Modeling syntactic properties of MWEs in LFG

Jakub Waszczuk, Agata Savary

Modeling syntactic properties of MWEs in LFG [WG2]

Jakub Waszczuk, Agata Savary
Université François-Rabelais Tours, 3 place Jean-Jaurès, 41000 Blois, France
{jakub.waszczuk@etu.,agata.savary@}univ-tours.fr

1 Introduction

This paper describes preliminary investigations on how to model syntactic properties of different types of MWEs within the framework of LFG. While there are already several works which address this topic (Attia, 2006; Asudeh et al., 2013; Patejuk, 2014), we are particularly interested in answering the question of which types of MWEs can be described at the level of the lexicon, and which (if any) require corresponding descriptions at the level of phrase-structure rules.

2 Structurally regular MWEs

Structurally regular MWEs can be successfully handled by conventional grammar rules but, due to their idiomatic meaning, should be represented as elementary language units. In contrast to 'free' structures, such expressions often adopt additional lexical and syntactic requirements, which in LFG can be conveniently represented with f-structures assigned to corresponding MWE descriptions.

Agreement: An MWE can introduce additional agreement constraints, as in the *NP vider DET sac* ‘to express NP's secret thoughts‘, lit. ‘NP empty DET bag‘ syntactically flexible French expression, where the possessive determiner DET embedded in the direct object of the verb *vider* ‘empty‘ must agree in person and number with the subject NP (Abeillé and Schabes, 1989). Otherwise, the idiomatic meaning is lost, e.g. *ils ont vidé son sac* should be only interpreted as semantically compositional ‘they have emptied his bag‘. This expression can be represented by the following idiomatic description assigned to the verb *vider*:

$$\begin{array}{ll} (\uparrow \text{ OBJ PRED FN}) =_c \text{'sac'} & \\ (\uparrow \text{ OBJ SPEC POSS}) = f & \quad (1) \\ (f \text{ PERS}) = g & (f \text{ NUM}) = h \\ (\uparrow \text{ SUBJ PERS}) = g & (\uparrow \text{ SUBJ NUM}) = h \end{array}$$

Modifiers: Modification requirements can be described at the level of f-structures as well. (Patejuk, 2014) shows how to account for four types of such requirements in Polish, two of which we consider in the description below.

Syntactically flexible expression *to spill the beans* ‘to reveal a secret‘ doesn't impose any lexical or syntactic constraints on the kind of modifiers it can take. This effect can be achieved by not including any additional constraints (apart from the lexical ones and the requirement that *beans* occur in plural, $(\uparrow \text{ OBJ NUM}) =_c \text{pl}$) in the functional description corresponding to the verb *spill*.

The French sentential idiom *NP casser sa pipe* ‘to die, lit. NP break his/her pipe‘ doesn't accept any internal modifications, just as its English analog *to kick the bucket* (both expressions are fixed and exhibit non-decomposable semantics). This can be specified by requiring that the f-structure corresponding to the verb's object is not paired with any adjunct, $\neg(\uparrow \text{ OBJ ADJUNCT})$.

Passivization: We look at passivization as a prominent example of a syntactic transformation which may be blocked by MWEs, e.g. the semi-fixed expression *to kick the bucket* ‘to die‘. The *bucket kicked by him* nominal phrase doesn't retain the idiomatic meaning. This is typically the case with opaque expressions with a completely non-decomposable meaning.

LFG posits separate active and passive verb lexical entries, therefore the grammar developer needs to make sure that the passive entry will not be generated for the idiomatic meaning of the verb *kick*. This is, however, a matter related to grammar development and not to the underlying formalism.

3 Structurally idiosyncratic MWEs

While structurally regular MWEs can be typically defined with functional descriptions at the level of the lexicon, structurally irregular ones require corresponding descriptions at the level of phrase-structure rules as well.

Verb-particle constructions: Verb-particle constructions constitute a well known example of expressions which should not be handled by productive grammar rules. Only specific pairs, e.g. *look up* ‘search for a reference‘ but not *look at* (in which case *at* is unambiguously a preposition), can be interpreted as MWEs. In the English ParGram grammar this issue is handled by an already existing mechanism dedicated to subcategorization frames. The standard, transitive VP rule accepts a particle either before or after the direct object of the verb

(VP → V PART OBJ | V OBJ PART, in simplified terms), while unification over the PRT-FORM feature, defined in both particle and verb lexical entries (e.g. (↑ PRT-FORM) = 'up' for the particle *up*) guarantees that only appropriate verb-particle constructions are recognized.

A certain inelegance can be noticed in the solution given above. First of all, it is broken into two different parts and the grammar does not provide any explicit evidence that the two parts are in fact strongly related and describe one and the same linguistic phenomenon. Secondly, on the level of c-structure rules, verb-particles are treated as productive constructions, which is counter-intuitive and does not reflect aptly their linguistic properties. Finally, it shows that in LFG there is no standard way of referring to orthographic forms or lemmas of individual MWE elements (note that PRED is sometimes used for the same purpose).

