PolylexFLE : une base de données d'expressions polylexicales pour le FLE

Amalia Todirascu[1] Marion Cargill[1] Thomas François[2]
(1) LiLPa, Université de Strasbourg, 22, rue René Descartes, 67084 Strasbourg, France
(2) CENTAL, UCLouvain, Place Montesquieu, 3, bte L2.03.02, 1348 Louvain-la-Neuve, Belgique
todiras@unistra.fr, mcargill@unistra.fr, thomas.francois@uclouvain.be

RÉSUMÉ

Nous présentons la base PolylexFLE, contenant 4295 expressions polylexicales. Elle est integrée dans une plateforme d'apprentissage du FLE, SimpleApprenant, destinée à l'apprentissage des expressions polylexicales verbales (idiomatiques, collocations ou expressions figées). Afin de proposer des exercices adaptés au niveau du Cadre européen de référence pour les langues (CECR), nous avons utilisé une procédure mixte (manuelle et automatique) pour annoter 1098 expressions selon les niveaux de compétence du CECR. L'article se concentre sur la procédure automatique qui identifie, dans un premier temps, les expressions de la base PolylexFLE dans un corpus à l'aide d'un système à base d'expressions régulières. Dans un second temps, leur distribution au sein de corpus, annoté selon l'échelle du CECR, est estimée et transformée en un niveau CECR unique.

ABSTRACT

PolylexFLE : a database of multiword expressions for French L2 language learning

We present the PolylexFLE database, containing 4295 polylexical expressions. It is integrated into a platform to support learning of verbal polylexical expressions (idioms, collocations or fixed expressions). In order to propose exercises adapted to the level of the European Framework of Reference for Languages (CEFR), we used a mixed approach (manual and automatic) to annotate 1098 expressions according to the CEFR levels. The paper focuses on the automatic procedure that first identifies the expressions from the PolylexFLE database in a corpus using a regular expression-based system. In a second step, their distribution in this corpus, labelled according to the CEFR scale, is estimated and transformed into a single CEFR level.

MOTS-CLÉS : expressions polylexicales vebales, niveau CECR, TAL pour la didactique du FLE.

KEYWORDS: verbal multiword expressions, CEFR level, NLP for French L2 language learning.

1 Contexte et motivation

Les expressions polylexicales (EP) constituent une classe d'objets linguistiques qui inclut les expressions idiomatiques, les expressions figées et des collocations. La définition exacte des EP reste discutée, mais plusieurs chercheurs s'accordent à les identifier comme "des séquences de mots, dont le sens est plus ou moins compositionnel, caractérisés par des propriétés morpho-syntaxiques, syntaxiques, sémantiques" (Baldwin & Kim, 2010). Ces unités seraient stockées directement en mémoire, ce qui rend leur traitement cognitif plus rapide (Pawley & Syder, 1983), du moins dans le cas de natifs. Les apprenants d'une langue étrangère éprouvent, quant à eux, bien des difficultés

dans l'acquisition et le traitement des EP. Leur maitrise des EP se situe souvent bien en-deçà de leurs connaissances lexicales générales (Bahns & Eldaw, 1993) et ils tendent à effectuer des traductions mot à mot de ces expressions, ignorant le sens figuré de ces expressions. De récentes études montrent pourtant qu'une bonne maitrise des EP améliore la compréhension en lecture (Kremmel *et al.*, 2017).

À ces difficultés d'acquisition vient s'ajouter le fait que les plateformes en ligne proposent rarement des exercices visant directement l'apprentissage des EP. Pour le français langue étrangère (FLE), *Bonjour de France* et *Le point du FLE* constituent à cet égard des exceptions. Autre exemple, une plateforme comme *The Writing Mentor* [1] qui vise à soutenir le développement des compétences de production écrite via des annotations collaboratives et un apprentissage par feedback (Hamel *et al.*, 2016), ne se focalise pas sur les expressions polylexicales. Quant aux plateformes de création d'activités pour l'apprentissage des langues [2], celles-ci se concentrent seulement sur certaines facettes du vocabulaire (élargissement par synonymie, reformulations), mais les EP ne font que rarement l'objet de corrections ou de retours proposés à l'utilisateur. Toutefois, *Language Muse* propose des mots composés ou des verbes à particules en anglais pour améliorer la lecture des apprenants (Madnani *et al.*, 2016). C'est pourquoi, nous avons cherché à développer une plateforme dédiée à la problématique des EP dans le cadre du projet SimpleApprenant, qui est décrit à la section 2.

Dans ce cadre, un problème nous est rapidement apparu : la rareté des ressources proposant une base d'EP adaptée à un usage pédagogique. En effet, afin de proposer des exercices adaptés au niveau des apprenants, il convient de disposer d'une base de données dans laquelle le niveau de difficulté des EP est signalé, de préférence en rapport avec l'échelle du Cadre européen commun de référence pour les langues (CECR), publié par le Conseil de l'Europe (2001) afin de structurer le secteur de l'enseignement des langues étrangères au niveau européen. Or, il existe également très peu de plateformes d'apprentissage où les exercices et les ressources portant sur les EP sont annotées selon le niveau CECR.

Plusieurs ressources sont disponibles pour l'apprentissage des expressions idiomatiques et des collocations. Ainsi, la *Base Lexicale du Français* (Verlinde *et al.*, 2006) propose une description détaillée des contextes syntaxiques et morpho-syntaxiques des EP, des traductions et des exemples tirés de corpus. Le projet *DIRE Autrement* (Hamel & Milicevic, 2007) propose quant à lui les collocatifs les plus fréquents pour certaines expressions et des exercices permettant de mettre en correspondance des expressions et des définitions. Le projet PARSEME-FR a crée un corpus annoté en EP (Ramisch *et al.*, 2018), mais adopte une classification très détaillée des expressions. Ces travaux ne font toutefois pas de liens entre les EP envisagées et l'échelle du CECR.

