


## Chapter 2

# Description of Pomak within IDION: Challenges in the representation of verb multiword expressions

✉ Stella Markantonatou<sup>a</sup>, ✉ Nikolaos T. Kokkas<sup>b</sup>,  
✉ Panagiotis G. Krimpas<sup>b</sup>, ✉ Ana O. Chiril<sup>a</sup>, Dimitrios  
Karamatskos<sup>a</sup>, Nicolaos Valeontis<sup>a</sup> & ✉ George Pavlidis<sup>a</sup>

<sup>a</sup>Institute for Language and Speech Processing, ATHENA Research Center,  
Greece <sup>b</sup>Democritus University of Thrace, Greece

Pomak is a non-standardised, endangered language variety of the East South Slavic dialect continuum. This article presents an online resource of 165 Pomak verbal multiword expressions collected via fieldwork. The resource has been developed with IDION, which is a web-based environment for the documentation of a wide range of multiword properties. The following information is encoded in this resource: lemma form of the expression, variants (if attested), definition in Pomak and translation in other languages, gloss, usage examples for 60 verb multiword expressions, morphosyntactic analysis in the Universal Dependencies framework as well as certain lexical relations among multiword expressions and verb alternations (if attested). Observations on the collected material that are not encoded in the Pomak edition of IDION but are presented in the article concern the types of verbal multiword expressions found in the data (light verb constructions, idioms) and the occurrence of very similar expressions in Modern Greek. The contents of Pomak-IDION are openly available; they belong to a set of resources of Pomak (corpus, morphological and syntactic models, embeddings, lexica) that have been developed as a case study of the Philotis project, which provides technological support for the documentation of living languages.

Stella Markantonatou, Nikolaos T. Kokkas, Panagiotis G. Krimpas, Ana O. Chiril, Dimitrios Karamatskos, Nicolaos Valeontis & George Pavlidis. 2024. Description of Pomak within IDION: Challenges in the representation of verb multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 39–72. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998633 



## 1 Introduction

This article presents a freely available online resource<sup>1</sup> that documents aspects of Pomak verbal multiword expressions (henceforth VMWEs). The resource, which will be referred to as Pomak-IDION, is intended to be useful to human users and Natural Language Processing (henceforth NLP) practitioners.

Pomak-IDION is a rare resource in a world where VMWE databases of endangered languages are sparse. Piirainen (2005), who has offered a very precise picture of idiom research in Europe, noted that idiom data existed for some standard European languages. She reported that minor languages and dialects were completely ignored (with a couple of exceptions). Of course, progress has been made over the years; however, databases with detailed information on idiom data for (not only European) endangered languages are still really few. Even the term “less resourced” language does not really describe the situation of endangered languages such as Pomak. An example is the report of Ní Loingsigh & Ó Raghallaigh (2016) on the development of an idiom database for Irish, which is a less-resourced European language in this respect. However, even this database draws on republished substantial work. On the contrary, there is no republished work on Pomak idioms.

Pomak is a living, endangered and non-standardised East South Slavic language variety with few written resources in various scripts. Pomak-IDION belongs to a set of state-of-the-art resources for this language (lexica, corpus, treebank, morphological and syntactic models, embeddings) that have been developed in the framework of Philotis. The Philotis project<sup>2</sup> has developed an infrastructure to facilitate the development of state-of-the-art NLP resources of living languages. The Pomak treebank has been annotated according to the Universal Dependencies formalism (henceforth UD) (Markantonatou et al. 2023).<sup>3</sup>

165 Pomak VMWEs have been collected via fieldwork. Both idioms and light verb constructions (henceforth LVCs) have been identified in this material. Several Pomak VMWEs have literal equivalents in Modern Greek. This fact suggests the presence of contact phenomena since trilingualism (Pomak, Greek, Turkish) is widespread in the Pomak community; native speakers of Greek, on the other hand, rarely speak Pomak.

To the best of our knowledge, this is the first systematic encoding of Pomak VMWEs that can be useful both to the human user and to NLP practitioners. Considerable field work was required for this task since, although usages of the

---

<sup>1</sup><https://pomak.idion.athenarc.gr/admin>

<sup>2</sup><https://philotis.athenarc.gr/>

<sup>3</sup><https://universaldependencies.org/>

VMWEs abound in everyday speech, they are extremely rare in the little available Pomak textual legacy. Lexical semantic relations among VMWEs, such as synonymy, are even more difficult to identify in the available corpora. We were fortunate enough to enjoy the cooperation of the Pomak community who embraced this effort and offered oral evidence.

We begin this discussion by introducing the Pomak language variety. In §2 we provide information about Pomak and the existing resources: §2.1 describes the corpus of Pomak and §2.2 the script and the orthography used in the corpus. The same orthography and script have been used in Pomak-IDION. Basic information about Pomak morphology and syntax is presented in §2.3 and §2.4 respectively and the UD Pomak treebank is introduced. §3 describes the collection of linguistic material about Pomak VMWEs.

Next, we present some observations on the collected material. §4 summarises the syntactic patterns observed in the collected material as sequences of Part of Speech (henceforth PoS) UD tags. In the collected material, both LVCs (§4.1) and idioms (§4.2) were identified; possible manifestations of contact phenomena between Pomak and Modern Greek are also addressed in these sections. More material about Pomak LVCs and idioms is provided in Appendix A and Appendix B respectively.

The Pomak VMWEs were encoded with the web-based IDION database which we introduce in §5. In this section, we briefly discuss issues to which state-of-the-art databases of VMWEs have to provide a response. Then, in light of this discussion, we explain our choices regarding the basic design principles of IDION.

There are two interfaces to the database: one available to users who access the database only to look up a VMWE (henceforth external users) and one used by registered linguists who document the VMWEs (henceforth encoders). Here, we present the interface available to external users. §6 is divided into subsections each describing the searches that are available with Pomak data. Searches supported by fuzzy matching retrieve VMWEs in lemma form (§6.1). Once the desired VMWE has been identified, other searches are available: variants (if any has been attested), gloss, translations (§5.1), usage examples and their translations (§6.2), morphosyntactic analysis of the variants and the usage examples in the UD framework (§6.3), and lexical relations among VMWEs (some of them of semantic nature), if attested (§6.4).

## 2 About Pomak

Pomak (endonym: Pomácky, Pomácko, Pomácku or other dialectal variants) is a non-standardised East South Slavic language variety. Apart from Greece, it is spoken in parts of Bulgaria and East Thrace (Türkiye) and in the places of Pomak diaspora (Constantinides 2007: 35). In Greece, it is spoken by about 35,000 people inhabiting the Rhodope Mountain area mainly (Adamou & Fanciullo 2018). The Pomak dialect continuum has been influenced by Greek and Turkish due to extensive bilingualism or trilingualism.

Pomak scores low on all six factors of language vitality and endangerment proposed by UNESCO (Brenzinger et al. 2003): there is little written legacy with only symbolic significance for the speakers of Pomak, the language is not taught at school, it is used mainly in family settings, and the dominant languages, namely Greek and Turkish, begin to penetrate family settings.

### 2.1 Textual resources of Pomak

Sporadic transcriptions and recordings of Pomak folk songs and tales have been published over the last 80 years (Theocharides 1996a,b) as well as a very few modern texts; these mostly include journalistic texts, translations from Greek and English into Pomak (Karahóga 2017), and material for teaching Pomak to Greeks as a second language (Kokkas 2004). In addition, descriptive works on Pomak morphology and grammar have been published (Papadimitriou 2013, 2008, Sandry 2013). Selected parts of this textual material have been included in a corpus of about 140,000 words, which will be made available for research purposes by Philotis. The corpus is presented here because it is the largest searchable collection of texts in Pomak and has been used as a source of VMWE instances while developing Pomak-IDION.

Table 1 shows the text genres included in the corpus and the size of the respective texts in words. Where possible, the geographical origin of the texts is also given as a hint to the Pomak variant appearing in the text.

The morphological and syntactic analysis adopted in this work draws on the approach to Pomak language that was developed in the Philotis project and is outlined in Karahóga et al. (2022) and Markantonatou et al. (2023). A Pomak treebank has been made available on the UD treebank repository along with the relevant detailed documentation.<sup>4</sup>

---

<sup>4</sup><https://universaldependencies.org/qpm/index.html>

Table 1: Pomak corpus: Type, size and geographical origins of texts.

