Representing Multiword Expressions in Lexical and Terminological Resources: Testing the Standards

Gyri Smørdal Losnegaard
Carla Parra Escartín
E-mail: Gyri.Losnegaard@uib.no; Carla.Parra@uib.no

In our paper "Representing Multiword Expressions in Lexical and Terminological Resources: An Analysis for Natural Language Processing Purposes" [1], we addressed the formal representation of multiword expressions (MWEs) for Natural Language Processing (NLP) purposes. Particularly, we concentrated on requirements to be met by the standards used to represent MWEs when we aim at creating resources which are reusable, interoperable and integrated into other NLP applications and workflows.

As pointed out in that paper, despite the efforts carried out by different projects and initiatives to encourage the usage of standards in the creation of new lexica and termbases, such standards did not seem to have been properly assessed for the representation of all kinds of possible entries. In fact, MWEs seemed to have been somewhat disregarded, resulting in a lack of mechanisms for representing all the required information to correctly represent them and subsequently re-use such resources in NLP.

As a follow up of our paper, we are currently testing and evaluating the two standards that seemed best to represent MWEs: TEI and LMF. In order to assess their suitability for representing MWEs, we test whether or not these standards allow us to add all the information we identified as prerequisites in [1]:

1. Type level (mandatory)
    (a) Part of Speech
    (b) PoS standard
    (c) Meaning
    (d) The number of component words

2. Type level, extended description (optional)
    (a) Canonical (base) form
    (b) Level(s) of idiosyncracy
    (c) Translational correspondences
    (d) Language variety
3. Token level (optional)
    (a) Part of Speech (PoS)
    (b) Lemma
    (c) Grammatical features

This exercise will also allow us to assess the implementability of our own prerequisites list and the way we envisaged it: a scalable, modular system.

Since different types of MWE have different structures and different representation requirements, we have chosen two types of MWEs for this preliminary analysis. The test expressions have been selected from our own PhD projects: a study of translational correspondences between German nominal compounds and Spanish noun phrases, and a study of Norwegian MWEs and their integration in an LFG grammar.

We represent two test MWEs in both LMF and TEI and we then assess to what extent they meet our requirements. Finally, we compare both standards and conclude which we deem best for representing MWEs for our purposes.

References

[1] 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. In I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, and M. Tulik, editors, *Electronic lexicography in the 21st century: thinking outside the paper. Proceedings of the eLex 2013 conference*, pages 338–357, Tallinn, Estonia, 17-19 October 2013. Trojina, Institute for Applied Slovene Studies/Eesti Keele Instituut (Ljubljana/Tallinn).