Annotation d'expressions polylexicales verbales en français

Marie Candito, Mathieu Constant, Carlos Ramisch, Agata Savary, Yannick Parmentier, Caroline Pasquer, Jean-Yves Antoine

Annotation d'expressions polylexicales verbales en français

Marie Candito[1] Mathieu Constant[2] Carlos Ramisch[3] Agata Savary[4]
Yannick Parmentier[5] Caroline Pasquer[4] Jean-Yves Antoine[4]
(1) Université Paris Diderot, LLF, Paris, France
(2) Université de Lorraine, ATILF, CNRS, Nancy, France
(3) Aix Marseille Université, CNRS, LIF, Marseille, France
(4) Université François Rabelais Tours, LI, Blois, France
(5) Université d'Orléans, LIFO, Orléans, France
marie.candito@linguist.univ-paris-diderot.fr, Mathieu.Constant@univ-lorraine.fr,
carlos.ramisch@lif.univ-mrs.fr, agata.savary@univ-tours.fr, yannick.parmentier@univ-orleans.fr,
caroline.pasquer@etu.univ-tours.fr, Jean-Yves.Antoine@univ-tours.fr

RÉSUMÉ

Nous décrivons la partie française des données produites dans le cadre de la campagne multilingue PARSEME sur l'identification d'expressions polylexicales verbales (Savary *et al.*, 2017). Les expressions couvertes pour le français sont les expressions verbales idiomatiques, les verbes intrinsèquement pronominaux et une généralisation des constructions à verbe support. Ces phénomènes ont été annotés sur le corpus French-UD (Nivre *et al.*, 2016) et le corpus Sequoia (Candito & Seddah, 2012), soit un corpus de 22 645 phrases, pour un total de 4 962 expressions annotées. On obtient un ratio d'une expression annotée tous les 100 tokens environ, avec un fort taux d'expressions discontinues (40%).

ABSTRACT

Annotation of verbal multiword expressions in French

We describe the French part of the annotated data produced for the multilingual PARSEME shared task on the identification of verbal multiword expressions (Savary *et al.*, 2017). The annotated verbal expressions for French are idioms, inherently reflexive verbs and a generalization over support verb constructions. These were annotated on the French-UD corpus (Nivre *et al.*, 2016) and the Sequoia corpus (Candito & Seddah, 2012) for a total of 22,645 sentences and 4,962 annotated expressions. On average, we obtain around 1 expression every 100 tokens with a high ratio of discontiguities (40%).

MOTS-CLÉS : Expressions polylexicales verbales, annotation, corpus.

KEYWORDS: Verbal multiword expressions, annotation, corpora.

1 Introduction

Les expressions polylexicales (ci-après EP), telles que ***pomme de terre***, ***tout à coup***, ***prendre*** *une* ***décision***, ***se comporter*** ou ***avoir l'air*** [1], sont des objets linguistiques constitués d'au moins deux composants (éléments se comportant comme un mot par ailleurs, ou bien des éléments qui ne peuvent

1. Nous adoptons la convention de mettre en gras les composants *lexicalisés* d'une EP, c'est-à-dire ceux toujours réalisés par les mêmes lexèmes, par opposition aux modifieurs non requis et aux arguments requis mais pouvant être choisis dans une large classe sémantique, qui eux apparaîtront simplement en italique, comme *il* ***a l'air*** *pressé, ils n'****ont*** *pas* ***l'air*** *mécontents.*