Plus proche de ce qui nous intèresse, *EmoProf*, intégré dans la base lexicale *EmoBase* (Diwersy *et al.*, 2014), propose des séquences didactiques pour les professeurs de FLE ciblant des expressions polylexicales liées au vocabulaire des émotions (verbes, noms, adjectifs). Les séquences didactiques sont classées par niveau (Cavalla *et al.*, 2013), mais ces séquences ne sont pas directement exploitables par une application de TAL. De même, il existe les référentiels pour le FLE de Beacco *et al.* (2004) qui incluent des EP dans les descriptions lexicales des niveaux du CECR. À nouveau, le format (papier) de cette ressource n'est pas exploitable dans un contexte de TAL. Il existe bien la ressource FLELex (François *et al.*, 2014) qui précise, pour plus de 2000 EP, leur distribution sur les 6 niveaux du CECR. Celle-ci est directement exploitable pour le TAL, mais intègre surtout des EP nominales. Au niveau des EP verbales, pourtant cruciales dans le contexte de l'apprentissage du FLE, on ne trouve dans FLELex que des formes à base des verbes *faire* (ex. *faire part*, *faire obstacle*) et *avoir* (ex.

1. https://mentormywriting.org/
2. https://languagemuse.org/

avoir peur, *avoir faim*).

Dans cet article, nous proposons une nouvelle base de données pédagogique pour les expressions polylexicales, PolylexFLE, dédiée aux EP verbales. Elle est intégrée dans la plateforme SimpleApprenant dont les exercices sont majoritairement dédiés à l'apprentissage des EP. Nous présentons tout d'abord cette plateforme à la section 2, avant de décrire le contenu de PolylexFLE et le processus de collecte des EP à la section 3. Ensuite, nous détaillons à la section 4 la façon dont un niveau de compétence du CECR a pu être associé à 1098 EP à l'aide d'une approche mixte : manuelle (basée sur des vocabulaires de référence) et automatique (basée sur le traitement automatique de corpus pédagogiques). Par rapport aux travaux existants, notre base PolylexFLE présente deux avantages : les EP y sont annotées avec leur niveau CECR et le format est directement utilisable pour le TAL.

2 Le projet SimpleApprenant

La base de données PolylexFLE a été développée dans le cadre du projet SimpleApprenant [3], qui a pour objectif d'aider les apprenants du FLE à améliorer leurs compétences écrites et leur connaissance des EP (Todirascu & Cargill, 2019). Dans ce but, une application web qui repose sur des outils et des ressources TAL a été développée, proposant 3 fonctionnalités :

— améliorer les connaissances des EP à travers des exercices, étalonnés en fonction du niveau CECR.
— suggérer des EP en relation avec un mot introduit par un utilisateur.
— corriger automatiquement au niveau typographique, lexical ou syntaxique, un texte écrit par un apprenant, à l'aide d'un module TAL, SimplifyYourFrench.

Comme la plateforme dispose des profils des utilisateurs, qui renseignent entre autres leur niveau CECR, elle peut leur proposer des exercices et des EP adaptés à leur niveau. Pour ce faire, il a cependant été nécessaire de mettre au point une base d'EP, PolylexFLE, qui renseigne le niveau CECR de chaque EP. Nous détaillons, à la section suivante, la méthode de conception de cette base lexicale.

3 La conception de PolylexFLE

La base PolylexFLE comporte, dans son ensemble, 4295 entrées (sous forme de lemmes), associées à leurs patrons syntaxiques, thématiques (par exemple : couleurs, parties du corps), traductions, définitions (extraites automatiquement du Wiktionnaire) et phrases en contexte. Parmi toutes ces entrées, 1098 EP sont également associées à un niveau CECR.

Les 4295 expressions présentes dans la base ont été extraites du Lexique-Grammaire (Gross, 1994; Laporte *et al.*, 2008), ressource lexicale présentant des verbes et des EP verbales associées à leurs informations syntaxiques, morpho-syntaxiques et sémantiques. Nous avons sélectionné des expressions et leurs contextes syntaxiques, sous forme de patrons syntaxiques. Nous avons filtré ces expressions selon les définitions proposées par (Baldwin & Kim, 2010). Ainsi, nous intégrons dans la base des expressions qui sont composées d'au moins deux unités lexicales, reliées par des dépendances syntaxiques, qui présentent au moins des spécificités syntaxiques, sémantiques ou morphologiques (Constant *et al.*, 2017). De ce fait, nous avons identifié trois catégories d'EP :

3. https ://simpleapprenant.huma-num.fr/SimplifyYourFrench/accueil.jsp

— les expressions idiomatiques, qui ont un sens figuré, non déductible à partir des sens de chaque unité : ex. *perdre pied* (perdre la confiance en soi), *jeter l'éponge* (abandonner). Ces expressions sont caractérisées par l'absence du déterminant pour le nom (*perdre pied*, mais pas **perdre les pieds*), ou la préférence pour un déterminant précis (*faire le clown* mais pas **faire les clowns*), l'impossibilité de modifier le nom (**perdre pied gauche*) ou l'impossibilité d'utiliser un adverbe entre le verbe et le nom (**avoir toujours d'autres chats à fouetter*) et l'impossibilité de passivation. Ces expressions posent le plus de problèmes aux apprenants, car il est parfois difficile de trouver la traduction correcte ou une expression équivalente.
— les collocations manifestent une préférence lexicale forte (*poser une question*, mais pas **demander une question*), dans un champ lexical restreint (*hisser le pavillon/le drapeau*), leur sens est plutôt compositionnel. La variabilité syntaxique de ce type d'EP est importante : la modification du verbe est possible (*prendre rapidement des mesures* ; la modification du nom également (*prendre des mesures drastiques*) ; enfin le passage à la diathèse passive est acceptable. Les éléments de la collocation résistent aux tests de substitution (remplacer le nom ou le verbe par un synonyme). Signalons que notre définition de la classe des collocations est différente de celle adoptée par le projet PARSEME-FR pour la campagne d'annotation des expressions verbales (Ramisch *et al.*, 2018) où les collocations sont des mots qui cooccurrent fréquemment (*lire un livre/un article/des PDFs/une carte/des BDs/un fichier*).
— les expressions figées, comprenant les expressions dont le verbe est conjugué (*être sans espoir pour*, *avoir droit à*, *être d'accord*), mais dont l'objet est totalement fixe et lexicalisé (le déterminant est fixe et le nom n'accepte pas des modifications morphologiques, des adjectifs ne peuvent pas se combiner avec le nom). Dans la catégorie des expressions figées, nous incluons les expressions ayant une valeur pragmatique, étiquétées comme « pragmatème » (Tutin *et al.*, 2015) : *ça va s'arranger, le soleil brille, il fait chaud*. Pour ces expressions, on ne peut pas insérer un déterminant, un adjectif ou une relative. Gross (1993) parle de groupe nominal figé, d'adverbe figé, etc. On inclut la préposition si elle indique un des arguments de l'expression. Ces expressions représentent des difficultés pour un apprenant du FLE, car une partie est fixe et les contraintes syntaxiques et morpho-syntaxiques sont importantes.

