An LMF Model for a French-Romanian Collocation Dictionary

WG1 : Lexicon-Grammar Interface

Amalia Todirascu
LiLPa University of Strasbourg
22, rue René Descartes, BP 80010
67084 Strasbourg cedex
todiras@unistra.fr

We present a LMF model for a multilingual collocation dictionary, available for French and for Romanian. The dictionary contains 250 bilingual Verb+Noun collocations. The dictionary contains rich morpho-syntactic information and subcategorisation information. We use the various LMF mechanisms to model our collocation dictionary.

Multilingual collocation dictionaries are valuable resources for NLP systems, for machine translation or for second language learning. These dictionaries contain several categories of linguistic information such as the collocation definitions, the syntactic and semantic properties of the collocations or their subcategorisation patterns, useful for parsing or for human translators. Even if several collocation extraction systems (Constant et al, 2013), (Heid and Ritz, 2005), (Nissim, Zaninello, 2013) help building multilingual MWE dictionaries, this task is still time-consuming : the candidates extracted from corpora should be manually validated. For this reason, many projects aiming to design MWE dictionaries adopt standards for lexical resources such as Text Encoding Initiative (TEI) or Lexical Markup Framework (LMF) (Francopoulo, 2013) to improve data accessibility, to reuse existing lexical resources. Some authors discuss the differences between the two standards according to Multi Word Expressions (MWE) (Romary, 2013), (Parra Escartin et al, 2013). LMF is the standard adopted by several monolingual MWE dictionaries. (Laporte et al, 2013) transforms the Lexicon-Grammar tables into LMF format, using subcategorisation and semantic mechanisms provided by the standard. DuELME, available for Dutch (Odijk, 2013) is converted into LMF format. The project adapts the specific MWE package to represent classes of MWE determined by their morphosyntactic properties.

In our project, we adopt the LMF standard (ISO/TC 37/SC 4). LMF provides a complete framework to model monolingual and multilingual NLP lexicons. LMF is based on UML, proposing several packages to model the dictionary structure and the detailed lexical information. Indeed, the core package of the standard describes the lexicon as a list of lexical entries. Simple lexical entries or multiword lexical entries are described as *LexicalEntries* or *ListOfComponents*. LMF uses additional packages for morphology, for representing MWE contextual information. Additionally, the Syntactic package represents specific subcategorisation information for each lexical entry. Semantic package provides classes to represent the sense of the lexical entries, semantic relations such as synonymy or hyperonymy, examples and semantic predicates and arguments. Multilingual (translation) information is represented via a pivot (described in the multilingual package such as *Sense Axis*).

In our bilingual dictionary (Todirascu et al, 2008) we consider that collocations are composed of two or several lexical units, with specific syntactic and semantic behavior. Their sense is often non-compositional and they present an important degree of syntactic variability (Gledhill, 2009, Hausmann, 2004).

The Verb+Noun collocations are characterized by specific contextual properties such as preference for a specific category of determiner (the determiner is always zero in *faire face*'to face' but always definite *in faire l'objet'*is subject of'), for some prepositions (*take into account*) or for a specific voice. Each entry represents the collocation (the lemmas of the verb and of the noun) with its global properties (subcategorisation patterns, preference for prepositions), their senses and the samples extracted from corpora (with their frequency). In each lexical entry, we explicitly represent noun's properties: preference for definite or zero determiner, preference for singular or for plural number (both in Romanian and in French), preference for accusative or dative case of the object (only for Romanian). The properties of the verbs represented in the dictionary are the preference for some specific arguments (marked by a given preposition), the preference for a specific list of words and the preference for active or passive voice.

Our LMF model adapts (Odijk, 2013) model for DuELME. MWE representation uses the LMF's class *ListOfComponents. MWEPattern, MWELex* and *MWENode* model the relations between the components of the collocations. The strong morphological preferences of nouns and of verbs are represented by using *MorphologicalPattern* class. The *Semantic* package represents collocation definition and examples. Syntactic properties of collocations are represented by some classes from the Syntactic package (*Syntactic Behavior, Subcategorisation Frame* and *Syntactic Argument)*. *Sense Axis* (from the multilingual package) is useful to represent translation equivalent matching between the monolingual parts of the dictionary. Finally, we discuss the differences between French and Romanian data.

References

Constant, M., Sigogne A., Watrin, P. (2013) Stratégies discriminantes pour intégrer la reconnaissance des mots composés dans un analyseur syntaxique en constituants. *Traitement Automatique des Langues*. Vol. 54(1). 24 pp.

Francopoulo, G. (2013) *LMF: Lexical Markup Framework*, ISTE / WILEY, ISBN: 978-1-84821-430-9

Gledhill, C. (2009) « Vers une analyse systémique des locutions verbales, constructions verbo-nominales et autres prédicats complexes », in D. Banks, (ed.) *La Linguistique systémique fonctionnelle et la langue française*, Paris, L'Harmattan.

Heid, U. et Ritz J. (2005). Extracting collocations and their contexts from corpora, *Actes de Conference on Computational Lexicography and Text Research*, Budapest

Laporte, E., Tolone, E., Constant, M., 2013, Conversion of Lexicon-Grammar Tables to LMF: Application to French. In Francopoulo, G. (ed.) *LMF: Lexical Markup Framework*, ISTE / WILEY, 2013

Parra Escartín, C., Losnegaard, G. S., Samdal, G.I.L, Patiño García, P, 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.

ISO/TC 37/SC 4, Language resource management - Lexical markup framework (LMF), http ://lirics.loria.fr/documents.html, 2007

Nissim,M, Zaninello, A. (2013) A Repository of Variation Patterns for Multiword Expressions, Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 101–105, Atlanta, Georgia, 13-14 June 2013.

Odijk, J., 2013, DUELME: Dutch Electronic Lexicon of Multiword Expressions, In Francopoulo, G. (ed.) *LMF: Lexical Markup Framework*, ISTE / WILEY, 2013

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

Todirascu, A., Heid, U., Stefanescu, D., Tufis, D., Gledhill, C., Weller M., Rousselot F. 2008. « Vers un dictionnaire de collocations multilingue » Cahiers de Linguistique, Université de Louvain