Lexicographic tool & resource for verb MWEs

STELLA MARKANTONATOU¹, ERI KOLETTI², ELPINIki MARGARITí², PANAGIOTIS MINOS¹, AIMILIA STRIPELI², GEORGIOS ZAKIS², NIKI SAMARIDí³

¹INSTITUTE FOR LANGUAGE AND SPEECH PROCESSING/“ATHENA” RIC, MARKS@ILSP.GR, PMINOS@GMAIL.COM

²NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS, ERKG7@YAHOO.GR, ELPIMARGARITí@GMAIL.COM, ASTRIPELI@GMAIL.COM, GEORGIZAK@GMAIL.COM,

³NSAMARIDí@GMAIL.COM
The Lexicographic tool & resource: (1)

- Combines a wide range of linguistic information on MG verb Multiword Expressions (MWEs).
- Addressed both to the human user and to NLP applications.
- For its XML editing a custom-made Java desktop application based on the NetBeans Rich-Client Platform (RCP) framework has been developed.
- Each entry constitutes a detailed description of a MWE (we understand a MWE as a string with no compositional meaning).
- Currently contains 225 MWEs (plus 75 under editing).
- Users interact with the lexicographic tool & resource through an auto-generated preview, summarizing all the information about the entry.
The Lexicographic tool & resource: (2)

Organized in 7 sections (tabs):

1. “Preview” tab
2. “General” tab
3. “Forms” tab
4. “Use” tab
5. “Diagnostics” tab
6. “Corpus” tab
7. “Relations” tab

For each section a detailed description is given below.
The “Preview” tab

### Semantics

**Greek:**
αφήνω την κουβέντα στη μέση

**English:**
to leave a discussion incomplete/not finished

### Use

<table>
<thead>
<tr>
<th>Example</th>
<th>Parole</th>
<th>Transcript</th>
</tr>
</thead>
<tbody>
<tr>
<td>[... ]</td>
<td>asked</td>
<td>rotissa</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td>tn</td>
</tr>
<tr>
<td>cashier</td>
<td>who</td>
<td>tama, i</td>
</tr>
<tr>
<td>me</td>
<td>me</td>
<td>opia</td>
</tr>
<tr>
<td>answered</td>
<td>this</td>
<td>mou</td>
</tr>
<tr>
<td>Sko</td>
<td>...</td>
<td>apantine</td>
</tr>
<tr>
<td>is</td>
<td>...</td>
<td>os</td>
</tr>
<tr>
<td>«Ενημερώσω σας»</td>
<td>«enumerosame toues pelates»</td>
<td>«enumerosame toues pelates»</td>
</tr>
<tr>
<td>«καθόλου...»</td>
<td>«mas gia to problima...»</td>
<td>«mas gia to problima...»</td>
</tr>
<tr>
<td>και</td>
<td>and</td>
<td>left</td>
</tr>
<tr>
<td>άφησε</td>
<td>left</td>
<td>talk</td>
</tr>
<tr>
<td>πρόβλημα</td>
<td>problem</td>
<td>NoCmFtSgAc</td>
</tr>
<tr>
<td>the DET. AdDNeSgAc</td>
<td>in the</td>
<td>NoCmFtSgAc</td>
</tr>
<tr>
<td>NoCmFtSgAc</td>
<td>talk</td>
<td>NoCmFtSgAc</td>
</tr>
</tbody>
</table>

### Translation:

[...] I made a question to the cashier, who gave me the answer that “we let our customers know about the problem...” and suddenly she interrupted her phrase.

**Source:**
http://www.philenews.com/el-gr/koumnia-anagnostis-echeis-logo-435/186584-me-ekaman-rezhi-ependi-don-echo-met...

### Examples

**Example:** Κάθε φορά που άφησε μια κουβέντα στη μέση, θα καθορίζεται από τη σίγα η τα γείτονα του να κρίνεται από δύο μέντες στον αέρα, άφησε τα μέτρα τους να καρφώνονται πάνω του, να στομάχει την κουβέντα στη μέση κατα το παρακολουθούν μέχρι να ακούσται στον καρπαφίτα που θα το τοποθετούσε μαζί με τα άλλα σε μια απότομη ακέραιο το ναυά.
The “General” tab provides:

1. the meaning of the MWE examined
2. comments that the editor may want to add.

The meaning is given in two languages: **English** & **Greek**

For instance, the meaning of the MWE “κόβω τα φτερά κάποιου” ("cut someone’s wings") is written as:

**Greek:** αποθαρρύνω κάποιον

**English:** to discourage someone, to demotivate someone
If an expression has two or more distinct meanings each meaning is encoded separately as a special MWE entry.

