Expressions polylexicales verbales : étude de la variabilité en corpus

Caroline PASQUER[1]
(1) LI - Laboratoire d'Informatique de l'Université de Tours
BDTLN - Bases de données et traitement des langues naturelles
caroline.pasquer@etu.univ-tours.fr

RÉSUMÉ

La reconnaissance et le traitement approprié des expressions polylexicales (EP) constituent un enjeu pour différentes applications en traitement automatique des langues. Ces expressions sont susceptibles d'apparaître sous d'autres formes que leur forme canonique, d'où l'intérêt d'étudier leur profil de variabilité. Dans cet article, nous proposons de donner un aperçu de motifs de variation syntaxiques et/ou morphologiques d'après un corpus de 4441 expressions polylexicales verbales (EPV) annotées manuellement. L'objectif poursuivi est de générer automatiquement les différentes variantes pour améliorer la performance des techniques de traitement automatique des EPV.

ABSTRACT

Verbal MWEs : a corpus-based study of variability

Identification and processing of multiword expressions (MWEs) are required by many natural language processing applications. MWEs can appear in other forms than their canonical form, hence the interest of studying their variability profile. In this paper, we present an overview of syntactic and / or morphological variation patterns based on a corpus of 4441 manually annotated verbal MWEs. Our goal is to automatically generate variants to improve the performance of MWE automatic processing techniques.

MOTS-CLÉS : Expressions polylexicales, figement, variabilité.

KEYWORDS: Multiword expressions, MWE, variability.

1 Introduction

Les expressions polylexicales (EP ou *multiword expressions, MWE* en anglais) sont des groupements de mots dont le sens individuel "ne permet pas d'interpréter l[a] combinaison" (Gross, 1982) comme dans l'expression *casser du sucre sur le dos* où il n'est pas plus question de *sucre* que de *dos*. Cette spécificité (nommée *non-compositionnalité*) pose problème pour différentes applications de traitement automatique des langues (par exemple l'impossibilité de traduction automatique "mot à mot" : *#to break sugar on one's back* [1]).

La détection des EP est une question d'autant plus importante qu'elles sont très fréquentes – 40 % des mots du journal *Le Monde* font partie d'EP (Gross & Senellart, 1998). Jackendoff (1997) estime d'ailleurs que chaque langue contient autant d'EP que de mots isolés. De plus, il est impossible

1. Nous adoptons la convention suivante : le symbole # désigne un glissement de sens : *prendre congé* (= *dire au revoir*) versus # *prendre des congés* (= *des vacances*) ; l'astérisque * indique une agrammaticalité (**il faut que je viens*).

d'en établir une liste exhaustive car de nouvelles expressions apparaissent sans cesse, qu'il s'agisse d'entités nommées polylexicales (noms de personnes, d'organisations, etc.) ou de néologismes obtenus en réutilisant le lexique disponible (*canon à électron* cité par Gross (1982)). Baldwin *et al.* (2004) indiquent qu'un défaut de détection des EP est responsable de 8 % d'erreurs de parsing. La prise en compte du caractère évolutif des EP justifie les nombreux travaux consacrés à leur détection automatique.

On peut supposer que, si la variabilité des EP est qualitativement et quantitativement hétérogène, cette variabilité est plus élevée pour les EP verbales (EPV) dont les emplois peuvent différer de leur forme canonique (répertoriée dans un lexique par exemple) en raison de discontinuités, passivation, etc. : ***adresser** des **reproches** → des **reproches** semblables ont d'abord été **adressés** à certaines sectes*.

Le système le plus performant en matière de détection d'EPV parmi les sept testés dans le projet PARSEME (décrit en section 4) ne reconnaît que 55 % de variantes d'EPV déjà vues durant l'entraînement, soit presque 2 fois moins que pour des formes de surface identiques (90 % détectées). On émet l'hypothèse que la description des motifs de variabilité des EPV permettra le rapprochement de variantes quelque soit cette forme de surface. De plus, la connaissance des variantes possibles mettra en lumière les accords interdits dans les EPV afin de différencier une lecture littérale d'une lecture idiomatique (*je **vide** mon **sac/#ton sac***).

Ainsi, cet article a pour objet l'analyse et la caractérisation de la variabilité (morphologique et/ou syntaxique) d'EPV attestées en corpus. Cette analyse pourra être mise à profit pour différentes applications : génération de variantes à partir de la forme canonique ou, à l'inverse, fusion de variantes avec leur forme canonique. Ces deux tâches permettront d'améliorer la couverture des systèmes de détection d'EP. D'autres applications sont possibles telles que l'annotation automatique d'EP dans des corpus arborés ou dans des sorties de parseurs, l'*entity linking* ou la résolution de coréférences. La mise en évidence de spécificités de certains types d'EPV (constructions à verbe support par exemple) pourrait aussi permettre un traitement sémantique adéquat.

Nous revenons (section 2) sur la définition des EP et faisons un état de l'art sur leur variabilité (section 3). Le corpus est présenté en section 4 et la méthodologie d'exploitation des données en section 5. Nous détaillons la variabilité interne (morphologique) des EPV en section 6 et la variabilité externe (insertions) en section 7. Enfin, nous nous intéressons en section 8 aux dépendances syntaxiques dans des expressions de type verbe-(déterminant)-nom (abrégées sous la forme VB-(DET)-NOM) pour en extraire des profils de variation.

2 Définition et classification des EP

Malgré les controverses liées à la définition des EP (Villavicencio *et al.*, 2007; Savary, 2008), Savary (2008) observe que les linguistes s'accordent pour qualifier d'EP les expressions satisfaisant les 3 critères suivants :

— le fait de contenir au moins 2 unités lexicales ("mots"),
— dénotation unique et constante,
— idiosyncrasie morphologique, syntaxique, distributionnelle ou sémantique.

Sag *et al.* (2002) distinguent ainsi les **expressions institutionnalisées** des **phrases lexicalisées**. Les phrases institutionnalisées (ou collocations) ne satisfont pas le critère de non-compositionnalité et se caractérisent par la tendance des locuteurs à associer certains mots de façon préférentielle (par exemple : *aimer à la folie*).