apparaître de manière autonome, comme *fi* dans ***faire fi***) et caractérisés par un certain degré d'idiosyncrasie au niveau morphologique, syntaxique ou sémantique (Baldwin & Kim, 2010). Les EP verbales (EPV), c'est-à-dire ayant un verbe comme tête syntaxique, posent des défis particuliers pour la modélisation comme pour le traitement automatique, essentiellement du fait d'une variabilité syntaxique en général plus forte que pour d'autres EP. Elles peuvent avoir des interprétations idiomatiques ou littérales (p. ex. *le roi* ***ouvre*** *alors* ***les yeux*** *sur son geste* vs. *la victime parle et ouvre les yeux*). Des EPV de même structure syntaxique peuvent appartenir à des catégories différentes (p. ex. ***avoir l'air*** est une expression idiomatique, tandis que ***prendre la fuite*** est une construction à verbe support). Certains composants peuvent être partagés par plusieurs EPV, en particulier lors d'une coordination (p. ex. *ils* $\textbf{\textit{ont}}_{1,2}$ *le* $\textbf{\textit{pouvoir}}_1$ *et le* $\textbf{\textit{devoir}}_2$ *de voter*), d'une relativisation (p. ex. *il veut* $\textbf{\textit{avoir}}_1$ *la* $\textbf{\textit{perception}}_{1,2}$ *qu'*$\textbf{\textit{ont}}_2$ *les entreprises*) ou de l'imbrication complète d'une EP au sein d'une autre (p. ex. *ils* $\textbf{\textit{font}}_{1,2}$ $\textbf{\textit{l'}}_{1,2}\textbf{\textit{objet}}_{1,2}$ d'une $\textbf{évaluation}_2$). Enfin, selon la convention de tokenisation utilisée, une EPV peut parfois correspondre à un seul token (p. ex. ***contre-indiquer***).

Tandis que l'étude linguistique des EP françaises de tout type a une longue tradition (Gross, 1986; Mel'čuk *et al.*, 1988), l'annotation en corpus des EPV n'avait pas été réalisée à large échelle. Nous présentons le fruit d'un travail qui tente de combler cette lacune. Il s'agit d'une tâche d'annotation d'EPV effectuée dans le cadre du réseau européen PARSEME [2], et plus spécifiquement de sa campagne d'évaluation [3] d'outils de reconnaissance automatique d'EPV (Savary *et al.*, 2017). Cet article se concentre sur la production des données françaises.

2 Travaux antérieurs

Le traitement des EP en TAL est l'objet d'une vaste littérature, et plusieurs typologies à visée universelle ont été proposées (Sag *et al.*, 2001; Heid, 2008; Baldwin & Kim, 2010; Mel'čuk, 2010; Tutin *et al.*, 2015), mais aucune n'a été effectivement mise à l'épreuve d'une annotation multilingue. En effet, si l'annotation d'EP et plus précisément d'EPV a été l'objet de nombreuses initiatives (voir Savary *et al.* (2017) pour un résumé), la campagne PARSEME part du constat d'une grande hétérogénéité dans les types d'EP annotées, les critères retenus et les formats utilisés.

En outre, concernant plus spécifiquement le français, s'il existe une forte tradition de ressources lexicales incluant des EP, avec en particulier le DELAC (Courtois *et al.*, 1997), leur annotation en corpus est plus rare. Le French Treebank (Abeillé *et al.*, 2003) constitue un projet pionnier en la matière. Il comprend environ 20 000 phrases annotées pour la morphologie et la syntaxe, avec une part non négligeable d'EP annotées (Abeillé & Clément, 2006), dont des EPV. Plus précisément, les EPV retenues sont celles comportant un mot inexistant par ailleurs (***faire fi***), ou une syntaxe irrégulière avec en particulier les composés V N sans aucun déterminant possible (p. ex. ***faire face*** est annoté, mais pas *avoir peur*, cf. on a aussi *avoir une peur immense*). Si un mécanisme a bien été prévu pour gérer d'éventuelles discontinuités, il a été en pratique très peu utilisé. Les combinaisons clitiques-verbes ont été volontairement écartées, ainsi que les constructions à verbe support.

D'autres corpus arborés ne comprennent que des mots composés grammaticaux, contigus. C'est le cas des deux corpus arborés utilisés comme base de l'annotation PARSEME, Sequoia et French-UD. Plus récemment, Tutin *et al.* (2015) ont produit une typologie fine d'EP, incluant des EPV, et leur annotation sur un corpus français d'environ 45 000 tokens.

2. http://www.parseme.eu
3. http://multiword.sourceforge.net/sharedtask2017

3 La campagne PARSEME

La campagne PARSEME citée supra est le fruit de l'effort collaboratif de 18 équipes nationales (Savary *et al.*, 2017). Deux phases d'annotation pilote, menées pour 15 langues, chacune suivie d'un retour d'expérience et d'améliorations de la méthodologie, ont permis d'aboutir à :

— une typologie universelle d'EPV, validée par annotation pilote multilingue, qui laisse néanmoins une place pour des types spécifiques à des sous-ensembles de langues,
— un guide d'annotation sous forme d'arbres de décision fondés sur des tests linguistiques,
— une infrastructure consistant en une plate-forme d'annotation, des scripts d'homogénéisation, des outils d'insertion automatisée d'exemples d'EPV dans le guide d'annotation,
— des mesures et outils d'évaluation des systèmes d'identification d'EPV, les systèmes participants étant évalués sur l'identification seulement, et pas sur la catégorisation d'EPV.