Text types	Words	Geographical origins
Folk tales	43,817	Aimonio, Glafki, Dimario, Echinós Myki, Pachni, Oreó
Language description	19,524	mixed
Journalism	25,236	Myki
Translations into Pomak	24,208	Myki, Pachni
Folk songs	18,434	mixed
Proverbs	550	mixed
Other	5,325	Myki
Total	137,094	

## 2.2 Pomak script and orthography

A variety of scripts and orthographies have been used so far in the Pomak textual legacy, ranging from Bulgarian-based Cyrillic to Modern Greek to an English-based Latin alphabet. Homogenisation of these texts in order to form a processable corpus required the adoption of a common script and a common orthography. To this end, the Latin-based alphabet devised and proposed by Ritvan Karahóga and Panagiotis G. Krimpas (henceforth K&K alphabet), which has a language resource-oriented accented version and a non-accented all-purpose version, has been used to transliterate the corpus semi-automatically and for the documentation of Pomak VMWEs in IDION.

The K&K alphabet has been developed to satisfy the following requirements (Karahóga et al. 2022): use of Unicode to ensure portability of the alphabet, phonetic transparency, easily learned representations of sounds (ensured by the use of similar diacritics for the same articulation sounds and the absence of digraphs) and, finally, consistent spelling not affected by predictable allophony. It should be noted that the K&K alphabet is based on the Pomak variety spoken in the area of Myki but can also partially serve as an all-variety script by allowing various predictable pronunciations of the same graph depending on the variety.

The orthographic tradition of other Slavic language varieties was taken into consideration if it did not contradict distributional and phonological evidence. For instance, certain interrogative, indefinite and negative pronouns, conjunctions and adverbs are spelled as a single word in most Slavic languages but, in the adopted Pomak orthography, are spelled as two words, e.g., *at kak* for *atkák*

‘since’, *ní kutrí* for *níkutrí* ‘nobody’ because the first word can be independently identified as a preposition or particle, and the second as an interrogative pronoun or adverb e.g., *at* ‘from; out of’, *kak* ‘how; as; like’.

### 2.3 Pomak morphology at a glance

Pomak common and proper nouns, determiners, adjectives, pronouns, participles and some of the numerals are morphologically marked for gender, number, case and (in)definiteness. The opposition *Animate* vs. *Inanimate* is overt with the nominative case of masculine plural adjectives, participles and 3rd person plural pronouns and rarely with masculine singular nouns, where it is found as residual morphological genitive/accusative. Pomak has three genders, namely masculine, feminine and neuter, and four cases, namely nominative, dative/genitive, accusative and vocative; the morphological dative case has assumed the functions of the historical dative and genitive cases, so we speak of dative/genitive case and use a notation reminiscent of this fact in glossing the Pomak examples. With possessive determiners both the number of the possessor and of the possessed object are encoded. Like most Balkan languages, Pomak has a rich inventory of diminutive and augmentative forms of nouns, adjectives, adverbs, and certain passive participles.

Pomak is special among East South Slavic languages in that, although it uses a tripartite enclitic definite article *-s*, *-t*, *-n* (Adamou & Fanciullo 2018, Constantinides 2020, Krimpas 2020) like Macedonian, this article is of the *-s*, *-t*, *-n* rather than *-v*, *-t*, *-n* type and occurs not only with nominals, but also with deictic adverbs as a deictic and definiteness marker, denoting:

- Proximity to the speaker, e.g., *čulákos* ‘the man close to the speaker’.
- Proximity to the listener, e.g., *čulákot* ‘the man close to the listener’.
- Distance from both the speaker and the listener, e.g., *čulákon* ‘the man who is away (or out of sight) from both the speaker and the listener’.

Verbs have finite and non-finite forms. There are three types of non-finite verb forms: converbs, participles and (residual) infinitives. The residual, i.e., Proto-Slavic, infinitive forms the prohibitive imperative when following the particles *na/ne* and *namój* (sing.)/*namójte* (pl.) ‘not’, e.g., *namój barzá* ‘do not rush’. Interestingly, Pomak has another, innovative form of infinitive, which may be called *the morphologically reduplicated infinitive*. This residual infinitive of a small number of imperfective verbs is repeated to form fixed multiword expressions that

denote the continuous/monotonous/rhythmic repetition of a motion, e.g. *čúktiti čúktiti* ‘hit and hit’.

Finite verbs are always marked for mood, number and person. Verbs in the indicative mood are marked for tense, either past or present. *Som* ‘be’ is the auxiliary verb used to form perfect verb tenses and the passive voice. Future tenses are formed with the indeclinable auxiliary particle *še* ‘will’, which historically derives from the verb meaning ‘want’.

## 2.4 Pomak syntax at a glance

Pomak is a nominative-accusative language, where subjects are typically marked with the nominative case and objects with the accusative; in addition, some verbs select objects in the dative/genitive case. Indirect objects are marked with the dative/genitive case, which is morphologically based on the Slavic dative case. As in other Slavic languages, ethical datives abound. The strong and the weak forms of the personal pronoun may co-occur in a sentence (clitic doubling).

Markers such as *óti*, *da*, *če*, *ta* introduce subordinated clauses that function as verb dependents and markers such as *akú*, *kugá*, *pak*, *za da*, *za to*, *óti* introduce clauses that function as adverbial modifiers. There is a question particle *li*, e.g., *dojděš li* ‘do you come?’.

With respect to word order, Pomak is a primarily SVO language with rather flexible word order, given its highly inflectional nature. Adjectives typically come before the noun, although the reverse is also possible, especially for emphasis or in literary contexts. Possessives are actually datives of the unstressed (enclitic) personal pronoun. The rules governing the word order of clitics in a clause, as well as the word order within a clitic cluster, are similar to those of Bulgarian, Macedonian and Serbo-Croat: a single clitic is always the second element of its clause; multiple clitics are arranged in the following order: auxiliary > clitic-in-dative-case > clitic-in-accusative-case but, if the auxiliary is 3rd person singular, the order changes into clitic-in-dative-case > clitic-in-accusative-case > auxiliary. Pomak is a pro-drop language, which means that pronominal subjects are normally used for clarity, emphasis, or literary purposes since verb endings normally provide information about the “number” and “person” of the syntactic subject. Given that infinitives are no longer in use in Pomak (except for the residual and reduplicative infinitives mentioned above), the so-called “Balkan subjunctive” (*da* particle + finite verb in the case of Pomak) has replaced the old Slavonic infinitive (much like Modern Greek, Albanian, Romanian, Bulgarian, Macedonian and, to a lesser extent, Serbian and Bosnian).

### 3 Material collection

Pomak VMWEs were collected mainly through interaction with native speakers of Pomak in the framework of Philotis. The collection of VMWEs was accomplished by Nicolaos Kokkas, one of the authors, who is fluent in Pomak. The native speakers who contributed to this study represented the variants of Pomak spoken in the following villages in the region of Xanthi: Bára (Greek name Στήριγμα ‘Stírigma’), Bašájkovo (Greek name Μάνταινα ‘Mándena’, Demirǵík (Greek name Δημάριο ‘Dimáριο’), Púlevo (Greek name Προσήλιο ‘Prosilío’).

Targeted interaction with native speakers involved (recorded) interviews and collection of written material. The speakers were two men and two women of secondary and tertiary education level, whose ages ranged between 20 and 50 years. During the interviews specific VMWEs were discussed. To collect written material, during the period September-December 2022, each week the speakers received a short list of VMWEs which they discussed in their community and enriched with semantically related VMWEs, namely synonyms and antonyms<sup>5</sup> (if they could identify any) and usage examples. Written material was collected in the form shown in Table 2 (<> indicates translation into the respective language, e.g. <Greek>: translation into Greek, [Pomak] any original text in Pomak and (Gloss) the gloss of the Pomak VMWE in Modern Greek or English).

Table 2: Form used to collect evidence about Pomak VMWEs.

VMWE	[Pomak]	<Greek>	<English>
Definition	[Pomak]		(Gloss)
Synonyms	[Pomak]	<Greek>	
Opposites	[Pomak]		
Examples	[Pomak]	<Greek>	<English>
ID:	Interviewed speaker(s):		Date:

The forms were further filled with material from the Pomak corpus that was searched for in-context usages of the VMWEs in a variety of texts (however, little material was collected in this way). Recordings of Pomak contemporary speech obtained in 2022 provided more VMWE instances of usage. The authors of this chapter have encoded the collected material.

<sup>5</sup>In IDION, the term *opposites* is preferred rather than the term *antonyms* for reasons explained in §5.6



## 4 A closer look into Pomak VMWEs

In this section, we will take a closer look at the structure of the collected Pomak VMWEs.