Correlative conjunctions: A similar solution is used in LFG to model correlative conjunctions such as *either _ or _* and *both _ and _*. They are handled by a coordination rule with an optional 'preconjunction'. Every preconjunction (*both*, *either*, ...) specifies (as a value of the COORD-FORM feature) with which conjunction it combines and unification guarantees that only corresponding pairs are recognized.

Phrasal configuration: In the two aforementioned cases there is no need to introduce new phrase-structure rules to handle the corresponding types of MWEs, but this is because appropriate machinery (for handling subcategorization frames and coordinations, respectively) already exists.

(Asudeh et al., 2013) provide an example of a Swedish traversal construction, e.g. *Sarah armbågade sig genom mängden* 'lit. S. elbowed SELF through crowd.DEF', which is distinguished by the requirement for the presence of a verb, a weak reflexive (coindexed with the subject), and a directional PP. It also exhibits a certain word-order peculiarity, which can be seen in expressions containing verb-particles: the particle follows the direct object of the verb, while normally it would adjoin to the verb. Thus, the authors claim, the syntactic structure of the expression can be most elegantly modeled by a dedicated c-structure rule.

In non-configurational languages MWEs may impose idiosyncratic preferences on the order of their constituents. In the Polish expression *doręczyć NP do rąk własnych* 'deliver NP as hand delivery, lit. deliver NP to hands own' (Patejuk, 2014), the *rąk własnych* configuration is much preferred over the alternative, *własnych rąk*. Moreover, the second variant can be assigned an idiosyncratic interpretation only if it is extended with a possessive specifier, e.g. ***jego** własnych rąk* *'**his** own hands'*. Otherwise, *własnych 'own'* can be only coindexed with the subject, which contradicts the internal semantics of this MWE. Such conditional word-order freezing can be modeled with the help of an *f-precedence* operator, which makes it possible to constraint the order of individual constituents (*rąk* and *własnych*, in this case) of the underlying expression (Mahowald, 2011).

In another Polish expression, *od przybytku* NP_0 *głowa nie boli* 'possessing a bigger quantity of NP_0 is never a problem, lit. from increase of NP_0 head not hurts', every permutation of PP (*od przybytku* NP_0), NP (*głowa*) and VP (*nie boli*) is possible, but the canonical PP NP VP is by far the most common one. To the best of our knowledge, LFG does not provide any means to define or, more importantly, make use of such word-order preferences. Optimality Theory (OT) allows to define preferences among concurrent configurations (Mahowald, 2011), but in OT only the optimal analyses are preserved while all the non-optimal ones are rejected during the parsing process. As a result, the mechanism is not appropriate for modeling soft, quantitative preferences.

Ungrammatical constructions: An idiomatic expression *all of a sudden* 'suddenly' has a highly irregular structure which is not recognized by the English ParGram grammar rules. This expression could be modeled with a dedicated c-structure rule. However, it seems that 'ungrammatical' constructions of this kind are typically fixed, which makes them easy to handle as words-with-spaces at the preprocessing stage of the parsing process.

An expression *nie wszystko złoto co się świeci* 'what seems ideal is not necessarily so, lit. not everything gold which shines' lacks the verb predicate which is required in regular Polish sentences. The expression could be handled as word-with-spaces but, when the numerous possible expressions based on this phrase (*nie wszystko opał co się pali* 'lit. not everything combustible which burns', to give an example) are considered, it becomes obvious that a lexicalized grammar rule (*S → nie wszytko NP co się VP*, in simplified terms) is indispensable to elegantly model this phrasal template.

References

Anne Abeillé and Yves Schabes. 1989. Parsing Idioms in Lexicalized TAGs. In *Proceedings of the Fourth Conference on European Chapter of the Association for Computational Linguistics*, EACL '89, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ash Asudeh, Mary Dalrymple, and Ida Toivonen. 2013. Constructions with lexical integrity. *Journal of Language Modelling*, 1(1):1–54.

Mohammed A. Attia. 2006. Accommodating Multiword Expressions in an Arabic LFG Grammar. In *Proceedings of the 5th International Conference on Advances in Natural Language Processing*, FinTAL'06, pages 87–98, Berlin, Heidelberg. Springer-Verlag.

Kyle Mahowald. 2011. An LFG Account of Word Order Freezing. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG11 Conference, Hong Kong*, pages 381–400, Stanford, Ca. CSLI Publications.

Agnieszka Patejuk. 2014. Handling MWEs in walenty, a new valence dictionary for Polish. Poster presented at the 2nd PARSEME general meeting, 10–11 March, Athens, Greece.