Les corpus résultants [4] couvrent 18 langues et comportent au total près de 5,5 million de mots. Les 60 000 EPV annotées, ainsi que le guide d'annotation, sont diffusés sous deux versions de la licence libre Creative Commons (CC BY et CC BY-NC-SA) [5].

4 Schéma d'annotation

L'annotation des EPV du français repose sur le guide d'annotation [6] de la campagne PARSEME (Savary *et al.*, 2017). Une EPV y est définie comme une EP dont la forme canonique a pour tête syntaxique un verbe, et dont la distribution est celle d'un verbe, d'un syntagme verbal ou d'une phrase [7]. Les formes non canoniques (p. ex. au passif, ou avec un composant extrait) sont également annotées. Le schéma d'annotation permet d'annoter des expressions verbales imbriquées, ainsi que des expressions partageant certains éléments (p. ex. *Luc* ***prend***$_{1,2}$ *une* ***douche***$_1$ *puis un* ***bain***$_2$).

L'annotation pour le français utilise toutes les catégories prévues sauf celle des verbes à particule : les expressions verbales idiomatiques (ID, p. ex. ***avoir lieu***, ***faire partie***), les constructions à verbe support (CVS, p. ex. ***faire*** *une* ***proposition***) et les verbes intrinsèquement pronominaux (SeV, p. ex. ***s'évanouir***). Une quatrième catégorie « Autre » a également utilisée marginalement pour le cas d'EP de type verbes coordonnés (p. ex. ***aller et venir***) ou des verbes dont la structure interne est irrégulière (p. ex. ***court-circuiter***). L'annotation d'une expression s'effectue en plusieurs étapes. Premièrement, pour chaque verbe, les annotateurs repèrent d'après leurs connaissances linguistiques s'il y a une idiosyncrasie potentielle dans la composition du verbe avec un ou plusieurs autres composants. La deuxième étape filtre plus précisément quels sont les composants qui entreraient effectivement dans l'expression candidate (p. ex. est-ce que les déterminants sont à inclure car figés). C'est ensuite la troisième étape qui permet de trancher sur le statut effectif d'EPV de l'expression candidate, et sur sa catégorie. Elle est réalisée en suivant deux arbres de décision, dont nous donnons un aperçu seulement, pour insister sur la marche à suivre très précise fournie aux annotateurs.

Le premier arbre de décision fonctionne comme une série de conditions suffisantes d'EP (on sort de l'arbre dès qu'une condition est vérifiée), comme le fait qu'un des composants ne puisse exister de

4. https://gitlab.com/parseme/sharedtask-data/tree/master
5. En ce qui concerne les données françaises : (i) les annotations en EPV sont diffusées sous CC-BY v4, (ii) les annotations morpho-syntaxiques pré-existantes sont diffusées sous licence CC BY-NC-SA v4 pour la partie UD, et LGPL-LR pour la partie Sequoia.
6. http://parsemefr.lif.univ-mrs.fr/guidelines-hypertext
7. Ainsi, par exemple, les noms composés comportant des verbes (p. ex. *porte-parole*) ne sont pas des EPV.

manière autonome (p. ex. ***prendre la poudre d'escampette***), ou qu'une transformation régulière, de type substitution lexicale, modification syntaxique ou morphologique, provoque une inacceptabilité ou un changement de sens inattendu. Par exemple, la forme idiomatique de ***jeter l'éponge*** ne se passive pas alors que c'est le cas pour sa forme libre ; de même, remplacer *éponge* par *serpillière* n'est pas autorisé dans son sens idiomatique (*#jeter la serpillière*). Par ailleurs *éponge* est invariable en nombre (*#jeter les éponges*).

À ce stade, si une condition suffisante a été trouvée, le candidat est considéré comme EPV, et le deuxième arbre de décision permet de préciser son type. Par exemple, on annotera « Autre » une expression où l'on n'a pas un seul verbe fonctionnant comme tête syntaxique. Toutefois, même si une condition suffisante n'a pas été trouvée avec le premier arbre, dès lors que le candidat a la forme d'une CVS, des tests spécifiques aux CVS doivent être appliqués.

