Object-oriented lexical encoding of multiword expressions: Short and sweet

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

Received: April 2020 / Accepted: July 2020
Published on line: December 2020

Object-oriented lexical encoding of multiword expressions: Short and sweet

Agata Savary
University de Tours, France
agata.savary@univ-tours.fr

Simon Petitjean
Heinrich Heine Universität Düsseldorf, Germany
petitjean@phil.uni-duesseldorf.de

Timm Lichte
University of Tübingen, Germany
timm.lichte@uni-tuebingen.de

Laura Kallmeyer
Heinrich Heine Universität Düsseldorf, Germany
laura.kallmeyer@phil.uni-duesseldorf.de

Jakub Waszczuk
Heinrich Heine Universität Düsseldorf, Germany
jakub.waszczuk@phil.uni-duesseldorf.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.

Keywords: multiword expressions, metagrammar, LTAG, XMG, FrenchTAG

Résumé

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 termes 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.

Mots-clefs : expressions polylexicales, méta-grammaire, LTAG, XMG, FrenchTAG

1. Introduction

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, Petitjean, Savary & Waszczuk, 2019). A defective property excludes a literal interpretation of the MWE, e.g. 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 #[1] (Constant, Eryiğit, Monti, van der Plas, Ramisch … Todirascu 2017).

[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.

When characterizing MWEs, some authors (Grégoire, 2010; Przepiórkowski, Hajnicz, Patejuk & Woliński, 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, Arad Greshler, Melnik and Wintner, 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 done only once, rather than repeatedly for each MWE lexicon entry. As shown in Section 2, no previous work seems to have addressed this challenge in a satisfactory way.

In this paper, 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, Duchier, Gardent, Le Roux & Parmentier, 2013; Petitjean, Duchier & Parmentier, 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 evaluate the proposal on a dataset based on the PARSEME corpus of VMWEs (Savary, Candito, Barbu Mititelu, Bejček, Cap … Vincze, 2018). The experiment shows 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).

The paper is organized as follows. We discuss the state of the art in lexical encoding of MWEs in computational lexicons and grammars (Sect. 2). We introduce our formalisms and tools (Sect. 3 and 4). We describe the original FrenchTAG metagrammar (Sect. 5) and we explain the methodology of making it MWE-aware (Sect. 6). We discuss the evaluation protocol and results (Sect. 7). Finally, we conclude and give directions for future work (Sect. 8).

2. Related work

Lexical encoding of MWEs has a long linguistic tradition, notably in French with Gross (1986) and Mel'čuk, Arbatchewsky-Jumarie, Dagenais, Elnitsky, Iordanskaja … Mantha (1988). They assume that units of meaning are located at the level of elementary sentences (i.e. predicates with their arguments) rather than of words, and MWEs, especially verbal, are special instances of predicates in which some arguments are lexicalized (i.e. MWEs can be syntactically described in the same framework as single words). Those works paved the way towards systematic syntactic description of MWEs, but were not directly applicable to NLP due to insufficient formalization (Constant and Tolone, 2010).

With the growing understanding of the challenges which MWEs pose to NLP, a large number of NLP-dedicated MWE lexicons have been created for many languages (Losnegaard, Sangati, Parra Escartín, Savary, Bargmann & Monti, 2016), some of which account for selected morpho-syntactic properties. They can be classified along two axes: (i) formalization of the lexicon-grammar interaction, (ii) existence of factorization mechanisms.

(i) *Formalization of the lexicon-grammar interaction*. This axis introduces a division between works which account for *continuous* MWEs only (i.e. those whose components are adjacent in text) and possibly *discontinuous* ones.

In case of continuous MWEs, the properties to describe mostly remain local (Savary, 2008), and there is no need to account for the grammar of a language in a comprehensive way. Notably, one needs to point at the components impacted by inflection or lemmatization of the whole MWE. For instance, when inflecting the English MWE ***attorney general*** for number, the first or the second component may vary (***attorneys general***, ***attorney generals***). When lemmatizing a French MWE ***mémoires vives*** 'live.FEM.PL memories.FEM.PL' ⇒ 'random access memories', both components need to be converted into singular but not into masculine ***mémoire vive*** 'live.FEM.SING memory.FEM.SING' ⇒ 'random access memory'). To express such constraints, finite-state-related formalisms, possibly enriched with unification, are often used (Karttunen, Kaplan & Zaenen, 1992; Breidt, Segond & Valetto, 1996; Oflazer, Çetinoglu & Say, 2004; Silberztein, 2005; Savary, 2009), and the associated tools sometimes allow to generate all possible surface realizations of a MWE. These extensional descriptions can then be straightforwardly matched as chunks against corpus occurrences.

Conversely, the description of discontinuous MWEs, most prominently of VMWEs, usually calls for more or less explicit reference to a full-fledged grammar, because of interactions between MWEs and external elements. For instance, the MWE in (2) has a compulsory but non-lexicalized modifier of the noun *fall*, which can be realized by syntactically complex nominal phrases (*John* ***broke*** *his secretly adored office mate's* ***fall***). Such long-distance dependencies have been covered with two objectives in mind: theory-independence and integration with computational grammars.

Firstly, it was postulated that MWE encoding, which is a labor-intensive task, should be done within a framework which is as neutral as possible with respect to particular syntactic theories. One assumes the existence of general grammar rules (of the language under study), whose observance a native lexicographer is able to verify. The description of a MWE is then done in such a way that only its idiosyncratic properties (i.e. those not conforming to the regular grammar) are encoded, while the regular ones are assumed implicitly. In Grégoire (2010), Przepiórkowski et al. (2014) and McShane, Nirenburg and Beale (2015), the idiosyncratic properties are represented directly at the level of morpho-syntax. Although these lexicons suffer from insufficient formalization (Lichte et al., 2019), they could be successfully applied to parsing after *ad hoc* conversion to particular grammar formalisms (Patejuk, 2015). By contrast to these three approaches, (Pausé, 2017), reminiscent of (Nunberg, Sag & Wasow, 1994), considers that apparent irregularities of morpho-syntactic behavior in MWEs are explicable by regular (rather than idiosyncratic) syntactico-semantic rules. What is idiosyncratic, however, and needs to be encoded in a lexicon, is the mapping between the MWE components and nodes of the semantic network inherent to the lexicon.

Secondly, a range of computational grammars accommodate some types of MWEs directly in their lexicons, focusing on the representation of the non-compositional or partly compositional meaning of MWEs. In Head-driven Phrase Structure Grammar (HPSG), Sag, Baldwin, Bond, Copestake and Flickinger (2002), Copestake, Lambeau, Villavicencio, Bond and Baldwin (2002) and Villavicencio, Copestake, Waldron and Lambeau (2004) represent decomposable English MWEs (*to* ***spill*** *the* ***beans***)

by paraphrasing (*spill* ⇒ 'reveal', *the beans* ⇒ 'a secret') and MWEs with opaque semantics (***to kick the bucket***) by separate semantic predicates. Bond, Ho and Flickinger (2015) additionally focus on co-indexation mechanisms needed to represent possessed idioms such as (3)-(4). Finally, Herzig Sheinfux et al. (2015) adopt a largely compositional analysis of MWEs in Hebrew by introducing dedicated lexicon entries for each lexicalized component of each MWE. In Lexical Functional Grammar (LFG), Attia (2006) parses Arabic continuous semi-fixed MWEs (***traffic light***) as single tokens, while syntactically compositional but semantically non-compositional MWEs ('fiery bike' ⇒ 'motorbike') are handled by the grammar via lexical selection rules, similarly to the HPSG approaches. Also, Dyvik, Losnegaard and Rosén (2019) account for a large range of MWEs in a Norwegian LFG grammar. Totally fixed MWEs are represented as words-with-spaces in the lexicon, in syntactic trees (C-structures) and in feature structures (F-structures). Morpho-syntactically flexible MWEs, notably verbal ones, are represented similarly to regular constructions, but with more lexical and morphological constraints in the lexicon entries for the head verb and its arguments. Syntactic flexibility in MWEs is simply handled by the regular grammar rules (it is not quite clear though how restrictions in syntactic variability patterns are represented). Utterances with syntactically flexible MWEs receive regular C-structures but their F-structures contain atomic predicates for semantically non-compositional MWEs. A formally very different account is found in Lexicalized Tree-Adjoining Grammar (LTAG). Since the elementary structures of LTAG, the elementary trees, correspond to an "extended domain of locality", even the surface structure of discontinuous MWEs can be directly represented within the lexicon (Abeillé & Schabes, 1989; Abeillé & Schabes, 1996; Vaidya, Rambow & Palmer, 2014; Lichte & Kallmeyer, 2016). This sort of approach is therefore most similar to the words-with-spaces approaches in HPSG and LFG, yet making more structure available and including slots rather than spaces. As a conclusion, there exist, on the one hand, generic lexical MWE resources which suffer from the lack of sufficient formalization, and, on the other hand, perfectly formalized solutions but restricted to particular grammar formalisms.

(ii) *Existence of factorization mechanisms*. Along this axis, the challenge to address is the proliferation of idiosyncrasy profiles of MWEs. The idea is to factorize descriptions, i.e. to divide larger detailed descriptions into smaller and more generic ones that can be reused. Some MWE lexicons do not introduce generalization of the MWE behavior (Al-Haj, Itai & Wintner, 2014). Some do it via inflection codes (Savary, 2009), syntactic patterns (Pausé, 2017), equivalence classes (Grégoire, 2010), macros (Przepiórkowski et al., 2014), or type hierarchies (Herzig Sheinfux et al., 2015), with a limited degree of recursiveness (i.e. a description – code, pattern, class, macro or type – can, for the sake of non-redundancy and modularity, recursively refer to another description but the length of possible reference chains is low). The metagrammatical approach by Jacquemin (2001) addresses the morphological, syntactic and semantic variation of French MWEs in a factorized way. There, canonical forms of MWEs are represented as fully lexicalized CFG-like rules with feature structures and unification, while

MWE variants are covered by metarules, but the proposal is restricted to continuous terminological compounds.

In view of this state of the art, it seems that a non-redundant lexical encoding of MWEs, which would account for both continuous and discontinuous MWEs, as well as their scale-wise regularity, has not yet received a satisfactory solution. The goal of this paper is to take steps towards such a solution in a metagrammatical and relatively theory-independent framework. We focus on lexical and morpho-syntactic properties of MWEs, and present a proof of concept in the context of LTAG.

3. Introduction to LTAG

This section provides a short overview of LTAG, the target formalism used in this paper for parsing natural language sentences. LTAG is a particular case of a Tree Adjoining Grammar (TAG), so this more generic formalism will be introduced first.

In brief, a TAG consists of a finite set of *elementary trees*. Larger trees can be derived via two composition operations, substitution and adjunction, which replace a leaf or an internal node, respectively, with a new tree. Examples of these two operations are given in Figure 2 and are formalized below. LTAG is a lexicalized version of a TAG, i.e. in which every elementary tree contains at least one lexical item in a leaf. The idea is to represent each predicate together with its valency (nodes corresponding to its arguments) in the same tree.

