Chapter 2

Annotating light-verb constructions for Human Language Technologies: The PARSEME-el corpus

Voula Giouli[a]

[a]Aristotle University of Thessaloniki and Institute for Language and Speech Processing, ATHENA RC, Greece

Light-verb constructions (LVCs) are idiosyncratic lexical items, pervasive in many languages. Being complex-verb predicates, they comprise a verb that is light in that it contributes little or no meaning to the phrase and a predicative noun, that is, a noun that has semantic arguments. LVCs—like other Multiword Expressions (MWEs)—are still an obstacle to many natural language processing tasks. Therefore, the existence of quality datasets is a prerequisite for their efficient processing. This chapter introduces a Modern Greek corpus annotated for MWEs, including LVCs. The chapter details the annotation methodology, the guidelines, challenges, and results, highlighting Greek LVC properties. The corpus is available for research via LINDAT/CLARIAH-CZ under a Creative Commons License.

Ως μία κατηγορία πολυλεκτικών εκφράσεων (ΠΛΕ), οι δομές με υποστηρικτικό ρήμα, δηλαδή περιφραστικά ρηματικά κατηγορήματα που αποτελούνται από ένα απολεξικοποιημένο ρήμα και ένα κατηγορικό ουσιαστικό, αποτελούν πρόκληση για διάφορες εφαρμογές Επεξεργασίας Φυσικής Γλώσσας. Τα σώματα κειμένων αποτελούν προϋπόθεση για την αυτόματη αναγνώρισή τους σε κείμενο. Στο κεφά-λαιο αυτό παρουσιάζεται σώμα κειμένων της Νέας Ελληνικής, το οποίο φέρει επισημείωση κατάλληλη για την αναγνώριση ΠΛΕ —μεταξύ των οποίων και δομών με υποστηρικτικό ρήμα. Παρουσιάζεται η μεθοδολογία χειροκίνητης επισημείωσης, με έμφαση στις προδιαγραφές, οι προκλήσεις και τα αποτελέσματα της έρευνας. Το σώμα κειμένων είναι διαθέσιμο στην ερευνητική κοινότητα μέσω του αποθετηρίου LINDAT/ CLARIAH-CZ με άδεια χρήσης Creative Commons.

Voula Giouli. 2024. Annotating light-verb constructions for Human Language Technologies: The PARSEME-el corpus. In Victoria Beatrix Fendel (ed.), *Support-verb constructions in the corpora of Greek: Between lexicon and grammar?*, 33–61. Berlin: Language Science Press. DOI: 10.5281/zenodo.14017923

1 Introduction

Support- or light-Verb constructions[1] have been the focus of attention in natural language processing (NLP henceforth) under the umbrella term Multi-Word Expressions (MWEs henceforth). The latter term encompasses a large variety of linguistic phenomena that range from nominal compounds (i.e., ***cat's eye***), phrasal verbs (i.e., ***give up***, ***take off***), multiword terms (i.e., ***black hole***, ***lithium chloride***), and multiword Named Entities (i.e., ***United Kingdom***, ***United Arab Emirates***) over light-verb constructions (i.e., ***give*** a ***lecture***, ***take*** a ***shower***), to idiomatic expressions (i.e., ***spill the beans***).

According to Sag et al. (2002: 190), MWEs are "idiosyncratic interpretations that cross word boundaries (or spaces)" thus posing challenges to downstream NLP applications. These challenges are due to their lexical, syntactic, semantic, and even pragmatic idiosyncrasies (Gross 1982, Baldwin & Kim 2010). In this regard, considerable effort has been made within the research community to efficiently process them in running text and thus to improve the accuracy of downstream NLP tasks, for example dependency parsing (Nivre & Nilsson 2004), probabilistic parsing (Arun & Keller 2005, Korkontzelos & Manandhar 2010, Constant et al. 2019), or applications such as Machine Translation (Ren et al. 2009, Carpuat & Diab 2010, Bouamor et al. 2012, Zaninello & Birch 2020). Other applications that benefit from automatic Verbal Multi-Word-Expression (VMWE henceforth) identification include automatic text simplification (Kochmar et al. 2020, Gooding et al. 2020, Shardlow et al. 2021), social media mining (Maisto et al. 2017), abusive and offensive language detection (Caselli et al. 2020), and language learning and assessment (Paquot 2019).

In this context, their classification in linguistically grounded categories is useful —a task that poses serious theoretical as well as practical difficulties. Verbal fixed or idiomatic expressions (VIDs henceforth), that is, word sequences that constitute a distinct semantic unit or a complex lexical unit are characterised as having a compound phonological, lexical, and morphological structure and a non-compositional meaning (Gross 1982). Similarly, support-verb or light-verb constructions (LVCs henceforth), that is word combinations that consist of a support or light verb and a predicative noun, that is, a noun that has semantic arguments, are ambiguous and variable across texts.

To facilitate training and testing of tools for MWE processing in running text, datasets are needed that model their properties - especially for languages other

[1]The dataset is accessible via the LINDAT/CLARIAH-CZ repository under a Creative Commons Licence: http://hdl.handle.net/11372/LRT-5124.

than the well-resourced ones including English, and even French, German, Spanish, and Chinese. In this regard, considerable effort has been made within the research community to model them in language resources —both lexica and corpora —in a way that facilitates their robust computational treatment (Constant et al. 2019).

This chapter presents work aimed at developing a corpus of Modern Greek[2] annotated with LVCs in the context of modelling VMWEs in running text. Note that we opt for the term light-verb construction as opposed to the term Support-Verb Construction which is used in the title of the volume since it corresponds to the notation adopted in our annotation scheme. The focus will be on the multilingual setting within which the annotation was performed, the typology of VMWEs that applies to Modern Greek, and the criteria set for classifying candidate VMWEs including LVCs; we will further discuss the methodology adopted for reliably annotating our corpus and the results obtained in terms of the types and properties of LVCs identified in the corpus. We will also report on the inter-annotator agreement focusing on the fuzzy or ambiguous instances that fall in between categories posing, thus, a challenge with regard to their identification.

Our contribution is twofold: on the one hand, we briefly present a multilingual – and, thus, to a great extent universal – annotation scheme, and on the other hand, we present the application of this generic scheme to Modern Greek, focusing on LVCs.

The chapter is structured as follows. In Section 2, we present the rationale and scope of our work and we report on the initiative within which corpus annotation took place, including the definition of a light verb (and light-verb construction); in Section 3, we give an account of previous work on light-verb constructions in Modern Greek. We will then present the Greek corpus in Section 4 focusing also on the typology defined and the annotation methodology adopted (Section 5). In Section 6, we discuss our findings in the corpus, and finally, in Section 7, we conclude.