Sag *et al.* (2002) analysent la variabilité (nommée *flexibilité* dans leur terminologie) des phrases lexicalisées – autrement dit des EP – en les classant par ordre de flexibilité croissante :

- Les **expressions figées** (par exemple *revenons à nos moutons*) n'admettent aucune variation morphosyntaxique : impossibilité d'insertion ou de flexion (#*revenez encore à mon mouton*).
- **Expressions semi-figées** : ces expressions présentent des contraintes strictes quant à l'ordre des mots ou leur composition mais autorisent certaines variations telles que le choix du déterminant ou la flexion temporelle : ***je tire / tirerai mon épingle du jeu***. Sag *et al.* (2002) utilisent le principe de compositionnalité sémantique défini par Nunberg *et al.* (1994) – le sens global d'une expression est lié à celui des éléments qui la composent – pour distinguer les idiomes sémantiquement décomposables et non-décomposables. Les *idiomes non-décomposables* (Sag *et al.*, 2002) font partie des expressions semi-figées et se caractérisent par leur opacité sémantique. Cette opacité impliquerait une variabilité limitée, par exemple une impossibilité de passivation : *#son épingle a été tirée du jeu*.
- **Expressions syntaxiquement flexibles** : figurent notamment ici les constructions à verbe support et les idiomes décomposables par équivalence sémantique (***spill the beans*** → *spill = reveal ; beans = secret(s)*). La distinction idiomes décomposables/non-décomposables est cependant remise en question (Abeillé, 1988; Sheinfux *et al.*, 2017) en raison de contre-exemples d'idiomes non-décomposables et pourtant flexibles.

Malgré ces contre-exemples, l'intérêt de la classification de Sag *et al.* (2002) est de considérer la flexibilité (ou variabilité) comme un critère essentiel de classification des EP. Nous adoptons la terminologie utilisée lors de la campagne d'annotation [2] du corpus PARSEME pour 18 langues (Savary *et al.*, 2017), selon laquelle les constituants requis et non substituables d'une EP sont qualifiés d'éléments *lexicalisés*. Ils seront signalés en caractères italiques gras dans les exemples. Nous adoptons également la typologie des EPV couvrant :

- Les constructions à verbe support, désormais abrégées **LVC**, de l'anglais *Light Verb Construction* : leur particularité réside dans le fait que ce n'est pas le verbe qui remplit la fonction de prédicat de la phrase, mais un nom prédicatif (***faire allusion***).
- Les idiomes, désormais abrégés **ID** : leur sens n'est pas compositionnel et ils possèdent souvent une double lecture littérale / idiomatique (***tourner la page***) nécessitant une désambiguïsation contextuelle. La classification des EPV dans le corpus relève d'un accord entre les 18 langues participant au projet et ne coïncide pas systématiquement avec celle traditionnellement admise (par exemple ***faire partie*** est un ID et non pas une LVC).
- Les verbes intrinsèquement réflexifs, désormais abrégés **IReflV**, de l'anglais *Inherently Reflexive Verbs*. Ils satisfont l'une des trois conditions suivantes : soit ils possèdent un sens différent du verbe seul (hors pronom), comme ***se rendre*** / *rendre*, soit ils n'existent pas sans *se* : ***se prélasser*** / **prélasser*, soit – tout en conservant le même sens – ils possèdent un cadre de sous-catégorisation spécifique (***se confesser*** *de X* / *confesser X*).

3 État de l'art sur la variabilité des EP

Quoique des travaux aient porté sur la question des EP dès les années 1980 (Gross, 1982), Sag *et al.* (2002) l'ont remise sur le devant de la scène avec leur classification des EP. Dès lors, les EP ont été examinées sous différents points de vue, qu'il s'agisse de la nature des EP (nominales, verbales, etc.) ou de certains types de variations : morphologiques (Nissim & Zaninello, 2013), syntaxiques ou sémantiques (Jacquemin, 2001). Ces différentes approches répondent à des objectifs différents :

2. Guide d'annotation PARSEME : `http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext/`

— La **description lexicale** : Lichte *et al.* (2017) ont répertorié différents formalismes aptes à rendre compte des variantes des EPV. Le codage Walenty a par exemple permis de constituer le *Polish Valence Dictionary* comportant 1774 verbes. 8000 EPV y sont décrites avec leur cadre valenciel et leurs motifs de variabilité. Ce formalisme nécessite des ajustements pour couvrir la diversité des variantes possibles (Przepiórkowski *et al.*, 2014) surtout si on souhaite l'extrapoler à une langue différente du polonais pour lequel il a été initialement conçu (par exemple : absence de déclinaisons en français mais présence de déterminants).
— L'**analyse de variabilité en corpus** : Nissim & Zaninello (2013) se sont intéressés à la variation interne d'EP nominales en italien et proposent une méthode de recherche flexible grâce à des patrons de variation pour améliorer la détection des EP. Pour le français, Tutin (2016) a établi des niveaux de variabilité pour les 30 EPV (ID et LVC) les plus fréquentes de type VB-(DET)-NOM (détails en section 8).
— La **prise en compte de la variabilité pour les applications** : dans le cas d'EP nominales, l'outil FASTER (Savary & Jacquemin, 2003) permet la reconnaissance de termes appartenant à une liste pré-établie en détectant les variantes des termes en corpus et en extrayant les variations par des métarègles morpho-syntaxiques et syntaxico-sémantiques.

Notre travail s'inscrit dans la continuité de ces approches en visant une description en corpus de la variabilité des EPV.

4 Corpus

Le corpus utilisé provient de l'identification manuelle d'EPV effectuée dans le cadre du projet PARSEME [3] (*PARSing and Multi-word Expressions*). Un guide d'annotation des EPV a été élaboré pour 18 langues avec des arbres de décisions permettant d'identifier puis de classer les EP. Une description détaillée de l'annotation et du corpus PARSEME est donnée par Candito *et al.* (2017). Tous les exemples mentionnés dans cet article proviennent de ce corpus à l'exception des exemples 3, 4 et 5 indiqués comme provenant du Web.