Formally, a TAG (Joshi & Schabes, 1997) is a tree-rewriting system. Given disjoint sets of terminal and non-terminal symbols, denoted by Σ and $\mathcal{N}$, respectively, trees manipulated by TAG contain symbols from $\mathcal{N}$ in non-leaf nodes and those from $\Sigma \cup \mathcal{N}$ in leaf nodes. If a leaf contains a non-terminal, it is marked either with a down arrow ($\downarrow$) or an asterisk ($*$). TAG trees further divide into *initial* or *auxiliary*. In an initial tree, all non-terminal leaves are marked with down arrows. An auxiliary tree has a selected non-terminal leaf called a *foot*, marked by an asterisk ($*$). Its non-terminal symbol must be the same as the one in the root of the tree. Figure 1 shows a set of trees over terminals $\Sigma = \{Jean, il, la, porte, alors, prend\}$ and non-terminals $\mathcal{N} = \{ADV, CL, D, N, S, V, VN\}$. Trees α_1 to α_4 and α_7 to α_{11} are initial, while trees β_5 and β_6 are auxiliary.

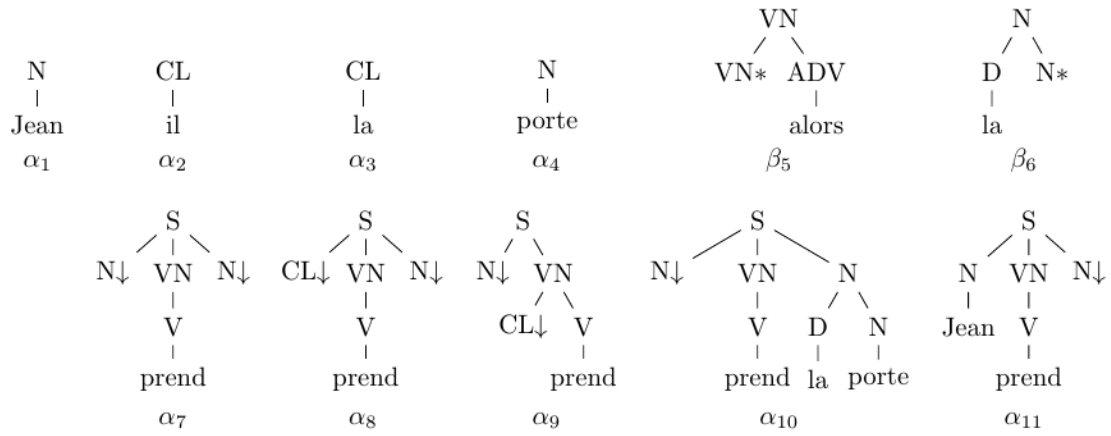

Figure 1. Sample TAG trees, initial: α1 to α4 and α_7 to α_{11}, and auxiliary: β_5 and β_6

Two tree rewriting operations are defined in TAG: *substitution* and *adjunction*. A substitution can extend an existing tree at one of its leaves. Namely, it replaces a non-terminal leaf marked with an arrow (↓) in one tree with any tree having the same non-terminal symbol in its root. Given the set of trees in Figure 1, Figure 2(a) shows a substitution of node N in α_7 with α_1. An adjunction extends an existing tree at any node, possibly internal. Namely, given an auxiliary tree *t*, and any tree *t'*, adjunction replaces *t'*s foot by a subtree *t"* in *t'* and then inserts this modified *t* in place of *t"* in *t'*, provided that the root non-terminals in *t* and *t"* are identical. Figure 2(b) shows and an adjunction of β_5 into α_{11} at node *VN*.

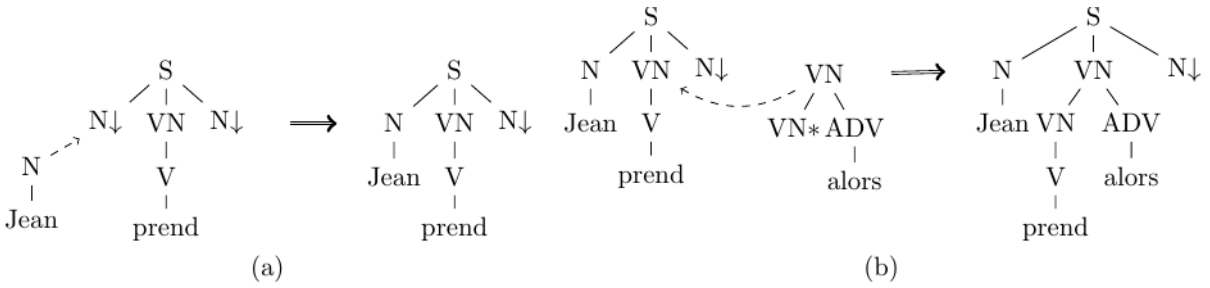

Figure 2. Sample TAG substitution (a) and adjunction (b).

Given the above definitions, a Tree Adjoining Grammar is a 5-tuple $(\Sigma, \mathcal{N}, \mathcal{I}, \mathcal{A}, \mathcal{S})$ where $\mathcal{I}$ is the set of *elementary initial trees*, $\mathcal{A}$ is the set of *elementary auxiliary trees*, and $\mathcal{S}$ is the start non-terminal. For instance, given the trees in Figure 1, a sample TAG grammar $\mathcal{G}_1 = (\Sigma_1, \mathcal{N}_1, \mathcal{I}_1, \mathcal{A}_1, \mathcal{S}_1)$ over $\Sigma_1 = \{$*Jean*, *il*, *la*, *porte*, *alors*, *prend*$\}$ and $\mathcal{N}_1 = \{$*ADV*, *CL*, *D*, *N*, *S*, *V*, *VN*$\}$ could have $\mathcal{I}_1 = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_7, \alpha_8, \alpha_9, \alpha_{10}\}$, $\mathcal{A}_1 = \{\beta_5, \beta_6\}$, and $\mathcal{S}_1 = \mathcal{S}$. Then tree α_{11} would be initial but not elementary, and could be obtained by substitution, as shown in Figure 2(a).

Given a particular TAG $\mathcal{G} = (\Sigma, \mathcal{N}, \mathcal{I}, \mathcal{A}, \mathcal{S})$, a *derivation* is a sequence of tree rewriting operations involving trees from $\mathcal{I} \cup \mathcal{A}$. The last tree in this sequence is called the *derived tree*. The *derivation tree* shows which elementary trees were combined and how. Sentence *Sent* can be derived by $\mathcal{G}$ if there is a derivation whose derived tree has $\mathcal{S}$ in its root and whose sequence of leaves is equal to *Sent*.

Figure 3 illustrates the results of two possible derivations of a sentence with grammar $\mathcal{G}_1$. These derivations lead to the same derived tree (a) but two different derivation trees (b) and (c).[2]

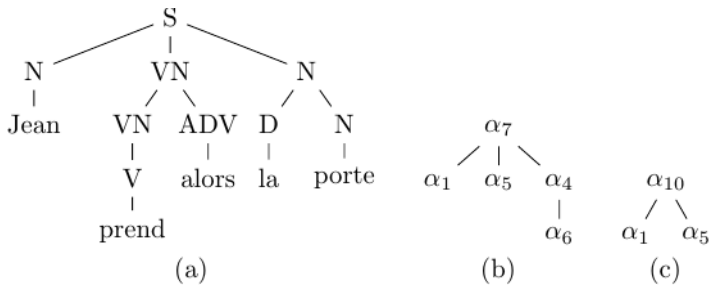

Figure 3. Two possible derivations of the sentence *Jean prend alors la porte* 'Jean takes then the door' with the same derived tree (a) and two different derivation trees (b) and (c).

A Lexicalized Tree Adjoining Grammar (LTAG) is a TAG in which every elementary tree (from $\mathcal{I}$ and from $\mathcal{A}$) has at least one terminal leaf, called an *anchor*. If several terminal leaves occur, one of them is selected as anchor and the others are called *co-anchors*. For instance, the leaf containing *prend* 'takes' in tree α_{10} in Figure 1 could be the anchor, while the leaves containing *la* 'the' and *porte* 'door' would be co-anchors. The TAG $\mathcal{G}_1$ described above is an LTAG.

For the sake of efficiency, implemented LTAGs are often "delexicalized", i.e. elementary trees are separated from their terminal anchor leaves. Nodes above the eliminated (co-)anchors are marked with a diamond (◊). For instance, the elementary initial trees α_7 to α_{10} from Figure 1 are stored as the delexicalized trees from Figure 4. Such delexicalized trees are called *templates*, because they can be shared by many anchors. To account for syntactic variability, templates are grouped into *families*. For instance, a family $\mathcal{F} = \{\alpha_7^d, \alpha_8^d, \alpha_9^d\}$ represents some syntactic realizations of the subject and the object of a transitive verb, e.g. *Jean prend la porte* 'Jean takes the door', *Il prend la porte* 'He takes the door' and *Jean la prend* 'Jean takes it'. Template families receive identifiers and are stored in the grammar proper. Terminal anchor leaves are stored in the lexicon, where they are assigned identifiers of the template families in which they initially appeared (under the ◊-marked nodes).

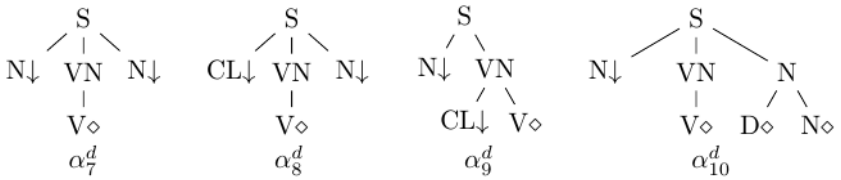

Figure 4. Delexicalized trees corresponding to the LTAG trees α_7 to α_{11} from Figure 1.

It is important to mention that TAGs are unification grammars. Every node in a TAG tree can receive one or two feature structures. They encode constraints under which substitutions or adjunctions can

[2] Derivation trees also contain addresses of the nodes at which substitutions and adjunctions take place. These addresses are omitted in this paper for the sake of brevity.

take place in the given node. When producing derivations, feature structures of appropriate nodes in the combined trees are unified (Vijay-Shanker & Joshi, 1988) and a given derivation is acceptable only if all unifications it implies can be performed. In this paper, we neglect unification for the sake of brevity.

In this work, LTAG is used as the target formalism, compiled from a metagrammar, because it has several advantages for representing and parsing MWEs (Abeillé & Schabes, 1989), like *prendre la porte* 'to take the door' ⇒ 'to be forced to leave'. One of them is that MWEs receive individual elementary trees in an LTAG (like α_{10} in Fig. 1), which elegantly expresses the fact that they are semantically non-compositional units, while still having a (possibly flexible) syntactic structure. Secondly, feature structures in such trees allow one to express long-distance dependencies between different MWE components, for instance the fact that the possessive determiner in examples (3)-(4) must agree with the subject in person and number. Thirdly, literal and idiomatic readings of MWEs obtain the same derived trees (since they have the same syntax) but different derivation trees (since the have different interpretations), as shown in Figure 3.[3] Fourthly, discontinuities in MWEs are easily handled in derivation. For instance, any number of adverbs like *alors* 'then' in β_5 can be inserted by adjunction between the verb *prend* 'takes' and the object *la porte* 'the door' in tree α_{10}, similarly to Figure 2(b).

4. From a metagrammar to parsing

Despite the advantages of LTAG described in the previous section, large-coverage LTAG grammars are usually huge and hard to maintain, even when MWEs are not yet accounted for. Making them MWE-aware can easily lead to prohibitive numbers of elementary trees and tree families, due to the variety of the existing idiosyncrasy profiles. What is more, such labor-intensive MWE descriptions are TAG-specific and cannot be reused for other grammatical formalisms.

Metagrammars were developed to cope with the proliferation of grammar rules. The main idea is to:

- use a new formalism which can be automatically compiled to several existing grammatical formalisms but which is more abstract than these target formalisms;
- factorize tree descriptions, i.e. divide them into smaller tree fragments, which can then be combined under well-defined constraints to form complete syntactic trees;
- perform tree factorization in such a way that the tree fragments are linguistically motivated (i.e. they represent elementary linguistic properties) and reusable across various target trees.

