WG1/WG2: Generating LF|G/XLE MWE entries from IDION (a theory neutral lexical DB)

Stella Markantonatou (ILSP/"Athena" RIC), Georgios Zakis and Elpiniki Margariti (University of Athens) and Panagiotis Minos (TEI of Athens)

The lexical DB IDION is addressed to the human user and to NLP systems and encodes various types of information, formal information included, on (Modern Greek (MG)) verb MWEs (Markantonatou et al, 2015). Morphological features are exhaustively described with the ILSP-PAROLE compatible tagset. MWEs are marked for passivisation and diathesis alternation phenomena. The encoding of the MWE structure is more or less flat and theory neutral except for (i) the phrasal categories in (1) that are used to denote free constituents of the MWE (Table 1, column ID1) (ii) the marking of the verbal heads in main and subordinate clauses (Table 1, column ID4) and, (iii) the marking of binding (Table 1, column ID4) and control phenomena.

(1) NP-NOM/NP-NOM-anim/NP-NOM-nonanim; NP-GEN/NP-GEN-anim/NP-GEN-nonamim; NP-ACC/NP-ACC-anim/NP-ACC-nonamim; VP

As opposed to other MWE DBs (such as DUELME (Grégoire, 2010)), IDION assumes an expressive standardized language for encoding morphological information that suits better to morphologically rich languages. Furthermore, IDION only indexes sequences of fixed parts (Table 1, column ID5) and does not encode their phrasal structure; this encoding suits to flexible word order languages, like MG, because fixed sequences of fixed parts (often called 'words with spaces' (WWS)) may spread across "phrasal constituents" in a MWE that would otherwise allow for word order permutations (MG: leo to_psomi_psomaki = call the bread little-bread 'I starve', where 'little bread' could be thought as predicated of 'bread'; these structures allow for more than one word order but the particular MWE does not).

To demonstrate IDION's availability for parsing and as a case study, we present how an LFG/XLE lexicon is developed with a program that reads off information from IDION. We assume the LFG/XLE MWE parser described in Samaridi et al. (2014) that treats WWSs as strings of words connected with underscores. The algorithm for developing the XLE lexicon is shown schematically here with the verb MWE in (2) that contains three WWSs, one with an "object" function and a bound pronoun dependent on it, an analytical preposition consisting of an adverb and a preposition proper and one with a prepositional "object" function and a bound pronoun dependent on it(Table 1). The brackets in (2) indicate phrasal units and the bold-faced strings indicate fixed sequences of fixed words (WWSs).

(2) (ego) echo [**to mialo** mou] [**pano apo to kefali** mou] ='I am thoughtless.'
I have the brain mine.1SG.GEN over from the head mine.1SG.GEN

Columns ID1-ID5 are a copy of the IDION FORMS tab and show exactly how the encoding of the morphosyntactic properties of the MWEs are encoded. ID1 assigns a phrasal label to the free parts and a morphological label to the fixed parts of the MWE. ID2 lists the lemmata of the fixed parts, ID3 lists the fixed parts of the MWE as well as control constraints on verbs, ID4 lists the precise morphological constraints on the parts of the MWE along with binding constraints (noted on Table 1 as 'GBC') and ID5 lists the indexes of fixed sequences of fixed MWE parts. Columns A0-A5 represent the steps in the lexicon developing procedure. We developed a Java application that implements the algorithm. The same application is used to both convert IDION to an XLE lexicon and evaluate the contents of IDION.

Of the six steps of the lexicon developing algorithm, steps 2, 5 and 6 return lexical entries that are stored in XLE. The steps apply if necessary, for instance a MWE may not contain an indexed part or a subordinated clause. A brief description of the 6 steps follows:

1. A0: the algorithm transcribes the PAROLE tagset to feature-value pairs.
2. A1: the algorithm defines WWSs from the indexed parts of the MWE (encoded on ID5). WWSs are stored as new predicates in XLE; they are assigned a part of speech, and they may subcategorise for Grammatical Functions. On Table 1 we see the three WWS entries that are developed to parse (2): two 'nominal' WWSs (NWWS) that subcategorise for a POSS, namely PRED 'to_mialo(POSS)' and PRED 'to_kefali(POSS)', and a 'prepositional' WWS (PWWS) that subcategorises for an OBJ, namely PRED'pano_apo(OBJ)'. WWSs inherit any morphological and binding constraints attached to their parts.

3. A2: PPs, AdvPs and S1s (an S introduced with a complementizer) are developed from WWSs and the other lexical parts of the MWE. On Table 1 a PP is developed out of the prepositional and nominal WWSs and on Table 2 an S1 is developed out of a COMPL(ementizer) and a verb head. Again, constraints on morphology, binding and control that were attached to the constituents of the phrases are carried over.

4. A3: GF labels: Nominal constituents in the nominative/accusative case are assigned the label SUBJ/OBJ respectively. S1s are assigned the COMP/XCOMP label depending on the existence of control constraints; PPs are assigned the OBLθ GF and AdvPs the PCOMP GF. Control and binding constraints are interpreted.