Expressions verbales idiomatiques (ID). Les expressions satisfaisant une condition suffisante d'EP sont typées ID dans les cas suivants :

— le verbe de l'expression n'a pas un et un seul dépendant syntaxique lexicalisé (p. ex. ***prendre le taureau par les cornes***, ***il est question***)

— le verbe de l'expression a un et un seul dépendant syntaxique lexicalisé, et ne satisfait pas les critères de CVS (cf. ci-dessous).

Constructions à verbes support (CVS). Une EPV candidate est annotée CVS si toutes les conditions suivantes sont satisfaites :

— Le candidat a la forme verbe v plus un dépendant nominal n direct ou *via* préposition régie.

— Le nom n a un de ses sens habituels, il a au moins un argument sémantique, et il décrit un événement ou un état (*décision*, *courage*) [8].

— Le verbe v n'ajoute aucune sémantique qui ne soit déjà présente dans le sens du nom, mis à part la sémantique des marques de flexion, et l'indication de quel est l'argument sémantique du nom qui est réalisé comme sujet du verbe. Ainsi on annotera aussi bien *donner un ordre* que *recevoir un ordre*. On n'annotera pas *X prend la responsabilité [de]* (car inchoatif), *X donne la migraine à Y* (car causatif), *X termine sa promenade* (aspectuel), alors qu'on annotera ces mêmes noms avec respectivement les verbes *avoir, avoir* et *faire*. **Il est important de noter que ce critère est à la fois une restriction de la notion habituelle de CVS, qui comprend des CVS aspectuelles ou causatives (p. ex. Gross (1993)), et une généralisation de la notion de verbe support, car on n'impose pas que le verbe support ait perdu sa sémantique habituelle**. On peut toutefois noter une faiblesse dans le traitement actuel, pour les cas où le verbe n'a pas un de ses sens habituels, mais a un sens causatif ou aspectuel, comme par exemple ***attirer l'attention***. On doit alors coder ID car le verbe est idiosyncratique, et pas CVS car le verbe n'est pas sémantiquement neutre.

— On doit pouvoir former un groupe nominal (GN) en ajoutant le sujet de v au GN de n, et ce GN étendu doit pouvoir référer à l'éventualité décrite par la version avec verbe support. Par exemple, avec *Luc* ***prend*** *une* ***décision***, on parle de *la décision de Luc*. Avec *Luc* ***donne*** *l'***ordre** *de partir à Paul*, la réduction est plus difficile telle quelle (? *L'ordre de Luc à Paul de*

8. Un relecteur fait remarquer que ce critère purement sémantique est sans doute superflu. En réalité, il a été utilisé dans la campagne comme condition nécessaire pour un nom d'avoir un argument sémantique, autre qu'un possesseur. La contrainte d'un argument sémantique écarte par exemple les noms atmosphériques : *La pluie tombe* n'est actuellement pas annoté, l'annotation se concentrant sur les cas où il y a divergence syntaxe-sémantique, avec un argument syntaxique du verbe qui est en réalité un argument sémantique du nom.

partir rapidement était un peu rude), mais acceptable si l'on pronominalise le sujet (*son ordre à Paul de partir rapidement était un peu rude*).

— Dans la forme canonique de la CVS, le sujet de v correspond à un actant de n, d'où l'impossibilité de réaliser cet actant à la fois au sein du complément nominal et comme sujet de v (**Paul reçoit la visite de Pierre à Jacques*), sauf à devoir interpréter une comparative (*Paul **fait** la **promenade** de Luc*) [9].