[3] Lichte and Kallmeyer (2016) propose a different solution in which ambiguity of interpretations is expressed at the level of semantics (represented with frames) rather than of syntactic derivations.

For instance, tree α_7 from Figure 1 expresses several properties of transitive verbs simultaneously:

- A canonical (nominal) subject precedes the verbal nucleus.
- A canonical object follows the verbal nucleus.
- The morphology of the verbal nucleus percolates from the main verb (this is expressed by feature structures, neglected here).

In a metagrammar, these properties can be expressed separately by different tree fragments. Figure 5 shows the tree fragment representing the first property above. Note that this property is also expressed by tree α_9, so the tree fragment from Figure 5 can be reused. This intuitive example will be enlarged and formalized in Section 5.

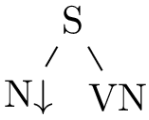

Figure 5. Tree fragment, shared by α_7 and α_9 from Figure 1, and representing the fact that a canonical subject precedes the verbal nucleus.

In this work, we are interested in the metagrammatical framework called eXtensible MetaGrammar (XMG) (Crabbé et al., 2013; Petitjean et al., 2016), inspired by Candito (1999). It provides description languages and dedicated compilers for generating a wide range of linguistic resources.[4] Descriptions are organized into *classes*, as in object-oriented programming. Classes have encapsulated name spaces and inheritance relations may hold between them. The crucial elements of a class are *dimensions*. They can be equipped with specific description languages and are compiled independently, thereby enabling the grammar writer to treat the levels of linguistic information separately.

In this paper, we use the `<syn>` and `<iface>` dimensions. `<syn>` holds tree constraints that express dominance and precedence relations among nodes. Nodes may carry (untyped) feature structures. `<iface>` is an interface dimension where (untyped) feature structures are used to share information between other dimensions and classes.[5] The general structure of a class is given on the left of Listing 1.

```
class classname
import someOtherClass[]
export ?someVariable
declare ?someVariable {
  <syn>{ ... };
  <iface> { ... } }
```

```
class CanSubject
export ?VN
declare ?VN {
  <syn>{node [cat=s]{
         node [cat=n]
         node ?VN[cat=vn]} }}
```

Listing 1. Structure of XMG classes (left) and a <syn> dimension example (right).

[4] http://xmg.phil.hhu.de/.

[5] An alternative notation for the <iface> dimension uses the *= operator as shown in the right column of Listing 5.

The more concrete class `CanSubject` on the right of Listing 1 describes the canonical position of a subject `N` on the left of the verbal nucleus `VN`. This class corresponds to the tree fragment given in Figure 5. The `declare` and `export` instructions are responsible for the creation of variables, and the control of their scopes: the first one defines the variables which can be used in the class (identifiers for nodes, for instance), while the second one provides the list of variables (local, by default) which can be accessed from other classes. The `import` instruction is used to build the class hierarchy: the tree description contained in the imported class is added to the one of the current class. In addition, the variables which are exported by the imported class are added to the environment of the current class.

In order to promote reusability of linguistic descriptions, XMG comes with compilers (implemented as constraint solvers) for several syntactic formalisms. Here, we focus on LTAG, for the reasons explained in Section 3. We use LTAG grammars compiled from XMG metagrammars as input to the TuLiPA parser (Parmentier, Kallmeyer, Maier, Lichte & Dellert, 2008; Arps & Petitjean, 2018), applied in the experiments reported on in Section 7.

5. FrenchTAG: A French XMG metagrammar

FrenchTAG (Crabbé, 2005) is a syntactic XMG implementation of the reference grammar of French by Abeillé (2002).[6] It contains 285 XMG classes,[7] including 96 families, which compile into 9045 TAG tree templates. Its main focus are verbs. It defines about 40 verbal subcategorisation frames, and – for each frame – the allowed diatheses (active, passive, middle, reflexive, impersonal, etc.) and argument realizations (canonical, clitic, extracted, omitted, etc.). Listing 2 shows an extract of the classes describing transitive verbs like *ouvrir* 'open' and *prendre* 'take'.

Classes can be combined by a disjunction or a conjunction, marked by | and ; , respectively. A disjunction means that the invoking class can realize one of the classes it invokes. In Listing 2, the `Subject` class is a disjunction of various realizations of a subject: canonical (*Jean* 'Jean'), clitic (*il* 'he.SUBJ'), relative (*Jean qui ouvre la porte* 'Jean who opens the door'), etc. Similarly, a direct `Object` can be canonical (*la porte* 'the door'), clitic (*la* 'it.OBJ'), relative (*la porte que Jean ouvre* 'the door which Jean opens'), reflexive (*Jean s'ouvre* 'Jean opens himself'), etc.

A conjunction means that all the invoked classes have to be combined and unified to form the invoking class. Here, the `dian0Vn1Active` class combines, by conjunction, any realization of a subject and an object with a verb in active voice.

---

[6] FrenchTAG later evolved into SEMTAG with a unification-based compositional semantic dimension (Gardent, 2008).

[7] Only 246 of them occur in the FrenchTAG version adapted to the current version of XMG.

Class `n0Vn1` is a disjunction of diatheses: active (*Jean ouvre la porte* 'Jean opens the door'), passive (*la porte est ouverte par Jean* 'the door is opened by Jean'), short passive (*la porte est ouverte* 'the door is opened'), etc. Class `n0ClV` represents inherently reflexive verbs discussed in Section 5.2.

As announced in Section 4, each class describes a more or less abstract set of tree fragments, which can be made more specific by adding constraints in the invoking classes. For instance, `CanonicalSubject` invoked by `Subject` corresponds to the tree fragment in Figure 6(a). As discussed in Section 4, this fragment can be shared by many target trees. Moreover, `dian0Vn1Active` yields several combinations of tree fragments (a)-(f), including tree (i) (*Jean l'ouvre* 'John opens it') obtained from (a), (d) and (e) unified along the verbal spine `S`→`VN`→`V`. Note that trees (g)-(i) are identical to trees α_7^d, α_8^d and α_9^d from Figure 4.

```
1  class Subject {
2    CanonicalSubject[]
3    | CliticSubject[]
4    | RelativeSubject[]
5    % More alternatives }
6  class Object {
7    CanonicalObject[]
8    | CliticObjectII[]
9    | RelativeObject[]
10   | reflexiveAccusative[]
11   % More alternatives }
```

```
1  class dian0Vn1Active {
2    Subject[];
3    activeVerbMorphology[];
4    Object[] }
5  class n0Vn1 {
6    dian0Vn1Active[]
7    | dian0Vn1Passive[]
8    | dian0Vn1ShortPassive[]
9    % More alternatives }
10 class n0ClV{
11   Subject[];
12   reflexiveVerbMorphology[] }
```

Listing 2. FrenchTAG classes for the realizations of a subject, a direct object, an active diathesis, all diatheses of transitive verbs, and inherently reflexive verbs.

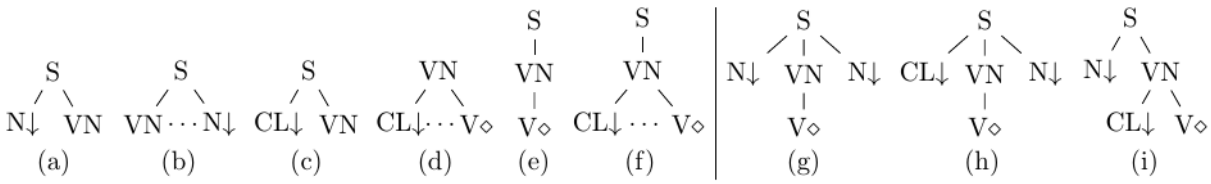

Figure 6. Tree fragments described by the XMG classes from Listing 2: canonical subject (a) and object (b), clitic subject (c) and object (d), active (e) and reflexive (f) verb morphology. The corresponding LTAG trees are shown in (g-i). Feature structures are omitted. The dots in (b), (d) and (f) represent a possibly non-immediate precedence of nodes.

5.1 Hierarchy of classes

Like in other object-oriented modelling formalisms, object classes in XMG form a hierarchy. When class `C2` appears in the body of class `C1` then `C1` is lower than `C2`. If a conjunction of `C2` and `C3` appears in `C1`, then `C1` inherits all properties of both `C2` and `C3`, possibly adding new ones. If `C2` and `C3` are invoked in `C1` with a disjunction, then `C1` has two different realizations, inheriting the properties of `C2` or `C3`, respectively (and possibly adding new properties). In FrenchTAG all classes are defined so that a conjunction and a disjunction are not used simultaneously in one class. Therefore, we can represent the hierarchy of classes in FrenchTAG as a graph a fragment of which is given in Figure 7. The uppermost `TopLevelClass` is the most abstract, while the lowermost one, here `n0Vn1`, is the most

concrete. Straight and wavy lines denote conjunctive and disjunctive inheritance, respectively. Conjunctive inheritance is used to combine descriptions, which then compile into bigger tree fragments, as explained above for `dian0Vn1Active`, while disjunctive inheritance allows to express alternatives.

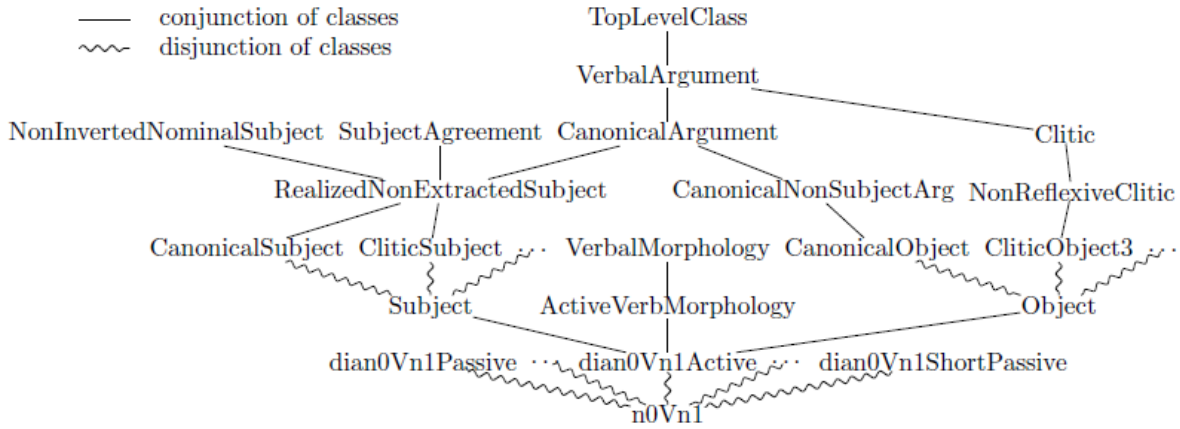

Figure 7. Extract of the XMG class hierarchy in the original FrenchTAG metagrammar.

