Lexical encoding of multiword expressions in XMG

Agata Savary, Simon Petitjean, Timm Lichte, Laura Kallmeyer, Jakub Waszczuk

Lexical encoding of multiword expressions in XMG

Agata Savary[1] Simon Petitjean[2] Timm Lichte[3]
Laura Kallmeyer[2] Jakub Waszczuk[2]
(1) University of Tours, France
(2) Heinrich Heine Universität Düsseldorf, Germany
(3) University of Tübingen, Germany
first.last@univ-tours.fr
last@phil.uni-duesseldorf.de
first.last@uni-tuebingen.de

ABSTRACT

Multiword expressions (MWEs) exhibit both regular and idiosyncratic properties. Their idiosyncrasy requires lexical encoding in parallel with their component words. Their (at times intricate) regularity, on the other hand, calls for means of flexible factorization to avoid redundant descriptions of shared properties. However, so far, non-redundant general-purpose lexical encoding of MWEs has not received a satisfactory solution. We offer a proof of concept that this challenge might be effectively addressed within eXtensible MetaGrammar (XMG), an object-oriented metagrammar framework. We first make an existing metagrammatical resource, the FrenchTAG grammar, MWE-aware. We then evaluate the factorization gain during incremental implementation with XMG on a dataset extracted from an MWE-annotated reference corpus. This paper is part of a larger publication to appear (Savary et al., 2020).

RÉSUMÉ

Codage lexical d'expressions polylexicales en XMG

Les Expressions polylexicales (EP) possèdent des propriétés à la fois régulières et idiosyncratiques. Leur idiosyncrasie requiert un codage lexical au même titre que celui des mots qui les composent. D'autre part, leur régularité (parfois complexe) nécessite des moyens de factorisation afin d'éviter des descriptions redondantes des propriétés partagées. À ce jour, il n'existe pas de solution idéale pour le codage lexical généraliste et non redondant des EP. Dans cet article nous présentons une preuve de concept que ce défi pourrait être relevé dans le cadre de XMG (eXtensible MetaGrammar), qui est un formalisme métagrammatical orienté-objet. Nous montrons comment une ressource métagrammaticale existante, FrenchTAG, peut être étendue pour couvrir les EP. Nous évaluons le gain en terme de factorisation de cette ressource lors de son développement incrémental. Cette expérience est menée sur un jeu de données extrait d'un corpus de référence annoté en EP. Cet article est extrait d'un publication plus large à venir (Savary et al., 2020).

MOTS-CLÉS : expressions polylexicales, métagrammaire, XMG.

KEYWORDS: multiword expressions, metagrammar, XMG.

Multiword expressions (MWEs) are combinations of words which encompass heterogeneous linguistic objects such as idioms (IDs : *to* ***pull*** *one's* ***leg***), compounds (*a* ***hot dog***), light verb constructions (LVCs : *to* ***pay*** *a* ***visit***), inherently reflexive verbs (IRVs : ***s'apercevoir*** 'perceive oneself'⇒'realize' in French), rhetorical figures (***as busy as a bee***), or named entities (*the* ***Sea of Tranquility***). Their

most pervasive and challenging feature is their non-compositional semantics, i.e. the fact that their meaning cannot be deduced from the literal meanings of their components, and from their syntactic structures, in a way deemed regular for the given language. For this reason, as well as because of their pervasiveness in texts, MWEs constitute a major challenge in semantically oriented NLP applications.

But MWEs also exhibit unexpected behavior on other levels of linguistic analysis including the lexical, morphological and syntactic ones. These properties can be *defective* or *restrictive* (Lichte et al., 2019). A defective property excludes a literal interpretation of the MWE, e.g. (EN) *a* ***lesser yellowlegs*** 'a shore bird species' cannot be understood literally because of the lack of number agreement between the determiner and the head noun. A restrictive property reduces the number of possible surface realizations of the MWE with respect to the literal reading. For instance in example (3), the possessive determiner has to agree with the subject, otherwise the expression can only be understood literally as in *#John crossed her fingers.* [1] Since defective and restrictive properties help distinguish literal from idiomatic readings of MWEs, their description and processing are important both for linguistic modeling and for NLP applications, including MWE identification (Constant et al., 2017).

When characterizing MWEs, some authors (Grégoire, 2010; Przepiórkowski et al., 2014) oppose the *regular* behavior of "free" phrases (i.e. those obeying the rules of a "regular" grammar), like (1), to the *idiosyncratic* behavior of MWEs, like (2)–(4).