2 Rationale and scope

Despite being a phenomenon pervasive in many languages, MWEs present lexical, syntactic, semantic, and even pragmatic idiosyncrasies (Gross 1982, Baldwin & Kim 2010), in a way that is not uniform across languages. This is particularly

[2]Modern Greek (EL) —henceforth simply Greek —is the official language spoken in Greece and Cyprus (1453- .).

true for VMWEs of all types, which, by default – like their simple-word counterparts – are used to denote the event, state of affairs, or action conveyed in utterances or text segments. As a result, their robust identification and classification in running text is of paramount importance for downstream NLP applications. Similarly to VIDs, LVCs pose challenges to NLP across the following lines:

- their meaning is semi-compositional in that it cannot be computed simply based on the meaning of their parts and the way they are combined. For example, the LVC (en) to ***give*** a ***stare*** does not imply a *giving* event but rather a *staring* one. This is possibly a pitfall for natural language understanding tasks, mainly those that involve the semantic interpretation of sentences, for example, event identification and Information Extraction;
- there is hardly any cross-lingual equivalence between LVCs, thus rendering their automatic translation problematic. As shown in (1) and (2) the predicative nouns (el) **απόφαση** *apofasi* 'decision' and its translational equivalence (en) ***decision*** select different light verbs in the two languages, namely (el) **παίρνω** *perno* 'take' and (en) ***make*** respectively. The same holds for the German LVC (de) ***Vortrag halten*** (lit. 'to hold a lecture') 'to lecture' and its English counterpart (en) ***to give a lecture***; here, word-order discrepancies are also attested.

  (1) ***παίρνω*** ***απόφαση***
  perno *apofasi*
  take.PRS.1SG decision.SG.ACC
  'to decide'

  (2) to ***make*** a ***decision***
  'to decide'

- when it comes to corpus occurrences, they appear in a variety of surface forms, including long-distance dependencies, as shown in (3) and (4):

  (3) the ***effort*** he ***made*** to remain calm

  (4) he ***gave*** himself one last word of ***advice***.

- moreover, besides an idiosyncratic meaning or reading, literal occurrences of MWEs are also attested —a phenomenon referred to as the *literal-idiomatic ambiguity* (Savary et al. 2019); a case of such ambiguity is shown in (5) and (6).

(5) Mary ***took*** a ***photo*** of the kids playing

(6) He ***took*** the ***photo*** I left on the table.

In this respect, the automatic identification of LVCs in running text is hindered despite the sound linguistic criteria that have been defined. Therefore, our corpus was developed in the framework of PARSEME,[3] a collective effort to create multilingual harmonised language resources, namely annotated corpora and dedicated tools that would serve as a workbench for training and evaluating tools for the robust identification of VMWEs in running text (Savary et al. 2017) and for as many languages as possible.

Over the years, the corpus has been expanded and made available to the research community via frequent releases (Savary et al. 2018, Ramisch et al. 2018, 2020, Savary et al. 2023). Ultimately, the goal was to build a universal framework of VMWE detection taking into account the special characteristics of each language. The working hypothesis, therefore, was that given a universal framework for annotating a linguistic phenomenon in corpora, the idiosyncrasies of discrete languages can be captured. The annotation of the Greek section of the PARSEME initiative seeks to test whether this hypothesis holds.

2.1 The setting: annotation scope

The task of annotating VMWEs in texts can be defined across three axes: (a) spotting all the occurrences of VMWEs in the texts, (b) marking their lexicalised elements as opposed to the non-lexicalised ones, and (c) assigning a tag to the VMWE identified that signals the category it falls into. Therefore, the task is conceived of as a classification one and, in this context, LVC is one of the categories that are foreseen in our typology and the relevant annotation scheme.

Although the exact definition of an LVC varies in the literature, we use the following operational definition: an LVC is a verb-complement pair in which the verb serves as the syntactic head of the phrase, but contributes no lexical meaning and is, therefore, "light"; by contrast, the semantic content of the phrase is retrieved from the complement, being, thus, the semantic head of the expression. The verb is semantically "bleached" contributing to the whole only morphological person, number, tense, and morphological aspect; on the contrary, the complement is a *predicative noun*, that is, one that denotes an event or state, as shown

[3] Parsing and multi-word expressions. Towards linguistic precision and computational efficiency in natural language processing (PARSEME) IC1207.

in (7); the noun is sometimes headed by a preposition whereas, less often, the complement is an adjective as in (8) and (9) respectively.

(7) ***κάνω*** ***ερώτηση***
kano *erotisi*
make.PRS.1SG question.SG.ACC
'to ask'

(8) ***προβαίνω*** ***σε*** ***διαγραφή*** χρεών
proveno *se* *διαγrafi* *chreon*
proceed.PRS.1SG to delisting.SG.ACC debt.PL.GEN
'to delist debts'

(9) ***κάνω*** ***γνωστό***
kano *γnosto*
make.PRS.1SG known.SG.ACC
'to make known'

Two are the main issues to be taken into account here: (a) the definition of a predicative noun, i.e., a noun that is used to predicate the whole phrase, and (b) the operational definition of the light verb. We will elaborate further on the annotation scheme and the framework within which our work is placed in the next sections.

2.2 Annotation scheme

As in any annotation project, the most critical component of our linguistic annotation project was the definition of the annotation scheme that defines the labels and associated features to be linked with the appropriate annotation unit (Ide 2017). This was not a trivial task for our project, —a task that was further hindered by the need to cover languages from different language families. To overcome this obstacle, an experimental procedure was adopted: a set of unified annotation guidelines across many languages from various genera were elaborated which were, then, tested against each language separately.

The outcome was the definition of a VMWE typology that provides the following categories of VMWEs: (a) *Light-verb constructions* (LVCs), which comprise a light verb and a predicative noun or adjective (sometimes headed by a preposition); (b) *Verbal Idioms* (VIDs) which are primarily identified based on semantic properties, i.e., non-compositionality, but also on the grounds of their lexical, syntactic, and pragmatic idiosyncrasies; (c) *Verb-Particle Constructions* (VPCs),

which comprise a verb head and a particle; (d) *Inherently Reflexive Verbs* (IRVs), that is, constructions comprising a verb head and a reflexive pronoun that bear a non-compositional meaning (i.e., (en) to ***find oneself*** in a difficult situation); and (e) *Multi-Verb Constructions* (MVCs), i.e., constructions with two verb heads, for example, (en) to ***let go***, to ***make do***.

In our annotation scheme, LVCs are further distinguished into two subcategories, namely, LVCs in which the verb is semantically totally *bleached* (LVC.full), as in (10), and LVCs in which the verb adds a *causative meaning* to the noun (LVC.cause), as shown in (11).

(10) to ***give*** a ***lecture***

(11) to ***grant*** someone ***rights***
to ***give*** someone a ***headache***

Similarly, the category of VPC is also divided into two subcategories, namely, *fully non-compositional VPCs* (VPC.full), in which the particle changes the meaning of the verb, as opposed to *semi non-compositional VPCs* (VPC.semi), in which the particle adds a partly predictable but non-spatial meaning to the verb; examples of both subcategories are provided in (12) and (13) respectively.

(12) to ***do in***

(13) to ***eat*** something ***up***

Of these, LVCs and VIDs are universal categories, in the sense that they are valid for all the languages participating in the initiative. Similarly, VPCs, IRVs, and MVCs are quasi-universal categories, in the sense that they are valid for some language groups or languages but non-existent or very exceptional in others.

The project also allows languages to define their own, language-specific categories, defined for a particular language in a separate documentation. Finally, to give an account of structures of the type ***to come across*** and ***to rely on***, the optional, experimental category *Inherently Adpositional Verb* (IAV) has been proposed, which (if admitted by a given language) would be considered in the post-annotation step.

The guidelines provide an ordered set of linguistic tests that need to be applied in a series; these tests are visualised as a diagram – called a decision tree – that helps annotators move through its paths to identify and categorise VMWEs —especially in difficult or ambiguous cases. [4] The tests are accompanied

[4]The latest guidelines can be found here: https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.3/?page=050_Cross-lingual_tests/010_Structural_tests__LB_S_RB_.

by language-specific examples, whereas language-specific guidelines are also set for specific cases. Each language or language variety is marked in a different colour or shade. The Greek examples appear in pink.