5. A4. Using the A3 GF labels and control constraints, if any (Table 2), the head verb predicate of a COMP/XCOMP is defined and stored in XLE.

6. A5. Using the A3 GF labels, the head verb predicate of the verb MWE is stored in XLE along with all the constraints carried over so far.

<table>
<tr><th>ID1</th><th>ID2</th><th>ID3</th><th>ID4</th><th>ID5</th><th>A0</th><th>A1</th><th>A2</th><th>A3</th><th>A5</th></tr>
<tr><td>NP-NOM-anim</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>SUBJ
SUBJ ANIM +</td><td></td></tr>
<tr><td>LEMMA</td><td>echo</td><td></td><td>Vb Head</td><td></td><td></td><td></td><td></td><td></td><td rowspan="10">PRED ‘echo(SUBJ, OBLpano_apo)’
SUBJ ANIM +
OBJ PRED ‘to_mialo(POSS);
POSS CAT PnWe
((*)Binding Constraints)

PRED ‘pano_apo (OBJ)’
OBJ PRED ‘to_kefali POSS)’
POSS CAT PnWe
((*)Binding Constraints)</td></tr>
<tr><td>LEMMA</td><td>o</td><td>to</td><td>AtDfNeSgAc</td><td>1</td><td></td><td rowspan="2">NWWS PRED ‘to_mialo(POSS);
Case Ac
POSS CAT PnWe
GBC NP-NOM-anim</td><td rowspan="2"></td><td rowspan="2">OBJ
OBJ PRED ‘to_mialo(POSS);
POSS CAT PnWe
((*)Binding Constraints)</td></tr>
<tr><td>LEMMA</td><td>mialo</td><td>mialo</td><td>NoCmNeSgAc</td><td>1</td><td>Cat No Gen Ne...</td></tr>
<tr><td><u>PnGe</u></td><td></td><td></td><td>PnGeWe
GBC NP-NOM-anim</td><td>1</td><td></td></tr>
<tr><td>LEMMA</td><td>pano</td><td>pano</td><td>AdXxBa</td><td>2</td><td></td><td rowspan="2">PWWS
PRED ‘pano_apo’</td><td rowspan="5">PP
PRED ‘pano_apo (OBJ)”
OBJ
PRED_’to_kefali POSS)’
Ptype pano_apo
POSS CAT PnWe</td><td rowspan="5">OBL pano_apo
PRED ‘pano_apo (OBJ)’
OBJ PRED ‘to_kefali POSS)’
POSS CAT PnWe
((*)Binding Constraints)</td></tr>
<tr><td>LEMMA</td><td>apo</td><td>apo</td><td>AsPpSp</td><td>2</td><td></td></tr>
<tr><td>LEMMA</td><td>o</td><td>to</td><td>AtDfNeSgAc</td><td>3</td><td></td><td rowspan="3">NWWS
PRED_’to_kefali POSS)’
POSS CAT PnWe
Case Ac
GBC NP-NOM-anim</td></tr>
<tr><td>LEMMA</td><td>kefali</td><td>kefali</td><td>NoCmNeSgAc</td><td>3</td><td>Cat No Gen Ne...</td></tr>
<tr><td><u>PnGe</u></td><td></td><td></td><td>PnGeWe
GBC NP-NOM-anim</td><td>3</td><td></td></tr>
</table>

Table 1. Generating LFG/XLE entries for (2) starting from the IDION encoding.
(*)Binding Constraints: OBJ POSS NUM= SUBJ NUM; OBJ POSS PERS = SUBJ PERS

Transforming IDION information to XLE entries was possible without affecting the original grammars. The conversion procedure can be applied to the future versions of IDION that will contain a larger population of MWEs thus indicating that the IDION's enrichment procedure can be completely dissociated from application development (here an LFG/XLE parser). On the other hand, the conversion process is used as a tool for checking the syntactic validity of the contents of IDION. The conversion procedure showed that highly idiomatic information, for instance, the exact type of WWSs (whether nominal, prepositional or other), would be better added manually to make sure that conversion will always work even if an unknown type of WWS occurs.

References

Grégoire, N. (2010). “DuELME: a Dutch electronic lexicon of multiword expressions”. In: Language Resources and Evaluation 44.1-2, pp. 23–39.

Markantonatou, S., E. Koletti, Elpiniki Margariti, P. Minos, E. Stripeli, G. Zakis, and N. Samaridi. (2015). *Lexical resource for free subject verb MWEs*. Presented in *Modern Greek MWEs 2015, a thematic workshop at the 12th International Conference on Greek Linguistics (ICGL 12)*, 16-18 September 2015, Freie University, Berlin.

Samaridi, N. and S. Markantonatou. (2014) Parsing Modern Greek verb MWEs with LFG/XLE grammars. *The 10th Workshop on Multiword Expressions (MWE 2014), Workshop at EACL 2014* (Gothenburg, Sweden), April 26-27, 2014

ILSP-PAROLE compatible Tagset:http://nlp.ilsp.gr/nlp/tagset_examples/tagset_en/index.html