(1) *John broke my mug*

(2) *John* ***broke*** *his/our* ***fall*** 'John made his/our fall less forceful'

(3) *John* ***crossed*** *his* ***fingers*** 'John hoped for good luck'

(4) *John* ***held*** *his* ***tongue*** 'John refrained from expressing his view'

Some others point out that regularity is a matter of scale rather than a binary phenomenon (Gross, 1988; Herzig Sheinfux et al., 2015). We take the latter stand, and extend it by assuming that the degree of regularity is a feature of linguistic properties on the one hand, and of MWEs on the other hand (Lichte et al., 2019). Firstly, the more (resp. less) objects share a certain property, the more it is regular (resp. idiosyncratic). For instance, allowing a possessive determiner in a Verb-Det-Noun construction is more regular than imposing that it agrees with the subject, because the former applies to (1)–(4), while the latter is limited to (3)–(4). Still the latter is not fully irregular since it is shared by many expressions. Secondly, in (3), while the direct object of the verb *to cross* is lexicalized (has to be realized by the lexeme *finger*), the subject is not. While the noun does not admit adjectival modifiers (*#He crossed his long fingers.*), passivization is allowed (***fingers crossed***). While the noun has to occur in plural, the verb can be inflected freely, etc. Thus, this MWE combines more regular properties (e.g. a free subject) with more idiosyncratic ones (e.g. a lexically and morphologically fixed object). Also, the MWE in (4) has the same properties (with the number of the noun fixed to singular instead of plural) except that passivization is not allowed (*#His tongue was held*). Therefore, the degree of regularity of the MWE in (3) can be considered higher than of the one in (4).

Because MWEs exhibit (more or less) idiosyncratic properties, their modeling has to include lexical encoding, i.e. MWEs should become separate lexical entries, additionally to their single-word components. The main challenge is then to account for the irregularity of a MWE, while avoiding redundancy, i.e. repeated description of common properties. For instance, the subject-possessive agreement is shared by (3)–(4) and many other MWEs, so its formalization should preferably be

1. The hash symbol # signals the loss of the idiomatic reading. Lexicalized components of a MWE, i.e. those always realized by the same lexemes, are marked in boldface.

done only once, rather than repeatedly for each MWE lexicon entry. Our state-of-the-art studies have shown that no previous work seems to have addressed this challenge in a satisfactory way.

In this work, we aim at providing a proof of concept that non-redundant lexical encoding of MWEs can be effectively achieved in an object-oriented metagrammar-based approach. We use XMG (Crabbé et al., 2013; Petitjean et al., 2016), a declarative constraint-based description language in which more or less regular tree structures are modeled via a hierarchy of classes. Higher (more abstract) classes encode more elementary and less constrained structures. Lower (more specific) classes combine higher ones and impose new constraints on these combinations. Both single-word lexemes and MWEs are then expressed as lexical entries assigned to particular low-level classes (usually leaves) of this class hierarchy. The description is independent of a particular grammatical framework but XMG comes with metagrammar compilers into several formalisms including Tree Adjoining Grammar (TAG). We therefore test our proposal on FrenchTAG (Crabbé, 2005), a pre-existing XMG resource which implements a large fragment of a reference grammar of French (Abeillé, 2002). We show how FrenchTAG can be adapted and extended so as to accommodate a small subset of verbal MWEs (VMWEs) of different syntactic structures and of varying degrees of syntactic flexibility. We evaluated the proposal on a dataset based on the PARSEME corpus of VMWEs (Savary et al., 2018). The experiment showed that adding MWE descriptions to a general grammar can be done elegantly by introducing interface constraints in pre-existing classes (to account for restrictive properties), and by adding some new classes (to account for defective properties and for various syntactic structures of lexicalized verbal arguments).

This abstract is part of a larger publication to appear (Savary et al., 2020).

Références

Abeillé, A. (2002). *Une grammaire électronique du français*. CNRS Editions.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword Expression Processing : A Survey. *Computational Linguistics*, 43(4) :837–892.

Crabbé, B., Duchier, D., Gardent, C., Roux, J. L., and Parmentier, Y. (2013). XMG : extensible metagrammar. *Computational Linguistics*, 39(3) :591–629.

Crabbé, B. (2005). *Représentation informatique de grammaires d'arbres fortement lexicalisées : le cas de la grammaire d'arbres adjoints*. PhD thesis, Université Nancy 2.

Grégoire, N. (2010). DuELME : a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2) :23–39.

Gross, G. (1988). Degré de figement des noms composés. *Langages*, 90 :57–71. Paris : Larousse.

Herzig Sheinfux, L., Arad Greshler, T., Melnik, N., and Wintner, S. (2015). Hebrew verbal multiword expressions. In Müller, S., editor, *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, pages 122–135, Stanford, CA. CSLI Publications.

Lichte, T., Petitjean, S., Savary, A., and Waszczuk, J. (2019). Lexical encoding formats for multiword expressions : The challenge of "irregular" regularities. In Parmentier, Y. and Waszczuk, J., editors, *Representation and parsing of multiword expressions : Current trends*, pages 1–33. Language Science Press, Berlin.

Petitjean, S., Duchier, D., and Parmentier, Y. (2016). XMG 2 : Describing description languages. In Amblard, M., de Groote, P., Pogodalla, S., and Retoré, C., editors, *Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996-2016) - 9th International Conference, LACL 2016, Nancy, France, December 5-7, 2016, Proceedings*, volume 10054 of *Lecture Notes in Computer Science*, pages 255–272.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., and Woliński, M. (2014). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Savary, A., Candito, M., Mititelu, V. B., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S. R., Eryiğit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C. P., van der Plas, L., QasemiZadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., and Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V., editors, *Multiword expressions at length and in depth : Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press., Berlin.

Savary, A., Petitjean, S., Lichte, T., Kallmeyer, L., and Waszczuk, J. (2020+). Object-oriented lexical encoding of multiword expressions : Short and sweet. *Lexique*, forthcoming.