Le corpus d'entraînement français de PARSEME comporte 17 880 phrases résultant de la fusion des corpus Séquoia (Candito & Seddah, 2012) et UD (Nivre *et al.*, 2016) – correspondant à 450 221 tokens – et contient 4 441 occurrences d'EPV uniques [4]. 2,2 % des tokens font partie d'EPV, ce qui laisse supposer que la majorité des EP ne sont pas verbales si l'on se réfère au pourcentage de 40 % d'EP relevé par Gross. Quoique de taille réduite, ce corpus est sans équivalent dans l'espace francophone. Il offre l'intérêt de fournir des informations détaillées établies manuellement, qu'il s'agisse d'étiquetage morphosyntaxique ou des relations de dépendances syntaxiques formulées avec les tagsets de *Universal Dependencies* [5] et de Sequoia [6]. Ces deux types d'étiquetages ont été harmonisés. Le chevauchement se produisant lorsqu'un même élément est commun à plusieurs EPV est pris en compte (dans la phrase *Les **demi-finales** et la **finale** se **jouent** à Copenhague*, 2 EPV sont annotées : ***jouer demi-finales*** et ***jouer finale***). Les trois types d'EPV (ID, LVC, IReflV) y sont présents dans des proportions relativement équilibrées (Table 1). L'unique mention **OTH** (OTHER) correspond à l'expression ***aller et venir***. Si l'on s'intéresse aux LVC et aux ID (en dehors des impersonnels : ***il faut, il s'agit, ...***), la majorité des patrons (72 %) est du type VB-(DET)-NOM.

3. http://typo.uni-konstanz.de/parseme/
4. Le corpus contenait initialement 4 462 occurrences mais 21 occurrences ont été supprimées en raison de la présence de quelques phrases en double.
5. http://universaldependencies.org
6. https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia

Type d'EPV	ID	LVC	IReflV	OTH	TOTAL
nombre d'occurrences	1777 occ.	1357 occ.	1306 occ.	1 occ.	4441 occ.

TABLE 1 – Répartition des différents types d'EPV dans le corpus.

ID				LVC					
avoir lieu	95 occ.	***faire partie***	88 occ.	***jouer rôle***	33 occ.	***faire appel***	32 occ.	***prendre mesure***	11 occ.
faire l'objet	27 occ.	***mettre fin***	23 occ.	***avoir besoin***	28 occ.	***avoir droit***	17 occ.	***marquer but***	10 occ.
faire face	20 occ.	***porter nom***	19 occ.	***jouer match***	16 occ.	***faire apparition***	15 occ.	***avoir effet***	10 occ.
prendre part	19 occ.	***tenir compte***	17 occ.	***signer contrat***	15 occ.	***disputer match***	15 occ.		
mettre un terme	13 occ.	***prendre nom***	13 occ.	***poser question***	14 occ.	***faire référence***	13 occ.		
donner lieu	12 occ.	***donner naissance***	10 occ.	***rendre hommage***	12 occ.	***lancer appel***	11 occ.		
avoir mal	9 occ.	***faire preuve***	9 occ.	***avoir chance***	11 occ.	***avoir tendance***	11 occ.		
TOTAL ID : 374 occ. (21 % des ID du corpus)				TOTAL LVC : 274 occ. (20 % des LVC du corpus)					

TABLE 2 – Liste des EPV de type VB-(DET)-NOM les plus fréquentes en corpus.

La Table 2 en montre les 648 occurrences les plus fréquentes.

Si l'on s'intéresse aux éléments lexicalisés pour identifier des patrons d'EPV [7] :

- **IReflV** : 1299 occ. Après vérification manuelle de 3 cas non conformes en raison d'erreurs d'annotation, 100 % des IReflV sont de type PRON-VB ou VB-PRON (***souvenez-vous***).
- **LVC** : 1298 occ. (95 % des LVC) sont du type VB-NOM ou NOM-VB et 30 occ. (2 %) du type VB-PREP-(DET)-NOM, d'où notre étude principalement focalisée sur le patron VB-NOM/NOM-VB.
- **ID** : on observe davantage de variation dans les patrons (84 différents). Le plus fréquent ne représente que 31 % (549 occ.) : il s'agit de VB-NOM *(**signer** son **retour***). La catégorie VB-DET-NOM, où DET est lexicalisé, correspond à 13 % des occ. (233 occ.). Enfin, 180 occ. (10 %) sont du type VB-PREP-(DET)-NOM (***appeler à secours, revenir à la raison***).

Hormis pour les IReflV dont la structure en PRON-VB est prévisible, le patron ne permet pas de distinguer les ID des LVC : le schéma VB-(DET)-NOM par exemple figure dans ces deux catégories. De même, le verbe présent dans l'EPV ne suffit pas à lui seul à distinguer les ID des LVC puisque les verbes majoritairement utilisés (*avoir* et *faire*) y ont une distribution similaire : respectivement 19 % et 16 % des verbes de ces deux catégories.

5 Méthodologie

L'annotation manuelle des EPV met en évidence les éléments lexicalisés. Le regroupement des EPV en fonction des éléments lexicalisés – en neutralisant leur ordre d'apparition ***jouer** un **rôle** / le **rôle** est **joué*** – conduit à l'obtention de 1584 entrées distinctes. La limite de cette approche concerne la possibilité que des constituants identiques renvoient à des EPV différentes. Ce risque est cependant très limité puisque sur les 31 EPV étudiées en détail (Table 2), seule l'entrée (*faire ;appel*) illustre ce cas de figure :

Ex. 1. *faire appel* = 'contester en justice' (4 occ.) → *Les deux pilotes déclassés **firent appel** Ø, appel jugé en janvier 2007.* (corpus PARSEME)

Ex. 2. *faire appel (à)* = 'recourir à' (28 occ.) → *Je vous déconseille vivement de **faire appel** à cette agence.* (PARSEME)

7. VB = verbe ; PRON = pronom ; DET = déterminant ; PREP = préposition ; ADJ = adjectif ; ADV = adverbe ; CONJ = conjonction ; AUX = auxiliaire ; REL = relatif

Si les éléments annotés mettent en évidence la structure syntaxique figée des EPV (section 6), les éléments non annotés (éléments insérés entre les éléments lexicalisés) sont également instructifs. En se restreignant aux EPV de type VB-(DET)-NOM (en raison de leur fréquence élevée), on obtient un premier aperçu de la variabilité reposant à la fois sur la fréquence et sur la nature des insertions linéaires (section 7).

Nous avons également mis à profit les dépendances syntaxiques fournies dans le corpus. Nous avons développé un système de requêtes (Figure 1) pour montrer les différents types de dépendances d'un nom ou d'un verbe lorsqu'il fait partie (ou non) d'une EPV. Pour cela, l'utilisateur spécifie le verbe ou le nom recherché et éventuellement un type de dépendances. Les phrases correspondantes s'affichent en indiquant s'il s'agit ou non d'une EPV. Dans chacun de ces cas, la nature et la fréquence des différents types de dépendances sont mentionnées.