Tools for handling the data, for the visualisation of the annotations, or for the semi-automatic inspection and manual validation of the data have also been made available to the corpus developers (language leaders). Using these tools ensures to a great extent the quality of the annotations performed.

To render the corpus as uniform as possible across all the participating languages, the pre-processing at the levels of lemma, part-of-speech (POS) tagging, and dependency annotation adheres – for most of the languages – to the Universal Dependencies (UD) guidelines (Nivre et al. 2020). Ultimately, conformance to a widely accepted annotation scheme ensures the development of harmonised corpora.

After all, the primary motivation that guided the creation of this highly multilingual corpus was to boost the VMWE-aware technology across languages. Therefore, a suite of Shared Tasks, that is, competitions for tools aimed at the identification and classification of VMWEs have been organized, and as one might expect the datasets developed have been used as training and testing data. The outcome of this effort is a rich ecosystem, an infrastructure that is as universal as possible taking also idiosyncrasies of each language into account.

3 Previous work: LVCs in Greek

Since initially introduced in the work of Jespersen (1965) for English, the notion of a light verb, that is, a verb that is void of lexical meaning, and therefore its predicational contribution in structures like the ones depicted in (14) is not that of a main verb, has received a lot of attention cross-linguistically. In English, the verbs *have*, *give*, *take*, *make*, *do*, and *get* inter alia, enter in constructions with predicative nouns to form the so-called light-verb constructions.

(14) ***have*** a ***try*** / a ***look*** / a ***shave***
give a ***glance*** / a ***look*** / a ***hint***
make a ***bolt*** / a ***plunge*** / a ***try***

Support- or light-verb constructions have received a lot of attention within the linguistic and computational linguistic community. Arguably, light verbs (and LVCs) are in nature a universal phenomenon, exhibiting, however, several idiosyncrasies in each language in terms of lexical, syntactic, and semantic properties (Grimshaw & Mester 1988, Butt 2003, 2010).

The first systematic attempts towards providing a formal definition of support- or light-verb constructions are found in the works of Gross (1982) and Giry-Schneider (1987) – among others – within the Lexicon-Grammar framework. In an attempt to create a universal Deep-Syntactic paraphrasing system, Mel'čuk (1982, 1996, 2004) tries to describe support or light verbs in the lexicon in terms of Lexical Functions based on French data; later on, he defines lexemic collocations (i.e., ***pay a visit***) as one of the universal categories of phraseological expressions based also on evidence from Russian (Mel'čuk 2023).

In this regard, LVCs are a well-studied area in theoretical linguistics. Our work builds on the findings of previous work on MWEs and LVCs in Modern Greek. Within the Lexicon-Grammar framework introduced by Gross (1975), the properties of VMWEs in Modern Greek were defined initially by Fotopoulou (1993) who developed Lexicon-Grammar tables in which lexical, syntactic, and distributional properties of Greek VIDs were encoded. Within the same framework of Lexicon-Grammar, Moustaki (1995) gives an account of the so-called "frozen" expressions with the support verb (el) **είμαι** *ime* 'to be' in Modern Greek, focusing on structures with prepositions and/or predicative nouns in the genitive or dative cases, and providing their properties at the levels of morphology and syntax.

Along the same lines, support verb constructions in Modern Greek with (el) *'δίνω' dino* 'to give', and (el) *παίρνω perno* 'to take' are presented in Tsolaki (1998). Based on the assumption that the semantic nature of different classes of nominal predicates controls the presence of different kinds of intensifying support verbs and that support verbs intensify a different parameter when they actualise an action, Gavriilidou (2004) gives an account of LVCs in Greek that denote emotion.

Previous studies have set the criteria for the identification of LVCs, and have revealed their properties (Sklavounou 1994, Sfetsiou 2007) also from a computational perspective. Cross-language comparative studies seek to capture the universal nature of LVCs (Fotopoulou & Giouli 2018). In this context, and in an attempt to develop Lexical Resources for NLP applications, Fotopoulou & Giouli (2015) try to develop a battery of formal linguistic tests to delineate support-verb constructions from verbal idiomatic expressions, and to apply them to Greek and French data, focusing on ambiguous cases. These formal tests (i.e., substitution, modification, coordination, etc.) help us classify VMWEs with verbs that are not normally considered light, as LVCs. Thus, verbs like (el) *τρέφω trefo* 'to feed' and (el) *χαίρω chero* 'to enjoy' are considered light when combined with predicative nouns denoting emotion or stance, as shown in (15) and (16.

(15) ***τρέφω*** ***ελπίδες***
trefo *elpides*
feed.PRS.1SG hope.PL.ACC
'to have hope, to hope'

(16) ***χαίρω*** ***σεβασμού***
chero *sevasmu*
enjoy.PRS.1SG respect.SG.GEN
'to be respected'

4 Corpus description

In contrast to corpora for other languages, the development of the Greek corpus spans consecutive releases due to a lack of substantial (human) resources. Over the years, the corpus has been gradually enhanced and enriched, and consecutive editions were released. The main design criteria for the textual material —set up for all languages centrally —were that texts should be written in the original rather than be translated and should be free from copyright issues, so as to be distributed under an open license.

The corpus comprises two main sub-corpora: (a) a collection of texts manually collected from various sources on the web; and (b) a part of the Greek Dependency Treebank (GDT henceforth) (Prokopidis & Papageorgiou 2017). The first sub-corpus was compiled manually by collecting raw data manually from the electronic version of major Greek newspapers (ΚΑΘΗΜΕΡΙΝΗ, ΠΡΩΤΟ ΘΕΜΑ, TA NEA, Athens Voice, etc.), news portals as well as Wikipedia articles; moreover, texts from news blogs (gova stileto, tromaktiko, etc.) and life-style and gossip news texts (espresso, etc.) were also collected; the latter bear a rather informal register. We managed to cover a variety of text genres, including newswire texts, press releases, opinion and popular science articles in various domains like medicine, physics, finance, etc., whereas the GDT also includes parliamentary debates.

The so-collected textual data were pre-processed at the lemma, POS and dependency annotation levels; all these annotations were performed automatically using UDpipe (Straka & Straková 2017) and the latest models for the Greek language. Due to time and scope constraints, no manual annotation of the preprocessing stages has been performed. To somehow remedy this shortcoming and further enrich our corpus with data manually annotated at the aforementioned levels of linguistic analysis, we also included part of the Greek Dependency Treebank that has been manually annotated and rendered compatible with

the Universal Dependencies initiative (Nivre et al. 2020). In essence, this is our so-called GOLD part of the corpus – GOLD in all levels. It should be noted that within the NLP community, the term GOLD STANDARD – or simply GOLD – corpus refers to quality text collections manually annotated, usually by experts. The annotation at both the VMWE level and at the levels of POS and dependency annotation can be viewed via Grew-match (Guillaume 2021) a dedicated tool for visualising and querying annotated corpora, as shown in Figure 1.