5.2 MWEs in FrenchTAG

FrenchTAG mainly aims at modelling the syntactic behaviour of French verbs and accounts for only a small number of MWEs. Prominently, it covers some inherently reflexive verbs (IRVs), in which the reflexive clitic *se* 'oneself' either is inherent to the verb (i.e. the verb never occurs alone, as in ***s'abstenir*** 'refrain') or markedly changes the meaning and/or the subcategorization frame of the verb (as in ***se passer*** *de quelque chose* 'pass oneself of sth' ⇒ 'do without sth'). Although the clitic in an IRV occupies the syntactic place of a direct object, it is not a semantic argument of the verb and does not alternate with non-reflexive objects, unlike for transitive verbs. For instance, the verb *ouvrir* 'open' assigned to `n0Vn1` from Listing 2 exhibits various object realizations including a canonical object (*Jean ouvre la porte* 'Jean opens the door') and a reflexive one (*Jean s'ouvre* 'Jean opens himself'). These alternatives are covered by the tree fragments (b) and (d) from Figure 6, which – together with the conjunct classes for the subject and the verb – compile into the LTAG trees (g), (h), (i), etc. Conversely, IRVs are assigned to the `n0ClV` class from Listing 2, which imports the `reflexiveVerbMorphology` class. The latter leaves no choice of the object realization but directly includes the reflexive object preceding the verb. The tree fragment it describes is identical to Figure 6(f), but additionally imposes – via feature structures – number and person agreement between the reflexive clitic *CL* and the verb *V*.

This mechanism displays one possibility of representing restrictive properties of MWEs: new classes are created by combining only those alternatives which are allowed by a given type of MWEs. Here, class `n0ClV` directly combines a free subject with a verb taking a clitic but not any other form

of object. In Section 6, we present a new approach to MWE encoding, aiming at limiting the number of new classes to be created.

5.3 Delexicalization and compilation

FrenchTAG, following the XTAG (XTAG Research Group, 2001) architecture, separates the grammatical description from the lexicon via delexicalization, discussed in Section 3. The lexicon covers lemmas and their inflected forms, as shown in Listings 3 and 4. Families (`fam`) assigned to the lemmas refer to grammar classes, i.e. sets of tree fragments. During parsing, the grammar is anchored with the lexicon, i.e. lexicon entries are linked with anchor nodes (marked with ◊) of the tree templates, provided that anchoring constraints are fulfilled. The latter are based on unification of feature structures (FSs) of two types. Firstly, the FSs attached to the ◊-marked anchor nodes are unified with the FSs of the lexicon entries to be anchored. For instance, the FS (not shown here) in the $V\Diamond$ node in Figure 6(g) is unified with the features (cat, mode, etc.) in the entry for *ouvre* 'opens' in Listing 4. Secondly, *interface FSs* can be used to eliminate some alternatives, as discussed in the following section.

```
class LemJean {
<lemma> {
 entry <- "Jean";
 cat <- n;
 fam <- propn}
}
class LemIl {
<lemma> {
 entry <- "il";
 cat <- cl;
 fam <- CliticT}
}
```

```
class LemOuvrir {
<lemma> {
 entry <-"ouvrir";
 cat <- v;
 fam <- n0Vn1}
}
class LemAbstenir {
<lemma> {
 entry <-"abstenir";
 cat <- v;
 fam <- n0ClV}
}
```

```
class LemPorte{
<lemma> {
 entry <- "porte";
 cat <- n;
 fam <- noun }
}
class LemLe {
<lemma> {
 entry <- "le";
 cat <- d;
 fam <- stddet }
}
```

Listing 3. Extract of the FrenchTAG lexicon with 6 lemmas.

```
class il {
<morpho> {
 morph <- "il";
 lemma <- "il";
 cat <- cl
 func <- suj;
 refl <- -;
 pers <- 3;
 gen <- m;
 num <- sg }}
```

```
class ouvre {
<morpho> {
 morph <- "ouvre";
 lemma <- "ouvrir";
 cat   <- v;
 mode <- ind;
 pers <- 3;
 num <- sg }}
```

```
class la {
<morpho> {
 morph <- "la";
 lemma <- "le";
 cat   <- d;
 gen <- f;
 num <- sg }}
```

```
class porte {
<morpho> {
 morph <- "porte";
 lemma <- "porte";
 cat   <- n;
 det <- -;
 gen <- f;
 num <- sg }}
```

Listing 4. Extract of the FrenchTAG lexicon with 4 inflected form.

Compiling an XMG metagrammar $\mathcal{M}$ into an LTAG boils down to finding minimal tree models which fulfill constraints expressed in $\mathcal{M}$. In Figure 6, the tree fragment (d) imposes that the clitic CL precedes the verb V possibly indirectly (which is signaled by dots between the sibling nodes). Since no other fragment imposes a third node between the CL and V, the minimal model is the one with a direct precedence.

If the target formalism selected for XMG compilation is LTAG, then the outcome is an LTAG grammar, with trees like those from Figure 4. This automatically generated grammar is then directly used to parse an input sentence, producing derived and derivation trees, as in Figure 3.

6. Enriching a metagrammar with MWEs

XMG offers elegant, fully formalized and powerful factoring mechanisms, which enable largely nonredundant grammar engineering. However, FrenchTAG, which is one of its most advanced use cases, only encodes a small number of MWE types and does it with a limited degree of factorization (cf. Sect. 5.2). Here, we describe another approach which aims at a reasonable factorization due to an extensive use of *interface filters* (or *interface FSs*) both in the grammar and in the lexicon.

6.1 Disjunctive classes with interface filters

Consider the VMWE in (5). On the one hand, it shares some properties with transitive verbs (class `n0Vn1`). For instance, its verb inflects freely, and its subject is unconstrained (canonical, clitic, etc.) and agrees with the verb as shown in (5)-(8). On the other hand, it exhibits restrictive properties. If the idiomatic meaning is to be retained, the verb cannot be passivized (9), and its object cannot be replaced by a synonym (10), cliticized (11), extracted (12) or modified (13).

(5) *Jean **prend la porte*** 'Jean takes the door' ⇒ 'Jean leaves (because he is forced to)'
(6) *Jean/il **prend la porte*** 'Jean/he takes the door' ⇒ 'Jean/he leaves'
(7) *Jean qui **prend la porte*** 'Jean, who takes the door' ⇒ 'Jean, who leaves…'
(8) ***Prenez la porte!*** 'Take.2.PL the door!' ⇒ 'Leave!'
(9) #*La porte est prise par Jean* 'The door is taken by Jean'
(10) #*Jean prend la sortie* 'Jean takes the exit'
(11) #*Jean la prend* 'Jean takes it'
(12) #*La porte que Jean prend* 'The door that Jean takes'
(13) #*Jean prend la grande porte* 'Jean takes the big door'

In order to account for these properties, we enrich the grammar with new classes some of which are presented in Listing 5. The `mweSubject` class is a disjunction of a free subject (represented by the pre-existing `Subject` class from Listing 2) and a lexicalized subject (`mweSubjectLexStruct`). Each of these alternatives is augmented by an interface FS (`[subj=free]` or `[subj=lex]`), i.e. a feature structure attached to the whole tree fragment rather than its particular node. Similarly, the new class `mweObject` is an alternative between a free and a lexicalized object (`mweObjectStruct` from Listing 6 explained below). Further, `mwedian0Vn1Active` is a MWE-aware version of `dian0Vn1Active` from Listing 2. It reuses the pre-existing `activeVerbMorphology` class but offers the choice between a free and a lexicalized subject and/or object. Finally, the `mwen0Vn1` class is similar to `n0Vn1` from Listing 2, but each of its diathesis alternatives is MWE-aware and receives an interface FS with the `dia` feature whose name conforms to this diathesis. The idea of the interface

FSs is to express restrictive properties, e.g. the `obj=lex` and `dia=active` features in lines 6 and 12 in Listing 5 will also be used in the lexicon entry (cf. Section 7) of the VMWE from example (5), which will let us eliminate reformulations like in examples (9) and (10).[8]

```
1  class mweSubject {
2   Subject[] *= [subj=free]
3   | mweSubjectLexStruct[] *= [subj=lex]}
4  class mweObject {
5   Object[] *= [obj=free]
6   | mweObjectLexStruct[] *= [obj=lex]}
7  class mwedian0Vn1Active{
8   mweSubject[];
9   activeVerbMorphology[];
10  mweObject[] }
11 class mwen0Vn1 {
12  mwedian0Vn1Active[] *= [dia=active]
13  |mwedian0Vn1Passive[] *= [dia=passive, passivetype=full]
14  |mwedian0Vn1ShortPassive[] *= [dia=passive, passivetype=short]}
```

Listing 5. MWE-aware classes based on those from Figure 2 and augmented with interface constraints.

```
1 class mweObjectLexStruct
2  import mweObjectLex[]
3  declare ?lexNP {
4  {   ?lexNP = mweLexDetLexNoun[ObjDetNode,ObjNode] *= [objstruct=lexDetLexN]
5    | ?lexNP = mweNoDetLexNoun[ObjNode] *= [objstruct=lexN]
6    | ?lexNP = mweLexNounLexAdj[ObjNode,ObjAdjNode] *= [objstruct=lexNLexAdj] };
7  ?LexObj = ?lexNP.xLexNPTop }
```

Listing 6. (Simplified) MWE-aware classes describing lexicalized objects of various syntactic structures.

The `mweObjectLexStruct` class from Listing 6 is an alternative of `mweLexDetLexNoun`, `mweNoDetLexNoun` and `mweLexNounLexAdj`, three classes which describe lexicalized noun phrases of the *Det-Noun, Noun* or *Noun-Adj* structures, respectively, as in (5), *Jean* ***fait face*** *au problème* 'Jean makes face to the problem' ⇒ 'Jean deals with the problem' and *Jean* ***fait profil bas*** 'Jean makes low profile' ⇒ 'Jean keeps quiet and discreet'. The corresponding tree fragments are depicted in Figure 8(b)-(d). The `mweObjectLex` class, inherited by `mweObjectLexStruct`, is similar to `Object` in Listing 2 up to substitution marking of the object nodes. For instance, one of the diatheses described by `mweObjectLex` is the one in Figure 8(a), which is identical to Figure 6(b) up to the substitution mark of the *N* node. This allows us to unify precisely this node with the root (`xLexNPTop`) of one of the trees in Figure 8(b)-(d), so as to ensure that the object is lexicalized. Note also that the nodes in the three object realizations bear names: `ObjDetNode`, `ObjNode` and `ObjAdjNode`, which are referred to from lexicon entries (see below). XMG compilation of the `mweObjectLexStruct` class for the first alternative (`mweLexDetLexNoun`) boilds down to unifying the tree fragments from Figure 8(a) and (b), as well as from Figure 6(a) and (e), which results

[8] Reformulation (12) is blocked by constraints expressed in another class, skipped here, from which the mweObject class inherits. Reformulation (13) has not yet received a satisfactory solution due to reasons inherent to the XMG implementation to which we turn in Section 8.

in the LTAG tree from Figure 8(e). The compilation of the two other alternatives (`mweNoDetLexNoun` and `mweLexNounLexAdj`) yields two additional LTAG trees, not shown here, in which the fragments (c) or (d) are used instead of (b).

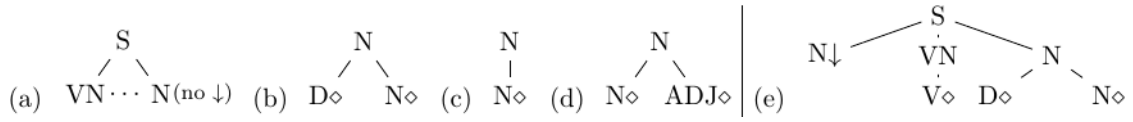

Figure 8. Tree fragments defined by the MWE-aware classes from Listing 6: (a) a canonical object with no substitution node (part of the `mweObjectLex` class), (b--d) a lexicalized nominal phrase with a fixed *Det-Noun*, *Noun* and *Noun-Adj* structure (`mweLexDetLexNoun`, `mweNoDetLexNoun` and `mweLexNounLexAdj` classes, respectively). The LTAG tree compiled from (a), (b), as well as Figure 6 (a) and (e), is shown in (e).