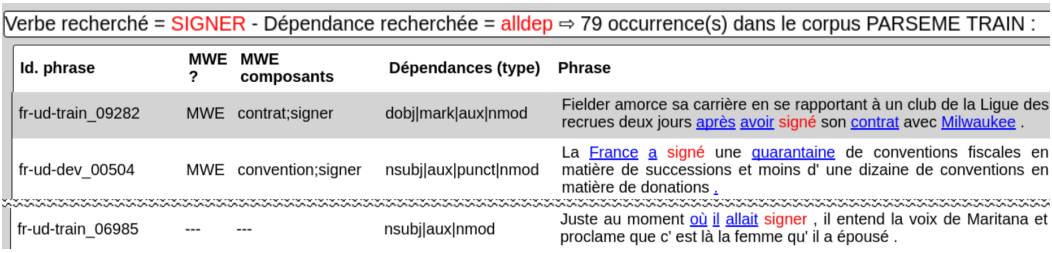

FIGURE 1 – Capture d'écran du concordancier : recherche de toutes les dépendances (en bleu) du verbe SIGNER (en rouge). Le type de dépendances s'affiche dans une infobulle au passage de la souris.

6 Variation morphologique des constituants

Nous nous intéressons à la variation dite interne des EPV, c'est-à-dire à la capacité de flexion de ses constituants (nom ou verbe). L'étendue des flexions autorisées nous renseigne en effet sur le degré de variabilité d'une EPV (***porter ses fruits***/#*porter son fruit*).

6.1 Variabilité nominale

L'étude de la variation flexionnelle du nom a été restreinte aux EPV de type VB-(DET)-NOM les plus fréquentes (Table 2) en raison de leur forte représentativité dans le corpus. Cet examen porte sur 17 LVC (274 occ.) et 12 [8] des 14 ID (356 occ.). La taille du corpus est suffisante pour identifier des motifs de variabilité, mais le recours à des ressources complémentaires est nécessaire pour confirmer l'invariabilité. Ainsi, 12 LVC présentent une variabilité du nom dans notre corpus et 3 autres dans un moteur de recherche sur Internet (Ex. 3, 4, 5). Seules 2 LVC (***faire appel***, ***avoir*** *(le)* ***droit***) sont effectivement morphologiquement rigides.

> **Ex. 3.** *Elle* ***fait*** *également une* ***apparition*** *au festival* (PARSEME) / *Cet artiste [...] a déjà* ***fait*** *des* ***apparitions*** *au Caveau de la Gare* (Web)

8. Le déterminant figé dans ***mettre un terme*** et ***faire l'objet*** signale une absence de flexion nominale.

Ex. 4. *l'occasion de lui* ***rendre*** *un dernier* ***hommage*** (PARSEME) **/** [...] *et leur* ***rend*** *les derniers* ***hommages***. (Web)

Ex. 5. ***J'ai*** *plutôt* ***tendance*** *à penser qu'il s'agit d'un coup monté.* (PARSEME) **/** ***J'ai*** *des* ***tendances*** *à être claustrophobe* (Web).

De la même façon, un examen des 11 ID apparemment figés met en évidence une flexibilité nominale de l'ID ***prendre nom***, d'où une variabilité nominale de 9,6 % des occurrences de type ID alors qu'elle est de 82 % pour les LVC.
Les contre-exemples mentionnés ne remettent pas en question la représentativité du corpus puisque 15 LVC (224 occ.) sur les 17 LVC et 11 ID (343 occ.) sur les 12 ID ont une flexibilité observée dans le corpus conforme à ce que l'on observe à plus grande échelle.

6.2 Variabilité verbale

Nous observons deux types de variabilité verbale dans le corpus :

— **Flexion personnelle et temporelle**
Parmi les cas d'invariabilité verbale, figurent des restrictions de flexion personnelle : verbes uniquement fléchis aux 3èmes personnes (*il(s)* ***se déroule(nt)/#***je me déroule*), voire uniquement à la 3ème personne du singulier. Parmi les ID, on dénombre ainsi 518 (29 %) tournures impersonnelles (***il y a, il faut, il s'agit, il pleut des cordes, sera-t-il question***,...). Par ailleurs, certaines EPV ont une flexion temporelle limitée (***qu'importe/qu'importait/#****qu'importa*) ou impossible (***honni soit qui mal y pense***).

— **Préfixation du verbe**
Le corpus comporte 7 EPV dans lesquelles le verbe est préfixé avec *r(e)-*, par exemple ***(re)donner raison*** ou ***(re)faire appel***. S'il s'agit bien ici de variantes, d'autres EPV nécessitent au contraire ce préfixe (***revenir à la raison / #****venir à la raison*). La question de la double lecture littérale / EPV se pose également dans ***relever la tête*** puisqu'au sens littéral *lever* et *relever* sont des variantes alors que le préfixe est obligatoire dans l'EPV. Prêter attention à ce phénomène pourrait permettre, à partir d'un lexique préétabli d'EPV, d'en identifier des variantes pour des verbes a priori très productifs comme *faire* ou *donner*.

S'intéresser à la variabilité des EP ne se restreint pas à la variabilité de ses constituants : il faut aussi prendre en compte la variabilité externe, c'est-à-dire les discontinuités d'éléments lexicalisés qui constituent un problème majeur pour certains outils TAL (étiquetage séquentiel).

7 Discontinuités d'éléments lexicalisés

Les discontinuités d'éléments lexicalisés nous informent sur le degré de variabilité d'une EPV : plus elles sont nombreuses et variées, plus l'EPV est flexible. Par exemple, dans les phrases (6) et (7), toutes les deux du type *prendre* + NOM, la seconde tolère moins de variations :

Ex. 6. *il* ***prend*** *encore / son / un grand* ***essor*** (PARSEME : annotation alternant entre ID et LVC).

Ex. 7. *il* ***prend*** *encore / *sa / *une grande* ***conscience*** (PARSEME : ID).

