# Reading materials
# PARSEME Hands-on Workshop
# on Lexical Encoding

**Required reading**

*Francopoulo and George (2013)*: a technical introduction to the LMF framework, the core lexicon model and extensions of the model (packages). These packages are: the *Morphology extension*, the *Machine-Readable Dictionary extension*, the *NLP syntax extension*, the *NLP semantic extension*, the *mulitilingual notation extension*, the *NLP morphological pattern extension*, and the *NLP multiword expression pattern extension*. Workshop attendants should read this paper in order to get an understanding of LMF and to familiarise themselves with the architecture behind it.

*Odijk (2013a)*: describes the conversion of an electronic lexicon for Dutch MWEs into LMF-compatible format (DUELME-LMF). The chapter introduces the DUELME model and compares it to the LMF core package and the NLP MWE pattern extension. Some deviations from the standard are argued necessary or recommended, and some additional recommendations are proposed to extend the LMF model.

*Annex M: NLP multiword expression patterns extension* and *Annex N: NLP multiword expression patterns examples*[1]. Appendices to the official ISO LMF document *("ISO 24613:2008 Language Resource Management – Lexical Markup Framework" 2008)[pp.73-76]*.

*Odijk (2013b)*: an account of the construction of the DUELME database of Dutch MWEs with emphasis on the MWE identification method and the lexical representation of the MWEs and their properties. The paper also gives an introduction

---

[1]*EncodingExample1-LMF-throw-somebody-to-the-lions.xml* is an XML representation of the MWE pattern example that is shown as a class diagram in Figure N.1 (p.76).

to the Equivalent Class Method (ECM) that forms the basis for the organisation of the MWEs in DUELME.

*Grégoire (2010)*: an account of the design and implementation of DUELME. This includes the representation choices and decisions made, and the main description categories. The paper briefly presents the *parametrised* Equivalent Class Method and further describes the identification, classification[2] and description of the data.

**Complementary reading (recommended)**

*Baldwin and Kim (2010); Sag et al. (2001)*: seminal paper on the treatment of MWEs in NLP.

*Grégoire (2007a)*: a more detailed account of the representation of fixedness categories in DUELME.

*Grégoire (2007b)*: a more detailed account of the parametrized equivalent class method (ECM).

*The DUELME documentation package: Pattern descriptions, Encoding protocol, XML data – User documentation.* (Available with license, accompanies the DUELME lexicon data files which will be provided later).

*"ISO 24613:2008 Language Resource Management – Lexical Markup Framework" 2008*: Official ISO LMF document.

# References

Baldwin, Timothy and Su Nam Kim (2010). "Multiword Expressions". In: *Handbook of Natural Language Processing, Second Edition*. Ed. by Nitin Indurkhya and Fred J. Damerau. ISBN 978-1420085921. Boca Raton, FL: CRC Press, Taylor and Francis Group.

Francopoulo, Gil and Monte George (2013). "Model Description". In: *LMF Lexical Markup Framework*. Ed. by Gil Francopoulo. Wiley Online Library. Chap. 2, pp. 19–40.

---

[2]MWEs in DUELME are represented according to the fixedness categories fixed, semi-fixed and syntactically flexible *(Baldwin and Kim, 2010; Sag et al., 2001)*.

Grégoire, Nicole (2007a). *Report on the lexical representation of subclasses of MWEs*.

— (2007b). *STEVIN-IRME WP2, deliverable 2.1 Report on formalizing and elaborating the parameterized Equivalence Class Method for Dutch*. URL: `http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.206.479`.

— (2010). "DuELME: a Dutch electronic lexicon of multiword expressions". In: *Language Resources and Evaluation* 44.1-2, pp. 23–39.

"ISO 24613:2008 Language Resource Management – Lexical Markup Framework" (2008). In: ISO Geneva.

Odijk, Jan (2013a). "DUELME: Dutch Electronic Lexicon of Multiword Expressions". In: *LMF Lexical Markup Framework*. Ed. by Gil Francopoulo. Computer engineering and IT series. Wiley Online Library. Chap. 9, pp. 133–143.

— (2013b). "Identification and Lexical Representation of Multiword Expressions". In: *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*. Ed. by P. Spyns and J.E.J.M Odijk. Theory and Applications of Natural Language Processing. Berlin/Heidelberg: Springer, pp. 201–217. URL: `http://link.springer.com/content/pdf/10.1007\%2F978-3-642-30910-6_12`.

Sag, Ivan A. et al. (2001). "Multiword Expressions: A Pain in the Neck for NLP". In: *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pp. 1–15.