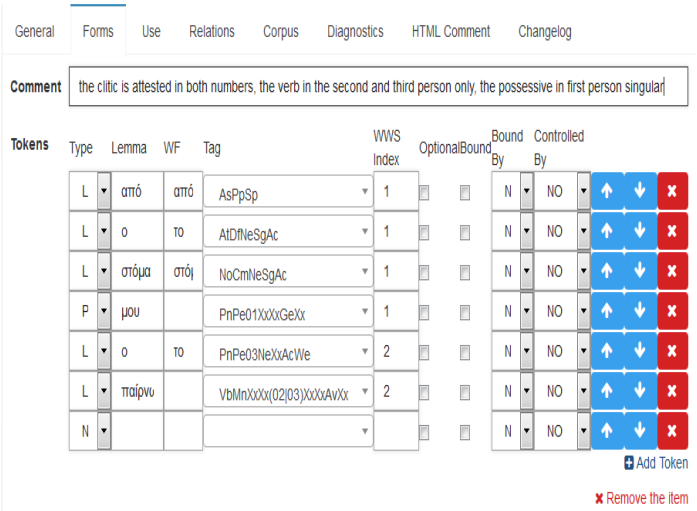

Figure 1 | Representation of (1) for NLP purposes

Seven syntactic flexibility tests are also exemplified with examples retrieved from the corpus. These examples concern the following:

1. Whether the subject of the MWE is free or fixed;
2. The possibility of word order permutations;
3. The insertion of a modifying XP (=NP, AdjP, AdvP, …) among the lexicalised parts of the VMWE (3);
4. The passivisation of the verb;
5. The alternation of dative genitive with a free noun phrase in genitive or a prepositional phrase introduced with the prepositions *se* or *apo* (4);
6. The cliticisation of the fixed object NP (5); and
7. The possibility of causative-inchoative alternation phenomena (6).

(3) *mu* *espases* *pali* *ta* *nevra*

me.DATGEN break.2nd.SG again the nerves.ACC

'you grated on me again'

(4)	*mu*	*spas*	*ta*	*nevra*	-	*spas*	*ta*	*nevra*	*mu*
	me.DATGEN	break.2nd. SG	the	nerves. ACC	-	break.2nd. SG	the	nerves.ACC	mine. POSS

'you grate on me'

(5)	*mu*	*ta*	*espases*	*ta*	*nevra*
	me.DATGEN	the	break.2nd.SG	the	nerves.ACC

'you grated on me'

(6)	*mu*	*espasan*	*ta*	*nevra*
	me.DATGEN	break.3rd.PL	the	nerves.NOM

'I got irritated a lot'

Seen from a different point of view, apart from being annotated for their acceptance and their literal interpretation, corpus examples are also annotated for a list of syntactic phenomena. Therefore, the IDION corpus can be used as a potential resource for empirical NLP systems.

1.3 Collection of annotated examples

Because there are no sizable corpora of texts of Modern Greek, examples are retrieved from the web through specific search; about 8 examples are offered per VMWE, and sometimes more than that when a point needs to be illustrated. 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, according to the seven syntactic flexibility diagnostics mentioned in 2.2, and whether it accepts emphasis (emphasis on modern Greek verbal multiword expressions is a work in progress; for more details see section 4); in short, the corpus contains examples annotated for acceptability and for the phenomena they exemplify. In case the MWE accepts both a literal and a fixed interpretation, then examples are provided to illustrate the literal interpretation of the expression. 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 (Fellbaum and Geyken 2005, Gregoire 2010, DuELME, The Berlin Idiom Project). Apart from the aforementioned patterns of emphasis, we also plan on enhancing IDION with the ability of encoding the frequency of occurrence of the various VMWE alternants.

1.4 Semantics

IDION documents a set of semantic relations among VMWEs (Table 1). These semantic relations include:

1. The definition of VMWE;
2. The case of polysemy;
3. The existence of opposite VMWEs;
4. The existence of synonymous VMWEs;
5. The existence of semantic pairs;
6. VMWEs in Possessive and Stative relations, as well as verb alternations;
7. Polarity, style and emphasis in the VMWEs (work in progress).

Online dictionaries and lexicographic databases, such as Algemeen Nederlands Woordenboek, WordNet, provide synonyms and opposites for the lemmas included in the database.

We have talked about the definition of VMWEs and polysemy in section 2.1. The *Opposite* relation is encoded for VMWE pairs with opposite meanings; for example, for the expression *afino to pedio elefthero*, Lit. I leave the ground free, 'I allow somebody to do what they want' we provide the opposite expression *den afino kapion na kani vima*, Lit. I do not let somebody make a step, 'I am confining somebody'. On the other hand, the *Synonymous* relation is used for VMWE pairs which share the same meaning; for instance, for the expression *vazo ta dio podia se ena papoutsi*, Lit. I put both the feet of somebody in one shoe, 'I bring sb to heel' we provide the synonymous expression *vazo se taksi kapion*, Lit. I put somebody into order, 'I bring somebody to heel'. Furthermore, we devised the term *Semantic pair* to denote pairs of morphologically unrelated predicates that stand in a causative/non-causative relation (9). Moreover, the *Stative* relation is encoded 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'), whereas the *Possessive* relation is encoded 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'). Last but not least, 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.