Nous proposons dans ce qui suit de décrire les insertions linéaires observées afin d'en tirer des conclusions sur la variabilité des différents types d'EPV. On appellera fenêtre d'annotation, matérialisée

par les crochets (Ex. 8 et 9), l'ensemble des éléments compris entre le premier et le dernier élément annoté. Les éléments non annotés dans la fenêtre d'annotation sont qualifiés de discontinuités et sous-entendent une plage de variabilité. Dans ce qui suit, le nombre de discontinuités correspond au nombre total d'insertions, qu'elles soient contiguës (Ex. 8) ou non (Ex. 9) :

Ex. 8. *C'est ce que [**mettent** à mon avis **en lumière**] quelques propositions d'amendement.*(PARSEME)

Ex. 9. *Le président du tribunal [**procède** alors **à** la **lecture**] des chefs d'accusation.* (PARSEME)

Sur 4441 EPV, 2772 sont continues (62 %), et 1097 (25 %) ont une discontinuité d'un seul élément. Autrement dit, l'absence de discontinuités et l'insertion d'un seul élément couvrent à elles-seules 87 % du corpus. La Figure 2 montre bien que les ID sont légèrement moins figés de ce point de vue que les IReflV, mais bien davantage que les LVC. En effet, sur l'ensemble des LVC, seules 281 occ. (20,7 %) sont continues.

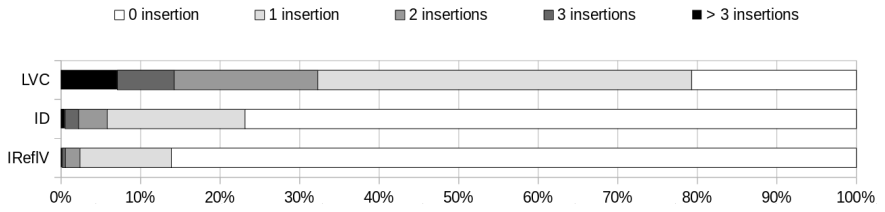

FIGURE 2 – Distribution des discontinuités dans les EPV du corpus.

Dans le cas des insertions uniques (Table 3), les différences sont significatives :

- 79 % de verbe *être* pour les IReflV. Il s'agit d'une variation aspectuelle régulière (par exemple *l'Empire [**s'est écroulé**]*),
- 66 % d'articles pour les LVC,
- 44 % d'adverbes et 28 % d'articles pour les ID. Cela rejoint le constat de Gross (1988) concernant la possibilité d'insertion d'adverbes dans les phrases figées (en dehors de quelques exceptions : *j'**aime** *vraiment **autant** vous le dire*). Quoique l'insertion d'un adverbe puisse être considérée comme une variation régulière (Tutin, 2016), la proportion d'adverbes dans les ID y est proportionnellement dix fois plus élevée que dans les LVC. Comme le déterminant est moins souvent lexicalisé dans les LVC, l'adverbe y est souvent associé. C'est ce que l'on observe pour les insertions de 2 éléments : sur 72 adverbes insérés, 4 insertions sont de type ADV-ADV et 42 de type ADV-DET. Par ailleurs, l'insertion moins fréquente de déterminant dans les ID est une propriété révélatrice car "l'extraction d'un nom est généralement interdite lorsque le déterminant est figé, et permise lorsque le déterminant et les modifieurs sont libres" (Laporte, 1988).

Un examen des éléments annotés met en évidence des négations obligatoires (***ne pas payer de mine***). Les insertions fournissent aussi un aperçu de la variabilité syntaxique. Dans les exemples de la Table 4, le clivage (*c'est [...] que/ c'est [...] qui*) et la relativation sont repérées par l'insertion d'un relatif et classées manuellement.

A l'instar de ce que l'on observe avec le clivage, l'utilisation libre du nom de l'EPV en dehors de

Catégorie de l'élément inséré	Exemple	ID occ.	ID %	LVC occ.	LVC %	IReflV occ.	IReflV %
ADJ	***transfert** bancaire **effectué***	3	0.97	15	2.35	1	0.67
ADV	***fit** notamment **campagne***	136	**44.16**	27	4.23		
REL	***transformations** que **subit***			3	0.47		
DETERMINANT :							
article	***faire** un **tour***	88	**28.57**	424	**66.46**		
démonstratif	***conclu** cette **union***			17	2.66		
indéfini	***joue** aucun **rôle***	5	1.62	27	4.23		
numéral	***effectué** deux **mandats***			12	1.88		
possessif	***attirer** votre **attention***	23	7.47	49	7.68		
(SEMI-)AUXILIAIRE :							
être	***en** est **arrivée***	6	1.95	19	2.98	119	**79.33**
avoir	***Il** aura **fallu***	13	4.22				
modal : *devoir, sembler,..*	***il** peut **pleuvoir***	10	3.25				
NOM	***traité** Euratom, **signé***			2	0.31		
PREP	***mesures** à **prendre***	1	0.32	30	4.70		
PRONOM :							
démonstratif	***prendre** cela **au sérieux***	2	0.64				
personnel	***Il** me **semble***	5	1.62				
en	***il y** en **a***	4	1.30			16	**10.67**
y	***s'y fier***					13	8.67
le	***il** le **faut***	4	1.30			1	0.67
passif SE	***match** se **dispute***			13	2.04		
t euphonique	***Y a** -t -**il***	8	2.60				
TOTAL		308	100	638	100	150	100

TABLE 3 – Nature des insertions dans le cas d'une seule insertion. Les pourcentages les plus significatifs sont mis en gras.

celle-ci (Ex. 10) nous renseigne sur sa capacité à fonctionner sémantiquement de façon autonome et laisse supposer une EPV de type LVC.

Ex. 10. *Les deux pilotes déclassés **firent appel**, appel jugé en janvier 2007*

Les passivations potentielles sont détectées grâce au patron NOM-VB. La présence du verbe ÊTRE non annoté dans la fenêtre d'annotation indique une passivation avec auxiliaire. Le repérage des EPV passivées sans AUX se fonde sur la terminaison du verbe. La passivation en SE (*le **match** se **dispute***) est identifiée par l'insertion de ce pronom. Les occurrences trouvées ont fait l'objet d'une vérification manuelle.

