P. Minos (TEI/Athens), S. Markantonatou (ILSP/RC "Athena"), G. Zakis (UoA), E. Margariti (UoA)

Generating LFG/XLE MWE entries from IDION (a theory neutral lexical DB)

WG1/WG2

THE RESOURCE

IDION http://idion.ilsp.gr/:
- a lexical DB
- encodes multiple aspects of information on MWEs: origins, semantics, morphosyntactic information, corpus examples, syntactic properties, diathesis alternation phenomena, lexical relations, usage examples
- addresses both the human user and the machine

THE ISSUE

We discuss the **transcription of IDION contents to a lexical resource for parsing**; as a case study we use an LFG/XLE lexicon that contains fixed sequences of fixed elements [4], known as Words With Spaces (WWS) [3], marked with underscores.

(1) Ανάβω όλα τα λαμπάκια σε κάποιον
anavo ola ta labakia se kapion
turn-on.1SG all.ACC the ligth-bulbs.ACC to smbd
'I make someone furious.'

Syntactic tests such as word order permutations and XP interpolation show that the MWE in (1) contains the WWS **όλα_τα_λαμπάκια** (all the light-bulbs).

MORPHOSYNTACTIC INFO IN IDION IS (almost) THEORY NEUTRAL

A theory neutral morphosyntactic encoding is **cost effective**:
-it is user-friendly for a larger number of encoders
-it reduces the number of errors
-it uses standard morphological encoding (PAROLE)
-it uses **syntactic notions of wide acceptance** (rather than notions specific to a particular theory)

The price to pay for the theory neutral encoding is that IDION cannot be used as a lexical resource directly by any syntactic formalism; some work of syntactic 'interpretation/translation' is left for an application that reads information off IDION and feeds the parser.

But this is the case for all encodings, whether bound to a theory or not.

Encoding in IDION
- provides a fully fledged morphological description
- provides a rather flat phrasal description of MWEs where only free constituents may have a phrasal status (2) while fixed constituents are listed as words
-avoids to assign syntactic functional information, for instance whether the fixed parts have some syntactic function.

(2) The IDION vocabulary of phrasal category symbols:
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

Component type	Lemma	Wordform	PAROLE tag	WWS index
NP-NOM				
LEMMA	ανάβω		Vb-Head	
WF	όλος	όλα	AjBaNePlAc	1
WF	ο	τα	AtDfNePlAc	1
WF	λαμπάκι	λαμπάκια	NoCmNePlAc	1
LEMMA	σε	σε	AsPpSp	
NP-ACC				

Table 1. Morphosyntactic encoding of (1) in IDION

REFERENCES

[1] Grégoire, Nicole. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. In: *Language Resources and Evaluation 44.1-2, pp. 23–39*
[2] Gross, Maurice. 1988a. Les limites de la phrase figée. *Langage 90, pp. 7-23*
[3] Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake & Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Gelbukh, Alexander (ed.), *Computational linguistics and intelligent text processing. Proceedings of the Third International Conference, CICLing [Conference on Intelligent Text Processing and Computational Linguistics] 2002, Mexico City, Mexico, February 17-23, 2002, pp. 1-15*. Berlin & Heidelberg: Springer-Verlag.
[4] Samaridi, Niki and Stella 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

MORAL

- Improvements, for instance fixed subject verb MWEs, seem to require only minimal additions/modifications.
- The development of the algorithm was not costly (it took about 8 days). Minimal modifications may suffice for other formalisms or variations of the LFG/XLE one.

IDION, although theory neutral, is useful for parsing purposes
the linguistic knowledge it encodes seems to be sufficient

As compared to other DBs with more elaborate syntactic information, such as DUELME [1], IDION's encoding schema has required much less development time while, probably, the procedure of MWE encoding does not present more difficulties.

Reusability of grammatical resources is a well-known issue, so there may be a point in minimising resource development time (design and encoding).

A WORKING EXAMPLE: The 4-step algorithm applies on the representation in Table 1

LFG employs abstract syntactic concepts and a very specific formalism → generating an LFG/XLE lexicon from IDION is a proof-of-concept for IDION's usability for parsing purposes.

Step 1. WWS detection
WWSs are identified with the WWS Index. The entries indexed with the same number are considered parts of a single token WWS. Morphological information is provided under PAROLE tag. The rules generate nominal (3), prepositional and verbal WWS. (R1) yields (3).

(R1) If WWS is of the form *Aj No/At Aj No/Aj At No/At No No/At No/No No* return the lexical entry: *wws* NWWS * (^ PRED)='*wws*' . Modify Table 1: collapse cells, use the first No tag .

(3) όλα_τα_λαμπάκια NWWS * (^ PRED) = 'όλα_τα_λαμπάκια', NoCmNePlAc

Step 2. PPs detection
(R2) If there is a LEMMA token marked as preposition followed by an NP-ACC/NP-GEN create a new phrasal constituent named PP which replaces the previous two and bears the constraint (^ PFORM)='preposition_type'. (R2) yields (4).

(4) σε NP, (^ PFORM)='σε'

Step 3. Select Vhead and check for verb arguments
V heads are marked so (PAROLE tag). Using LEMMA, the algorithm incrementally creates the verb lexical entry (5).

(5) ανάβω V * (^ PRED)=ανάβω<...>'

(R3) If there is a phrasal constituent marked as **NP-NOM/NP-NOM-anim/NP-NOM-nonanim,** assign the verb the **SUBJ GF** and the constraints **(^ SUBJ CASE) =c nom {(^SUBJ SEM ANIM)=+/-};**

(R4) If there is a nominal **WWS (NWWS) in ACC case** (case is marked in the 4th Column), assign the verb the **OBJ GF** and the constraint **(^ OBJ CASE) =c acc;**

(R5) If there is a phrasal constituent marked as **PP, (^ PFORM)='preposition_type'**, assign the verb the **OBL GF** and the constraint **(^OBL PFORM) =c 'preposition_type'.**

Rules (R3), (R4),(R5) yield (6).
(6) (ανάβω V * (^ PRED)=ανάβω<(^SUBJ)(^OBJ)(^OBL)>'
(^ SEM) = "make someone furious"
(^ SUBJ CASE) =c nom
(^ OBJ CASE) =c acc
(^ OBL PFORM) =c 'σε'

Step 4. Semantic information
Add semantics to the VbHead lexical entry: **(^SEM)="string_of_the_MWE_meaning"**

The XLE entries produced correspond to the idiomatic parts of the MWE while the non-idiomatic parts are received from the lexicon of the general language.

ανάβω V * (^ PRED)=ανάβω<(^SUBJ)(^OBJ)(^OBL)>'
(^ SEM) = "make someone furious"
(^ SUBJ CASE) =c nom
(^ OBJ CASE) =c acc
(^ OBL PFORM) =c 'σε'

όλα_τα_λαμπάκια NWWS * (^ PRED) = 'όλα_τα_λαμπάκια'