In this article, we will make frequent use of the term *lexicalised components of a MWE* which was introduced in Savary et al. (2018: 94). These are the components with either fixed form or fixed lemma. Apart from the lexicalised components, VMWEs have free components that set them apart from proverbs; still, these free components are subject to strong semantic and morphosyntactic constraints. Throughout this article, the lexicalised components of the VMWEs that are used as examples are typed in bold. The slots in the VMWE that should be filled with free arguments are indicated by means of pronouns in regular script.

Based on the collected data, a set of observations have been made; these observations are not available in the existing literature on Pomak and/or on Pomak idiomaticity:

- Pomak uses LVC constructions.
- The set of light verbs identified in the Pomak data is very similar to those of other European languages (see §4.1).
- The pattern VERB+NOUN is very frequent in LVCs and in idioms (1).
- Pomak VMWEs demonstrate verb alternation phenomena (see §6.4).
- Several Pomak idioms have literal equivalents in Modern Greek. This observation could possibly contribute to a wider study of idiomaticity in the Balkan languages.

The syntactic patterns of the lexicalised components of the collected VMWEs are listed below as PoS sequences. When a literally equivalent Greek VMWE exists, this is introduced with the prefix “GE” (Greek Equivalent) next to the English translation of the Pomak VMWE. Of the syntactic patterns (1–8), (1) has been attested in both idioms and LVCs and the other patterns in idioms only. It should be noted that all these patterns are in use in non-idiomatic Pomak. Throughout this text, the infinitive is not used in the English glosses of verbs because both Pomak and Modern Greek use the verb’s 1SG.PRES.IND. form as its lemma form.

- (1) VERB + NOUN (LVCs; certain idioms)
- (2) VERB + ADJECTIVE  
**stánavom fukará**  
 become.1SG.VERB poor.ADJ.SG.NOM  
 ‘I become poor’ Verb: *fukarjásavom* ‘I become poor’
- (3) VERB + ADPOSITION + NOUN  
**astánavom na mǎsto**  
 remain.1SG.VERB at.ADP place.NOUN.SG.ACC  
 ‘I die instantaneously’
- (4) VERB + ADPOSITION + ADJECTIVE + NOUN  
**astánavom sas atvórena ustá**  
 remain.1SG.VERB with.ADP open.ADJ.SG.FEM.ACC mouth.NOUN.SG.FEM.ACC  
 ‘I remain speechless’ GE: *μένω με το στόμα ανοιχτό*
- (5) VERB + NOUN + ADPOSITION + NOUN  
**atvárem belǎ na glavóso**  
 open.1SG.VERB trouble.NOUN.ACC on.ADP head.NOUN.SG.ACC  
 ‘I cause problems to myself’ GE: *βάζω μπελά στο κεφάλι μου*
- (6) VERB + ADJECTIVE + ADPOSITION + NOUN  
**právem bannóga čórna ad sópa**  
 do.1SG.VERB somebody black.ADJ.ACC from.ADP beating.NOUN.SG.ACC  
 ‘I beat someone hard’ GE: *κάνω μαύρο στο ξύλο κάποιον*
- (7) VERB + ADPOSITION + NOUN  
**klávom nǎeko faf óči**  
 put.1SG.VERB something in.ADP eye.NOUN.DEF.PL.ACC  
 I crave for something
- (8) VERB + ADVERB  
**glódom kríve bannóga**  
 look.1SG.VERB away.ADV somebody  
 ‘I glare at somebody’

Word order permutations can be observed in the collected material. Here one sees, e.g., that non-lexicalised variable indirect objects may come either after all the lexicalised parts of the VMWE, or immediately after the verb of the VMWE.

- (9) a. **dávom kolájene bannómu** OR **dávom bannómu kolájene**  
 give.1SG eases somebody.GEN  
 ‘I greet somebody’

- b. **dávom habér**      **bannómu**      OR **dávom bannómu habér**  
 give.1SG news.NOUN somebody.GEN  
 ‘I inform somebody’

#### 4.1 Pomak LVCs in the collected material

LVCs were first introduced by Jespersen (1965) and since then they have attracted a lot of attention (e.g., Baldwin & Kim 2010, Laporte 2018). LVCs consist of a verb and a nominal complement, possibly introduced by a preposition. Savary et al. (2018) list a set of diagnostics for setting LVCs apart from idioms with a VERB+(PREPOSITION)+NOUN syntactic structure: the noun has one of its original senses and denotes an event or a state; the verb only contributes morphological features, such as tense, mood, person and number; the noun can head an NP containing all the syntactic arguments of the verb and denoting the same event or state as the LVC; and, the overall construction is subject to semantic and syntactic uniqueness constraints.

Here, we identify as LVCs those VERB+NOUN formations that can be replaced by (are synonymous with) verbs that are morphologically related to their noun (10); such structures seem to satisfy the LVC diagnostics listed above. Among the Pomak verbs used as light verbs are *dávom* ‘I give’, *právem* ‘I do’, *stánavom* ‘I become’, *stórevom* ‘I make’, *zímom* ‘I take’. More examples of Pomak LVCs are listed in Appendix A.

- (10) a. **dávom**      **izét**  
 give.1SG.VERB pain.NOUN.SG.ACC  
 ‘I torture’ Verb: *izettóvom* ‘I torture’
- b. **stánavom**      **fukará**  
 become.1SG.VERB poor.ADJ.SG.NOM  
 ‘I become poor’. Verb: *fukarjásavom* ‘I become poor’
- c. **stórevom**      **izméte**  
 do.1SG.VERB service.NOUN.PL.ACC  
 ‘I do the housework’. Verb: *izmetóvom* ‘I serve’
- d. **zímom**      **emín**  
 take.1SG.VERB oath.NOUN.SG.ACC  
 ‘I take an oath’. Verb *eminledísavom* ‘I take oath, I vow’

The Pomak corpus has provided some usage examples of LVCs (11):

- (11) a. Nimó                      ma právi rezíl.  
do-not.2SG.VERB me do    infamous.ADJ  
'Do not humiliate me.'
- b. Čuláekon zíma                      karáre                      annók déne da  
man.the take.3SG.VERB decision.NOUN.PL.ACC one day that  
íde                      da nájde                      Alláha.  
go.3SG.VERB that find.3SG.VERB Allah  
'One day, the man makes the decision to go and find Allah.'

#### 4.2 Idioms occurring in both Pomak and Modern Greek

In our collection of 165 Pomak VMWEs, we traced 55 VMWEs that have literal equivalents in Modern Greek. We consider two VMWEs as *literally equivalent* if they consist of translationally equivalent lexicalised parts for the same non-compositional meaning. These data may present an interesting aspect of language contact phenomena between Greek and Pomak or, perhaps, an instance of wider linguistic interactions in the Balkans or other parts of Europe (Piirainen 2005, Krimpas 2022). More VMWEs of this type are listed in Appendix B. Some pairs of equivalent Pomak and Modern Greek VMWEs are exemplified in (12).

- (12) a. i. **ablízavom si pórstovene**  
lick.1SG.VERB I.PRON finger.NOUN.DEF.PL  
'I find the food delicious'
- ii. GE: γλείφω τα δάχτυλά μου  
**glifo ta dachtila mou**  
lick.1SG.VERB the.ART.PL.ACC finger.NOUN.PL.ACC my  
'I find the food delicious'
- b. i. **čéftom balíkoso**  
chisel.1SG.VERB wound.NOUN.DEF.PL.ACC  
'I open old wounds'
- ii. GE: ξύνω πληγές  
**ksino pliges**  
chisel.1SG.VERB wound.NOUN.PL.ACC  
'I open old wounds'
- c. i. **klávom dvéne nógy na annó**  
put.1SG.VERB two.NUM.DEF foot.NOUN.DUAL.ACC on.ADP one.NUM  
**amenýe**  
shoe.NOUN.SG.ACC  
'I try to control somebody's life'

- ii. GE: βάζω τα δυο πόδια κάποιου σε ένα παπούτσι  
**vazo ta dio podia**  
 put.1SG.VERB the.ART.PL.ACC two.NUM feet.NOUN.PL.ACC  
 kapiou se ena papoutsi  
 someone.GEN in.ADP one.NUM shoe.NOUN.SG.ACC  
 ‘I try to control somebody’s life’
- d. i. **sečé mi akýlos**  
 cut.3SG.VERB I.DET.GEN brain.NOUN.DEF.SG.NOM  
 ‘I am intelligent’
- ii. GE: κόβει το μυαλό μου  
**kovi to mialo mou**  
 cut.3SG.VERB the.ART.SG.NOM brain.NOUN.SG.NOM my  
 ‘I am intelligent’

## 5 Issues in VMWE documentation: The IDION approach

Modern MWE databases are expected to provide information that can be used both by people who study or use a language and in NLP (Grégoire 2010, Losnegaard et al. 2016). Gantar et al. (2018) compare seven dictionaries and NLP databases and list the MWE properties they document, namely: (i) variants (ii) definition (iii) morphology of MWE components (iv) contiguity of MWE components (v) phrase structure (vi) usage example. In what follows, we discuss these properties and how they are treated in IDION. Furthermore, we extend our discussion to additional information about VMWEs that is encoded in IDION and includes a variety of semantic properties and the full morphosyntactic description of VMWE lemmas and usage examples according to the UD framework.