Verbes intrinsèquement pronominaux (seV). Les verbes pronominaux sont des combinaisons d'un verbe v et d'un clitique réflexif (en français *se*, *me*, *te*, *nous*, *vous*, que l'on note ci-dessous SE), ayant différents statuts possibles. Le phénomène existe dans de nombreuses langues, dont pour la campagne, les langues romanes, les langues slaves, l'allemand et le suédois. Un sous-guide d'annotation a été mis au point pour ces cas, sous la forme d'un arbre de décision. Ont été considérés comme des EPV les cas intrinsèquement impersonnels, c'est-à-dire pour lesquels il n'existe pas une relation régulière avec une version sans SE du verbe. N'ont donc pas été annotés les cas de vrais réfléchis ou réciproques, les moyens ou « à agent fantôme » (p. ex. *une telle vitre se casse avec un marteau*), ou les neutres (p. ex. *la branche s'est cassée d'un coup*). En revanche, nous avons annoté les cas de verbes v n'apparaissant jamais sans le clitique SE (p. ex. ***se suicider***), ou dont le clitique SE modifie de manière imprévisible le comportement de v, sur le plan syntaxique (p. ex. ***s'apercevoir** de Y*, cf. **X aperçoit Z de Y*) ou sémantique. Pour trancher ce dernier cas, difficile, le guide d'annotation utilise un critère d'implication logique : si « X v Y » $\Rightarrow$ « Y SE v » alors l'expression candidate n'est pas considérée comme une EPV. Par exemple *le clown égaye les enfants* $\Rightarrow$ *les enfants s'égayent.*

5 Méthodologie d'annotation

Une fois le guide multilingue PARSEME stabilisé, la phase d'annotation proprement dite a été réalisée sur deux mois, en parallèle pour les différentes langues, avec l'outil FLAT [10] (van Gompel & Reynaert, 2013). Six personnes ont annoté le français (les auteurs de cet article). Par manque de temps et de moyens, il n'a pas été possible de réaliser une double annotation suivie d'une adjudication : chaque portion de corpus a été annotée par une seule personne, sauf un extrait pour le calcul de l'accord (cf. section 6). Pour compenser la perte de qualité potentielle (erreurs d'inattention et incohérences du fait d'interprétations différentes des consignes d'annotation), nous avons utilisé différents outils :

— Pendant la phase d'annotation, des questions sur l'interprétation du guide ou des demandes de précision pouvaient être adressées et débattues via le gestionnaire de tickets gitlab, à différents niveaux (langue, groupe de langues, toutes les langues).

— Parallèlement, les annotateurs ont maintenu une liste de cas précis tranchés collectivement. Contrairement à la résolution de conflits via adjudication, une telle organisation ne permet pas de garantir la cohérence des différentes décisions entre elles, cohérence normalement évaluée par la capacité des annotateurs à converger sur la seule base du guide d'annotation. Elle permet bien cependant de limiter les incohérences d'annotation pour un même phénomène.

— Après l'annotation simple, nous avons utilisé un outil de repérage automatique de bruit et de silence. Il extrait la liste des EPV annotées, avec pour chaque EPV les occurrences annotées, et de possibles occurrences oubliées avec une recherche approchée bruitée. Un parcours manuel

9. À noter que ne sont pas pris en compte les rares cas où le nom prédicatif est sujet (*les applaudissements de la foule crépitèrent*, (Jousse, 2010)).

10. `http://github.com/proycon/flat`, `http://flat.science.ru.nl`

de cette liste a permis de rapidement corriger des incohérences.
— Enfin une dernière étape a utilisé la liste des cas tranchés négativement pour repérer automatiquement et supprimer des cas annotés à tort.

6 Description et évaluation du corpus

Les EPV ont été annotées sur deux corpus préexistants, annotés pour la morphologie et la syntaxe en dépendances : la partie française du corpus Universal Dependencies [11] (Nivre *et al.*, 2016), qui comprend 16 447 phrases françaises extraites au hasard de Google News, Blogger, Wikipedia et des avis de consommateurs ; et le corpus Sequoia [12] (Candito & Seddah, 2012), qui comprend 3 099 phrases issues de l'Est Républicain, de rapports de l'Agence Européenne du Médicament, de Wikipedia et d'Europarl. Pour la campagne PARSEME, les 500 premières EPV ont été réservées comme corpus d'évaluation des systèmes participants (*test*). Le restant du corpus a été considéré comme corpus d'entraînement pour les systèmes (*train*) [13].

	#Phrases	#Tokens	#EPV	#ID	#SeV	#CVS	#Autre
Complet	19 547	486 005	4 962	1 905	1 418	1 633	6
Train	17 880	450 221	4 462	1 786	1 313	1 362	1
Test	1 667	35 784	500	119	105	271	5

TABLE 1 – Statistiques sur le corpus annoté divisé en corpus d'entraînement (*train*), corpus d'évaluation (*test*) et corpus complet : nombre de phrases, de tokens, nombre total d'EPV, suivi d'un découpage par catégorie d'EPV.