4 Identification des niveaux de difficulté

Les expressions que nous avons extraites du Lexique-Grammaire ne sont pas annotées en niveaux CECR. Pour ajouter cette information à notre liste d'expressions, nous avons utilisé une approche mixte : manuelle, dans laquelle les niveaux sont obtenus à partir de ressources pédagoqiques et automatique, qui utilise des techniques de TAL.

L'approche manuelle est facile à décrire. Nous avons consulté des vocabulaires de référence pour l'apprentissage du FLE listant des expressions ainsi que, pour chacune de ces expressions, le niveau CECR auquel elle est supposée apprise. Il s'agit des référentiels de Beacco *et al.* (2004, 2008), ainsi que de manuels de FLE (Rey, 2007). Nous avons ainsi retrouvé 535 de nos expressions dans ces sources et leur avons attribué le même niveau CECR que celui renseigné dans ces sources. Nous avons renseigné ces niveaux dans la base.

Toutefois, une majorité de nos EP n'étaient pas reprises dans ces sources et nous avons dû mettre au point une méthodologie empirique basée sur corpus afin de leur attribuer automatiquement un niveau de difficulté. Cette technique consiste à identifier ces EP dans un corpus de textes (décrit à la

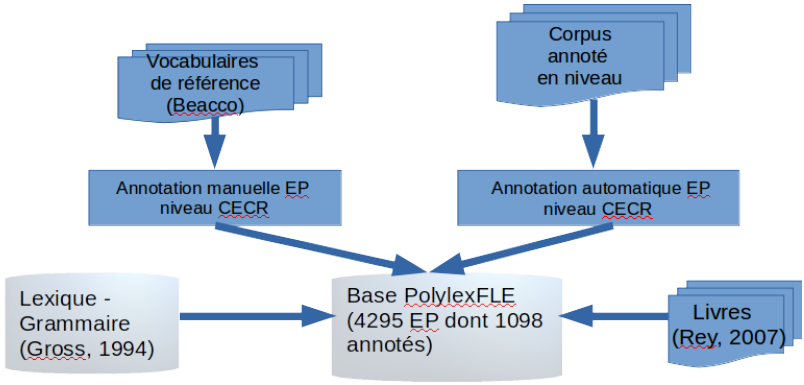

FIGURE 1 – Méthodologie d'annotation du niveau CECR

section 4.1) de FLE dont le niveau CECR est connu, ce qui permet de calculer leur distribution de fréquence par niveau du CECR. Ces étapes sont respectivement décrites aux sections 4.2 et 4.4, tandis que la qualité de l'identification automatique des expressions est évaluée à la section 4.3. Une fois ces distributions de fréquence obtenues, celles-ci sont transformées en un niveau unique (*cf.* section 4.4). Si le niveau de l'EP n'est pas renseigné dans la base, ce niveau unique calculé automatiquement sera ajouté. Si l'expression a déjà un niveau annoté manuellement, nous gardons ce niveau. Nous avons également effectué une évaluation préliminaire de la fiabilité de nos annotations CECR auprès d'apprenants du FLE, dont les résultats sont décrits à la section 5.

4.1 Description du corpus

Le corpus utilisé regroupe une quantité notable de textes extraits de manuels utilisés pour enseigner le FLE et associés à un niveau CECR. Ainsi, chaque texte se voit attribuer le même niveau du CECR que le manuel ou la ressource pédagogique dont il provient. Nous avons rassemblé plusieurs textes, regroupés en deux sous-corpus dont les caractéristiques divergent quelque peu. Le *corpus1*, collecté dans le cadre de cette étude, reprend des romans ou des contes très courts s'adressant aux apprenants du FLE. Il s'agit de textes écrits directement pour les apprenants, qui sont aussi accompagnés d'un lexique expliquant les mots et les termes les plus compliqués. Le *corpus2* correspond quant à lui au corpus décrit en détails dans François (2014). Il s'agit de textes directement extraits de manuels de FLE généralistes, destinés à des adultes ou grands adolescents et publiés après 2001. Le corpus comprend différents genres de textes, notamment des textes narratifs, informatifs, dialogiques et argumentatifs, mais aussi des petites annonces, des poèmes, des chansons, des recettes de cuisine, etc. La taille de chaque composante de ces deux corpus est reprise à la table 1. Pour notre étude, ces deux corpus ont été rassemblés au sein d'un corpus unique appelé *Corpus-M*, qui comprend 857 737 mots.

	A1	A2	B1	B2	C1	C2	Total
Corpus1	15 620	43 422	57 795	101 361	54 057	52 290	324 545
Corpus2	62 592	95 117	176 973	71 701	92 327	34 482	533 192
Corpus-M	78 212	138 539	234 768	173 062	146 384	86 772	857 737

TABLE 1 – Taille du corpus (en nombre de mots) par niveau du CECR

Pour réaliser l'annotation des EP selon l'échelle du CECR, nous examinons la distribution des EP dans ces deux corpus. Nous faisons l'hypothèse que l'expression trouvée dans un niveau CECR doit être connue par les apprenants ayant acquis ce niveau. Cela requiert tout d'abord d'être capable de détecter automatiquement ces EP verbales, ce qui constitue un sérieux défi, même au sein du domaine de l'extraction automatique d'EP. La section suivante détaille la technique mise au point dans ce but.

4.2 Méthode d'extraction des EP

Notre définition des EP répose sur les notions d'idiomaticité syntaxique, sémantique et morpho-syntaxique, selon Baldwin & Kim (2010) et Constant *et al.* (2017) : en plus des combinaisons statistiques fréquentes, les EP sont caractérisées par des liens syntaxiques et sémantiques et par des fortes préférences lexico-syntaxiques.