Pour ces différents paramètres (passivation, relativation, clivage), on observe une moindre variabilité des ID que des LVC (Table 4). Les ID ont également davantage tendance à lexicaliser les déterminants (20 fois plus souvent) ou les négations (10 fois plus).

type de variation	LVC occ.	LVC %	exemple LVC	ID occ.	ID %	exemple ID
passive AUX + VB	59	4.3	*son **rôle** est **joué***	5	0.3	*Le **ton** a d'ailleurs été **donné***
passive sans AUX	135	9.9	*le **rôle joué***	5	0.3	***compte-tenu***
passive SE + VB	45	3.3	*La **finale** se **joue***	0	0	
TOTAL passives	239	17.6		10	0.6	
relative	48	3.5	***tâche** qu'il a **accompli***	1 ?	0.06	***question** que nous devons nous **poser***
clivage	2	0.1	*c'est du bon **boulot** que vous m'avez **fait***	0	0	
$\geq$ 1 det lexicalisé	13	1	***avoir l'impression***	346	19.5	***tentent le tout pour le tout***
négation lexicalisée	1 ?	0.1	***ne pas avoir de prix***	20	1.1	***ne pas tarir***
TOTAL	1357	100		1777	100	

TABLE 4 – Certains aspects de variabilité et de figement des LVC vs ID sur l'ensemble du corpus

8 Analyse syntaxique de la variabilité

L'étude des insertions implique que les éléments en dehors de la fenêtre d'annotation ne sont pas pris en compte, comme l'adjectif dans l'exemple 11 :

Ex. 11. *les Byzantins **[remportent** une **victoire]** décisive*. (PARSEME)

Pour bénéficier d'une vue plus large des contraintes exercées sur les 29 EPV de type VB-NOM les plus fréquentes et n'ayant pas de déterminant figé, nous avons donc utilisé les relations de dépendance fournies dans le corpus, ce qui permet aussi une analyse plus aisée des insertions multiples (soulignées dans l'ex. 12)

Ex. 12. *La famille du cinéma français a **[rendu** vendredi à Paris un émouvant **hommage]** à l'un de ses plus brillants représentants* [...] (PARSEME)

Pour chaque expression testée, une seule occurrence de chaque type de dépendance du nom est comptabilisée. En effet, on s'intéresse à la capacité du nom à être modifié plutôt qu'au nombre de modifieurs [9]. On obtient 374 patrons de dépendances pour les noms inclus dans des ID et 434 dans des LVC, valeurs suffisamment proches pour permettre une comparaison. L'absence de dépendances est trois fois plus fréquente dans les ID que dans les LVC (48 % vs 13,6 %) et, dans une proportion similaire, le déterminant non lexicalisé est plus fréquent dans les LVC (30,4 % vs 7,5 %). Alors que le nom garde tout son sens dans les LVC, cela se produit rarement dans les ID (***faire office***) sans pour autant être impossible (***faire prisonnier***). Cela peut expliquer qu'il y ait plus fréquemment une absence de dépendances du nom dans les ID (***faire** *un grand **office***) que dans les LVC (***faire** une grande **révélation***).

En s'appuyant sur l'exploitation des dépendances, on peut :

- établir un classement des EPV de type VB-(DET)-NOM selon leur degré de variabilité
- identifier des comportements différents du NOM selon qu'il fait partie (ou non) d'une EPV,
- dégager les profils syntaxiques des variantes d'EPV de type VB-(DET)-NOM pour faciliter leur génération ou fusion automatique.

8.1 Classement des EP selon leur degré de variabilité

Tutin (2016) a étudié la variabilité des 30 EPV (ID et LVC) de type VB-(DET)-NOM les plus fréquentes en français. Les propriétés [10] les plus discriminantes sont la pluralisation du nom, les constructions relatives et passives (*le **rôle** est **joué** ; l'**attention prêtée***). 2 EPV ont un comportement inattendu : ***avoir recours*** (LVC) s'avère moins variable que ***avoir** du **mal*** (ID) [11]. Mais Tutin (2016) souligne l'ambiguïté de cette EP [12] (***avoir mal / avoir** le **mal** de X*).

Nous avons cherché à comparer les niveaux de variabilité définis par Tutin (2016) pour certaines EP avec les tendances observables en corpus. Il s'agit d'une comparaison partielle faute de disposer des mêmes EP ou de ne pas en avoir un échantillon représentatif (moins de 10 occ.), d'où l'absence des

9. Par exemple, la phrase *Les femmes*[nsubj] *jouent au football*[nmod] *depuis la fin*[nmod] *du XIXe siècle en Angleterre*[nmod] aura pour dépendances associées au verbe JOUER : nsubj|nmod et non nsubj|nmod|nmod|nmod.

10. Chacune des 5 propriétés satisfaites accroît le niveau de variabilité d'une unité : pluralisation du nom possible, déterminant (ou son absence) non figé(e), relativation, passivation, nom modifiable par un adjectif.

11. Dans notre corpus : ***avoir** (du) **mal**, **avoir** (également) beaucoup de **mal**, n'**avoir** aucun **mal**, **avoir** plus **mal***.

12. Cette ambigüité est également présente dans ***poser question / poser** une **question*** quoique seule la seconde figure dans notre corpus. Pour ***jouer** un **rôle***, toutes les occurrences sont comptabilisées.

Niveaux définis par Tutin (2016) →	Niveau 4	Niveau 4	Niveau 1	Niveau 1	Niveau 0	Niveau 0
EP	***jouer rôle***	***poser question***	***faire partie***	***tenir compte***	***faire appel***	***donner lieu***
Type	LVC	LVC	ID	ID	LVC	ID
VB-NOM	79 %	50%	100 %	71 %	100 %	100 %
NOM-VB	21 %	50%		29 %		
0 insertions	9 %	14 %	93 %	94 %	94 %	92 %
1 insertion	76 %	36 %	7%	6 %	6 %	8%
2 insertions	9%	36 %				
> 2 insertions	6%	14 %				
Nom variable	oui	oui	non	non	non	non
Passivation	oui	oui	non	oui	non	non
Relativation	oui	oui	non	non	non	non
nombre moyen par occurrence de types de dépendances différentes du nom (y compris absence de dépendance)	0,39	0,57	0,03	0,29	0,09	0,17
Total	33 occ.	14 occ.	90 occ.	17 occ.	32 occ.	12 occ.

TABLE 5 – Comparaison de la variabilité observée en corpus avec les nivaux établis par Tutin (2016). Les cellules grisées indiquent des invariabilités. Le niveau 0 correspond à la variabilité minimale et le niveau 4 à la variabilité maximale.

niveaux 2 et 3 dans la Table 5. On y observe que les résultats sont cohérents pour les degrés extrêmes de (non-)variabilité mais moins marqués pour le niveau 1 puisque, dans notre corpus, l'expression ***faire partie*** a un comportement similaire à celui de ***faire appel***. Cette tendance est corroborée par le nombre de dépendances différentes associées au nom figurant dans l'EPV puisque ***faire partie*** montre une plus faible variabilité que les 2 expressions du niveau 0.