La table 1 fournit la taille du corpus annoté en phrases et en tokens, et les nombres d'EPV annotées. La table 2 donne des informations sur la longueur des EPV en nombre de tokens et les discontinuités au sein des EPV. Si l'on se concentre sur le corpus complet, on constate qu'environ 4 EPV sur 10 sont des ID, 3 sur 10 sont des verbes pronominaux, et 1 sur 3 est une CVS. La catégorie Autre est marginale. On constate que le corpus de *test* est atypique : il comprend beaucoup plus de CVS en proportion, et les EPV y sont globalement plus discontinues que dans le corpus complet (seulement $42,9\%$ d'EPV continues dans le *test*, contre $60,0\%$ dans le corpus complet). Pour ce qui est de la longueur des EPV, les 3 quarts des instances comportent deux tokens ($77,4\%$), et un peu moins de 20% comportent 3 tokens. En étudiant ces mêmes indicateurs mais par type d'EPV, on constate que la plupart des EPV de longueur > 2 sont des ID ($45,5\%$ des IDs sont de longueur 3).

La même table 2 fournit des informations sur les discontinuités, calculées en nombre total de tokens apparaissant entre des composants d'EPV mais n'en faisant pas partie. Là encore, on constate de fortes disparités selon le type d'EPV. Par exemple, la proportion d'EPV sans aucune discontinuité est globalement de $60,0\%$, mais de $76,6\%$ pour les IDs, $85,8$ pour les SeV, et seulement de $18,1$ pour les CVS. Pour ces dernières, environ la moitié des annotations ont une discontinuité réduite à un seul token (souvent pour le déterminant du nom), et 8% ont une discontinuité de plus de 3 tokens.

11. version 1.4, `http://universaldependencies.org/`

12. version 7.0, `https://www.rocq.inria.fr/alpage-wiki/tiki-index.php?page=CorpusSequoia`. La tokenisation du corpus Sequoia a été automatiquement rapprochée de celle du corpus UD, en particulier pour les prépositions contractées et les mots composés grammaticaux (seuls composés annotés originellement dans Sequoia).

13. Il est apparu *a posteriori* que ce découpage correspond exactement à la partie médicale et la partie Est Républicain du corpus Sequoia. Cela explique des différences importantes de distributions des phénomènes entre *train* et *test*.

	Longueur de l'EPV			Longueur des discontinuités (en nombre de tokens)						
Corpus	Moy.	% lg=2	% lg=3	Moy.	DAM	% d=0	% d=1	% d=2	% d=3	% d>3
Tot.	2,28	77,4	18,8	0,68	0,81	60,0	26,4	7,9	2,7	2,9
Tot. ID	2,67	46,3	45,1	0,33	0,51	76,6	17,4	3,7	1,6	0,6
Tot. SeV	2,00	99,8	0,1	0,17	0,3	85,8	11,8	1,8	0,4	0,1
Tot. CVS	2,07	94,5	4,2	1,52	1,07	18,1	49,5	18,3	6,0	8,0
Train	2,29	77,1	19,2	0,65	0,80	61,9	25,0	7,5	2,8	2,8
Test	2,24	80,2	15,2	0,95	0,81	42,9	39,1	11,9	1,6	4,5

TABLE 2 – Longueur des EPV et longueur cumulée des discontinuités (en nombre de tokens) dans le corpus complet (en tout, et par type d'EPV), et dans les corpus d'entraînement (*train*) et d'évaluation (*test*). Col 1 : longueur moyenne. Col 2 et 3 : pourcentages d'EPVs de longueur 2 et 3. Col 4 et 5 : Long. moyenne et déviation absolue moyenne (DAM) des discontinuités. Col 6 : Pourcentage d'EPV sans discontinuité. Col 7 à 10 : Pourcentage d'EPV avec discontinuité de 1, 2, 3 et plus de 3 tokens.