IDION is a web environment for the rich documentation of MWEs. IDION allows for new editions, accessible from the same or a different site. So far, two editions have been created, one for Modern Greek VMWEs (Markantonatou et al. 2019) and one for Pomak VMWEs. The contents are available under a CC-BY-NC license.

### 5.1 Lemma form

In IDION, a *lemma form* of a VMWE contains:

- the components with a fixed form (fixed lexicalised components);

- the components whose lemma is fixed but whose form inflects; these are included in their lemma form or the form that best approximates the lemma convention (non-fixed lexicalised components), e.g., if the (head) verb of the VMWE appears in the second and third persons of all numbers, in the lemma form it is in the second person singular;
- the variables, such as free NPs functioning as subjects, direct/indirect objects or ethical genitives or datives of the MWE and free phrasal complements.

In addition, the lemma form takes into account the possibly fixed order of the MWE components; otherwise, it keeps the lexicalised components close to the verb. Such a typical order is the following: (lexicalised or free) subject, lexicalised verb, other lexicalised components (if any), (lexicalised or free) object. Attested usage instances of the VMWE with a different word order are separately listed in the CORPUS tab of the IDION-Pomak database (see §6.2) as manifestations of the syntactic flexibility of the VMWE.

Below, in order to better explain IDION's features we may resort to examples from Modern Greek since Pomak could provide only limited material.

## 5.2 Variants

Variants have to do with the lemma form of the MWEs. The lemma form is one of the two features of a MWE that have to be considered in order to create an entry in the database; meaning is the second feature. It turns out that the identity of the VMWE is established as a combination of a meaning with a non-empty set of lemma forms, the so-called variants (Vondříčka 2019). The issue of variants occurs because VMWEs are mutable entities of spoken, colloquial language. In other words, it is not the case that each VMWE lemma form corresponds to a different meaning and vice versa. For instance, all the lemma forms of the Modern Greek VMWE in (13) share the same meaning. These lemma forms are identical as regards the syntactic dependencies among the lexicalised parts that belong to content word categories, namely nouns, adjectives and verbs. In the same spirit, optional or mutually exclusive lexicalised non-content word components of the MWE may define new variants but not a new VMWE. In (13) four different lemma forms of the same VMWE result from the optionality of the article *τη* and the exclusive disjunction between *γύρω από το* and *στο*.

It should be clarified that the syntactically flexible usages of VMWEs (see §5.5) are not treated as VMWE variants in IDION.

- (13) a. GE: βάζω τη θηλειά γύρω από το λαιμό κάποιου  
**vazo ti thilia giro apo to lemo** kapiou  
 put.1SG the noose around from the neck somebody.GEN
- b. GE: βάζω θηλειά γύρω από το λαιμό κάποιου  
**vazo thilia giro apo to lemo** kapiou  
 put.1SG noose around from the neck somebody.GEN
- c. GE: βάζω τη θηλειά στο λαιμό κάποιου  
**vazo ti thilia sto lemo** kapiou  
 put.1SG the noose to.the neck somebody.GEN
- d. GE: βάζω θηλειά στο λαιμό κάποιου  
**vazo thilia sto lemo** kapiou  
 put.1SG noose to.the neck somebody.GEN  
 ‘I force someone to be involved in an unpleasant situation’

A lexicalised content word component of the VMWE may appear in both the singular and the plural with no consequences for the idiomatic meaning. This situation is not exactly rare but it is unpredictable and part of the idiomatic character of a VMWE. Variation in number may induce changes to other lexicalised components of the MWE, e.g., the singular and plural lexicalised subjects in (14a) and (14b) respectively induce agreement phenomena on the (lexicalised) verbs of the respective variants of the same VMWE.

- (14) a. GE: πήρε αέρα το μυαλό κάποιου  
**pire aera to mialo** kapiou  
 take.3SG.PAST air the.SG.NOM brain.SG.NOM somebody.GEN  
 ‘to get above oneself’
- b. GE: πήραν αέρα τα μυαλά κάποιου  
**piran aera ta miala** kapiou  
 take.3PL.PAST air the.PL.NOM brain.PL.NOM somebody.GEN  
 ‘to get above oneself’

(15) shows the VMWE in (13) with an ethical genitive<sup>6</sup> rather than a possessive one (Sailer & Markantonatou 2016). The ethical genitive alternation in (15) has to do with the morphosyntactic form of a variable and does not affect the meaning of the expression. Given this fact and the wide use of VMWEs with ethical genitives, in IDION lemma forms with an ethical genitive are listed as variants of

<sup>6</sup>Modern Greek: ethical genitive; Pomak: ethical dative.

the lemma forms exemplifying the other member of the alternation pair; the latter contains either an inalienable possession structure or a suitable prepositional phrase.

- (15) GE: του βάζω (τη) θηλεία [γύρω από το] / [στο] λαιμό  
          tou           vazo   (ti) thilia [giro apo to] / [sto] lemo  
          I.PRON.GEN put.1SG (the) snoose [around from the] / [to.the] neck  
          ‘I force someone to be involved in an unpleasant situation’

Variants are considered a challenging feature of MWEs (Vondřička 2019, Grégoire 2010, Villavicencio et al. 2004, Skoumalová et al. 2024 [this volume]) because criteria such as the ones presented above are required to decide which forms will be listed as variants under the same MWE entry and which ones will not. No general agreement on this issue has been achieved as yet. For instance, VMWEs are often members of sets of expressions that stand in various lexical and semantic relations as discussed in §5.6. In IDION, variants are members of a set of lemma forms of one VMWE. VMWEs that differ in lexicalised content words and/or semantically define separate entries in the database. IDION allows for the encoding of sets of lemma forms because it considers these variations important for the human user and for NLP.

In developing IDION, once a first decision about a meaning and form combination is made, encoders collect as many variants and syntactically flexible usage instances as possible from corpora and/or the web. This procedure may change the original decision about the identity of the VMWE. For instance, it may arise that there are more meanings out there corresponding to the same set of forms than originally expected, or that there are forms that cannot be considered variants of the documented VMWE, for instance, because their syntactic structure cannot be reduced to the structure of the documented one. Therefore, at the heart of IDION stands the collection of usage instances of the VMWE that determines the amount and the types of information on VMWEs to be documented in IDION. It should be stressed that IDION relies on actual usage examples, preferably collected from corpora and the web. There is room for encoding the intuitions of native speakers but these examples are kept to a minimum and are marked as such. Eventually, a non-empty set of variants is collected. The longest one is chosen as the “preferred variant” and represents the VMWE.

No contracted representations are used for the lemma forms, such as representations based on regular expressions; for a comprehensive discussion on VMWE representations see Lichte et al. (2019). Since we have drawn on limited lexicographic and financial resources, we preferred to invest in the collection and study



of usage examples. Furthermore, given the current NLP technology, morphosyntactic representations of both the lemma form of the VMWE and its usage examples can be obtained, edited, and searched for structural patterns with open-source tools, thus facilitating (steps of the) encoding without any additional machinery. In addition, usage examples constitute a valuable reusable resource for model development and that was a strong motivation because IDION is designed to support NLP. Finally, human users profit from usage examples because they illustrate usage particularities that can hardly be included in the definition of the meaning of the VMWE.

### 5.3 Meaning, glossing, translations

In IDION, the definition of the VMWE is a short text describing the meaning of the MWE. Definitions are in the language of the VMWE and contain compositional expressions only. Because the type of arguments a VMWE supports (the variables) is an important contribution to its meaning, in the definition pronouns like ‘someone’ and ‘something’ stand for nominal complements denoting humans and non-humans, if such constraints are imposed by the VMWE.