6.2 MWEs in the lexicon

These new classes and their interface FSs can be used in the lexicon. Thus, the MWE ***prendre la porte*** 'to take the door' ⇒ 'to be forced to leave' receives the lexical class `mweLemmePrendreLaPorte` from the upper left-hand column in Listing 7. The `entry` holds the head of the MWE, here *prendre* 'take', which anchors (in some inflected form) tree templates from the `mwen0Vn1` family (cf. Listing 5) specified in `fam`. A bunch of `filter` feature-value pairs help to make a selection of the appropriate tree templates in `mwen0Vn1`. Namely, the `filter` features of the lemma and the interface features of the LTAG tree are unified, and if the unification fails, so does the anchoring of the lexical item to the tree. For example, the `dia=active` filter selects the trees that correspond to active diathesis (line 12 in Listing 5), while the `subj=free` and `obj=lex` filters ensure that the subject and the object are free and lexicalized, respectively (lines 2 and 6 in Listing 5). The other lexicalized components of ***prendre la porte*** are specified by `coanchor`, where `ObjectDetNode` and `ObjNode` are node names from the `mweObjectLexStruct` class in Listing 6. These names, as well as the `objstruct` filter, ensure that only the right syntactic alternation is selected from `mweObjectLexStruct` (here `mweLexDetLexNoun`) and that the co-anchors *la* and *porte* attach to nodes `D` and `N` in Figure 8(e). Node names are also used in the `equation` part in order to pass on morphological features onto the referred nodes. These `equation` features and unified with the feature structures of the corresponding tree nodes.

The crucial advantage of interface filters is that no new classes have to be created for VMWEs which share the same syntactic structure but allow different syntactic alternations. The upper right-hand column of Listing 7 encodes another VMWE, ***prendre la parole*** 'take the speech' ⇒ 'take the floor', in which the direct object is composed of a lexicalized determiner *la* 'the' and a noun *parole* 'speech'. Contrary to ***prendre la porte***, this VMWE does allow the passive diathesis: ***la parole*** *a été* ***prise*** *par la représentante de la République Dominicaine* 'the speech was taken by the representative of the Dominican Republic' ⇒ 'the representative of the Dominican Republic took the floor'. Note that both VMWEs are assigned the same family, `mwen0Vn1`, however the latter does not include the

`dia` filter. This enables all three alternatives in lines 12-14 of Listing 5 to be used in parsing, which means that idiomatic interpretation of ***la parole*** *a été* ***prise*** 'the speech was taken' ⇒ 'someone took the floor' is allowed.

Further, the middle part of Listing 7 shows lexicon classes for two other VMWEs: ***faire appel*** 'make call' ⇒ 'appeal' and ***faire profil bas*** 'make low profile' ⇒ 'avoid attracting attention'. Like ***prendre la porte***, these VMWEs are assigned to the `mwen0Vn1` family and only allow the active alternation (cf. the `dia=active` filter) but the syntactic structures of their lexicalized objects are different. In ***faire appel*** the lexicalized noun *appel* 'appeal' with no determiner is encoded by the `objstruct=lexN` filter, which selects only line 5 from Listing 6, i.e. the tree fragment (c) from Figure 8. In ***faire profil bas***, the lexicalized noun *profil* 'profile' modified by an adjective bas 'low' is represented by `objstruct=lexNLexAdj` selecting line 6 from Listing 6 and the tree fragment (d) from Figure 8.

Finally, the lower part of Listing 7 shows the VMWE ***les carottes sont cuites*** 'the carrots are cooked' ⇒ 'there is no turning back'. This proverb can only appear in passive (#*quelqu'un a cuit les carottes* 'someone cook the carrots') and with no agent (#*les carottes sont cuites par Jean* 'the carrots are cooked by Jean'). This fact is encoded by the `dia=passive` and `passivetype=short` filters, which select line 14 in Listing 5.

Note that no additional grammar class was needed to account for all the diverse idiosyncrasy profiles of these VMWEs. This joint effect of precision and conciseness is described in the title of this paper as ***short and sweet*** 'pleasantly brief'. The aim of the evaluation presented in Section 7 is to check how well the conciseness can be preserved while new naturally occurring VMWEs are being added to the metagrammar.

6.3 Extended hierarchy of classes

Figure 9 shows an extract of the class hierarchy resulting from making FrenchTAG MWE-aware. Classes marked with underlined mwe prefixes, e.g. `mweObject`, were obtained from the original FrenchTAG classes by adding interface filters (and possibly restricting the number of alternatives), like in Listing 5.

```
class mweLemmePrendreLaPorte {
<lemma> {
 entry <- "prendre";
 cat   <- v;
 fam   <- mwen0Vn1;
 filter dia = active;
 filter subj = free;
 filter obj = lex;
 filter objstruct = lexDetLexN;
 coanchor ObjDetNode -> "la"/d;
 coanchor ObjNode -> "porte"/n;
 equation ObjNode -> gen=f;
 equation ObjNode -> num=sg }}
```

```
class mweLemmePrendreLaParole {
<lemma> {
 entry <- "prendre";
 cat   <- v;
 fam   <- mwen0Vn1;
 filter subj = free;
 filter obj = lex;
 filter objstruct = lexDetLexN;
 coanchor ObjDetNode -> "la"/d;
 coanchor ObjNode -> "parole"/n;
 equation ObjNode -> gen=f;
 equation ObjNode -> num=sg }}
```

```
class mweLemmeFaireAppel {
<lemma> {
 entry <- "faire";
 cat   <- v;
 fam   <- mwen0Vn1;
 filter dia = active;
 filter subj = free;
 filter obj = lex;
 filter objstruct = lexN;
 coanchor ObjNode -> "appel"/n;
 equation ObjNode -> gen=m;
 equation ObjNode -> num=sg }}
```

```
class mweLemmeFaireProfilBas {
<lemma> {
 entry <- "faire";
 cat   <- v;
 fam   <- mwen0Vn1;
 filter dia = active;
 filter subj = free;
 filter obj = lex;
 filter objstruct = lexNLexAdj;
 coanchor ObjNode -> "profil"/n;
 coanchor ObjAdjNode -> "bas"/adj;
 equation ObjAdjNode -> gen=m;
 equation ObjAdjNode -> num=sg; }}
```

```
class mweLesCarottesSontCuites {
<lemma> {
 entry <- "cuire";
 cat   <- v;
 fam   <- mwen0Vn1;
 filter dia = passive;
 filter passivetype = short;
 filter obj = lex;
 filter objstruct = lexDetLexN;
 coanchor ObjDetNode -> "les"/d;
 coanchor ObjNode -> "carottes"/n;
 equation ObjNode -> gen=f;
 equation ObjNode -> num=pl }}
```

Listing 7. Lexicon classes for five MWEs *prendre la porte* 'take the door' ⇒ 'to be forced to leave', *prendre la parole* 'take the speech' ⇒ 'take the floor', *faire appel* 'make call' ⇒ 'appeal', *faire profil bas* 'make low profile' ⇒ 'avoid attracting attention' and *les carottes sont cuites* 'the carrots are cooked' ⇒ 'there is no turning back'.

Classes with boxed names are new and encode the syntactic structure of lexicalized verbal arguments of MWEs, as in Listing 6.[9] All other classes remain unchanged with respect to Figure 7. Note in particular that whole subgraphs above the previously existing classes like `Subject` and `Object` remain intact. Also, the subgraphs consisting of new classes like `mweObjectLexStruct`,

[9] Note that `mweObjectLexStruct` inherits from classes both conjunctively and disjunctively: a tree fragment generated by this class combines a fragment encoded by the `mweObjectLex` class and a fragment generated by one class among `mweLexDetLexNoun`, `mweNoDetLexNoun` and `mweLexNounLexAdj`, as shown in Listing 6. The hierarchy in Figure 9 does not convey this combination precisely but only shows some of the 4 parent classes for `mweObjectLexStruct`.

`mweObjectLex` and `mweCanonicalObject` are relatively shallow: beyond three inheritance levels of MWE-specific classes they inherit from (unchanged) pre-existing classes (here `CanonicalNonSubjectArg`). This is a visual evidence that the degree of modifications needed to make the original metagrammar MWE-aware is relatively limited.

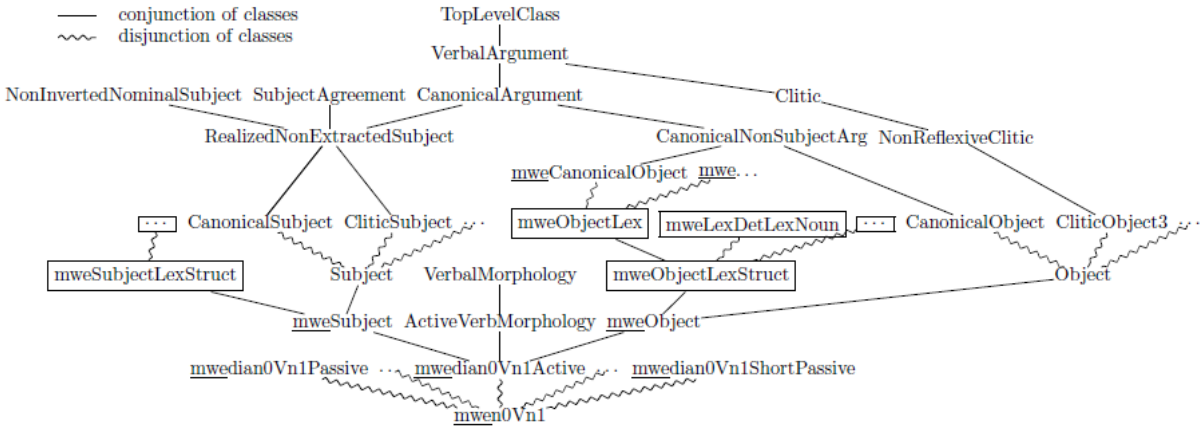

Figure 9. Extract of the XMG class hierarchy in the MWE-aware metagrammar, with the pre-existing classes from Figure 2, as well as augmented (mwe-prefixed) and newly created (boxed) classes.

6.4 Availability and a complete example

In the previous sections, we have shown how a French XMG metagrammar could be extended so as to cover MWEs. Henceforth, we refer to this MWE-aware version of FrenchTAG as mweFrenchTAG. The mweFrenchTAG source codes, together with its compiled TAG version, and usage instructions, are available under the Creative Commons Attribution v4 license.[10]

Appendix A at the end of this article shows a complete example of how mweFrenchTAG models the sentence *Jean **prend la porte*** 'Jean takes the door' ⇒ 'Jean leaves because he is forced to'. All lexicon entries necessary for this sentence are included in Listing 8. Then Table 2 shows: (i) the tree fragments corresponding to the XMG classes referred to from the lexicon entries, (ii) the TAG trees compiled from these fragments, (iii) the outcome of the derivation, with the same derived tree but different derivation trees, as previously illustrated in Figure 3.

7. Evaluation

This work is meant to provide a proof of concept that non-redundant lexical encoding of MWEs can be effectively achieved in XMG. Therefore, we do not aim at evaluating the coverage of the metagrammar. Instead, we wish to see: (i) to what extent the metagrammar allows us to cover the properties of MWEs of a variable degree of regularity, as they are present in real data, (ii) how far the

[10] https://gitlab.com/agata.savary/mwe-xmg/

metagrammar changes or extends, when new MWEs and phenomena are to be covered. To this end, we first prepared a dataset of manageable size based on a French reference corpus of VMWEs and divided it into two parts: DEV and TEST (see Sect. 7.1). We then adapted mweFrenchTAG to encode the MWEs present in DEV (stage 1) and TEST (stage 2), respectively, in order to see how many additional classes are needed to encode new MWEs (see Sect. 7.2) in the proposed framework.