Afin d'estimer la qualité de la méthodologie d'annotation et du corpus résultant, des extraits de corpus ont été doublement annotés. L'extrait pour le français comprend 1 000 phrases (24 666 mots) [14]. L'accord inter-annotateur (AIA) pour la tâche d'identification est évalué *via* une F-mesure, où les annotations du premier annotateur jouent le rôle de la référence. Une annotation est considérée correcte si elle couvre l'ensemble précis des éléments lexicalisés d'une EPV de la référence. Pour la catégorisation, on calcule un kappa de Cohen sur les EPV identifiées par les deux annotateurs avec les mêmes composants. On obtient pour l'extrait français $F = 0,819$ et $\kappa = 0,93$, soit un accord substantiel, parmi les 3 meilleurs dans l'ensemble des 12 langues concernées [15].

7 Conclusion

Nous avons présenté une ressource d'environ $5\,000$ instances d'expressions polylexicales verbales annotées sur environ $19\,500$ phrases en français, comprenant des expressions idiomatiques, des verbes intrinsèquement pronominaux et des constructions à verbe support. Les perspectives futures sont par exemple d'étudier la variation interne au corpus, selon les domaines des phrases annotées, ainsi que les taux d'ambiguïté. Une extension à tout type d'EP est également prévue.

Remerciements

Ce travail a été mené dans le cadre de l'Action COST PARSEME (IC1207), et du projet ANR PARSEME-FR (ANR-14-CERA-0001). Les auteurs remercient chaleureusement les organisateurs de la campagne, pour le travail mené sur le guide d'annotation, et pour toute l'infrastructure d'annotation.

14. Un des deux annotateurs de l'extrait doublement annoté a participé à la rédaction du guide.
15. Notons que la qualité des annotations a pu encore être améliorée *via* les outils de recherche de bruit et de silence (sec. 5).

Références

ABEILLÉ A. & CLÉMENT L. (2006). *Annotation morpho-syntaxique - Les mots simples, les mots composés - Corpus Le Monde.* Rapport interne, TALANA, Université Paris 7.

ABEILLÉ A., CLÉMENT L. & TOUSSENEL F. (2003). Building a treebank for french. In A. ABEILLÉ, Ed., *Treebanks*, p. 165–187. Dordrecht : Kluwer.

BALDWIN T. & KIM S. N. (2010). Multiword expressions. In N. INDURKHYA & F. J. DAMERAU, Eds., *Handbook of Natural Language Processing*, p. 267–292. Boca Raton, FL, USA : CRC Press, Taylor and Francis Group, 2 edition.

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*.

COURTOIS B., GARRIGUES M., GROSS G., GROSS M., JUNG R., MATHIEU-COLAS M., SILBERZTEIN M. & VIVÈS R. (1997). *Dictionnaire électronique des noms composés DELAC : les composants NA et NN. Rapport technique.* Rapport interne, LADL, Université Paris 7.

GROSS G. (1993). Trois applications de la notion de verbe support. *L'Information Grammaticale*, **59**(1), 16–22.

GROSS M. (1986). Lexicon-grammar : The Representation of Compound Words. In *Proceedings of the 11th Coference on Computational Linguistics*, COLING '86, p. 1–6, Stroudsburg, PA, USA : Association for Computational Linguistics.

HEID U. (2008). In *Phraseology. An interdisciplinary perspective*, chapter Computational phraseology. An overview, p. 337–360. John Benjamins Publishers : Amsterdam, Netherlands.

JOUSSE A.-L. (2010). *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. PhD thesis, Université de Montréal et Université Paris Diderot.

MEL'ČUK I. (2010). La phraséologie en langue, en dictionnaire et en taln. In *Conférence invitée de TALN 2010*, Montréal, Canada.

MEL'ČUK 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*, volume II of *Recherches lexico-sémantiques*. Presses de l'Univ. de Montréal.

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).

SAG I. A., BALDWIN T., BOND F., COPESTAKE A. & FLICKINGER D. (2001). Multiword expressions : A pain in the neck for nlp. In *In Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexique.

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)*, Valencia, Spain : Association for Computational Linguistics.

TUTIN A., ESPERANÇA-RODIER E., IBORRA M. & REVERDY J. (2015). Annotation of multiword expressions in French. In C.-P. GLORIA, Ed., *European Society of Phraseology Conference (EUROPHRAS 2015)*, Computerised and Corpus-based approaches to phraseology : monolingual and multilingual perspectives, p. 60–67, Malaga, Spain.

VAN GOMPEL M. & REYNAERT M. (2013). FoLiA : A practical XML format for linguistic annotation - a descriptive and comparative study. *Computational Linguistics in the Netherlands Journal*, **3**, 63–81.