For example, (a) means that someone was imprisoned, while (b) that someone has lost money and is in debt:

a. Ήθελε φακελάκια αυτός ο γιατρός και (επιτέλους) τον βάλανε μέσα!
   Wanted-V.3SG bribe-ACC this-PN the doctor-NOM and (at_last) him put-V.3PL.PAST in-ADV
   “This doctor wanted a bribe and (at last) he was imprisoned”
   (http://www.athensmagazine.gr/portal/athenstalk/51087)

b. Το σπίτι που έβαλε μέσα χοντρά τη Σάρα Τζέσικα Πάρκερ...
   The house-NOM which put-V.3SG.PAST in-ADV a_lot-ADV the Sarah Jessica Parker-ACC
   “This house costed Sarah Jessica Parker a lot of money, and now she is in debt”
   (www.womenonly.gr/gallery.asp?catid=37522&subid=2&pubid)
The “Forms” Tab (1)

- Provides for the exhaustive **morphological** and **syntactic** description of the MWE
- The encoding ➔ theory neutral ➔ aimed to serve as a basis for any type of parser.
- Standardized morphological tags are used (PAROLE [http://nlp.ilsp.gr/nlp/tagset_examples/tagset_en/](http://nlp.ilsp.gr/nlp/tagset_examples/tagset_en/)).

- **The encoded syntactic relations include:**
  - Information about **free** constituents
    - **Phrasal** information for constituents that are realized with **full phrasal structures**
    - **Lexical** information for constituents that are realized with **weak pronouns**
  - binding and control relations
  - delineation of fixed/semi-fixed strings
The “Forms” Tab (2)

- A tabular arrangement with **four columns** is used together with **controlled vocabularies** that help to minimize the number of errors.
- The label ‘**tokens**’ is used to cover both lexical and phrasal parts of a MWE.
- Each token occupies a line in the tabular format.

**The controlled vocabulary provided in Column 1 is:**

- **LEMMA**: declinable
- **WF** (WORDFORM): words that are not encountered in environments other than the MWE
- **COMPL**: designates the notion COMPLEMENTIZER (να, θα, που, …)
- **XP**: any completely free XP including an S following a complementizer
- **NP-NOM/ NP-NOM-anim/ NP-NOM-nonamim**
- **NP-GEN/ NP-GEN-anim/ NP-GEN-nonamim**
- **NP-ACC/ NP-ACC-anim/ NP-ACC-nonamim**
- **PnGe**: pronoun in genitive case
For instance, the MWE...

"αφήνω κάποιον στον τόπο"

let-1SG someone-NP.ACC at-PREP place-NP.ACC

"I kill someone"

...is encoded as:
We introduce a new form if we have variations of the MWE due to:

1. the occurrence of diminutives (ουρά-ουρίτσα)
2. slightly different versions of a lemma (ουράνια-μεσουράνια),
3. specified adjectives/modifiers
4. a variation of definite/indefinite determiner
5. a variation of prepositions

! → We do not introduce a new form in order to encode differences in animacy.
The “USE” Tab

- For each MWE entry, a characteristic example, along with the phonetic transcription, PAROLE annotation and English translation is provided.
- The glossed example is given with a tabular representation:

<table>
<thead>
<tr>
<th>Example</th>
<th>Η</th>
<th>Μαλένα</th>
<th>ἀφησε...</th>
<th>μπουκάλα</th>
<th>τὸν</th>
<th>Κάρλες</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parole:</td>
<td>The.AtDfFeSgNm</td>
<td>Mialena.NoPrFeSgNm</td>
<td>left.VbMnIdPa03SgXxPeAvXx</td>
<td>bottle.NoCmFeSgAc</td>
<td>the.AtDfMaSgAc</td>
<td>Karles.NoPrMaSgAc</td>
</tr>
<tr>
<td>Transcript:</td>
<td>I</td>
<td>Malena</td>
<td>afise...</td>
<td>boukala</td>
<td>ton</td>
<td>Karles.</td>
</tr>
</tbody>
</table>

**Translation:**
Malena walked out on Carles.

**Source:**
In this section are stored:

- **Both grammatical and ungrammatical strings** featuring MWEs.
  - Strings are directly linked to diagnostics providing data to support or challenge the assignment of properties to the MWE
  - Grammatical strings are drawn from the Hellenic National Corpus ([http://hnc.ilsp.gr/](http://hnc.ilsp.gr/)) and from Google.
  - Ungrammatical strings (and some grammatical) are evaluated by native speakers (introspection).