8.2 Dépendances de différents noms en EPV (ID) et hors EPV

Nous avons observé (section 7) que les dépendances du nom étaient différentes selon qu'il s'agissait d'un ID ou d'une LVC [13]. Lorsqu'un mot n'apparaît qu'au sein d'EPV (comme *mouron* dans ***se faire du mouron***), on le qualifie de *cranberry word* (Richter & Sailer, 2003). Ce phénomène étant rare, nous avons comparé les dépendances du nom lorsqu'il faisait partie du type d'EPV le plus contraint (ID) et lorsqu'il fonctionnait de façon autonome (hors EPV). Parmi les noms les plus fréquents dans les ID du corpus, *partie* et *objet* sont notamment présents dans les EPV ***faire partie/être de la partie*** et ***faire l'objet***. Lorsque ces noms font partie d'un ID, ils ont tendance (Table 6) à :

- moins souvent régir une préposition,
- moins souvent régir un modifieur adjectival,
- plus souvent régir un modifieur nominal.

S'il y a plus de modifieurs nominaux pour les noms figurant dans ces ID que lors d'un fonctionnement autonome, cela peut s'expliquer par leur structure syntaxique (que partagent 9 des 14 ID de la Table 2). En effet, ils requièrent un complément (pronom ou nom) : ***faire partie*** *de NOM,* ***faire l'objet*** *de NOM*.

Pour le nom *objet*, qu'il s'agisse ou pas d'une EP, la présence de déterminant se produit avec une fréquence assez similaire (quoique supérieure dans ce premier cas). Le déterminant est en effet lexicalisé dans l'EPV ***faire l'objet***. En revanche, l'association de *partie* avec un déterminant apparaît peu souvent dans une EPV compte-tenu de la sur-représentation de l'expression ***faire partie*** (94 %) interdisant sa présence.

13. La capacité du nom à être modifié ne fait pas partie des critères d'identification/classification des EP, dans le cas contraire l'intérêt de cette étude serait caduc.

catégorie d'un élément régi par le nom	*partie*		*objet*		Exemples
	hors EPV	EPV	hors EPV	EPV	
det	24.7 %	2.4 %	37 %	43.6 %	*ils **seront** bien **de la partie** (EPV)* *les deux parties de l' Empire*
nmod	20.8 %	82 %	6.2 %	35.2 %	*il **fait partie** du groupe (EPV)* *la plus grande partie du personnel*
amod	12.4 %	4.8 %	12.3 %	1.4 %	*il **fait partie** des plus anciens (EPV)* *en grande partie*
case	16.5 %	2.4 %	17.2 %	0 %	*ils **seront** bien **de la partie** (EPV)* *en grande partie*
TOTAL	232 occ.	84 occ.	37 occ.	31 occ.	

TABLE 6 – Dépendances des noms *partie* et *objet* en fonction de leur (non-)appartenance à une EPV. La terminologie *amod* (modifeur adjectival), *nmod* (modifieur nominal) et *case* (préposition) provient du tagset des *Universal Dependencies.*

8.3 Schémas de dépendance des différentes variantes syntaxiques

Nous cherchons ici à à établir des profils syntaxiques en dépendances pour différentes manifestations en surface des EPV. Cela permettra de générer ou normaliser automatiquement des variantes à partir de (ou vers) la forme canonique ou une autre variante. En utilisant un nouveau corpus parsé, nous pourrions ainsi faire des requêtes intégrant ces variantes pour accroître le nombre d'EPV détectées.

Nous nous sommes focalisés sur les 3 EPV (***jouer** un **rôle**, **signer** un **contrat /accord/partenariat**...), **jouer** la **finale***) offrant une variation syntaxique importante pour en décrire les régularités de construction. Parfois le verbe et le nom d'une EPV n'entretiennent pas, d'après les schémas de dépendances fournis, une relation de dépendance directe : dans ***signer** une quarantaine de **conventions***, le nom *quarantaine* s'interpose entre ***signer*** et ***conventions***. L'emploi de tels déterminants complexes s'observe 7 fois (***effectuera** une remarquable seconde partie de **saison**,...*).

L'identification de l'élément régisseur, des dépendances associées à celui-ci ainsi qu'à l'élément régi permettent d'identifier les différentes variantes syntaxiques. La Table 7 montre, pour chaque variante, la hiérarchie des éléments au sein de la chaîne de dépendance et leurs caractéristiques. Dans cette hiérarchie, on attribue la valeur 1 à l'élément régissant et cette valeur croît à chaque niveau de dépendances. A titre de comparaison, la fréquence des différentes variantes des 2090 occurrences de type VB-(DET)-NOM et NOM-VB dans le corpus s'avère relativement similaire.

Type de variante	3 EPV testées		% corpus	élément	éléments régis	
	Exemples	Fréquence	VB-(DET)-N et N-VB	régissant		
Niveau de dépendance				1	2	3
forme canonique	*Il **signe** le **contrat***	85,9 %	85,6%	**VB**	**NOM**-dobj	
passive avec AUX	*le **contrat** est **signé***	1,6 %	2,8 %	**VB**	**NOM**-subjpass + auxpass-AUX	
passive sans AUX	*le **contrat signé***	6,6 %	6,6 %	**NOM**	**VB**-acl	
passive en SE	*la **finale** se **joue***	2,9 %	2,1 %	**VB** + se-dobj-PRON	**NOM**-nsubj	
infinitif	*a un **rôle** à **jouer***	1,2 %	0,4 %	**NOM**	**VB**-acl	case (*à, de*)
clivage		0 %	0,05 %			
relative	*l'**accord** qu'il a **signé***	1,6 %	1,5 %	**NOM**	**VB**-acl :relcl	

TABLE 7 – Différents schémas de dépendances : forme canonique, passivation, relative,... acl = *adjective clause* ; acl :relcl = *relative clause modifier* ; dobj = objet direct ; auxpass = auxiliaire passif.