Nous annotons le corpus à l'aide de l'analyseur syntaxique Mind the Gap (Coavoux & Crabbé, 2017) et de Glàff (Hathout *et al.*, 2014). Ainsi, Mind the Gap applique une annotation syntaxique en dépendances que l'on peut exploiter pour extraire les expressions à partir du corpus. Cet analyseur identifie aussi certaines EP (en particulier, les locutions prépositionnelles, les noms propres), mais pas les EP verbales. Mind the Gap fait une analyse morphosyntaxique détaillée sans identifier les lemmes ; ceux-ci sont ajoutés à l'aide du dictionnaire Glàff sur la base de la catégorie lexicale identifiée par Mind the Gap.

L'identification des EP au sein du corpus est une tâche difficile, car les EP ne sont pas toujours rencontrées dans les textes sous la même forme que celle indexée dans la base. Notre approche est symbolique, utilisant les annotations morpho-syntaxiques, la lemmatisation, les dépendances syntaxiques et les patrons stockés dans la base. Parmi les trois catégories d'expressions (*cf.* section 3), les expressions idiomatiques et les expressions figées acceptent très peu de variations. Généralement, seul l'ajout d'un adverbe est possible : *il fait chaud/il fait très chaud* ; *il a jeté l'éponge/ il a jeté rapidement l'éponge*. Toutefois, les verbes qui composent les expressions peuvent se conjuger avec un auxiliaire. Les collocations, quant à elles, sont caractérisées par une grande variabilité syntaxique : variation du déterminant, insertion de modifieurs pour le nom ou pour le verbe, passivation, nominalisation.

Avant d'effectuer la recherche des expressions dans le corpus, nous générons toutes les variantes possibles pour chaque collocation, à l'aide des patrons syntaxiques provenant du Lexique Grammaire, qui sont représentés dans la base. Par exemple, pour la collocation *garder le silence*, nous utilisons le patron du Lexique-Grammaire associé :

<ENT>V_<ENT>Det1_<ENT>C1, ou V est le verbe (garder), Det1 est le déterminant (le) et C1 est le nom (silence), <ENT> est une balise qui sépare les éléments du patron.

A partir de ce patron, on génère les variantes possibles, en rajoutant des modifieurs adjectivaux (<ENT>V_<ENT>Det1_<ENT>C1_<ENT>Adj1), les modifieurs adverbiaux (<ENT>V_<ENT>Adv_<ENT>Det1_<ENT>C1) (*garder prudemment le silence*) ou les patrons qui indiquent la passivation (<ENT>Det1_<ENT>C1_<ENT>_être<ENT>VPP, VPP est le verbe au participe passé) (*le silence a été gardé*).

Parmi les variantes possibles, nous identifions également les nominalisations (*mise à jour*) ou les noms modifiés par un participe passé du verbe (*le silence bien gardé*). Pour les autres catégories d'expressions polylexicales, telles que les expressions idiomatiques et les expressions figées, nous ajoutons les auxiliaires (*j'ai eu d'autres chats à fouetter*) et les modifieurs adverbiaux (*il faisait*

tellement beau). Au terme de cette étape, ces variantes ont été ajoutées aux EP listées dans la base. La recherche peut alors être effectuée sur le corpus complet.

Ces patrons sont implémentés dans un système à base de règles, en Perl, utilisant des expressions regulières qui génèrent les variantes morpho-syntaxiques sur la base des patrons du Lexique Grammaire pour chaque candidat de la base. Ainsi, l'algorithme lit le texte phrase par phrase. Dès qu'un candidat est identifié dans la phrase courante, on vérifie les contraintes morpho-syntaxiques et syntaxiques associées dans la base. Ainsi, on vérifie les relations de dépendandance entre le nom et le verbe (complément d'objet direct ou indirect, sujet), la distance entre le verbe et le nom (limitée à 4 mots). De plus, on vérifie la catégorie lexicale des mots insérés entre le nom et le verbe (adjectifs ou adverbes, les éventuels determinants qui précedent le nom).

Évaluation. Des corpus annotés en EP sont disponibles, mais nous avons voulu évaluer la qualité de notre extraction directement sur des documents de FLE, dont les caractéristiques sont assez éloignées des textes utilisés dans les corpus d'évaluation classique. Nous avons dès lors sélectionné aléatoirement 20 textes représentatifs des niveaux A1 à C2 (A1 : 366 mots ; A2 : 724 mots ; B1 : 2 059 mots ; B2 : 2 008 mots ; C1 : 2 425 mots ; C2 : 453 mots) dans le *Corpus2*.

L'annotation des EP verbales de ce corpus d'évaluation a été réalisée à l'aide d'un guide d'annotation. Notre guide d'annotation diffère du guide d'annotation proposé dans le cadre de la campagne d'annotation des EP verbales par le projet PARSEME [4]. En effet, PARSEME propose une classification très détaillée des EP verbales, valable pour un grand nombre de langues : construction à verbes supports (*faire peur*), construction à verbes multiples (*to let go*), verbes à particules (*give up*), etc. Pour le français, les ressources et les outils créés dans le cadre du projet PARSEME FR permettent l'annotation des expressions polylexicales spécifiques. Ainsi, ces outils annotent les formes verbales pronominales (*se laver, s'assoir*), les expressions idiomatiques et des constructions à verbe support (*faire peur, prendre une décision*). Pour notre application, qui s'adresse aux apprenants de FLE, la classification est restreinte aux trois catégories présentées à la section 3 : expressions idiomatiques, collocations et expressions figées. Notre classification coincïde avec la classification proposée par PARSEME-FR en ce qui concerne les expressions idiomatiques. Les constructions à verbes supports peuvent être retrouvées parmi notre classe de collocations (si le nom est variable en nombre ou accepte plusieurs types de déterminant) ou parmi les expressions figées (si le nom est invariable). En revanche, les formes pronominales ne sont pas annotées dans notre projet.