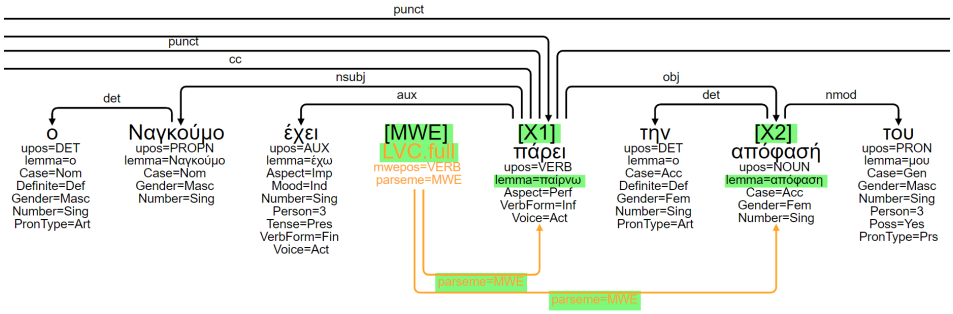

Figure 1: GrewMatch: the annotated VMWE is highlighted.

Using the tools that were made available for all the teams, we managed to improve the quality of the corpus by spotting discrepancies between annotators, and adjudicating as appropriate, ultimately providing annotations that are consistent throughout the corpus.

In the latest release (version 1.3) of the PARSEME corpus, the Greek section (PARSEME-el) amounts to 698,424 tokens or 26,175 sentences (Savary et al. 2023) in which a total of 8,508 VMWEs have been identified of which LVCs are the most frequently occurring category – see Table 1.

Since the corpora were primarily developed to be used as a dataset for the Shared Tasks, the corpus for each language was split into three subsets: the training, development, and evaluation subsets. The former is provided by the Shared Task organisers to the participants to train their MWE identification systems, whereas the development sub-corpus is used to perform model selection and fine-tuning; the evaluation of the systems is performed against the test sub-corpus. Splitting into the three sub-corpora is performed based on specific criteria, and in a way that ensures that there is a balance between the development and test parts of the corpus in terms of VMWEs not previously seen (Ramisch et al. 2020, Savary et al. 2023).

Table 1: The PARSEME-el corpus in numbers for the latest releases.

	Release 1.2	Release 1.3
LVC.full	4,696	5,293
LVC.cause	122	179
VID	2,297	2,842
VPC.full	119	143
MVC	48	51
Total	7,282	8,508

5 Annotation methodology

Like all the corpora for all the languages, the Greek corpus was manually annotated for VMWEs as per the guidelines. It should be noted that before annotation proper, a two-phase pilot annotation was performed: during pilot annotation phase 1, two trained linguists, native speakers of the Greek language with extensive experience in annotation tasks and VMWEs alike, worked towards the development and testing of the universal guidelines; during annotation pilot phase 2, extended annotation of naturally occurring text took place and resulted in the consolidation of the universal guidelines. After the guidelines were consolidated, language-specific examples were elaborated as appropriate to help annotators assess difficult or ambiguous cases.

Annotation proper was performed with the aid of the FoLiA Linguistic Annotation Tool (FLAT), a dedicated web-based multi-user and open-source annotation platform (van Gompel & Reynaert 2013). FLAT allows for the annotation of non-contiguous structures and is customised to support the file format adopted by PARSEME. Following the specifications set early in the lifecycle of the project, in this annotation task, all the occurrences of VMWE categories were annotated in the text, as shown in Figure 2. Over the years, expert annotators – all native speakers of the language – contributed to the task of annotation.

Initially, annotations were performed by each annotator separately; annotators then met to discuss difficult and ambiguous cases. After this initial training period was over, annotators worked separately.

However, the task of manually annotating data is always demanding and prone to all sorts of errors. We calculate the degree of inter-annotator agreement in order to assess the consistency or reliability of annotations made by different annotators for the same spans of text. This measure helps us understand the

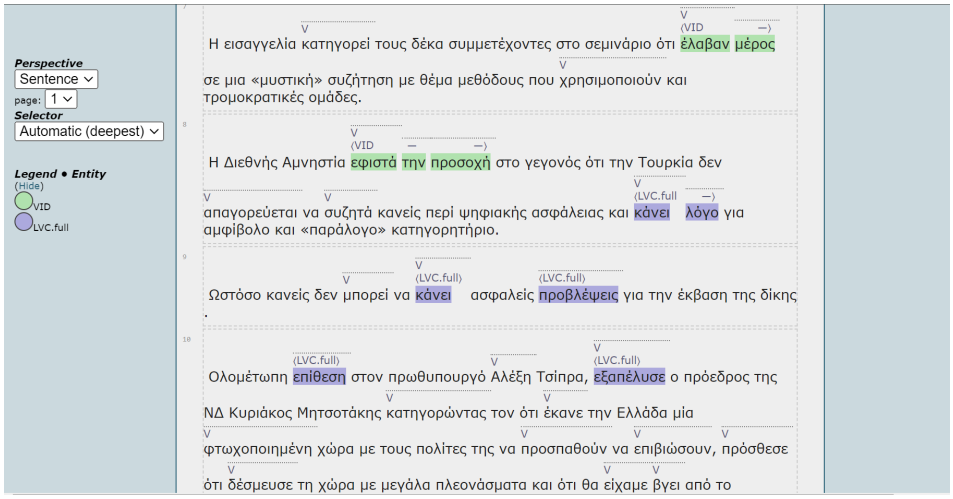

Figure 2: Annotating VMWEs in FLAT.

level of agreement between annotators when labeling data. Ultimately, it is a measure that shows the extent to which multiple annotators can make the same annotation decision for a certain category.

The inter-annotator agreement rate gives us an estimate of how clear the annotation guidelines are, how uniformly the team of annotators understood the guidelines, and how reproducible the annotation task is. High inter-annotator agreement indicates that annotators are interpreting the guidelines consistently and are reaching similar conclusions enhancing the reliability of the annotated data. On the other hand, low inter-annotator agreement suggests inconsistencies or discrepancies in the annotations, which may signal the need for clarifying guidelines or additional training for annotators to improve the quality of the annotations.

Therefore, to ensure the quality of the annotated corpus, a fragment of the data was annotated by all the team members who viewed the data independently. Then, the agreement between annotators was measured using a standard metric, namely Cohen's kappa co-efficient (Carletta 1996, Artstein & Poesio 2008) using the VMWEs for which annotators agree on the span of the VMWE.

The annotation *span* or *scope* is determined by the lexicalised or fixed elements that can form a separate word. Therefore, determiners, modifiers, auxiliaries, and paricles are included in the markable only if they are lexicalised. As shown in (17), the determiner (el) *την tin* 'the' and the pronoun (el) *μου mu* 'my' are not included in the span of the VMWE because they are not fixed (or integral) parts

of the expression. Identifying the lexicalised elements of an expression is not always a trivial task.

(17) ***πήρα*** την ***απόφασή*** μου
pira *tin* *apofasi* *mu*
take.PST.1SG the.SG.ACC hope.SG.ACC my.SG.GEN
'I made my decision, I decided.'

Additionally, we used the F-score metric, since it is particularly relevant in applications that are primarily concerned with the positive class. Note that in our annotation project, negative cases were not annotated.[5] The F-score measures a system's accuracy and is calculated as the harmonic mean of a system's precision and recall values. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'. In our case, the F-score is measured based on the annotations of pairs of raters. One rater is considered the one providing the GOLD annotation (as senior or expert annotator) and the other is the one providing the system's output. The F-score was 68.6 and Cohen's kappa was equal to 0.632 for the Greek data (Savary et al. 2018) – one of the best scores among the participating languages. In this way, the quality of our corpus is ensured.