9 Conclusion et perspectives

Ce travail confirme des comportements différents pour les EPV de type LVC et ID : les LVC ont une variabilité supérieure aux ID au niveau de la flexion nominale, de la relativation ou du clivage. La passivation y est aussi plus fréquente (17,6 % *versus* 0,6 %). De plus, les noms des LVC VB-NOM ont tendance à avoir une plus grande variété et quantité de dépendances que ceux des ID. Cependant, les noms employés dans les 2 ID examinés ont de 4 à 7 fois plus tendance à avoir des modifieurs nominaux qu'en dehors d'une EPV. Un examen à plus grande échelle sera nécessaire pour confirmer et affiner ce constat.
Les motifs de variabilité qui se dégagent pourront être mis à profit pour améliorer la détection d'EPV déjà connues. La typologie des insertions uniques et les schémas de variation syntaxique en dépendances permettront de générer ou normaliser automatiquement différentes variantes pour les EPV de type VB-(DET)-NOM. Comme il est peu probable que toutes les variations possibles d'une EPV donnée soient représentées dans notre corpus, il faudra recourir à d'autres ressources afin d'obtenir la fréquence d'apparition de ces variantes. Ces statistiques permettront d'attribuer aux variantes un caractère obligatoire, optionnel ou impossible : *il **attire** mon /*Ø/*deux **attention**(s) / mon **attention** est **attirée***. Pour les LVC, une étude conjointe des expressions contenant un même nom (ou un synonyme) en utilisant par exemple le Lexicoscope [14], serait sans doute judicieuse (*{**conclure / signer / sceller**} un {**accord / partenariat / alliance**, etc.}*).
Enfin, la recherche de dépendances interposées (***signer** une quarantaine de **contrats***) est un autre axe d'étude envisagé pour obtenir une vision globale de la variabilité dans l'objectif de favoriser la détection des EPV quelque soit leur forme de surface.

Références

ABEILLÉ A. (1988). Light verb constructions and extraction out of NP in a tree ad-joining grammar. *Papers of the 24th Regional Meeting of the Chicago Linguistics Society*.

BALDWIN T., BENDER E. M., FLICKINGER D., KIM A. & OEPEN S. (2004). Road-testing the English Resource Grammar over the British National Corpus. In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

CANDITO M., CONSTANT M., RAMISCH C., SAVARY A., PARMENTIER Y., PASQUER C. & ANTOINE J.-Y. (2017). Annotation d'expressions polylexicales verbales en français. In *TALN 2017*, Actes de TALN 2017, Orléans, France : Association pour le Traitement Automatique des Langues.

CANDITO M. & SEDDAH D. (2012). Le corpus Sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proceedings of TALN 2012*.

GROSS M. (1982). Une classification des phrases « figées » du français. *Revue québécoise de linguistique*, **11**(2).

GROSS M. (1988). Les limites de la phrase figée. *Langages*, **23**(90), 7–22.

GROSS M. & SENELLART J. (1998). Nouvelles bases statistiques pour les mots du français. In *Journée d'Analyse Statistique des Données Textuelles (JAD T)*, p. 335–349, Nice.

JACKENDOFF R. (1997). The architecture of the language faculty. *Linguistic Inquiry Monographs*.

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

14. http://phraseotext.u-grenoble3.fr/lexicoscope/

LAPORTE E. (1988). Reconnaissance des expressions figées lors de l'analyse automatique. *Langages*, **23**(90), 117–126.

LICHTE T., PETITJEAN S., SAVARY A. & WASZCZUK J. (2017). *Lexical encoding formats for multi-word expressions : The challenge of "irregular" regularities"*, In Y. PARMENTIER & J. WASZCZUK, Eds., *Representation and Parsing of Multiword Expressions*, p. 79–111. Language Science Press, à paraître.

NISSIM M. & ZANINELLO A. (2013). Modelling the internal variability of multiword expressions through a pattern-based method. In *ACM Transactions on Speech and Language Processing , Special issue on Multiword Expressions*, volume 10.

NIVRE J., DE MARNEFFE M.-C., GINTER F., GOLDBERG Y., HAJIC J., MANNING C. D., MCDONALD R., PETROV S., PYYSALO S., SILVEIRA N., TSARFATY R. & ZEMAN D. (2016). Universal dependencies v1 : A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* : European Language Resources Association (ELRA).

NUNBERG G., SAG I. A. & WASOW T. (1994). Idioms. *Language*, **70**, 491–538.

PRZEPIÓRKOWSKI A., HAJNICZ E., PATEJUK A. & WOLIŃSKI M. (2014). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, p. 83–91, Dublin, Ireland : Association for Computational Linguistics and Dublin City University.

RICHTER F. & SAILER M. (2003). Cranberry words in formal grammar. In C. BEYSSADE, O. BONAMI, P. CABREDO HOFHERR & F. CORBLIN, Eds., *Empirical Issues in Formal Syntax and Semantics 4*, volume 4, p. 155 – 171. Presses de l'Université de Paris-Sorbonne, Paris.

SAG I., BALDWIN T., BOND F., COPESTAK A. & FLICKINGER D. (2002). Multiword expressions : A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, p. 1–15, Mexico City,Mexico.

SAVARY A. (2008). Computational Inflection of Multi-Word Units, a contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, p. 1–53.

SAVARY A. & JACQUEMIN C. (2003). Reducing Information Variation in Text. *LNCS*, **2705**, 145–181.

SAVARY A., RAMISCH C., CORDEIRO S., SANGATI F., VINCZE V., QASEMIZADEH B., CANDITO M., CAP F., GIOULI V., STOYANOVA I. & DOUCET A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, p. 31–47, Valencia, Spain : Association for Computational Linguistics.

SHEINFUX L. H., GRESHLER T. A., MELNIK N. & WINTNER S. (2017). *Verbal MWEs : Idiomaticity and flexibility*, In Y. PARMENTIER & J. WASZCZUK, Eds., *Representation and Parsing of Multiword Expressions*, p. 5–38. Language Science Press, à paraître.

TUTIN A. (2016). Comparing morphological and syntactic variations of support verb constructions and verbal full phrasemes in French : a corpus based study. In *PARSEME COST Action. Relieving the pain in the neck in natural language processing : 7th final general meeting*, Dubrovnik, Croatia.

VILLAVICENCIO A., KORDONI V., ZHANG Y., IDIART M. & RAMISCH C. (2007). Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, p. 1034–1043, Prague, Czech Republic.