Le processus d'annotation s'est déroulé selon la procédure suivante. Dans un premier temps, trois annotateurs ont annoté les 20 textes selon le guide et nous avons confronté les résultats. Les annotations communes à au moins deux annotateurs ont été retenues et les autres cas ont été discutés par l'équipe pour la création du corpus de référence. Celui-ci ne contenait toutefois que 89 expressions verbales (et il a été utilisé comme test pour se mettre d'accord sur les critères d'annotation). Nous avons complété celui-ci avec un autre corpus de manuels FLE, sélectionnées à partir du corpus 1 et annoté suivant le guide mis à jour après la constitution du corpus de référence. Deux annotateurs ont annoté ce nouveau corpus et ont obtenu un accord inter-annotateur ayant un bon rappel, mais une faible précision (précision : 0,56, rappel :0,97, F-mesure : 0,71). Les expressions idiomatiques et les expressions figées sont souvent correctement annotées par les deux annotateurs. Les divergences entre annotateurs se situent surtout au niveau des collocations. Par exemple, les expressions trop spécifiques à un domaine (*mener une enquête*) ou des expressions qui sont plutôt des combinaisons libres de verbes et de noms (*limiter l'accès*) ont été annoté différemment selon les annotateurs. Au total, 271 EP verbales ont

4. https://typo.uni-konstanz.de/parseme/index.php/2-general/202-parseme-shared-task-on-automatic-identification-of-verbal-mwes-edition-1-1

été identifiées manuellement (41 expressions idiomatiques, 97 collocations, 133 expressions figées). Seulement 81 EP de ce corpus de référence se retrouvent dans notre base, ce qui montre les limites de sa couverture. Notre méthode d'extraction obtient quant à elle un bon rappel (0,77), mais une précision faible (0,43) et une F-mesure de 0,55. Il est assez difficile de comparer ces résultats avec d'autres systèmes et corpus annotés, car les catégories d'expressions et les critères de sélection ne sont pas les mêmes (Ramisch *et al.*, 2018).

Ces résultats sont explicables par le nombre de variantes générées pour les collocations. Cela augmente le risque d'identifier par erreur une suite de mots similaire à une EP, mais qui n'en est pas une (ex. *il a vu le jour en décembre* vs. *il a vu le jour se lever*). Par ailleurs, l'évaluation a été effectuée en ne considérant comme corrects que les cas de reconnaissance exacte des EP présentes dans la base (à l'exception de variations telles qu'un déterminant possessif ou une préposition). Les cas de reconnaissance partielle ont été considérés comme erronés.

Sur la base de cette évaluation, nous pouvons constater que la couverture de PolylexFLE reste limitée : le nombre d'expressions trouvées dans le corpus d'évaluation est très réduit. Par contre, nous avons observé que les annotateurs humains identifient un nombre important d'expressions absentes de notre base, ce qui laisse penser qu'une analyse de nos corpus pédagogiques de plus grande ampleur pourrait se révéler utile pour compléter notre base.

4.3 Comparaison avec des outils d'extraction automatique

Pour évaluer les résultats de notre système, nous comparons les résultats de notre extracteur avec Veyn (Zampieri *et al.*, 2018), outil développé dans le cadre du projet PARSEME-FR. Ce système adopte une approche à base de réseaux de neurones pour l'annotation des expressions polylexicales. Nous avons appliqué Veyn sur le corpus annoté en niveau qui a servi pour construire le corpus de référence (le corpus FLE de 20 textes, qui contient 89 expressions). Sur les 109 expressions annotées par Veyn, seules 8 expressions sont annotés par les deux outils, ce qui est assez surprenant.

Un premier problème qui se pose est la différence entre les catégories d'expressions polylexicales utilisées. Dans le cadre du projet PARSEME-FR, ce sont surtout les expressions idiomatiques et les formes verbales pronominales qui sont annotées. Contrairement à Veyn, qui annote des expressions à verbe support (verbes support avec ou sans nom prédicatif), nous ne détaillons pas ce type d'expressions. En revanche, notre système annote principalement des collocations. PARSEME n'annote pas les collocations, car ils adopent la définition de Sag *et al.* (2001) : les collocations sont des combinaisons fréquentes de mots. Notre définition des collocations se situe plutôt dans la lignée de travaux de Baldwin & Kim (2010) : les collocations ne présentent pas uniquement des combinaisons fréquentes de mots, mais ces mots doivent entretenir des liens syntaxiques et des préférences lexicales fortes, et présenter une variabilité importante (présence des modifieurs, variation des déterminants). Signalons aussi que sur les 109 candidats extraits par Veyn, 43 d'entre-elles sont des formes verbales pronominales, que notre outil ne reconnait pas. En conséquence de ces différentes divergences, on constate peu d'intersections entre les résultats de notre extracteur et ceux de Veyn. Les 8 expressions détectées en commun sont des expressions idiomatiques, la seule catégorie d'expressions véritablement commune aux deux outils.

Un autre problème est celui de la délimitation d'expressions. Ainsi, Veyn connaît parfois des problèmes quand il rencontre une expression polylexicale, car il sélectionne parfois toute la phrase jusqu'à la fin. Ainsi, 15 expressions sur 109 ont été mal délimitées par Veyn.

En résumé, les différences de catégories des EP sont très importantes entre les deux outils, ce qui explique ces résultats divergeants. Il est dès lors compliqué de comparer les résultats des deux systèmes sans une définition commune de ces catégories.

4.4 Des distributions à un niveau CECR unique

Dans la dernière étape de ce projet, qui consistait à attribuer un niveau CECR à chaque EP qui n'en comportait pas encore, nous avons suivi la méthodologie utilisée pour construire FLELex (François *et al.*, 2014). En résumé, une fois les expressions détectées dans les textes du *Corpus-M* décrit à la section 4.1, nous les comptons afin d'obtenir, pour chaque expression, un vecteur de fréquence selon la procédure suivante. Soit une expression E_i de notre collection, celle-ci est associée à un vecteur de fréquences $F_i = (f_{A1}, f_{A2}, f_{B1}, f_{B2}, f_{C1}, f_{C2})$ dans lequel les fréquences sont initialisées à 0. Ensuite, chaque texte du corpus étant associé à l'un des six niveaux du CECR, lorsqu'une expression cible y est trouvée, nous incrémentons de 1 la fréquence f_j du vecteur où j correspond au niveau CECR du texte. À la différence de François *et al.* (2014), nous utilisons uniquement la fréquence relative pour chaque niveau, sans la multiplier par une mesure de dispersion.