The representative lemma form is glossed and translated into a language other than that of the VMWE. Glosses are simple with no morphological and syntactic annotation since this information is given via the UD analysis of a lemma form, which is also made available in IDION. Glosses are addressed to the human user who can complement them with the UD analysis of the lemma form. On the translation level, MWEs with an equivalent meaning are preferred when they exist in the target language.

### 5.4 About morphology and syntax

The morphological and syntactic analyses of MWEs are necessary both for the precise definition of the MWE form and for supporting NLP. In rule-based NLP, an important task is the development of computational lexica of MWEs enriched with the full inflectional paradigms of the entries, e.g., Savary (2009) for compounds in several languages and Al-Haj et al. (2013) for Hebrew. This requires a morphological and a syntactic description of both the language system to which the MWE belongs and the particularities of each MWE.

The morphological and syntactic representation of the MWEs must be compatible with the formal language and the framework used by the NLP tool that they will support; this raises reusability concerns. For instance, databases aimed

at supporting phrase structure-based NLP (Grégoire 2010, Vondřička 2019) employ encoding schemes that allow for non-terminal nodes. To be reused in, for instance, the UD framework, which is compatible with several popular state-of-the-art non-rule-based NLP tools and is adopted in several state-of-the-art MWE databases including IDION (Skoumalová et al. 2024 [this volume], Leseva et al. 2024 [this volume], Osenova & Simov 2024 [this volume]), these encodings have to be adapted accordingly. This is because UD uses no non-terminal nodes, has its own metalanguage for morphosyntactic annotation and the analysis is encoded in CoNLL-U.<sup>7</sup> On the other hand, state-of-the-art NLP tools learn from data, so they can be possibly trained on the (adapted) inflectional paradigms of VMWEs or, alternatively, on appropriately annotated corpora of diverse and syntactically flexible usages of MWEs (Savary et al. 2019).

However, this need for many flexible usage instances proves to be hard for less-resourced languages, let alone for endangered ones. It is hard to construct corpora of spoken languages for which even a consensus on their alphabet and orthography has not yet been reached; Pomak is such an endangered language (Karahóga et al. 2022). Also, less-resourced languages with only a few corpora representing their spoken version can hardly provide syntactically flexible usages of MWEs. For instance, in the case of Modern Greek, which is a medium-resourced language according to the criteria proposed by Joshi et al. (2020), only the web (and not the published corpora) offers a reasonable amount of representative usage instances of most of the VMWEs, let alone their syntactically flexible ones.

## 5.5 Syntactic flexibility

The syntactic flexibility of the VMWE is documented separately with six diagnostics. Each diagnostic is exemplified with usages from the corpus. As a result, all the collected usage instances are marked for at least one syntactic phenomenon. The six diagnostics are briefly explained below:

- Subject-head verb flexibility: Can the VMWE accept different subjects? Can it appear in all persons/numbers/tenses/moods?
- Can word order variation phenomena be observed with this VMWE?
- Interpolation: Can adverbs, adjectives or even phrases occur in between the lexicalised components of the VMWE?

---

<sup>7</sup>CoNLL-U is the encoding scheme adopted by UD and the tools that process annotated corpora with the UD annotation scheme: <https://universaldependencies.org/v2/conll-u.html>

- Cliticisation of lexicalised nominal content word components.
- Passive voice: does the VMWE have both an active and a passive form?
- Ethical genitive (for Pomak, dative) alternation (see §5.2).

It has already been pointed out in §5.2 that flexible usages of the VMWE are not considered variants of the VMWE apart from the usages containing ethical genitives/datives.

## 5.6 Lexical (semantic) relations among VMWEs

Lexical semantic relations have found their way into state-of-the-art databases for MWEs (Leseva et al. 2024 [this volume], Giouli et al. 2024 [this volume]). In Skoumalová et al. (2024 [this volume]), the notion of “super lemma” approximates that of (several) lexical semantic relations from a different point of view. IDION documents a set of lexical (semantic) relations among VMWEs. In order for a relation to be defined, it has to be attested with a usage example. Here, we will discuss the following pairs: synonyms, opposites, Has\_causative (inverse: Has\_inchoative), Verb alternation that have been attested in our collection of Pomak VMWEs.

A comment is due on synonymy. Synonymy in IDION disregards stylistic differences such as +/-colloquial, +/-offensive and rather relies on a notion of *close semantic proximity* (Hüllen 2004: 39). It is well known that synonymy cannot be considered the linguistic equality relation because, in this way, synonyms would be the words or phrases capable of substituting each other in any context and such words or phrases hardly exist in any language. On the other hand, our everyday linguistic practice seems to consider synonymy a fact, e.g., when we explain the meaning of a word or a phrase using the language to which they belong (Hüllen 2004: 38).

VMWEs are also documented for opposites, that is VMWEs describing situations that cannot hold simultaneously for the same entities, e.g., one cannot at the same time be denoted by the subject of the (EN) VMWE *to kick the bucket* and the VMWE *to be alive and kicking*. Opposition in language is a multi-dimensional and much discussed phenomenon and a rich terminology has been devised for its description (Lyons 1977: 270–287). In IDION, we have chosen the term *opposites* because it seems to denote the general idea described above. We have not used the term *antonym* because it has been devised to describe a relation among gradable words (Lyons 1977).

## 5.7 Other relations among VMWEs

We now turn to the causative/inchoative alternation and the relation among VMWEs which in IDION is called *verb alternation relation*. Strictly speaking, the causative/inchoative and verb alternation relations are defined over verbs; VMWEs, on the other hand, are structures headed by verbs. We use these terms to describe relations among VMWEs with verb heads standing in the respective relations.

The causative/inchoative alternation has been discussed extensively in the literature. Haspelmath (1993: 90) describes the phenomenon that is defined over pairs of verbs as follows:

...it is a pair of verbs that express basically the same situations (generally a change of state, more rarely a going-on) and differ only in that the causative verb meaning includes an agent participant who causes the situation, whereas the inchoative verb meaning excludes a causing agent and presents the situation as occurring spontaneously.

He further distinguishes various morphological types of alternation, one of which is the *labile* type, where the same verb is used both in the inchoative and the causative sense.

In the literature, the term *verb alternation* has been used as a cover term for a large set of phenomena, whereby a verb supports different subcategorisation frames with relatively minor and systematic differences in meaning, such as the *spray-load* alternation and the passive voice. In IDION, a restricted use of the term *verb alternation* is made: practically, it is used for those verb alternations that have not been assigned their own label in the database, for instance, passivisation and causative/inchoative alternation have their own labels and the relevant VMWE pairs are not assigned the “verb alternation” label.

### 5.7.1 UD representation

In IDION, UD representations are provided for the variants and the corpus examples and offer full morphosyntactic analysis. At the moment, IDION adopts the standard UD approach according to which VMWEs are analysed in the same way as compositional structures (de Marneffe et al. 2021: 281). These UD representations are very useful to state-of-the-art NLP as training or fine-tuning material (Savary et al. 2019).

## 6 Pomak-IDION: The Pomak edition of IDION

In §5 we explained the main ideas regarding the documentation of VMWEs in IDION. In §1 we mentioned that IDION has two interfaces: one for encoders and one for external users. This section presents the information on Pomak VMWEs that can be retrieved from IDION.<sup>8</sup> At the same time, it presents the interface for external users. A description of the interface for encoders can be found in Markantonatou et al. (2019).

The properties documented in Pomak-IDION enable the searches described in this section and are summarised in Table 3. Searching facilities were designed to conform to (i) the “what you see is what you get”, or WYSIWYG concept,<sup>9</sup> and (ii) the ten heuristic criteria that describe a user-friendly interface for simplicity of use and navigation (Nielsen & Molich 1990).

Table 3: VMWE properties encoded in Pomak-IDION

1	Lemma form, definition orthographic variations	Pomak
2	Translations	English, Modern Greek
3	Codification for NLP	UD analysis (lemma form, variants)
4	Corpus	Usage examples by native speakers
5	Synonyms	Pomak VMWEs
6	Opposites	Pomak VMWEs

### 6.1 Fuzzy matching for VMWE retrieval

The Pomak VMWEs shown in (16) will serve as a working example.

- (16) a. *nǎko*      *mi*      *alóknava*      *dušó-no*  
          something me.DAT unburden.3SG soul-the.ACC  
          ‘something makes me feel relieved of anxiety’
- b. *alóknava*      *mi*      *dušá-sa*  
          unburden.3SG me.DAT soul-the.ACC  
          ‘I feel relieved of anxiety’

<sup>8</sup>It should be noted that only part of the encoding and search capabilities of IDION have been used in Pomak-IDION, since the required data, such as utterances demonstrating the syntactic flexibility of VMWEs cannot be easily obtained in the case of an under-resourced language (see discussion in §5.4).