Apart from LVCs, the Greek section of the PARSEME corpus bears annotations for verbal idioms, as well as verb-particle and multi-verb constructions. In Modern Greek, we retained the two universal VMWE categories, namely VIDs (verbal idioms) which have an entirely non-compositional meaning as in (18), and LVCs of both sub-categories. In this regard, cases in which the light verb contributes to the meaning of the whole only morphological features (i.e., tense, grammatical aspect, number, and person) are annotated as LVC.full as in (19); on the contrary, they are annotated as LVC.cause once the light verb is causative, in that it indicates that the subject of the verb is the cause or source of the event or state expressed by the predicative noun; these cases are expected to be less idiomatic than other VMWEs and can be understood as complex predicates with a causal support verb, as shown in (20).

(18) ***βάζω*** ***λάδι*** ***στη*** ***φωτιά***
vazo *ladi* *sti* *fotia*
put.PRS.1SG oil.SG.ACC to.the.SG.ACC fire.SG.ACC
'make things even worse'

[5] Given a candidate VMWE, a positive case is when it is considered idiomatic and is therefore annotated, whereas a negative case is when the same candidate is used literally.

(19) ***κάνω*** ***επίσκεψη***
kano *episkepsi*
make.PRS.1SG visit.SG.ACC
'to pay a visit, to visit'

(20) ***προκαλώ*** ***καταστροφή***
prokalo *katastrofi*
cause.PRS.1SG destruction.SG.ACC
'to cause destruction, to destroy'

Our language-specific annotation scheme includes two semi-universal categories, namely MVC (Multi-Verb Constructions) and VPC (Verb-Particle Constructions). MVCs in Greek are phrases that comprise two verbs, a *vector verb* that is the functionally governing verb (*V-gov*) and a *polar verb* that functions as the dependent verb (*V-dep*); in a dependency-based syntactic analysis, *V-gov* might be seen as the head and *V-dep* as the dependent and they have a shared subject. Ultimately, the two verbs function as a single predicate with non-compositional semantics, as shown in (21).

(21) ***απορώ*** ***και εξίσταμαι***
aporo *ke* *eksistame*
wonder.PRS.1SG and get-surprised.PRS.1SG
'to question myself'

As VPCs, on the other hand, we have annotated those verb + adverb constructions, in which the adverb shares characteristics with particles in languages like English, shown in (22).

(22) ***βάζω*** κάποιον ***μέσα*** / ***βάζω*** ***μέσα*** κάποιον
vazo *kapion* *mesa* / *vazo* *mesa* *kapion*
put.PRS.1SG someone.SG.ACC in / put.PRS.1SG in someone.SG.ACC
'to cause someone to go bankrupt'

As we have already mentioned, the annotation guidelines are universal but were adopted in a way that the idiosyncrasies of each language are taken into account. We opted for retaining the category of VPCs, based on linguistic tests that proved that the adverbs in question exhibit most, if not all, of the properties that particles in other languages have (Giouli et al. 2019).

As argued in Giouli et al. (2024), these adverbs are not morphologically derived from adjectives, and they have two distinct functions: as adverbs denoting time or

location, they are used as modifiers; when combined with prepositions, they form complex prepositions (Holton et al. 1997), for example (el) *μπροστά από brosta apo* (lit. 'in-front from') 'in front of', (el) *μέσα σε mesa se* (lit. 'in to') 'in', etc. Therefore expressions of the form (el) *πέφτω μέσα pefto mesa* (lit. 'fall in') 'to guess correctly' and (el) *βάζω μπρος vazo bros* (lit. 'put in-front') 'to start' were classified as VPCs.

In terms of their semantics, Greek VPCs were identified as non-compositional in meaning. As previously shown (Savary et al. 2019), these constructions are the most ambiguous. Depending on the context, they can be used literally and have a fully compositional meaning. In that case, they are not VMWEs. In the remainder, we will focus on the annotation of LVCs.

6 LVCs in the Greek section of the PARSEME corpus

6.1 The data

When it comes to annotation, there are two major questions that annotators need to tackle: (a) what to annotate, and (b) how to annotate. The former question – "what to annotate" – has to do with the linguistic phenomenon that we need to capture, which also comes with the extra flavour of "how much" to annotate. The latter brings to mind the question of the markable extent that is always crucial —especially when computational aspects are entailed. In other words, we need to specify the string length and the elements that must be annotated.

In the case of VMWEs in general (and LVCs in particular), we annotate as integral parts all lexicalised elements of the expression that form a separate word. We consider lexicalised those elements that have some sort of morphological, syntactic, or lexical idiosyncrasy or fixedness. For instance, determiners and modifiers of the predicative nouns are not lexicalised, and therefore, they are not part of the markable; similarly, auxiliaries or other dependents of the light verb are not included in the annotation, as shown in (23).[6]

(23)	ο	Ναγκούμο	έχει	***πάρει***	την	***απόφασή***
	o	*Nagoumo*	*echi*	*pari*	*tin*	*apofasi*
	the.SG.NOM	Nagoumo	have.PRS.3SG	take.INF	the.SG.ACC	decision.SG.ACC
	του.					
	tu					
	his.3.SG					
	'Nagoumo has decided'					

[6] According to the notation followed, the lexicalised elements of the expression that are marked in boldface are annotated.

The question of "what to annotate" is tackled by the annotation guidelines that we have already mentioned and the operational definition of LVCs provided. This definition obviously includes two elements as integral parts of an LVC: a verb head with void semantics (the syntactic head) and a predicative noun that serves as the semantic head of the expression.

This entails that phrases that comprise *aspectual variants* of light verbs, i.e., verbs that contribute an aspectual meaning to the expression once they substitute the light verb proper were not taken into account and not annotated – a decision that has received criticism (Fotopoulou et al. 2021). In theoretical linguistics, these aspectual variants are usually studied under the umbrella term of LVCs (Gross 1982, Giry-Schneider 1987). However, there are discrepancies between the two which we wish to keep for later study. In this respect, the expression (el) ***δίνω απάντηση*** *dino apantisi* (lit. 'give answer) 'to answer' is annotated as an LVC, whereas its aspectual variant (el) ***παίρνω απάντηση*** *perno apantisi* (lit. 'take answer') 'to receive an answer' is not.

Once again, the data prove the assertion that LVCs form a very productive category of highly idiosyncratic expressions, in that predicative nouns select their syntactic head instead of verbs selecting their dependents, see (24).

(24) ***παίρνω*** ***απόφαση*** / ****κάνω*** ***απόφαση***
perno *apofasi* / *kano* *apofasi*
take.PRS.1SG decision.SG.ACC / make.PRS.1SG decision.SG.ACC
'to make a decision, to decide'

In our corpus, the most frequently encountered light verbs are (el) *κάνω kano* 'to make, to do', *έχω echo* 'to have', *παίρνω perno* 'to take', and *δίνω dino* 'to give'. Other light verbs include (el) *ασκώ asko* 'to exert', *βάζω vazo* 'to put', *βγάζω vgazo* 'to take.out', *βγαίνω vgeno* 'to go.out', *θέτω θeto* 'to put, to set', *καταβάλλω katavalo* 'to give', *λαμβάνω lamvano* 'to get', *κρατάω kratao* 'to keep', *παρέχω parecho* 'to provide', *αναλαμβάνω analamvano* 'to undertake', *αποδίδω apodido* 'to give', *διαπράττω diapratto* 'to commit', *διενεργώ dienergo* 'to carry out', *διεξάγω διexαγο* 'to conduct', *εκπονώ ekpono* 'to conduct, to carry out', *εκτελώ ektelo* 'to execute, to carry out', and *έρχομαι erchome* 'to come'.