La table 2 donne un aperçu des vecteurs obtenus pour quelques expressions. On peut distinguer différents profils fréquenciels pour les expressions de notre liste. Certaines, comme *aller bien* ou *avoir (+nombre) enfant*, sont plutôt typiques des premiers niveaux du CECR et des situations de communication concrètes (ex. se décrire, faire connaissance, etc.) ; d'autres sont plutôt des expressions communément utilisées dans des situations de communication plus professionnelles (ex. *prendre en compte*, *faire partie*) et se retrouvent aux différents niveaux du cadre ; enfin, certaines expressions apparaissent clairement à des stades plus avancés du processus d'apprentissage (ex. *être en droit de*, *avoir tendance*) et correspondent à un usage plus soutenu de la langue. Le principal problème de cette approche est le nombre réduit d'occurrences de nos expressions, vu la petite taille du corpus, qui nuit à la robustesse de l'estimation de ces fréquences.

expression	A1	A2	B1	B2	C1	C2	total
avoir (+nombre) enfant	6	5	0	0	0	0	11
aller bien	71	111	86	2	2	0	272
faire partie	1	2	16	3	7	7	36
prendre en compte	0	1	1	3	2	2	9
être en droit de	0	0	0	0	5	1	6
avoir tendance	0	0	6	2	5	0	13

TABLE 2 – Exemples de vecteurs de fréquence obtenus pour quelques expressions de notre liste.

Une fois les distributions de fréquences estimées, nous les avons transformé en un niveau CECR unique selon les règles suivantes. Si l'EP a été observée au sein d'un seul niveau, nous proposons ce niveau par défaut. C'est le cas de l'expression *faire l'effet d'une bombe* qui n'est observée qu'au niveau C2. Si l'EP est présente dans plusieurs niveaux, nous calculons alors la fréquence relative maximale et nous proposons le niveau où se trouve la fréquence la plus élevée. Ainsi, pour l'expression *faire partie* (*cf.* table 2), le niveau retenu sera B1. Après ce calcul, nous vérifions si l'EP est déjà associée à un niveau CECR dans la liste des expressions obtenue manuellement. Quand ce n'est pas le cas, le niveau obtenu sur le corpus est ajouté à la base. Quand un niveau existe déjà, nous le comparons à celui calculé sur le corpus et conservons systématiquement le niveau obtenu manuellement. Si la

distribution estimée pour une EP est uniforme par niveau, alors la base ne sera pas mise à jour. Pour l'expression *faire partie* nous avons identifié 44,45% d'occurences présentes dans le niveau B1, 19,44 % d'occurences sont présentes dans le niveau C1 et C2 et 8,33 % dans le niveau B2, le reste de 5,55 % est présent dans le niveau A2 et de 2,77 % dans le niveau A1. Dans ce cas, nous avons sélectionné le niveau B1, qui était le plus représenté.

En respectant cette méthodologie, nous avons extrait et annoté automatiquement 580 expressions à partir du *Corpus1* et 506 EP à partir du *Corpus2*. Certaines expressions sont présentes dans les deux corpus et aussi dans la base. Au final, 1098 expressions sont annotées dans la base, à l'aide de l'annotation automatique et manuelle.

5 Évaluation de la qualité des annotations CECR

Au terme de ces deux étapes - manuelle et automatique - nous avons été capable d'attribuer une annotation CECR à 1098 EP parmi les 4295 entrées que comprend la base PolylexFLE. Notre approche comporte toutefois deux faiblesses. D'une part, les annotations de niveau provenant de plusieurs sources tantôt pédagogiques (Beacco, Rey), tantôt des données de corpus, elles courrent le risque d'être partiellement hétérogènes si les critères d'attribution d'une expression à un niveau CECR ne sont pas homogènes. D'autre part, le corpus utilisé pour l'étude étant relativement petit , l'estimation du vecteur de fréquence semble susceptible d'être trop peu robuste. C'est pourquoi, dans cette dernière section, nous avons effectué une expérience préliminaire de la fiabilité pédagogique des annotations CECR de la base PolylexFLE.

Pour réaliser cette évaluation, nous avons constitué un questionnaire composé de 30 expressions sélectionnées selon un échantillonnage aléatoire stratifié (de 4 à 7 EP par niveau du CECR). À l'aide d'un formulaire GoogleForm, il a été demandé à chacun de nos 26 participants volontaires de renseigner son niveau CECR avant de distinguer, parmi les trente expressions, celles qu'il/elle connaît et ne connaît pas. Nous avons ensuite aggrégé les annotations sur le statut connu/inconnu de chaque expression afin de déterminer s'il y avait accord entre le niveau des EP renseignées dans notre base et les connaissances effectives des apprenants.

La distribution des niveaux de compétence parmi nos participants est la suivante : 4 apprenants A1, 11 de niveau A2, 7 participants B1, seulement un B2, 2 C1 et un de niveau C2. Cette distribution n'est pas optimale pour évaluer les EP avancées (B2 à C2), mais il est souvent plus difficile de recruter des participants plus avancés.

Afin de vérifier dans quelle mesure les niveaux CECR de nos EPs sont correctement estimés, nous avons comparé le niveau de chaque participant (N_P) avec celui de l'expression (N_E) et avons agrégé les données de l'expérience de la façon suivante. Pour une EP donnée, nous avons collecté 26 jugements, exprimés par des apprenants dont le niveau varie. Nous les avons répartis au sein de trois classes :

- Inférieure : les apprenants pour lesquels $N_P < N_E$, c'est-à-dire dont le niveau de compétence est inférieur au niveau estimé de l'EP.
- Égale : les apprenants où $N_P = N_E$.
- Supérieure : les apprenants où $N_P > N_E$.