<sup>9</sup><https://www.merriam-webster.com/dictionary/WYSIWYG>

- c. **alóknava**      **mi**      **na dušó-no**  
 unburden.3SG me.DAT to soul-the.ACC  
 ‘I feel relieved of anxiety’

A VMWE can be retrieved with segments of its lemma form (Figure 1); this is a fuzzy matching facility that returns a, possibly empty, list of VMWEs in lemma form, each one with its definition in Pomak (Figure 2). Fuzzy matching is applied to all the variants of a VMWE; the reader may recall that the variants are listed in their lemma form and that the longest variant is used as the preferred one (see §5.1). However, few VMWEs come with variants in the Pomak edition of IDION given the way data was collected. Translations of the VMWE into other languages are accessible through the screen with the (fuzzy matching) search results.

Expression

alók

Insert an idiom or (consecutive) parts of it. Then select a language.

Pomak


Search

Figure 1: Search with fuzzy matching in IDION-Pomak.

Expression	Definition
alóknava mi dušasa	húbbe som
alóknava mi na dušono	húbovo mi stánava, rahatladísavom
næko mi alóknava dušono	næko mi stóreva hubbe

Figure 2: Searched with the string *alók* (Figure 1), IDION-Pomak returns 3 VMWEs.

When a VMWE is selected, a set of tabs pops up at the lower part of the screen. The first tab on the left provides access to the orthographic variants of the lemma form (if any exist). The second tab shows the gloss of the VMWE (Figure 3). More tabs are available and described in §6.2–§6.4.

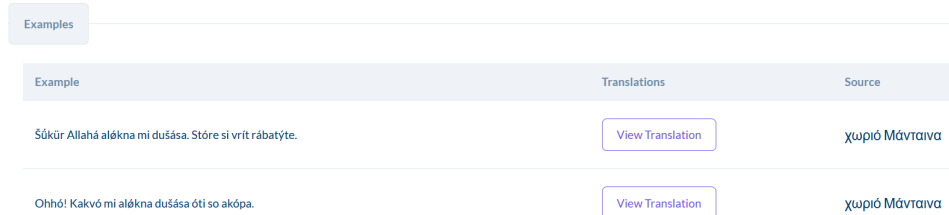


Word	Gloss
aléknava	relieves
mi	to me
na	to
dušóno	soul-the

Figure 3: Gloss of (16c)

## 6.2 Usage examples

The Corpus tab provides access to usage examples of the VMWE (Figure 4). For each usage example, a set of translations and the source of the example are available. The Source tab provides the name of the village of the speaker who contributed the respective usage example; for instance, in Figure 4 both usage examples have as their source the village of Mándena. Book references and URLs are normally used as sources of examples. However, the vast majority of Pomak usage examples of VMWEs were collected by means of interviews with native speakers (§3).



Example	Translations	Source
Šúkür Allahá aléknava mi dušása. Stóre si vrit rábatýte.	<a href="#">View Translation</a>	χωριό Μάντανα
Ohhó! Kakvó mi aléknava dušása óti so akópa.	<a href="#">View Translation</a>	χωριό Μάντανα

Figure 4: Usage examples of (16b).

### 6.3 UD analysis of the lemma form

The UD-analysis tab gives the graphical format (Figure 5) and the CoNLL-U format of the analysis of the variants of the lemma form according to the UD formalism; the CoNLL-U version of the UD analysis can be viewed and downloaded through the dedicated button. The UD analysis, together with the gloss of the lemma form (Figure 3), offer detailed structural information about the VMWE. The analysis draws on the approach to Pomak morphology and syntax that has been applied on the UD Pomak treebank; this approach is outlined coarsely in §2.2, §2.3 and §2.4.

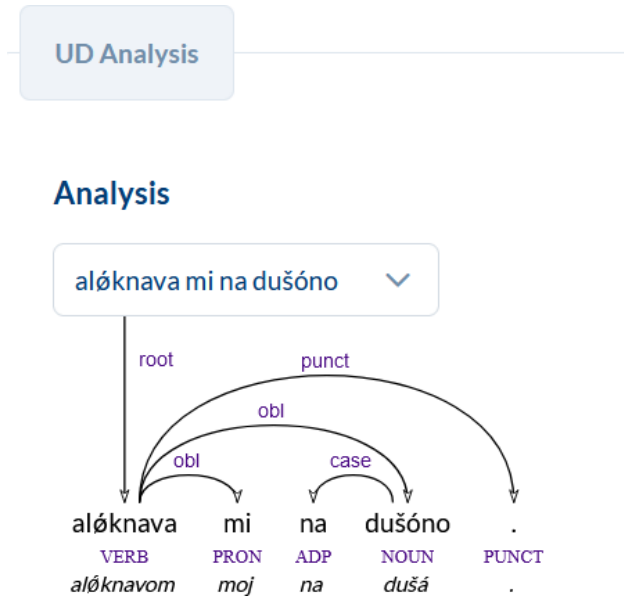


Figure 5: Graphical format of the UD analysis of (16c)

### 6.4 Lexical (semantic) relations: Other relations

In Figure 6 the synonyms and opposites of (16c) are given. In addition, there is a VMWE standing in the verb alternation relation with (16c).

Our data show that Pomak exemplifies the labile type of the causative/inchoative alternation (§5.6). The labels *Has\_causative* and *Has\_inchoative* are used to annotate pairs of VMWEs that stand in this relation. In Figure 7, the causative VMWE (16a) has the inchoative counterpart (16b). To our knowledge, this is the first time that verb alternation phenomena have been discussed for Pomak.



Open in new tab

Other Relations

Relation ↑↓	Expression
Opposite	atóžnavá mi na dušóno
Verb Alternation	náeko mi alóknava dušóno

Remove tab

Synonyms

pačúnnavom ad uvótre  
páda mi húbovo na dušóso  
alóknava mi na dušóno

Figure 6: Synonyms of (16c).

Other Relations

Relation ↑↓	Expression
Has_inchoative	alóknavá mi dušása
Verb Alternation	alóknava mi na dušóno
Opposite	izzéde mi dušóso

Figure 7: The Has\_inchoative relation defined on (16a).

## 7 The future

Pomak-IDION is a unique resource of an endangered living language. It belongs to the set of Pomak resources developed in the framework of the project Philotis.

Pomak-IDION offers material and motivation for a future thorough study of Pomak VMWEs, e.g., studies on the role of the triple enclitic deictic article in idioms, the syntactic flexibility properties of VMWEs, verb alternation phenomena, language contact phenomena observed with LVCs and idioms and studies on the semantics of idiomatic Pomak.

Enriching IDION, and Pomak-IDION, with the inflectional paradigms of the VMWEs is among our future plans. This presupposes the encoding of the idiosyncratic constraints that hold for a number of VMWEs, other than constraints on the lexicalised parts: for instance, a VMWE may never appear in the future tense or the 1st person but may fully inflect for all the other tenses and persons. Such constraints are not expressed by the UD representation of the lemma form and are only partially covered by the corpus material; at the moment, encoders keep notes in IDION describing these properties of the VMWEs.

The development of the Pomak edition of IDION has shown that it can accommodate detailed information on VMWEs of different languages. In the future, cross-edition relations between VMWEs may be added to IDION. So far, each edition has been independent of the others; as a result, switching between the respective editions is required in order to see two equivalent expressions in two different editions. The implementation of cross-edition relations is an interesting documentation capability that will facilitate comparative studies on idiomaticity and other linguistic activities such as teaching and translation.

## Abbreviations

GE	Modern Greek equivalent
LVC	Light verb construction
NLP	Natural Language Processing
K&K alphabet	Alphabet by R. Karahođa and P. G. Krimpas
PoS	Part of speech
UD	Universal Dependencies
VMWE	verbal multiword expression

## Acknowledgements

We acknowledge full support of this work by the project “PHILOTIS: State-of-the-art technologies for the recording, analysis and documentation of living languages” (MIS 5047429), which is implemented under the “Action for the Support of Regional Excellence”, funded by the Operational Programme “Competitiveness, Entrepreneurship and Innovation” (NSRF 2014–2020) and co-financed by Greece and the European Union (European Regional Development Fund).