Alternative light verbs also occur with the same predicative noun, often signalling a shift in the register. In most cases, pairs of verbs like *παίρνω perno* ('take') and *λαμβάνω lamvano* ('take'), or *κάνω kano* ('make') and *ασκώ asko* ('exert') are variants, the latter bearing a formal register, as attested in (25) and (26).

(25) ***παίρνω*** ***απόφαση*** / ***λαμβάνω*** ***απόφαση***
perno *apofasi* / *lamvano* *apofasi*
take.PRS.1SG decision.SG.ACC / take.PRS.1SG decision.SG.ACC
'to make a decision, to decide'

(26) ***κάνω*** ***κριτική*** / ***ασκώ*** ***κριτική***
kano *kritiki* / *asko* *kritiki*
make.PRS.1SG criticism.SG.ACC / exert.PRS.1SG criticism.SG.ACC
'to criticise'

Similarly, some sort of lexical variation is due to the predicative noun used – notably in the case of LVCs with loan words (neologisms) and terms. For instance, the predicative nouns (el) *εκφοβισμό ekfovismo* 'bullying' and *μπούλιγκ bullying* 'bullying' in (27) and (28) are synonymous – the latter being a loanword that has been adopted in Greek (target language) as a transliterated form of the term *bullying* in English (source language). The loanword is also attested in the corpus as non-transliterated, keeping, thus, the orthography of the source language.

(27) ***κάνω*** ***εκφοβισμό*** / ***ασκώ*** ***εκφοβισμό***
kano *ekfovismo* / *asko* *ekfovismo*
make.PRS.1SG bullying.SG.ACC / exert.PRS.1SG bullying.SG.ACC
'to bully'

(28) ***κάνω*** ***μπούλιγκ*** / ***κάνω*** ***bullying***
kano *bullying* / *kano* *bullying*
make.PRS.1SG bullying.SG.ACC / make.PRS.1SG bullying.SG.ACC
'to bully'

The phenomenon of *language mixing* which "is understood as involving lexical items and grammatical features from two languages that appear in one sentence, [...] can either be word internal, [...] or involve lexical elements of two languages", has been studied for bilingual speakers of many languages/language pairs, including Greek Alexiadou (2017: 166).

In our news corpus, this type of mixing is attested in texts that belong to specific domains. For instance, LVCs with loanwords such as (el) ***κάνω σουτ*** *kano sut* ('make shoot') 'to shoot' in the domain of SPORTS is used in parallel with the derived verb (el) *σουτάρω sutaro* 'to shoot'. Similarly, LVCs of the form (el) ***κάνω πρέσιγκ*** *kano pressing* ('make pressing') 'to press' are attested in the domain of FINANCE. Finally, LVCs of this type are abundant in the sub-corpus of lifestyle texts. In the next sections, we will elaborate on the linguistic properties of LVCs as they are attested in the corpus.

6.2 Linguistic properties of LVCs

Our data reveal the linguistic properties of LVCs. As in many other languages, most of our LVCs are morphologically related to a full verb that can 'replace' them without a significant change in meaning. Therefore, (el) ***δίνω υπόσχεση*** *δino iposchesi* 'to give a promise' can be replaced by the verb *υπόσχομαι iposchome* 'to promise'. According to the guidelines, this was the primary linguistic test used while annotating. Where a morphologically related verb was not found, we checked for a synonymous one to use. To this end, the linguistic tests of lexical substitution or lexical and phrasal paraphrasing were applied.

A high degree of variation was also attested in the corpus, namely morphological, syntactic, and lexical variation. As it has been noticed in many studies, for example, (Butt 2010), the predicative noun may be used in the plural:

(29) ***δίνω*** ***υπόσχεση*** / ***δίνω*** ***υποσχέσεις***
δino *iposchesi* / *δino* *iposchesis*
give.PRS.1SG promise.SG.ACC / give.PRS.1SG promises.PL.ACC
'to make a promise, to promise'

Syntactic variants of LVCs are also attested quite often in the corpus - the most frequent one being LVCs that enter in diathesis alternations (passive, causative-inchoative), as shown in (30) and (31).

(30) ***έλαβα*** μία δύσκολη ***απόφαση***
elava *mia diskoli* *apofasi*
take.PST.1SG one difficult apofasi.SG.ACC
'I made a tough decision'

(31) ***ελήφθησαν*** δύσκολες ***αποφάσεις***
elifθisan *diskoles* *apofasis*
take.PASS.PST.3PL difficult.PL.NOM apofasi.PL.NOM
'Tough decisions were made'

Note that in some cases, different verbs signal diathesis alternation. LVCs which comprise certain pairs of light verbs combined with the same predicative noun signal syntactic alternations (i.e., diathesis alternation, causative-inchoative alternation, etc.). This is mainly true for pairs of verbs like (el) βγάζω *vgazo* 'to take out' and (el) βγαίνω *vgeno* 'to be taken out', or (el) κάνω *kano* 'to do, to make' and (el) γίνομαι *ginome* 'to be made'. They predominately differ in the grammatical features and the syntactic function that the predicative noun assumes.

For example, the LVCs (el) ***βγάζω συμπέρασμα*** *vgazo symperasma* (lit. 'take-out.PRS.1SG conclusion.SG.ACC') 'to conclude' and (el) ***βγαίνει συμπέρασμα*** *vgeni symperasma* (lit. 'is-taken-out.3SG conclusion.SG.NOM') 'it is concluded' enter in the causative-inchoative alternation. In the former, the lexicalised element is the argument in object position (and following the rules of the language, it is realised as a Noun Phrase (NP henceforth) in the accusative case); in the latter, the predicative noun is the subject and is realised as an NP in the nominative, as shown in (32) and (33).

(32) Οι πολίτες ***βγάζουν*** τα ***συμπέρασματά***
I *polites* *vgazun* *ta* *simperasmata*
The.PL.NOM citizen.PL.NOM take.out.PRS.3SG the.PL.ACC conclusion.PL.ACC
τους.
tus
their3SG
'Citizens come to a conclusion.'

(33) ***Βγαίνει*** το ***συμπέρασμα*** ότι η χώρα
vgeni *to* *simperasma* *oti* *i* *chora*
go.out.PRS.3SG the.SG.NOM conclusion.SG.NOM that the country
κινδυνεύει.
kindinevi
is-in-danger
'It is concluded that the country is in danger.'

According to the universal guidelines, nominal groups (headed by nominal complements taken from the prototypical LVCs) with relative clauses are also annotated. As a matter of fact, the structure in (34) is also used as a test for deciding whether a candidate LVC should be annotated or not. The test is shown in the decision tree of the guidelines.

(34) η ***απόφαση*** που ***πήραμε***
i *apofasi* *pu* *pirame*
the.SG.NOM decision.SG.NOM that take.PST.1PL.PRES
'the decision we made'

LVCs in running text sometimes appear as constructions in which the predicative nouns share the same verb head, as shown in (35). These LVCs are annotated separately.