7.1 Dataset preparation

We used the French part of the PARSEME corpus of VMWEs in version 1.0 (Savary et al., 2018). For the 1,449 unique VMWEs, their idiomatic and potentially literal occurrences were extracted using the method proposed by Savary and Cordeiro (2018). By a potentially literal occurrence of a known VMWE we mean a co-occurrence of its lemmas which was not annotated as a VMWE. Such a co-occurrence may, indeed, be literal, as in *Jean prends la porte rouge et la charge dans le camion* 'Jean takes the red door and loads it onto the truck'. However, it is most often coincidental, as in *Jean prend la décision de peindre la porte* 'Jean takes the decision to paint the door'. The extraction resulted in 5,893 occurrences. From those, an evaluation dataset of manageable size was selected following two main criteria: frequency and syntactic variety. We first chose 14 most frequent VMWEs of various categories. For each chosen VMWE, we selected a subset of 4 occurrences while trying to preserve the syntactic diversity of the original set.[11] This led to an evaluation set of 56 occurrences in total, which was randomly divided into two disjoint subsets: DEV (50%) and TEST (50%).

We initially intended to use literal readings in order to demonstrate the correctness of mweFrenchTAG with respect to encoding restrictive properties of MWEs (cf. Sect. 1). For instance, a sentence like in example (5) should receive one derived tree but two different derivation trees (cf. Sect. 3) representing both its compositional and its idiomatic reading. Conversely, a sentence like in (9) should only get one derivation tree corresponding to its literal reading. However, it turned out that most occurrences extracted as potentially literal were in fact coincidental.[12] We therefore abandoned literal occurrences in this study and thus reduced DEV and TEST to 26 sentences each.

The original FrenchTAG is focused on verbal predicates and their arguments, while providing only basic coverage for other syntactic categories. It also has a rather scarce lexicon. Therefore, the 52 sentences were manually simplified so as to: (i) avoid negation, subordinate clauses and coordinations

[11] We repeatedly randomly selected a subset of 56 occurrences, 4 occurrences per VMWE, and selected the one with the largest syntactic variety. As a measure of the syntactic variety of a set of VMWE occurrences, we used the Shannon's entropy of the multiset in which each occurrence is replaced by its "syntactic configuration": a sequence of components, each component represented by (i) its POS, part-of-speech tag, and (ii) the set of outgoing dependency relation types limited to 2 (first two in the lexicographic order).

[12] This conforms to the observations about the scarceness of literal readings by Savary, Cordeiro, Lichte, Ramisch, Inurrieta and Giouli (2019b).

except when VMWEs occurred in them, (ii) reduce the non-lexicalized arguments of the verbs (to the head nouns, or to single-word proper names and adverbials). Finally, all contractions were extended (e.g. *du* → *de le, d'être* → *de être*), as required by the original FrenchTAG. For instance, the original sentences (14), (16) and (18) were reduced into (15), (17) and (19), respectively.

(14) *Quatorze célébrités ont* ***fait partie*** *du casting de cette deuxième saison.* 'Fourteen celebrities have been part of the casting of the second season.'

(15) *les célébrités ont* ***fait partie*** *de le casting* 'the celebrities have been part of the casting'

(16) *Tant que les rois wisigoths n'avaient exercé de pouvoir sur les populations autochtones qu'au nom de l'Empire romain et en vertu de titres tels que "maîtres de l'armée", l'unification doctrinale n'****avait*** *pas* ***lieu*** *d'être.* 'As long as Visigoth kings exercised power over indigenous populations only on behalf of the Roman Empire and by virtue of titles such as "army masters", there was no reason for the doctrinal unification.'

(17) *la unification doctrinale* ***avait lieu*** *de être* 'the doctrinal unification had place to be' 'there was a reason for the doctrinal unification'

(18) *Ce prelat accorda à l'abbaye une partie du chef du saint dont elle* ***porte*** *le* ***nom****.* 'This prelate granted the abbey part of the head relic of the saint whose name it bears.'

(19) *le saint dont elle* ***porte*** *le* ***nom*** 'the saint whose name it bears'

The simplified datasets, DEV-S and TEST-S, are cited in Appendix B.[13] The lexicon necessary for parsing them was automatically derived from the Unitex corpus processor.[14]

7.2 Evaluation proper

In phase 1 of the evaluation, we (manually) added to mweFrenchTAG the grammar and lexicon classes which were necessary to cover most VMWEs and their syntactic alternations occurring in DEV-S (additionally to a dozen of previously encoded examples). The new version of the metagrammar was automatically re-compiled to LTAG by the XMG compiler and the resulting LTAG grammar was used to parse the sentences from DEV-S with the Tulipa parser (cf. Sect. 4).

As shown in Table 1, after this phase, we observed an 18% increase in the number of XMG classes covering the grammar (i.e. not the lexicon), which is substantial but expected, since this phase included re-engineering the class hierarchy to make it MWE-aware (cf. Sect. 6).

In phase 2, we repeated the same actions for TEST-S. Only 4 new classes had to be added in order to cover the sentences from TEST-S. These classes cover:

[13] The original and simplified datasets, and the simplification rules, are available online at https://gitlab.com/agata.savary/mwe-xmg/evaluation

[14] http://unitexgramlab.org/

- a syntactic variant of the impersonal expression ***il y a*** 'it there has' ⇒ 'there is' expression, namely the case when the object of a 'has' is a common noun with no determiner (***il y a*** *contrôle* 'there is control');
- two syntactic variants of the impersonal expression ***il faut*** 'it must' ⇒ 'it is necessary', namely those where the verb *faut* 'must' takes a non-sentential, possibly clitic, object (***il*** *la* ***faut*** 'it she must' ⇒ 'she is necessary') or a genitive object (***il faut*** *de la pratique* 'it must of the practice' ⇒ 'practice is necessary');
- the causative reflexive construction (*Jean* ***se fait*** *tuer par Marie* 'Jean himself makes kill by Marie' ⇒ 'Jean gets killed by Marie').

These four new classes were defined as conjunctions of pre-existing FrenchTAG classes (i.e. not specific to MWEs). The overall increase in the number of classes between phases 1 and 2 is equal to 1.1% only. We expect this increase to be even lower in the next encoding iterations, as new MWEs should exhibit increasingly similar syntactic structures and properties to those already encoded. Experimental large-scale validation of this hypothesis is left to future work.

	FrenchTAG	mweFrenchTAG with MWEs from	
		DEV-S	DEV-S + TEST-S
classes	285	337 (+18%)	341 (+1.1%)
MWE lemmas	5	31 (+520%)	37 (+19%)

Table 1. The growth (in terms of the number of classes) of FrenchTAG as a result making it MWE-aware.

Some examples in DEV-S and TEST-S have not been successfully encoded so far. Example (39) contains a regular extraction whose type seems not to be covered by the original FrenchTAG. The MWEs in (44) and (69) are reflexive but have a behavior of a copula and their optimal representation requires more insight. Finally, the encoding of the MWE in (42), (43), (66) and (67) is partly satisfactory since the implementation of a compulsory but non-lexicalized modifier has not yet been solved in our approach. Note that the ratio of not (or partly) encoded MWEs decreases between phases 1 and 2.

8. Conclusions and perspectives

There is a fundamental tension between the flexibility of MWEs, that is their unforeseeable but possibly recurring "irregularities", and lexical encoding systems which ought to be denotationally precise and at the same time maximally theory-independent. In this paper, we have argued that XMG incorporates both the necessary flexibility and denotational rigor. To this end, we have shown how to extend the metagrammar underlying FrenchTAG, a large-coverage grammar of French, by grammar-specific classes to which the rather grammar-agnostic descriptions of MWEs can be linked. By evaluation

against a MWE-annotated reference corpus, we noticed a considerable drop in the growth of grammar size during incremental implementation, which seems to suggest that the redundancy of MWE is efficiently captured.

The methodology of making FrenchTAG MWE-aware, in brief, is based on 3 principles: (i) reusing pre-existing classes for (more) regular properties and augmenting some of them with interface filters, (ii) creating new classes for lexicalized arguments of various syntactic structures, (iii) creating MWE lexicon entries with co-anchors and interface filters constraining the syntactic alternations. The objective in this experiment was to keep the original FrenchTAG rules intact. It would, however, be interesting to intensively apply interface FSs to the whole metagrammar, including the classes covering regular rather than idiosyncratic behaviour. This might allow an even more compact and maintainable resource.

The current version of mweFrenchTAG is meant as a proof of concept that MWE encoding can be done both with high precision and with little redundancy. Thus, we currently cover only 37 verbal MWEs having the most frequent syntactic structures. The next steps are to manually encode new examples, so as to cover the large diversity of the syntactic structures occurring in French VMWEs. This work could be based on pre-existing linguistic analyses like (Pausé, 2017).

There are other work packages left for future work: (i) to extend the approach to challenging phenomena such as co-indexation; (ii) to make the XMG compiler and LTAG parser also handle compulsory or prohibited modification of anchor or co-anchor nodes; (iii) to allow for co-anchoring lemmas rather than fixed inflected forms; (iv) to introduce intermediate classes into the original hierarchy from Figure 3 which directly account for more co-anchoring phenomena.[15]

While verbal MWEs are particularly challenging (Savary et al., 2018), it goes without saying that MWEs of other types need to be covered as well. Two cases can be distinguished. First, predicative MWEs: nominal (***bras droit*** *de quelqu'un* 'someone's right arm' ⇒ 'someone's right-hand man'), adjectival (***inversement proportionnelle*** *à quelque chose* 'inversely proportional to sth'), adverbial (***à l'opposé*** *de quelque chose* 'in the opposite to sth' ⇒ 'contrary to'), etc., can be covered following the same methodology as with verbal MWEs, except that less morphosyntactic variants will have to be covered. We expect most classes already designed for lexicalized arguments in VMWEs to be reusable here. Second, other, non-predicative MWEs, nominal (***petit ami*** 'small friend' ⇒ 'boyfriend'), adjectival (***ivre-mort*** 'drunk-dead' ⇒ 'dead drunk'), adverbial (***à petit pas*** 'at small steps' ⇒ 'slowly'), etc. will have to be accounted for. Those can probably be handled by reusing metagrammar classes

[15] These classes should represent tree fragments in which non-terminal leaves can either be substitution nodes or unifiable with subtrees representing lexicalized arguments. In this way, redundancy could be avoided e.g. between tree fragments from Figure 6(b) and Figure 8(a).

like `mweLexDetLexNoun`, `mweNoDetLexNoun` and `LexNounLexAdj` in Listing 6, and by new classes covering more syntactic structures and following the same principles.

Once such a metagrammar with representative examples of various syntactic types of MWEs is operational, a lexicon could be induced automatically from corpora, both MWE-annotated and raw, and from machine readable MWE lexicons and lists. The objective would be to develop new generation lexicon induction methods in which MWEs are discovered together with (at least part of) their morpho-syntactic properties, as suggested by Savary, Cordeiro and Ramisch (2019a). These properties could be expressed with features and values like those appearing in the mweFrenchTAG lexicon entries in Listing 7. The advantage of coupling automatic discovery with a metagrammar would be to simultaneously ensure a large coverage of the MWE lexicon and a complete formalisation of the property labels used in this lexicon, as suggested by Lichte et al. (2019).