## Appendix A Pomak LVCs

- (17) a. **fátom**                      **nazára**  
 catch.VERB.1SG evil eye.NOUN  
 ‘I am jinxed’. Verb: *nazarjásavom* ‘I am affected by evil eye’
- b. **stánavom**                      **budalá**  
 become.VERB.1SG mad.ADJ  
 ‘I go crazy’. Verb: *pabudalævom* ‘I go crazy’
- c. **stánavom**                      **dløg**  
 become.VERB.1SG tall.ADJ  
 ‘I grow tall’. Verb: *izdløgnavom* ‘I grow tall’
- d. **stánavom**                      **gulæm**  
 become.VERB.1SG big.ADJ  
 ‘I grow big’. Verb: *nagulæmævom* ‘I grow big’
- e. **stánavom**                      **hazýr**  
 become.VERB.1SG ready.ADJ  
 ‘I get ready’. Verb: *hazyrladísavom so* ‘I get ready’
- f. **stánavom**                      **star**  
 become.VERB.1SG old.ADJ  
 ‘I grow old’. Verb: *sastarævom, stárem* ‘I grow old’
- g. **stánavom**                      **zengínin**  
 become.VERB.1SG rich.ADJ  
 ‘I become rich’. Verb: *zenginjásavom* ‘I become rich’
- h. **tavárem**                      **so**                      **græha**  
 load .VERB.1SG myself.PRON sin.NOUN  
 ‘I commit a sin’. Verb: *græhóvom* ‘I commit a sin’

## Appendix B Pomak idioms

- (18) a. **adbávem**            **hatýrane**  
destroy.VERB.1SG favour.NOUN  
'I refuse to satisfy somebody's wishes' GE: *χαλάω χατίρι*
- b. **atkáčem**            **jazýkate**  
sever.VERB.1SG tongue.the.NOUN  
'I make someone stop talking' GE: *κόβω τη γλώσσα κάποιου*
- c. **atvárem**            **ačise**  
open.VERB.1SG eyes.the.NOUN  
'I realize what is going on' GE: *ανοίγω τα μάτια μου*
- d. **fáta**            **gi**            **sas**            **annóš**  
catch.VERB.3SG them.PRON with.ADP once.ADV  
'he is bright' GE: *τα πιάνει με την μία*
- e. **fórnem**            **něko**            **na**            **pótene**  
throw.VERB.1SG somebody in.ADP street.the.NOUN  
'I kick out someone' GE: *πετάω κάποιον στον δρόμο*
- f. **glódom**            **tavánase**  
look.VERB.1SG ceiling.the.NOUN  
'I am absent minded' GE: *κοιτάω το ταβάνι*
- g. **hránem**            **zmíje**            **faf**            **skútase**  
feed.VERB.1SG snake.NOUN in.ADP bosom.the.NOUN  
'I befriend somebody who proves to be deceitful' GE: *τρέφω φίδι στον κόρφο μου*
- h. **izzéde**            **mi**            **dušóso**  
eat off.VERB.3SG.PST to/of-me.PRON soul.NOUN  
'it has distressed me' GE: *μου έφαγε την ψυχή*
- i. **je**            **mi**            **so**            **katá**            **vólek**  
be.AUX.1SG to/of-me.PRON REFL like.ADV wolf.NOUN  
'I am starving' GE: *πεινάω σα λύκος*
- j. **na**            **móžom**            **da**            **zgýbem**            **nagýse**  
not.PART can.VERB.1SG that.ADP move.1SG leg.NOUN.PL  
'I am exhausted, I am burnt out' GE: *δεν μπορώ να κουνήσω τα πόδια μου*

- k. **na pamína ad móse**  
not.PART go.3SG.VERB through.ADP my.the.SING.FEM.ACC  
**róky**  
hand.NOUN.PL  
'it does not depend on me' GE: *δεν περνάει από το χέρι μου*
- l. **na sésta so ad láfa**  
not.PART understand.3SG.VERB REFL from.ADP word.NOUN.PL  
'he is indifferent' GE: *δεν καταλαβαίνει από λόγια*
- m. **na zaznáje mu so ušána**  
not.PART sweat.3SG.VERB to-me.PRON REFL ear.NOUN  
'I don't give a damn' GE: *δεν ιδρώνει το αυτί μου*
- n. **pádom na móko**  
fall.VERB.1SG on.ADP soft place.NOUN  
'I escape unpunished' GE: *πέφτω στα μαλακά*
- o. **píjem bannómu karvtóno**  
drink.VERB.1SG somebody's.PRON blood.NOUN  
'I drain somebody's blood' GE: *πίνω το αίμα κάποιου*
- p. **púkom ad játo**  
break.VERB.1SG from.ADP food.NOUN  
'I eat excessively' GE: *σκάω από το φαγητό*
- q. **rábatem katá kúče**  
work.VERB.1SG like.ADV dog.NOUN  
'I work hard' GE: *δουλεύω σαν σκύλος*
- r. **sédom sas svózany róky**  
sit.VERB.1SG with.ADP crossed.ADJ arms.NOUN  
'I do nothing, remain inactive' GE: *κάθομαι με δεμένα χέρια*
- s. **videm bála déne**  
see.VERB.1SG white.ADJ day.NOUN  
'I get a break, I get ahead in life' GE: *βλέπω άσπρη μέρα*
- t. **zímom go ad ustána mu**  
get.VERB.1SG it.PRON from.ADP mouth.NOUN his.PRON  
'I take the words out of somebody's mouth' GE: *το παίρνω από το στόμα του*
- u. **katá vadíca go naúčem**  
like.ADV water.NOUN it.PRON learn.VERB.1SG  
'I learn something perfectly' G E: *μαθαίνω νεράκι κάτι*

- v. **korv**            **plújem**            OR **kyrv**            **hráčem**  
 blood.NOUN spit.VERB.1SG OR blood.NOUN spit.VERB.1SG  
 ‘I work hard to succeed’ GE: *φτύνω αίμα*
- w. **máhnnavot so**            **káto**            **dve**            **kápky**            **vódo**  
 look alike.3PL.VERB like.ADP two.NUM drop.NOUN.PL water.NOUN  
 ‘they are like peas in a pod’ GE: *μοιάζουν σα δυο σταγόνες νερό*

## References

- Adamou, Evangelia & Davide Fanciullo. 2018. Why Pomak will not be the next Slavic literary language. In D. Stern, M. Nomachi & B. Belić (eds.), *Linguistic regionalism in Eastern Europe and beyond: Minority, regional and literary microlanguages*, 40–65. Berlin: Peter Lang. <https://halshs.archives-ouvertes.fr/halshs-02105739>.
- Baldwin, Timothy & Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha & Fred J. Damerau (eds.), *Handbook of Natural Language Processing*, 2nd edn., 267–292. Boca Raton, FL: CRC Press.
- Brenzinger, Matthias, Akira Yamamoto, Noriko Aikawa, Dmitri Koundioubu, Anahit Minasyan, Arienne Dwyer, Colette Grinevald, Michael Krauss, Osahito Miyaoka, Osamu Sakiyama, María E. Villalón, Akira Y. Yamamoto & Ofelia Zepeda. 2003. *Language vitality and endangerment*. Document submitted to the International Expert Meeting on UNESCO Programme Safeguarding of Endangered Languages.
- Constantinides, Nicolaos Th. 2007. *Η πομακική πολιτισμική μονάδα στην ελληνική Θράκη από άποψη Παρευξείνιων Σπουδών: Σύντομη ιστορική επισκόπηση, γλώσσα, ταυτότητες*. Democritus University of Thrace. (MA thesis).
- Constantinides, Nicolaos Th. 2020. Συγκλίσεις και αποκλίσεις στην Πομακική της ελληνικής Θράκης αφορούσες τα πεδία της αοριστίας, της οριστικότητας και του τριμερούς προσδιορισμού υπό το πρίσμα μιας σύνθετης λαογραφικής θέωρησης. (‘Convergences and divergences in the Pomak of Greek Thrace concerning the fields of indeterminacy, finality and tripartite determination in the light of a complex folklore view’). *Mare Ponticum* 8(1). 56–76.
- de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. DOI: 10.1162/coli\_a\_00402.