(35) η κυβέρνηση ***έχει*** τη ***βούληση*** και την
i *kivernisi* *echi* *ti* *vulisi* *ke* *tin*
the.SG.NOM government.SG.NOM have.PRS.3SG the volition and the
ικανότητα
ikanotita
ability
'the government wants and can'

Insertion of other elements, for example, modifiers, and determiners, are a serious drawback not only to systems that seek to automatically identify VMWEs in text but also to human annotators. In effect, long-distance dependencies, that is, dependencies that need not hold between strictly linearly adjacent words or morphemes, are problematic to annotators as well. In most cases, LVCs are non-continuous constructions; sometimes, the elements of the LVC are completely discontinous.

6.3 Ambiguous cases

The distinction between LVCs and fixed or idiomatic expressions is not always straightforward and the limits between the two are often fuzzy. According to Fotopoulou & Giouli (2015) among others, there exists a scalar passage between the two types of VMWEs. The annotation guidelines provide robust linguistic tests that guide annotation. After all, the task of annotation - any annotation - is a deterministic one; decisions need to be made.

Sometimes, synonymous VMWEs fall into different categories based on the noun: if the noun is predicative, the expression is tagged as an LVC, as shown in the examples. We consider predicative a noun that denotes an event, a situation, or a sentiment, etc. (Gross 1975, 1982). VIDs, on the other hand, are defined as having a non-compositional meaning that cannot be deduced from the meaning of their parts (Gross 1982). According to this principle, the noun (el) *ρεζίλι rezili* 'ridicule' in (36) is predicative, whereas the noun (el) *ρόμπα roba* 'robe' in (37) is not.

(36) ***κάνω*** κάποιον ***ρεζίλι*** (LVC)
kano *kapion* *rezili*
make.PRS.1SG someone ridicule.SG.ACC
'to ridicule'

(37) ***κάνω*** κάποιον ***ρόμπα*** (VID)
kano *kapion* *roba*
make.PRS.1SG someone robe.SG.ACC
'to ridicule'

Literal occurrences of MWEs, also referred to as their literal readings or literal meanings, have received considerable attention equally from the linguistic and computational linguistic communities. In an experiment run for German, Greek, Basque, Polish, and Brazilian Portuguese, (Savary et al. 2019) almost 11.5% of the VMWE occurrences in the Greek corpus were found to be literal readings of the VMWE surface forms – a phenomenon referred to as the *literal-idiomatic ambiguity*.[7] Literal occurences of LVCs were not annotated.

7 Conclusion and outlook for future research

We have presented a corpus of Modern Greek that has been annotated for VMWEs within the framework of a highly multilingual initiative that currently covers 26 languages and language varieties. Before presenting our work, the definition of LVCs in our approach was given. Our work is primarily intended to serve applications in the field of natural language processing, where LVCs are generally treated under the umbrella term MWEs, and to prepare a corpus for Modern Greek that is compatible with multi-lingual initiatives. From another perspective, the corpus and the accompanying infrastructure can be used for the study of LVC-related phenomena.

Future work has already been envisaged towards enriching the corpus with new data and extending the annotation scheme to new grammatical categories, for example, nominal or adverbial MWEs. Of great importance in the future are the adjudication of the pre-processing levels, so as to have a corpus resource that is GOLD at all the levels of linguistic analysis. This will allow us – among other things – to provide the research community with a corpus that is usable for linguistic analyses.

[7]For a definition of the literal-idiomatic ambiguity, see Savary et al. (2019).

Abbreviations

FLAT	oLiA Linguistic Annotation Tool
IAV	Inherently Adpositional Verb
IRV	Inherently Reflexive Verbs
LVC	Light Verb Construction
MVC	Multi-Verb Constructions
MWE	Multiword Expression
NLP	Natural Language Processing
NP	Noun Phrase
POS	Part-of-Speech
UD	Universal Dependencies
VMWE	Verbal Multiword Expression
VID	Vebal Idiomatic Expression
VPC	Verb-Particle Construction

Acknowledgements

The work described in this chapter has been supported by the IC1207 PARSEME COST action. The author is also grateful to the annotators of the Greek section of the PARSEME corpus for their contribution. We thank the anonymous reviewers and the editor for their insightful comments, which led to a significant improvement in the original text.

References

Alexiadou, Artemis. 2017. Building verbs in language mixing varieties. *Zeitschrift für Sprachwissenschaft* 36(1). 165–192. DOI: 10.1515/zfs-2017-0008.

Artstein, Ron & Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596. DOI: 10.1162/coli.07-034-R2.

Arun, Abhishek & Frank Keller. 2005. Lexicalization in crosslinguistic probabilistic parsing: The case of French. In *Proceedings of ACL 2005*, 306–313.

Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Fred J. Damerau & Nitin Indurkhya (eds.), *Handbook of Natural Language Processing*, 2nd edn., 267–292. New York: Chapman & Hall/CRC.

Bouamor, Dhouha, Nasredine Semmar & Pierre Zweigenbaum. 2012. Automatic construction of a multiword expressions bilingual lexicon: A statistical machine translation evaluation perspective. In *Proceedings of COLING2012*, 95–107.

Butt, Miriam. 2003. The light verb jungle. In Gulsat Aygen, Claire Bowern & Conor Quinn (eds.), *Harvard Working Papers in Linguistics. Papers from the GSAS/Dudley House Workshop on light verbs*, vol. 9, 1–49.

Butt, Miriam. 2010. The light verb jungle: Still hacking away. In Mengistu Amberber, Brett Baker & Mark Harvey (eds.), *Complex predicates: Cross-linguistic perspectives on event structure*, 48–78. Cambridge: Cambridge University Press.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics* 22(2). 249–254.

Carpuat, Marine & Mona Diab. 2010. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. In *Proceedings of NAACL/HLT2010*, 242–245.

Caselli, Tommaso, Valerio Basile, Jelena Mitrović, Inga Kartoziya & Michael Granitzer. 2020. I feel offended, don't Be abusive! Implicit/explicit messages in offensive and abusive language. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the twelfth Language Resources and Evaluation Conference*, 6193–6202. Marseille, France: European Language Resources Association.

Constant, Mathieu, Gülşen Eryiğit, Carlos Ramisch, Mike Rosner & Gerold Schneider. 2019. Statistical MWE-aware parsing. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of Multiword expressions: Current trends*, 147–182. Berlin: Language Science Press.

Fotopoulou, Angeliki. 1993. *Une classification des phrases à compléments figés en grec moderne: étude morphosyntaxique des phrases figées*. Université Paris VIII. (Doctoral dissertation).

Fotopoulou, Angeliki & Voula Giouli. 2015. MWEs: Support/light verb constructions vs fixed expressions in Modern Greek and French. In Gloria Corpas Pastor and Johanna Monti and Violeta Seretan and Ruslan Mitkov (ed.), *Workshop on Multiword Units in Machine translation and translation technology*, 68–73. Malaga, Spain: Tradulex.

Fotopoulou, Angeliki & Voula Giouli. 2018. MWEs and the Emotion Lexicon: Typological and cross-lingual considerations. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 63–91. Berlin: Language Science Press.

Fotopoulou, Angeliki, Eric Laporte & Takuya Nakamura. 2021. Where do aspectual variants of light verb constructions belong? In Paul Cook, Jelena Mitrović, Carla Parra Escartín, Ashwini Vaidya, Petya Osenova, Shiva Taslimipoor & Carlos Ramisch (eds.), *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, 2–12. Association for Computational Linguistics.