Note that, like FrenchTAG, our metagrammar does not account for semantics. For instance, it is impossible to check whether the subject of ***prendre la porte*** 'to take the door' ⇒ 'to be forced to leave' is an animated entity or not. However, XMG provides a semantic dimension, which allows to specify semantic constraints, for example on the semantic arguments of verbs (Lichte & Petitjean, 2015). The semantic dimension could also help represent morpho-syntactic constraints imposed by MWEs. For instance, the subject-possessive agreement in examples (3)-(4) might be modelled as the effect of inheriting the constraints stemming from inalienable possession (Guéron, 2003) expressed in literal interpretations of these MWEs. The extension of mweFrenchTAG with a semantic dimension is one of our perspectives.

Finally, what needs to be done in the long run is to investigate how the MWE descriptions we propose can be reused with grammars from other frameworks, for example HPSG or LFG.

Bibliography

Abeillé, A., & Schabes, Y. (1989). Parsing idioms in lexicalized TAGs. In H. L. Somers & M. McGeeWood (Eds.), *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester* (pp. 1–9). https://www.aclweb.org/anthology/E89-1001.pdf.

Abeillé, A., & Schabes, Y. (1996). Non-compositional discontinuous constituents in Tree Adjoining Grammar. In H. Bunt & A. van Horck (Eds.), *Discontinuous Constituency* (pp. 279–306). Berlin: Mouton de Gruyter. .

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

Al-Haj, H., Itai, A., & Wintner, S. (2014). Lexical Representation of Multiword Expressions. In *Morphologically-complex Languages. International Journal of Lexicography*, *27*(2), 130–170. Oxford University Press. https://doi.org/10.1093/ijl/ect036.

Arps, D., & Petitjean, S. (2018). A parser for LTAG and frame semantics. In N. Calzolari (Conference chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 2223–2229). Miyazaki, Japan. European Language Resources Association (ELRA) https://www.aclweb.org/anthology/L18-1351.pdf.

Attia, M. A. (2006). Accommodating multiword expressions in an Arabic LFG grammar. In T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (Eds.), *Proceedings of the 5th international conference on Advances in Natural Language Processing, FinTAL'06* (pp. 87–98). Berlin: Springer. https://doi.org/10.1007/11816508_11.

Bond, F., Ho, J. Q., & Flickinger, D. (2015). Feeling our way to an analysis of English possessed idioms. In S. Müller (Ed.), *Proceedings of the 22nd International Conference on Head- Driven Phrase Structure Grammar* (pp. 61–74). Stanford, CA: CSLI Publications. http://web.stanford.edu/group/cslipublications/cslipublications/HPSG/2015/bhf.pdf.

Breidt, E., Segond, F., & Valetto, G. (1996). Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In *Proceedings of COLING-96, Copenhagen* (pp. 1036–1040). https://www.aclweb.org/anthology/C96-2182.pdf.

Candito, M.-H. (1999). *Organisation modulaire et paramétrable de grammaires électroniques lexicalisées*. Ph.D. Thesis, Université Paris 7.

Constant, M., & Tolone, E. (2010). A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In M. De Gioia (Ed.), *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie, volume 1 of Lingue d'Europa e del Mediterraneo, Grammatica comparata* (pp. 79–93). Aracne, April. ISBN 978-88-548-3166-7.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword Expression Processing: A Survey. *Computational Linguistics*, *43*(4), 837–892. MIT Press. https://www.aclweb.org/anthology/J17-4005.pdf

Copestake, A., Lambeau, F., Villavicencio, A., Bond, F., Baldwin, T., Sag, I. A., & Flickinger, D. (2002). Multiword expressions: linguistic precision and reusability. In *Proceedings of LREC 2002* (pp. 1941–1947). Las Palmas, Canary Islands – Spain. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2002/pdf/145.pdf

Crabbé, B., Duchier, D., Gardent, C., Le Roux, J., & Parmentier, Y. (2013). XMG: extensible metagrammar. *Computational Linguistics, 39*(3), 591–629. MIT Press. https://doi.org/10.1162/COLI_a_00144

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

Dyvik, H., Losnegaard, G. S., & Rosén, V. (2019). Multiword expressions in an LFG grammar for Norwegian. In Y. Parmentier & J. Waszczuk (Eds.), *Representation and parsing of multiword expressions: Current trends* (pp. 69–107). Language Science Press: Berlin. https://doi.org/10.5281/zenodo.2579036.

Gardent, C. (2008). Integrating a Unification-Based Semantics in a Large Scale Lexicalized Tree Adjoining Grammar for French. In D. Scott & H. Uszkoreit (Eds.), *22nd International Conference on Computational Linguistics, Proceedings of the workshop on Cognitive Aspects of the Lexicon: 24 August 2008, Manchester, UK* (pp. 249–256). New York, NY: Association for computing machinery. https://dl.acm.org/doi/10.5555/1599081.1599113.

Grégoire, N. (2010). DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, *44,* 23–39. https://doi.org/10.1007/s10579-009-9094-z.

Gross, M. (1986). Lexicon-grammar: The Representation of Compound Words. In *Proceedings of the 11th Conference on Computational Linguistics, COLING'86* (pp. 1–6). Stroudsburg, PA, USA: Association for Computational Linguistics. https://www.aclweb.org/anthology/C86-1001.

Gross, G. (1988). Degré de figement des noms composés. *Les expressions figées, sous la direction de Laurence Danlos. Langages, 90*, 57–71. https://doi.org/10.3406/lgge.1988.1991.

Guéron, J. (2003). Inalienable possession and the interpretation of determiners. In M. Coene & Y. D'hulst (Eds), *From NP to DP, volume 2 of Linguistik aktuell* (pp. 189–220). Amsterdam: J. Benjamins Pub. https://doi.org/10.1075/la.56.13gue.

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

Jacquemin, C. (2001). *Spotting and Discovering Terms through Natural Language Processing*. MIT Press.

Joshi, A. K., & Schabes, Y. (1997). Tree-adjoining grammars. In G. Rozenberg & A. Salomaa (Eds.), *Handbook of Formal* Languages (Vol. 3: *Beyond Words*; pp. 69–123). Berlin: Springer.

Karttunen, L., Kaplan, R. M., & Zaenen, A. (1992). Two-Level Morphology with Composition. In *Proceedings of COLING-92, Nantes* (pp. 141–148). France: ICCL.

Lichte, T., & Kallmeyer, L. (2016). Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions. In C. Piñón (Ed.), *Empirical Issues in Syntax and Semantics 11* (pp. 111–140). Paris: CSSP.

Lichte, T., & Petitjean, S. (2015). Implementing semantic frames as typed feature structures with XMG. *Journal of Language Modelling, 3*(1), 185–228. http://dx.doi.org/10.15398/jlm.v3i1.96.

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

Losnegaard, G. S., Sangati, F., Parra Escartín, C., Savary, A., Bargmann, S., & Monti, J. (2016). PARSEME Survey on MWE Resources. In N. Calzolari (Conference Chair), K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 2299–2306), Paris, France. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1364.pdf.

McShane, M., Nirenburg, S., & Beale, S. (2015). The Ontological Semantic treatment of multiword expressions. *Lingvisticæ Investigationes, 38*(1), 73–110. https://doi.org/10.1075/li.38.1.03mcs.

Melčuk, I., Arbatchewsky-Jumarie, N., Dagenais, L., Elnitsky, L., Iordanskaja, L., Lefebvre, M.-N., & Mantha, S. (1988). *Dictionnaire explicatif et combinatoire du français contemporain : Recherches lexico-sémantiques. II : Recherches lexico-sémantiques*. Montréal : Presses de l'Université de Montréal.

Nunberg, G., Sag, I. A., & Wasow, T. (1994). Idioms. In S. Everson (Ed.). *Language, 70*(3), 491–538.

Oflazer, K., Çetinoglu, Ö., & Say, B. (2004). Integrating Morphology with Multi-word Expression Processing in Turkish. In *Second ACL Workshop on Multiword Expressions, July 2004* (pp. 64–71). Barcelona, Spain. Association for Computational Linguistics.

Parmentier, Y., Kallmeyer, L., Maier, W., Lichte, T., & Dellert, J. (2008). TuLiPA: A syntax-semantics parsing environment for mildly context-sensitive formalisms. In *Proceedings of the Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+9)* (pp. 121–128). Tübingen, Germany.

Patejuk, A. (2015). *Unlike coordination in Polish: an LFG account*. Ph.D. Dissertation, Institute of Polish Language, Polish Academy of Sciences, Cracow.

Pausé, M.-S. (2017). *Structure lexico-syntaxique des locutions du français et incidence sur leur combinatoire*. Ph.D. Thesis, Université de Lorraine, Nancy, France.

Petitjean, S., Duchier, D., & Parmentier, Y. (2016). XMG 2: Describing description languages. In Amblard, M., de Groote, P., Pogodalla, S. & Retoré, C. (Eds.), *Logical Aspects of Computational Linguistics. Celebrating 20 Years of LACL (1996-2016) – 9th International Conference, LACL 2016* (pp. 255–272). Nancy, France. Berlin: Springer. https://doi.org/10.1007/978-3-662-53826-5_16.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., & 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)* (pp. 83–91), Dublin,

Ireland: Association for Computational Linguistics and Dublin City University. https://www.aclweb.org/anthology/W14-5811.pdf.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002) Multiword Expressions: A Pain in the Neck for NLP. In A. Gelbukh (Eds.) *Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science* (vol. 2276; pp 1–15). Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45715-1_1.

Savary, A., & Cordeiro, S. (2018). Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT 16), Jan 2018, Prague, Czech Republic* (pp. 64–72). https://www.aclweb.org/anthology/W17-7610.pdf.

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

Savary, A., Cordeiro, S., & Ramisch, C. (2019a). Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)* (pp. 79–91). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W19-5110.

Savary, A., Cordeiro, S., Lichte, T., Ramisch, C., Inurrieta, U., & Giouli, V. (2019b). Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. *The Prague Bulletin of Mathematical Linguistics*, *112*, 5–54.

Savary, A. (2008). Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology, 1*(2), 1–53. CSLI Publications.

Savary, A. (2009). Multiflex: A Multilingual Finite-State Tool for Multi-Word Units. In S. Maneth (Ed.), *Implementation and Application of Automata, volume 5642 of Lecture Notes in Computer Science* (pp. 237–240). Berlin: Springer. https://doi.org/10.1007/978-3-642-02979-0_27.

Silberztein, M. (2005). NooJ's dictionaries. In *2nd Language & Technology Conference (LTC'05), Poznań* (pp. 291–295).

Vaidya, A., Rambow, O., & Palmer, M. (2014). Light verb constructions with 'do' and 'be' in Hindi: A TAG analysis. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing* (pp. 127–136). Association for Computational Linguistics and Dublin City University. http://dx.doi.org/10.3115/v1/W14-5816.

Vijay-Shanker, K., & Joshi, A. K. (1988). Feature Structures Based Tree Adjoining Grammar. In *Proceedings of COLING* 88 (pp. 714–719). Budapest, Hungary.

Villavicencio, A., Copestake, A., Waldron, B., & Lambeau, F. (2004). Lexical Encoding of MWEs. In *ACL Workshop on Multiword Expressions: Integrating Processing, July 2004* (pp. 80–87). Association for Computational Linguistics. https://www.aclweb.org/anthology/W04-0411.

XTAG Research Group. (2001). *A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS-01-03*. Philadelphia, Pa.: University of Pennsylvania.