- Gantar, Polona, Lut Colman, Carla Parra Escartín & Héctor Martínez Alonso. 2018. Multiword expressions: Between lexicography and NLP. *International Journal of Lexicography* 32(2). 138–162. DOI: 10.1093/ijl/ecy012.
- 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*, 147–186. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998639.
- Grégoire, Nicole. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation* 44(1/2). 23–39. DOI: 10.1007/s10579-009-9094-z.
- Al-Haj, Hassan, Alon Itai & Shuly Wintner. 2013. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography* 27. 130–170. DOI: 10.1093/ijl/ect036.
- Haspelmath, Martin. 1993. More on the typology of inchoative/causative verb alternations. In Bernard Comrie & Maria Polinsky (eds.), *Causatives and transitivity* (Studies in Language Companion Series 23), 87–120. Amsterdam, Philadelphia: John Benjamins. DOI: 10.1075/slcs.23.05has.
- Hüllen, Werner. 2004. *A history of Roget's "Thesaurus"*. New York: Oxford University Press.
- Jespersen, Otto. 1965. *A Modern English grammar on historical principles*, vol. 6: Morphology. London: Allen & Unwin.
- Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali & Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 6282–6293. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.560. <https://aclanthology.org/2020.acl-main.560>.
- Karahóga, Ritván, Panagiotis G. Krimpas, Vivian Stamou, Vasileios Arampatzakis, Dimitrios Karamatskos, Vasileios Sevetlidis, Nikolaos Constantinides, Nikolaos Kokkas, George Pavlidis & Stella Markantonatou. 2022. Morphologically annotated corpora of Pomak. In *Proceedings of the fifth workshop on the use of computational methods in the study of endangered languages*, 179–186. Dublin. DOI: 10.18653/v1/2022.computel-1.22. <https://aclanthology.org/2022.computel-1.22>.
- Karahóga, Sebajdín. 2017. *Μεταφράσεις ελληνικής και αγγλικής ποίησης στην πομακική γλώσσα*. Ξάνθη: Πολιτιστικός Σύλλογος Πομάκων Ξάνθης.
- Kokkas, Nikolaos. 2004. *Uchem so Pomátsko: Μαθήματα πομακικής γλώσσας*, vol. A. Ξάνθη: Πολιτιστικό Αναπτυξιακό Κέντρο Θράκης.

- Krimpas, Panagiotis G. 2020. Η γλώσσα και η καταγωγή των Πομάκων υπό το φως της Βαλκανικής Ζώνης Γλωσσικής Επαφής. In Manolis Varvounis, Antonis Bartsiokas & Nadia Macha-Bizoumi (eds.), *Οι Πομάκοι της Θράκης: πολυεπιστημονικές και διεπιστημονικές προσεγγίσεις* (Μελέτες Λαογραφίας και Κοινωνικής Ανθρωπολογίας 7), 167–204.
- Krimpas, Panagiotis G. 2022. Ευρωγλωσσολογία, νεοελληνική γλώσσα και ευρωπαϊκή ολοκλήρωση. In Zoe Gavriilidou, Nikolaos Mathioudakis, Maria Mitsiaki & Asimakis Fliatouras (eds.), *Γλωσσανθοί: Μελέτες αφιερωμένες στην Πηνελόπη Καμπάκη-Βουγιουκλή*, 153–169. Athens: Herodotus, Democritus University of Thrace.
- Laporte, Éric. 2018. Choosing features for classifying multiword expressions. In Manfred Sailer & Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 143–186. Berlin: Language Science Press. DOI: 10.5281/zenodo.1182597.
- Leseva, Svetlozara, Verginica Barbu Mititelu, Ivelina Stoyanova & Mihaela Cristescu. 2024. A uniform multilingual approach to the description of multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 73–116. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998635.
- Lichte, Timm, Simon Petitjean, Agata Savary & Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of “irregular” regularities. In Yannick Parmentier & Jakub Waszczuk (eds.), *Representation and parsing of multiword expressions: Current trends*, 1–33. Berlin: Language Science Press. DOI: 10.5281/zenodo.2579033.
- Losnegaard, Gyri Smørdal, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann & Johanna Monti. 2016. PARSEME survey on MWE resources. In *10th international conference on Language Resources and Evaluation (LREC 2016)*, 2299–2306. Portorož. <https://hal.science/hal-01316351>.
- Lyons, John. 1977. *Semantics*, vol. 1. Cambridge University Press. DOI: 10.1017/CBO9781139165693.
- Markantonatou, Stella, Panagiotis Minos, George Zakis, Vassiliki Moutzouri & Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the joint workshop on multiword expressions and WordNet (MWE-WN 2019) at ACL 2019*, 130–134. Florence. DOI: 10.18653/v1/W19-5115.
- Markantonatou, Stella, Nicolaos Th. Constantinides, Vivian Stamou, Vasileios Arampatzakis, Panagiotis G. Krimpas & George Pavlidis. 2023. Methodological issues regarding the semi-automatic UD treebank creation of under-resourced



- languages: The case of Pomak. In *Proceedings of the sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, 27–35. Washington, D.C. <https://aclanthology.org/2023.udw-1.4>.
- Ní Loingsigh, Katie & Brian Ó Raghallaigh. 2016. Starting from scratch: The creation of an Irish-language idiom database. In Tinatin Margalitadze & George Meladze (eds.), *Proceedings of the 17th EURALEX International Congress*, 726–734. Tbilisi: Tbilisi State University.
- Nielsen, Jakob & Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'90)*, 249–256. DOI: 10.1145/97243.97281.
- Osenova, Petya & Kiril Simov. 2024. Representation of multiword expressions in the Bulgarian integrated lexicon for language technology. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 117–146. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998637.
- Papadimitriou, Panayotis. 2008. *Τα Πομάκικα: Συγχρονική περιγραφή μιας νότιας τοπικής ποικιλίας της αναλυτικής σλαβικής από τη Μύκη του Ν. Ξάνθης*. Thessaloniki: Kyriakides Bros.
- Papadimitriou, Panayotis. 2013. *Λαλιές Πομάκων της ελληνικής Ροδόπης: Περιφερειακή Αναλυτική Σλαβική και μουσουλμάνοι ομιλητές στη Νοτιοανατολική Ευρώπη*. Thessaloniki: Institute for Balkan Studies.
- Piirainen, Elisabeth. 2005. Europeanism, internationalism or something else? Proposal for a cross-linguistic and cross-cultural research project on widespread idioms in Europe and beyond. *HERMES: Journal of Language and Communication in Business* 18(35). 45–75. DOI: 10.7146/hjlc.v18i35.25816. <https://tidsskrift.dk/her/article/view/25816>.
- Sailer, Manfred & Stella Markantonatou. 2016. Affectees in MWEs: German and Modern Greek. In *Posters from the PARSEME 6th general meeting, 7–8 April 2016, Struga, North Macedonia*. <https://typo.uni-konstanz.de/parseme/images/Meeting/2016-04-07-Struga-meeting/WG1-MARKANTONATOU-SAILER-poster-1.pdf>.
- Sandry, Susan. 2013. *Phonology and morphology of Paševik Pomak with notes on the verb and fundamentals of syntax*. University College London, School of Slavonic & East European Studies. (MA thesis).
- Savary, Agata. 2009. Multiflex: A multilingual finite-state tool for multi-word units. In Sebastian Maneth (ed.), *Implementation and application of automata*, 237–240. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-02979-0\_27.

- Savary, Agata, Marie Candito, Verginica Barbu Mititelu, Eduard Bejcek, Fabienne Cap, Slavomir Ceplo, Silvio Ricardo Cordeiro, Gulsen Eryigit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaite, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartin, 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.14715.
- Savary, Agata, Silvio Cordeiro & Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the joint workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, 79–91. Florence. DOI: 10.18653/v1/W19-5110.
- Skoumalová, Hana, Marie Kopřivová, Vladimír Petkevič, Tomáš Jelínek, Alexandr Rosen, Pavel Vondříčka & Milena Hnátková. 2024. LEMUR: A lexicon of Czech multiword expressions. In Voula Giouli & Verginica Barbu Mititelu (eds.), *Multiword expressions in lexical resources: Linguistic, lexicographic, and computational perspectives*, 1–37. Berlin: Language Science Press. DOI: 10.5281/zenodo.10998631.
- Theocharides, Petros. 1996a. *Γραμματική της Πομακικής Γλώσσας*. Thessaloniki: Egiros.
- Theocharides, Petros. 1996b. *Πομακο-Ελληνικό Λεξικό / Πομάχτσκου-Ουρούμτσκου Λεκσικό*. Thessaloniki: Egiros.
- Villavicencio, Aline, Timothy Baldwin & Benjamin Waldron. 2004. A multilingual database of idioms. In *Proceedings of the fourth international conference on Language Resources and Evaluation (LREC'04)*, 1127–1130. Lisbon. <http://www.lrec-conf.org/proceedings/lrec2004/pdf/760.pdf>.
- Vondříčka, Pavel. 2019. Design of a multiword expressions database. *The Prague Bulletin of Mathematical Linguistics* 112. 83–101.