- Strings are classified as grammatical or ungrammatical with the use of a button which has three options to click:
  - **acceptable**
  - **unacceptable**
  - **???** (It is assigned to strings whose acceptability is questioned)
That’s how the corpus looks in our lexicographic tool & resource: (Notice that the source of every string is always provided) We think of the corpus as a future resource for research or machine learning.
The Diagnostics Tab (1)

- In the “Diagnostics tab”, we investigate:

1. whether a verb MWE has a **free subject or not** (dedicated diagnostic).
   - If different NPs trigger agreement on the verb a free subject exists, otherwise the subject is fixed.
   - The situation is shown with **corpus examples** that demonstrate **subject-verb agreement** with a variety of subjects.

2. **the number of constituents** that a MWE contains. For instance, the MWE below contains three constituents:

   [κάνω] [μαύρα μάτια] [να δω NP-ACC]
   [do]     [black eyes]     [to see NP-ACC]

   - The diagnostics ‘**admission of a free XP**’ and ‘**word order permutations**’ are used as constituency diagnostics.
We also investigate:

1. whether the MWE alternates between a free NP-GEN/ se-PP, apo-PP and a form with a Dative Genitive
2. whether the fixed parts of the MWE can be replaced with a clitic in the same predication
3. whether the MWE participates in the causative-inchoative alternation
4. whether the MWE passivizes

Each question is assigned a yes/no button and the ability to be exemplified with examples drawn from the Corpus Tab illustrating the phenomenon in question.
The Diagnostics Tab (3)

Free/fixd subject diagnostic:
Can different NPs trigger agreement on the verb?

Number of constituents within the VP
Does the MWE accept a free NP?

Word order permutations involving the subject as well

Cliticisation of the WME

Alternation of a free NP-GEN / free ac-PP / free one-PP with Dative Genitive

Causative inchoative alternation

* ta fterá tou kóvoun

Passive tone

Ta fterá tou kóvoun káneis perissóstero to apógyma tou zebrátiou me to "diplo" tou Panathinikou...
Cliticisation of the fixed parts

→ The value **YES** is assigned if cliticisation is possible in the context of the same MWE:

'Εβαλε την ουρά του κι ο… Τσιώλης!
Put the tail-ACC his-POSS and the Tsiolis-NOM!
Nai, tηn έβαλε.
Yes, this-PN.ACC put-V.3SG.PAST.

≠ The value **NO** is assigned if cliticisation is not possible in the context of the same MWE:

'Εταξε στην Ελένη λαγούς με πετραχήλια.
Promised in-PREP Helen-ACC rabbits-ACC with-PREP vestments-ACC.
*Nai, tous έταξε.
*Yes, these-PN.ACC promised.

→ The value ‘---’ is assigned when the cliticisation diagnostic is **irrelevant**, eg. when the fixed part is a PP.
Alternation of a free NP-GEN/ free se-PP/ free apo-PP with a form with a Dative Genitive

- The value **YES** is assigned if the alternation with a Dative Genitive is possible:

  Ο Γιώργος έκοψε τα φτερά της Ελένης.
  The George-NOM cut-V.3SG.PAST the wings-ACC of-POSS Helen-GEN.
  “George discouraged Helen”

  Της έκοψε τα φτερά ο Γιώργος.
  Her-DAT.GEN cut-V.2SG.PAST the wings-ACC the George-NOM.

≠ The value **NO** is assigned if the alternation with a Dative Genitive is not possible:

  Τρώω τα νύχια μου.
  Eat-1SG the nails-ACC my-POSS.GEN.
  “I’m anxious”

  *Μου τρώω τα νύχια.
  *DAT.GEN.3RD eat-V.1SG the nails-ACC.
The “Relations” Tab

- the **semantic relations** among MWEs are exhaustively stored.
- 5 types of relations:
  - Synonymous MWEs
  - Opposite MWEs
  - Semantic pair:
    - \[ \text{ανάβω το πράσινο φως} \] turn-on the green light
    - \[ \text{δίνω το πράσινο φως} \] give the green light
  - Verb alternations
  - ??? (this choice exists for relations that are not easy to characterize)
Open Issues

(1) Forms: the encoding of the **variants** should be elaborated

(2) **Polarity**

   The two meanings of \( \pi\alpha\rho\nu\omega \ \tau\alpha \ \beta\omega\nu\nu\alpha \)

(3) **Semantic organization** of the DB

(4) The DB has been designed to feed **any grammatical formalism**. This has to be evaluated yet.

→ Such is the case of the dative genitive that is aimed to be derived with rules from the without-the-dative-genitive encoded version.

We are planning to make the Lexicographic tool and resource publicly available in the near future.
Thank you!