Gavriilidou, Zoé. 2004. Verbes supports et intensité en grec moderne. *Lingvisticæ Investigationes* 27(2). 295–308.

Giouli, Voula, Vasiliki Foufi & Angeliki Fotopoulou. 2019. Annotating Greek VMWEs in running text: A piece of cake or looking for a needle in a haystack? In Maria Chondrogianni, Simon Courtenage, Geoffrey Horrocks, Amalia Arvaniti & Ianthi Tsimpli (eds.), *Proceedings of the 13th International Conference on Greek Linguistics (ICGL 13)*, 125–134.

Giouli, Voula, Vera Pilitsidou & Hephestion Christopoulos. 2024. A FrameNet approach to deep semantics for MWEs. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 147–168. Berlin: Language Science Press.

Giry-Schneider, Jacqueline. 1987. *Les prédicats nominaux en français. Les phrases simples à verbes supports*. Genève: Librarie Droz.

Gooding, Sian, Shiva Taslimipoor & Ekaterina Kochmar. 2020. Incorporating Multiword Expressions in phrase complexity estimation. In Núria Gala & Rodrigo Wilkens (eds.), *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, 14–19. Marseille, France: European Language Resources Association.

Grimshaw, Jane & Armin Mester. 1988. Light verbs and θ-marking. *Linguistic inquiry* 19. 205–232.

Gross, Maurice. 1975. *Méthodes en syntaxe: Régime des constructions complétives*. Paris: Hermann.

Gross, Maurice. 1982. Une classification des phrases « figées » du français. *Revue québécoise de linguistique* 11(2). 36–41.

Guillaume, Bruno. 2021. Graph Matching and Graph Rewriting: GREW tools for corpus exploration, maintenance and conversion. In Dimitra Gkatzia & Djamé Seddah (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 168–175. Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-demos.21.

Holton, David, Peter Mackridge & Irene Philippaki-Warburton. 1997. *Greek: A comprehensive grammar of the modern language*. London & New York: Routledge.

Ide, Nancy. 2017. Introduction: The Handbook of Linguistic Annotation. In Nancy Ide & James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, 1–18. Dordrecht: Springer Netherlands.

Jespersen, Otto. 1965. *A Modern English grammar on historical principles, part VI, morphology*. London: George Allen & Unwin Ltd.

Kochmar, Ekaterina, Sian Gooding & Matthew Shardlow. 2020. Detecting Multiword Expression type helps lexical complexity assessment. In *International Conference on Language Resources and Evaluation*.

Korkontzelos, Ioannis & Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *Proceedings of NAACL/HLT 2010*, 636–644.

Maisto, Alessandro, Serena Pelosi, Simonetta Vietri & Pierluigi Vitale. 2017. Mining offensive language on social media. In Roberto Basili, Malvina Nissim & Giorgio Satta (eds.), *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, 252–256. Torino: Accademia University Press. DOI: 10.4000/books.aaccademia.

Mel'čuk, Igor. 2004. Verbes supports sans peine. *Lingvisticæ Investigationes* 27(2). 203–217. DOI: 10.1075/li.27.2.05mel.

Mel'čuk, Igor. 2023. *General phraseology: Theory and practice*. Amsterdam: John Benjamins.

Mel'čuk, Igor. 1982. Lexical functions in lexicographic description. In *Proceedings of the VIIIth Annual Meeting of the Berkeley Linguistics Society*. Berkeley, California: University of California, Berkeley.

Mel'čuk, Igor. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In Leo Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, 37–102. Amsterdam & Philadelphia: John Benjamins.

Moustaki, Argyro. 1995. *Les expressions figées être prép C W en grec moderne*. Thèse de doctorat dirigée par Gross, Maurice. Université Paris VIII. (Doctoral dissertation).

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the twelfth Language Resources and Evaluation Conference*, 4034–4043. Marseille, France: European Language Resources Association.

Nivre, Joakim & Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Proceedings of methodologies and evaluation of Multiword Units in real-world applications (MEMURA)*, 39–46.

Paquot, Magali. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research* 35(1). 121–145. DOI: 10 . 1177 / 0267658317694221.

Prokopidis, Prokopis & Haris Papageorgiou. 2017. Universal dependencies for Greek. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, 102–106. Gothenburg, Sweden: Association for Computational Linguistics.

Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya & Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, 222–240. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Ramisch, Carlos, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh & Hongzhi Xu. 2020. Edition 1.2 of the PARSEME Shared Task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, 107–118. Association for Computational Linguistics.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu & Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. In *Proceedings of the ACL2009 Workshop on MWEs*, 47–54.

Sag, Ivan, Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Lecture notes in Computer Science. Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, vol. 2276, 189–206. Berlin & Heidelberg: Springer.

Savary, Agata, Cherifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev, Chaya Liebeskind, Hongzhi Xu, Sara Stymne, Tunga Güngör, Thomas Pickard, Bruno Guillaume, Eduard

Bejček, Archna Bhatia, Marie Candito, Polona Gantar, Uxoa Iñurrieta, Albert Gatt, Jolanta Kovalevskaite, Timm Lichte, Nikola Ljubešić, Johanna Monti, Carla Parra Escartín, Mehrnoush Shamsfard, Ivelina Stoyanova, Veronika Vincze & Abigail Walsh. 2023. PARSEME corpus release 1.3. In Archna Bhatia, Kilian Evang, Marcos Garcia, Voula Giouli, Lifeng Han & Shiva Taslimipoor (eds.), *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, 24–35. Dubrovnik, Croatia: Association for Computational Linguistics. DOI: 10.18653/v1/2023.mwe-1.6.

Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova & Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary & 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.1469555.

Savary, Agata, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta & Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics* 112(1). 5–54.

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova & Antoine Doucet. 2017. The PARSEME Shared Task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, 31–47. Valencia, Spain: Association for Computational Linguistics.

Sfetsiou, Vasileia. 2007. *Predicative nouns: Methods of analysis for electronic applications.* Aristotle University of Thessaloniki. (Doctoral dissertation).

Shardlow, Matthew, Richard Evans, Gustavo Henrique Paetzold & Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical complexity prediction. In Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot & Xiaodan Zhu (eds.), *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 1–16. Association for Computational Linguistics. DOI: 10.18653/v1/2021.semeval-1.1.

Sklavounou, Elsa. 1994. Support nouns: Application to the special lexicon of tennis. In Irene Philipakki-Warburton, Katerina Nicolaidis & Maria Sifianou (eds.),

Themes in Greek Linguistics. Papers from the 1st International Conference on Greek Linguistics, 515–520. Amsterdam & Philadelphie: John Benjamins.

Straka, Milan & Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual parsing from raw text to Universal Dependencies*, 88–99. Vancouver, Canada: Association for Computational Linguistics.

Tsolaki, Sophia. 1998. Support verb constructions. The support verb *δίνω*: A first approach. In *Studies in Greek linguistics. Proceedings of the 18th annual meeting of the department of linguistics, school of philology, faculty of philosophy, Aristotle University of Thessaloniki*, 473–486.

van Gompel, Maarten & Martin Reynaert. 2013. FoLia: A practical XML format for linguistic annotation – a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal* 3. 63–81.

Zaninello, Andrea & Alexandra Birch. 2020. Multiword Expression aware Neural Machine Translation. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC2020)*, 3816–3825.