Ensuite, nous avons calculé, au sein de chaune de ces trois classes, le pourcentage d'apprenants qui connaissait l'expression. La situation optimale devrait correspondre à un pourcentage de 0%

pour la classe inférieure (les apprenants ne sont pas encore supposés avoir étudié cette EP) ; à un pourcentage d'environ 50% pour la classe égale (certains apprenants auront déjà rencontré l'EP, tandis que d'autres pas) et 100% pour la classe supérieure (les apprenants d'un niveau de compétence supérieur devraient bien connaître cette expression). Les résultats obtenus sur nos 30 expressions sont toutefois différents de ces pourcentages idéaux (*cf.* Figure 2, quand X vaut 0). Les apprenants de niveau inférieur connaissaient tout de même nos expressions dans 45, 6% des cas ; ceux du niveau de compétence égal au niveau de l'EP la connaissaient dans 70% des cas, tandis que ceux de niveau supérieur ne les connaissaient que dans 76% des cas.

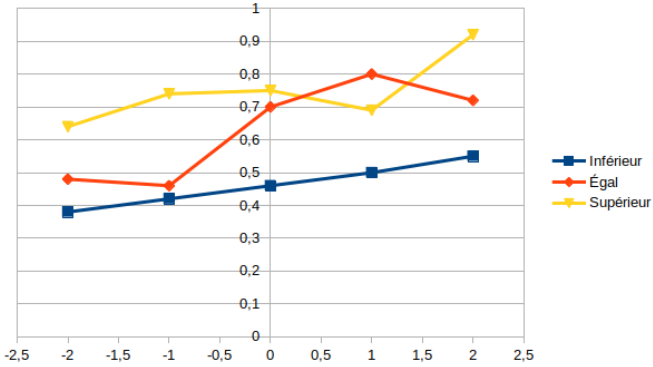

FIGURE 2 – Graphique montrant l'annotation optimale sur la base de notre échantillon

Ces résultats paraissent acceptables, même s'ils sont loins de correspondre à la situation idéale décrite ci-dessus. Afin d'obtenir une comparaison plus réaliste, nous nous sommes demandés quels auraient été ces pourcentages si nous avions classés différemment nos expressions. La Figure 2 montre les pourcentages obtenus dans 5 configurations différentes. Celle de base (X = 0) correspond à l'annotation de PolylexFLE. Dans cette configuration, l'expression *faire partie* est donc classée comme B1. Nous avons ensuite imaginé deux configurations dans lesquelles les expressions auraient été classées un (X = -1), voire deux (X = -2) niveaux en-dessous de la valeur rapportée dans PolylexFLE : ainsi, l'EP *faire partie* y serait respectivement classée comme A2 ou A1. Il y a aussi deux configurations dans lesquelles les EP auraient été classées un (X = 1), voire deux (X = 2) niveaux au-dessous de la valeur rapportée dans PolylexFLE (cad. pour l'EP *faire partie*, comme B2 ou C1). Nous pouvons observer sur la Figure 2 que la configuration qui se rapproche le plus des pourcentages optimaux est celle où la difficulté des EP est d'un niveau inférieur à ceux décrits dans PolylexFLE, en particulier pour la classe "Égale". Il est donc possible que les niveaux estimés dans PolylexFLE soient légèrement sur-évalués. Ce n'est pas totalement surprenant, puisque les expressions tirées de Beacco et ses collègues sont orientés vers la production, alors que notre ressource et notre expérience évaluent les connaissances en réception. Dès lors, le niveau de maîtrise d'une expression en production est logique supérieure à celui de sa maîtrise en réception. Une expérience de plus grande ampleur serait toutefois nécessaire pour confirmer ces résultats préliminaires.

6 Conclusion et perspectives

Pour favoriser l'apprentissage des expressions verbales polylexicales, nous avons construit une base comptant 4295 d'EP verbales, PolylexFLE, dont 1098 se sont vu attribuer un niveau CECR. Ce niveau

est soit obtenu par l'intermédiaire de ressources pédagogiques, soit calculé automatiquement à l'aide d'outils de TAL capables de repérer ces EP verbales dans un corpus lemmatisé et annoté en niveaux CECR à l'aide des patrons morpho-syntaxiques et d'en estimer la distribution. Pour améliorer la qualité de l'extraction, nous utilisons des informations issues d'une analyse syntaxique automatique afin de détecter les paires verbe-objet et vérifier s'il s'agit bien d'expressions polylexicales. Nous avons comparé notre méthode, qui utilise une base lexicale et un système à base de règles avec Veyn, l'outil d'extraction automatique des EP sur le même corpus. Toutefois, cette comparaison s'avère difficile car les catégories choisies dans le projet PARSEME et dans notre projet sont différentes. On constate que les expressions idiomatiques, la seule classe commune aux deux projets, sont reconnus par les deux systèmes. Il y a plus de différences entre collocations et constructions à verbe support. D'autre part, Veyn ne s'adresse pas explicitement aux apprenants et annote les formes verbales pronominales, qui ont un intérêt restreint pour les apprenants d'une langue. Nous n'avons pas pu comparer le corpus annoté dans le cadre du projet PARSEME FR, avec les mêmes textes annotés par notre methode d'identification, car les différences de catégories rendent la tâche très complexe.

La base PolylexFLE présentée dans cet article sera disponible en ligne pour la communauté scientifique via deux sources et sous licence Creative Common. La base PolylexFLE dans son ensemble (4295 entrées) sera rendue disponible sur le site du projet SimpleApprenant [5]. La plateforme qui intègre PolylexFLE sera également, à terme, évaluée par des apprenants de français dans plusieurs universités partenaires selon le protocole suivant : entrainement intensif, écriture des textes avec des expressions aléatoirement générées pour le niveau donné et évaluation du nombre d'erreurs. Par ailleurs, les 1098 entrées pour lesquelles une distribution de fréquence a été calculée se rapprochent très fortement des objectifs des ressources du projet CEFRLex (François *et al.*, 2014, 2016; Tack *et al.*, 2018; Dürlich & François, 2018) et sera rendue disponible sur le site du projet [6].

En ce qui concerne les perspectives ouvertes par cette étude, l'elément qui nous paraît le plus important est celui qui dérive de la méthodologie proposée. Celle-ci pourrait en effet être facilement adaptée à d'autres données, en particulier des données produites par des apprenants du FLE. Il serait alors possible de comparer les distributions de fréquence des EP du français en contexte de réception (ex. lecture) et de production. Une autre piste qui mériterait d'être approfondie est de modifier la fonction de discrétisation qui a servi à transformer les distributions de fréquence en un niveau unique. La procédure utilisée, à savoir la comparaison des fréquences par niveau, pourra être améliorée en normalisant les fréquences par une mesure de dispersion, suivant la méthodologie appliquée pour FLELex (François *et al.*, 2014).