Appendix

A. Complete example of encoding and parsing a sentence with a MWE

```
class LemJean {
<lemma> {
 entry <- "Jean";
 cat <- n;
 fam <- propn} }
class LemPrendre {
<lemma> {
 entry <-"prendre";
 cat <- v;
 fam <- n0Vn1} }
class LemPorte{
<lemma> {
 entry <- "porte";
 cat <- n;
 fam <- noun } }
```

```
class LemIl {
<lemma> {
 entry <- "il";
 cat <- cl;
 fam <- CliticT}}
class LemLe {
<lemma> {
 entry <- "le";
 cat <- d;
 fam <- stddeterminer
}}
class LemLeClitic {
<lemma> {
 entry <- "le";
 cat <- cl;
 fam <- CliticT}}
```

```
class mweLemmePrendreLaPorte {
<lemma> {
 entry <- "prendre";
 cat   <- v;
 fam   <- mwen0Vn1;
 filter dia = active;
 filter subj = free;
 filter obj = lex;
 filter objtype = canonical;
 filter objstruct = lexDetLexN;
 coanchor ObjDetNode -> "la"/d;
 coanchor ObjNode -> "porte"/n;
 equation ObjNode -> gen=f;
 equation ObjNode -> num=sg }}
```

```
class Jean {
<morpho> {
 morph <- "Jean";
 lemma <- "jean";
 cat <- n }}
class il {
<morpho> {
 morph <- "il";
 lemma <- "il";
 cat <- cl }}
```

```
class prend {
<morpho> {
 morph <- "prend";
 lemma <- "prendre";
 cat   <- v }}
class la {
<morpho> {
 morph <- "la";
 lemma <- "le";
 cat   <- d }}
```

```
class porte {
<morpho> {
 morph <- "porte";
 lemma <- "porte";
 cat   <- n}}
class laClitic {
<morpho> {
 morph <- "la";
 lemma <- "le";
 cat   <- cl }}
```

Listing 8. Extract of the (simplified) mweFrenchTAG lexicon with 7 lemmas and 6 inflected forms.

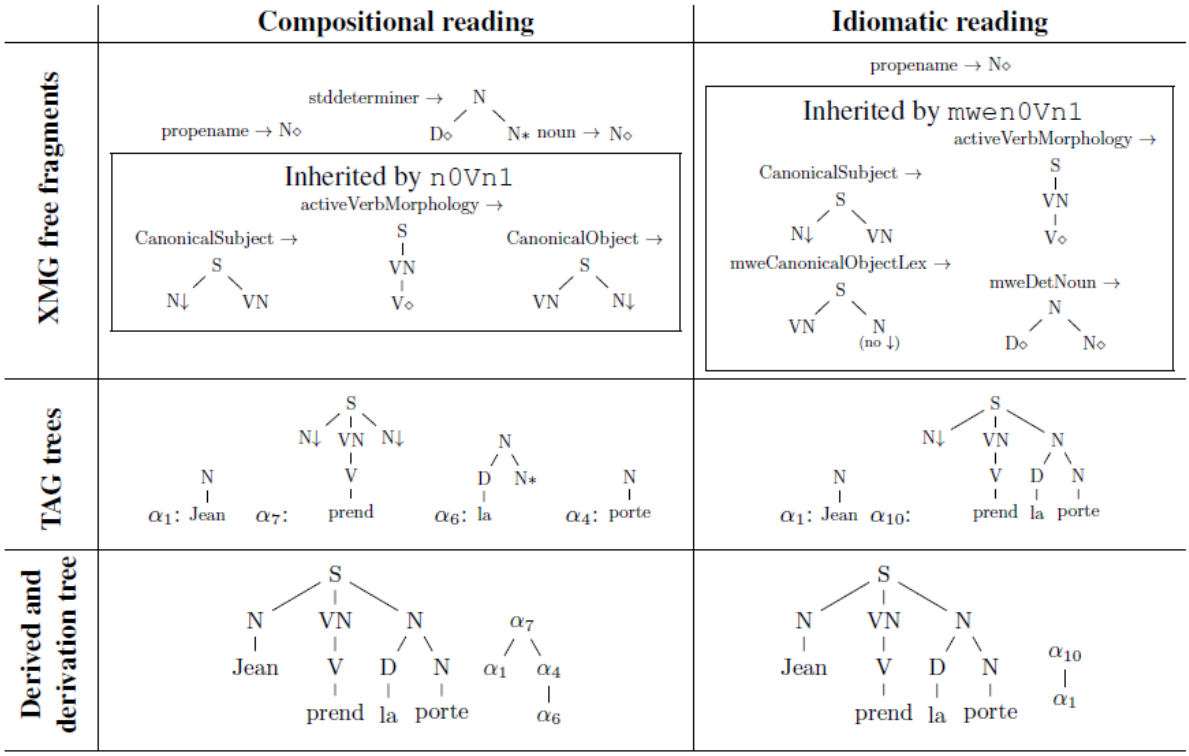

Table 2. Two syntactic analyses of the sentence *Jean **prend la porte*** 'Jean takes the door' ⇒ 'Jean leaves because he is forced to' with a compositional and an idiomatic reading.

B. Sentences from the evaluation corpus

DEV-S:

(20) *la unification **avait lieu** de être* 'the unification had place of be' ⇒ 'the unification had good reasons to take place'

(21) *une évolution **avait** alors **lieu*** 'an evolution had then place' ⇒ 'then an evolution took place'

(22) *ils **faisaient appel** à des charpentiers* 'they made appeal to carpenters' ⇒ 'they appealed to carpenters'

(23) *les pilotes **faisaient appel*** 'the pilots made appeal' ⇒ 'the pilots appealed'

(24) *je **faisais appel** à eux* 'I did appeal to them' ⇒ I appealed to them'

(25) *ils **faisaient appel*** 'they did appeal' ⇒ 'they appealed'

(26) *les commandants **faisaient** certainement **face** à un dilemme* 'the commanders made certainly face to a dilemma' ⇒ 'the commanders certainly had to deal with a dilemma'

(27) *elles **faisaient le objet** de un ordre* 'they did the object of an order' ⇒ 'they were the subject of an order'

(28) *ils **feront le objet** d'une campagne* 'they will do the object of a campaign' ⇒ 'they will be the object of a campaign'

(29) *le service **faisait** alors **le objet** de les critiques* 'the service did then the object of criticisms' ⇒ 'the service was subject to criticism'

(30) *elle **faisait partie** de son territoire* 'it did part of its territory' ⇒ 'it was part of its territory'

(31) *les célébrités ont **fait partie** de le casting* 'the celebrities did part of the casting' ⇒ 'the celebrities took part of the casting'

(32) ***il faut** le dire* 'it should it say' ⇒ 'it should be said'

(33) ***il faut*** *choisir un nom* 'it should choose a name' ⇒ 'one should choose a name'
(34) ***il se agit*** *de avoir une garantie* 'it itself acts of have a guarantee' ⇒ 'the point is to have a guarantee'
(35) ***il se agit*** *certainement de un organe* 'it itself acts certainly of an organ' ⇒ 'an organ is certainly concerned'
(36) ***il se agit*** *de le tube* 'it itself acts of the hit' ⇒ 'the hit is concerned'
(37) ***il se agissait*** *alors de colorier* 'it itself acted of color' ⇒ 'coloring was needed'
(38) ***il y a*** *de la concurrence* 'it there has of the competition' ⇒ 'there is competition'
(39) *mis* ***en examen****, Jean a été suspendu* 'put under investigation, Jean has been suspended' ⇒ 'Jean was suspended after having been indicted'
(40) *Jean est mis* ***en examen*** 'Jean is put in exam' ⇒ 'Jean is indicted'
(41) *les personnes qui sont* ***mises en examen*** 'the persons who are put in exam' ⇒ 'the persons who are indicted'
(42) *le saint dont elle* ***porte*** *le* ***nom*** 'the saint whose it bears the name' ⇒ 'the saint whose name it bears'
(43) *Paris qui* ***porte*** *aujourd'hui son* ***nom*** 'Paris which bears today his name' ⇒ 'Paris which bears its name today'
(44) *les visions* ***se font*** *intenses* 'the visions make themselves intense' ⇒ 'the visions become intense'
(45) *son tombeau* ***se trouve*** *ici* 'his tomb itself finds here' ⇒ 'his tomb is here'

TEST-S:

(46) *elle* ***avait lieu*** *ici* 'it had place here' ⇒ 'it took place here'
(47) *les championnats* ***ont lieu*** *ici demain* 'the championships have place here tomorrow' ⇒ 'the championships take place here tomorrow'
(48) *les figurines qui se* ***font face*** 'figures which themselves do face' ⇒ 'figures which face each other'
(49) *la charte qui doit* ***faire face*** *à le impérialisme* 'the charter which must do face to the imperialism' ⇒ 'the charter which must face the imperialism'
(50) *un buste qui* ***fait face*** *à le spectateur* 'a bust which does face to the spectator' ⇒ 'a bust which faces the spectator'
(51) *le futur* ***fait le objet*** *de un débat* 'the future makes the object of a debate' ⇒ 'the future is subject to debate'
(52) *elle* ***fait partie*** *de le groupe* 'she makes part of a group' ⇒ 'she is part of a group'
(53) *un monument qui* ***fait partie*** *de les sites* 'a monument which makes part of the sites' ⇒ 'a monument which is part of the sites'
(54) ***il*** *en* ***faut*** 'it thereof needs' ⇒ 'it is needed'
(55) ***il*** *la* ***faut*** 'it her needs' ⇒ 'she is needed'
(56) ***il y a*** *contrôle* 'it there has control' ⇒ 'there is control'
(57) *le* ***rôle*** *qui est* ***joué*** *par le personnage* 'the role which is played by the character'
(58) *la religion* ***joue*** *un* ***rôle*** *majeur* 'the religion plays a role major' ⇒ 'religion plays a major role'
(59) *des macrophytes* ***jouent*** *le* ***rôle*** *de barrière* 'macrophytes play the role of barrier' ⇒ 'macrophytes play the role of a barrier'
(60) *il* ***joue*** *un* ***rôle*** *important* 'he plays a role important' ⇒ 'he plays an important role'
(61) *un policier était* ***mis en examen*** *ici* 'a policeman was put in exam here' ⇒ 'the policeman was indicted here'

(62) *les habitants veulent y* ***mettre fin*** 'the inhabitants want there put end' ⇒ 'the inhabitants want to put an end to it'

(63) *un dialogue qui* ***met fin*** *à la occupation* 'the dialogue which puts end to the occupation' ⇒ 'the dialogue which puts an end to the occupation'

(64) *il fait* ***mettre fin*** *à cette pratique* 'he makes put end to this practice' ⇒ 'he has the practice stopped'

(65) *il décide de* ***mettre fin*** *à cette anarchie* 'he decides to put end to this anarchy' ⇒ 'he decides to put an end to this anarchy'

(66) *il verra le puits* ***porter*** *son* ***nom*** 'he will-see the well bear his name' ⇒ 'the well will bear his name'

(67) *la école* ***porte*** *son* ***nom*** *maintenant* 'the school bears his name now'

(68) *Jean* ***se fait*** *tuer par Marie* 'Jean himself makes kill by Mary' ⇒ 'Jean gets killed by Mary'

(69) *le terrain* ***se trouve*** *pris sous le feu* 'the site itself finds taken under the fire' ⇒ 'the site finds itself under fire'

(70) *Jean* ***se trouve*** *ici* 'Jean himself finds here' ⇒ 'Jean is here'

(71) *il est incroyable de* ***se trouver*** *ici* 'it is incredible of oneself find here' ⇒ 'it is incredible to be here'