Representing MWEs: testing the standards

Gyri Smørdal Losnegaard, Carla Parra Escartín
University of Bergen/Parseme WG1

Representation of MWEs using TEI and LMF

In a preliminary study [2] we concluded that the existing standards TEI and LMF seem suitable for encoding MWEs for NLP purposes, but perhaps too flexible to work as *de facto* standards. We have seen how the two standards meet our encoding requirements by trying to represent two sets of MWEs: the Norwegian idiom *katta i sekken* (cat.DEF in bag.DEF) "pig in a poke", and the German compund noun *Warmwasserbereitungsanlagen* "warm water production systems" and its Spanish translational equivalent, the NP *sistemas de preparación de agua caliente*.

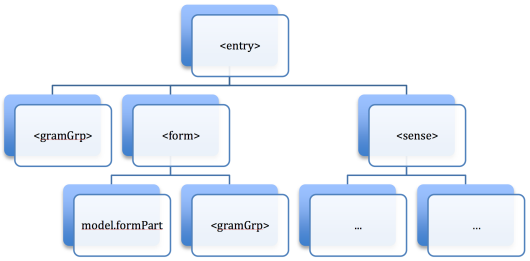
TEI

- Text Encoding Initiative (TEI).
- Well-documented.
- Specific module for dictionary encoding.
- Originally developed for Machine Readable Dictionaries.
- No available implementations of MWE databases for NLP use.

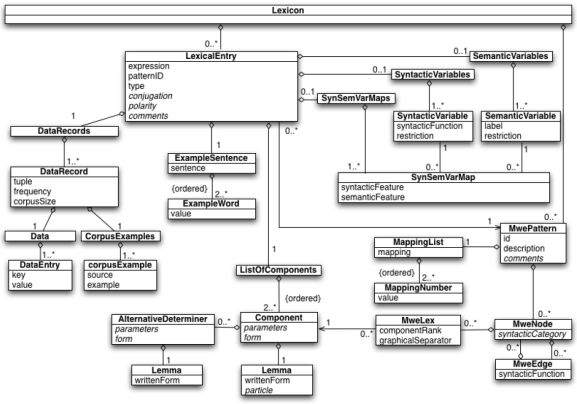
LMF

- Lexical markup framework (LMF).
- Combines the best designs and methods from NLP lexicons.
- Developed for NLP use, not for human users.
- Specific module for the representation of MWEs.
- An implementation of a MWE database exists [1].

The TEI dictionary class model



The DUELME-LMF class model



Encoding requirement checklist

MWE property	TEI	LMF
Encoding at MWE level (mandatory)		
PoS	(✓)	✓
PoS standard (tagset)	(✓)	(✓)
Meaning	(✓)	(✓)
Number of component words	✓	✓
Extended encoding at MWE level (recommended)		
Canonical form	(✓)	✓
Idiosyncracy levels (types of idiomaticity)	✗	✓
Translational correspondences	(✓)	(✓)
Encoding at word level (optional)		
PoS	✓	✓
Lemma	✓	✓
Grammatical features	✗	(✓)

(✓) = optional feature ✓ = not explicitly represented

Encoding with TEI: noteworthy

- Only *extensional* morphological description is possible (i.e., all word forms must be listed explicitly).
- Due to the lack of examples, the number of encoding decisions made this simple encoding task very challenging.
- Increasing interest within the TEI community towards the development of LMF-compatible lexicons [3].

Encoding with LMF: noteworthy

- No PoS attribute, but an *MWE pattern description* includes the syntactic type of the overall expression (NP, VP, etc.).
- Alternating lexical items (e.g. "**buy/get** a pig in a poke") can be represented as *lists*.
- The model is *extendable* with classes and attributes from the LMF core package.

Evaluation

- The DUELME-LMF model currently stands out as a best practice for the representation of MWEs for NLP use.
  - A comprehensive framework for the representation of MWEs in the European synthetic languages.
  - Should also be tested on agglutinative languages.
- Both TEI and LMF recommend the use of ISOcat data categories.
  - Apart from the DUELME implementation, there is no available data category selection (DCS) for the description of MWEs.
  - A need for the development of an ISOcat taxonomy for MWEs.

References

[1] Jan Odijk, *Duelme: Dutch electronic lexicon of multiword expressions*, LMF lexical markup framework (Francopoulo, Gil, ed.), Computer engineering and IT series, pp. 133–143, Wiley, 2013.

[2] Carla Parra Escartín, Gyri Smørdal Losnegaard, Gunn Inger Lyse Samdal, and Pedro Patiño García, *Representing Multiword Expressions in Lexical and Terminological Resources: An Analysis for Natural Language Processing Purposes*, Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference (Tallinn, Estonia) (I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tuulik, eds.), Trojina, Institute for Applied Slovene Studies (Ljubljana/Tallinn), October 2013, pp. 338–357.

[3] Laurent Romary, *TEI and LMF crosswalks*, Digital Humanities: Wissenschaft vom Verstehen (Stefan Gradmann and Felix Sasaki, eds.), Humboldt Universität zu Berlin, 2013.

Acknowledgements

The current research was financed by the EU under the Marie Curie Actions, FP7 People programme (grant agreement 238405).