(7)	*afino*	anavdo	kapion	-	*meno*	anavdos
	leave.1st. SG	speechless. ACC	somebody. ACC	-	stay.1s. SG	speechless. NOM

'I leave somebody speechless - I become speechless'

This set of relations will be exploited to define a network of VMWEs expressing a concept. Such concepts "emerge" from the synonyms sets in a bottom-up way, e.g., the

concept in (1) or of being let down exactly the moment when a desire is about to be satisfied, etc. (see section 4 for future plans on semantically enhancing the database).

3 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, which is structured in 7 tabs:

1. *General*; this tab contains core information on the entry, such as the lemma, the definitions and comments on the definitions both in Greek and English, as well as translations in English and French.
2. *Forms;* this tab contains the theory-independent morphosyntactic information which is provided for each MWE, specifically comments, tokens and word forms, PoS tags as well as tags for phrasal categories (NP, VP), an index for contingency, optionality, anaphor binding, subject control, as well as animacy (whether a free NP should denote an animate) (see section 2.1).
3. *Use;* this tab contains the tokenized form of a glossed corpus example, as well as its English translation and a link to its source. It is accompanied by PoS tags based on MG PAROLE and their phonemic transcription (see section 2.2).
4. *Relations;* this tab is used to encode the semantic relations between the MWEs. It includes the MWE which is related to the expression in view, as well as the six semantic relationships which are tested for each expression: synonymity, opposition, semantic pairs, stativity, possession, and verb alternates (see section 2.4).
5. *Corpus;* this tab contains the examples in the web-based corpus and a link to their source (a web address or the word *introspection*), an indicator for their acceptance as grammatical or not, and an indicator for literal interpretation (see section 2.3).
6. *Diagnostics;* this tab contains seven syntactic flexibility tests, which are further exemplified with real usage examples from the aforementioned corpus. These flexibility tests are: free or fixed subject, internal modification with free XP, word order permutations, cliticisation of the fixed object NP, causative - inchoative alternation, passivisation, and dative genitive alternation. (see section 2.2).
7. *Changelog;* the tab is used to encode information about the documentators and the documentation time.

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.

4 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.

Our future priorities include the following: further populating the IDION database, 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. In addition to the semantic relations, with our future work we plan to develop IDION in a way so as to apply the transitive property on the pairs of synonymous VMWEs. The editors will still need to check the validity of the provided synonymous VMWEs against appropriate contexts, since there is a lack of large corpora and lexicographic resources of MG that would provide a variety of synonyms for each VMWE.

Last but not least, we plan to provide IDION with the ability of encoding polarity, style and emphasis information (Gregoire 2010, Fotopoulou et al. 2014, DuELME, Polytropon). For style, the VMWE will be assigned one of the values *Formal, Colloquial, Offensive* (Christopoulou 2016). To encode polarity information, we plan to use three different values, (-) for VMWEs occurring in negative environments only, (+) for VMWEs occurring in positive environments only and *unspecified* otherwise. As regards to the feature of emphasis, to the best of our knowledge, this kind of information related 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 as a starting point (since a detailed view would require dedicated research), we plan to study 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 will assign the values (+/-) to the feature Emphasis.

Βιβλιογραφία

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/

Christopoulou, Katerina. 2016. "A Lexicological Approach to the Modern Greek Marginal Vocabulary." PhD Diss., University of Patras.

Fellbaum, Christiane 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

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

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

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

Grégoire, Nicole. 2010. "DuELME: a Dutch Electronic lexicon of Multiword Expressions." *Language Resources and Evaluation* 44(1-2):23–39.

Labropoulou, Penny, Elena Mantzari, and Maria Gavrilidou. 1996. "Lexicon-Morphosyntactic Specifications: Language Specific Instantiation." *PP-PAROLE, MLAP*, 63-386 (Report).

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

Minos, Panagiotis, Stella Markantonatou, George Zakis, and 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.

Osherson, Anne 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/.

Sarantakos, Nikos. 2013. "*"Logia toy Aera" and More than 1000 Fixed Expressions.*" Athens: Publications of the 21st Century.

Savary, Agata et al. 2018. "PARSEME Multilingual Corpus of Verbal Multiword Expressions." In *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*, edited by Stella Markantonatou, Carlos Ramisch, Agata Savary and Veronika Vincze, 87-147. Berlin: Language Science Press. DOI: 10.5281/zenodo.1471591

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

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