Remerciements

Le projet SimpleApprenant a été financé par le programme IdEx de l'Université de Strasbourg pour la période de juin 2017 à février 2019. Le site Web du projet SimpleApprenant est hébérgé par la TGIR Huma-Num. Une première version du script d'extraction a été dévelopé par Colm Stapleton. Nous remercions nos partenaires de l'Université d'Opole (Pologne) (Mme Magda Danko et M. Fabrice Marsac) et de l'Université de Chypre (Mme Fabienne Baider et Mme Marina Christofi) pour leur aide précieux pour la mise en place des évaluations pédagogiques.

5. https ://simpleapprenant.huma-num.fr/SimplifyYourFrench/accueil
6. http ://cental.uclouvain.be/cefrlex/

Références

BAHNS J. & ELDAW M. (1993). Should We Teach EFL Students Collocations ? *System*, **21**(1), 101–14.

BALDWIN T. & KIM S. N. (2010). Multiword Expressions. In *Handbook of natural language processing*, p. 267–292. Boca Raton, FL : CRC Press, Taylor and Francis Group, 2 edition.

BEACCO J.-C., BOUQUET S. & PORQUIER R. (2004). *Niveau B2 pour le français : un référentiel : utilisateur-apprenant indépendant*. Mayenne : Didier.

BEACCO J.-C., LEPAGE S., PORQUIER R. & RIBA P. (2008). *Niveau A1 pour le français : utilisateur-apprenant élémentaire*. Mayenne : Didier.

CAVALLA C., LOISEAU M., DIWERSY S., LASCOMBE V. & SOCHA J. (2013). EmoProf. In *Journées Lig-Lidilem*, Eybens (Grenoble), France.

COAVOUX M. & CRABBÉ B. (2017). Incremental discontinuous phrase structure parsing with the gap transition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 1259–1270, Valencia, Spain : Association for Computational Linguistics.

CONSEIL DE L'EUROPE (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris : Hatier.

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.

DIWERSY S., GOOSSENS V., GRUTSCHUS A., KERN B., KRAIF O., MELNIKOVA E. & NOVAKOVA I. (2014). Traitement des lexies d'émotion dans les corpus et les applications d'emobase. *Corpus*, **13**, 269–293.

DÜRLICH L. & FRANÇOIS T. (2018). EFLLex : A Graded Lexical Resource for Learners of English as a Foreign Language. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, p. 873–879.

FRANÇOIS T., GALA N., WATRIN P. & FAIRON C. (2014). FLELex : a graded lexical resource for French foreign learners. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, p. 3766–3773.

FRANÇOIS T., VOLODINA E., ILDIKÓ P. & TACK A. (2016). SVALex : a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, p. 213–219.

FRANÇOIS T. (2014). An analysis of a french as a foreign language corpus for readability assessment. In *Proceedings of the 3rd workshop on NLP for Computer-assisted Language Learning, NEALT Proceedings Series Vol. 22, Linköping Electronic Conference Proceedings 107*, p. 13–32.

GROSS M. (1993). Les phrases figées en français. *L'information grammaticale*, **59**, 36–41.

GROSS M. (1994). Constructing Lexicon-grammars. In R. ATKINS & A. ZAMPOLLI, Eds., *Computational Approaches to the Lexicon*, p. 213–263, Oxford : Oxford Univ. Press.

HAMEL M.-J. & MILICEVIC J. (2007). Analyse d'erreurs lexicales d'apprenants du fls : démarche empirique pour l'élaboration d'un dictionnaire d'apprentissage. *Canadian Journal of Applied Linguistics*, **10**(1), 25–45.

HAMEL M.-J., SLAVKOV N., INKPEN D. & XIAO D. (2016). Myannotator : A tool for technology-mediated written corrective feedback. *Traitement Automatique des Langues*, **57**(3), 119–142.

HATHOUT N., SAJOUS F. & CALDERONE B. (2014). GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, p. 1007–1012, Reykjavik, Iceland.

KREMMEL B., BRUNFAUT T. & ALDERSON J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, **38**(6), 848–870.

LAPORTE E., RANCHHOD E. & YANNACOPOULOU A. (2008). Syntactic variation of support verb constructions. *Lingvisticae Investigationes*, **31**(2), 173–185. DOI : 10.1075/li.31.2.04lap.

MADNANI N., BURSTEIN J., SABATINI J., BIGGERS K. & ANDREYEV S. (2016). Language MuseTM : Automated Linguistic Activity Generation for English Language Learners. In R. ATKINS & A. ZAMPOLLI, Eds., *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, p. 213–263, Berlin : ACL.

PAWLEY A. & SYDER F. (1983). Two puzzles for linguistic theory : nativelike selection and nativelike fluency. In J. RICHARDS & R. SCHMITT, Eds., *Language and Communication*, p. 191–225. London : Longman.

RAMISCH C., CORDEIRO S., SAVARY A., VINCZE V., MITITELU V., BHATIA A., BULJAN M., CANDITO M., GANTAR P., GIOULI V. *et al.* (2018). Edition 1.1 of the parseme shared task on automatic identification of verbal multiword expressions. In *The Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 222–240.

REY I. G. (2007). *La didactique du français idiomatique*. Editions Modulaires Européennes InterCommunication.

SAG I., BALDWIN T., BOND F., COPESTAKE A. & FLICKINGER D. (2001). Multiword expressions : A pain in the neck for nlp. In *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, p. 1–15.

TACK A., FRANÇOIS T., DESMET P. & FAIRON C. (2018). NT2Lex : A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL 2018)*.

TODIRASCU A. & CARGILL M. (2019). SimpleApprenant : a Platform to Assist French L2 Language Learners to Improve Writing Skills. In *Proceedings of the 27th EUROCALL conference (to appear)*, Louvain-La-Neuve, Belgique.

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)*, Computerized and Corpus-based Approaches to Phraseology : Monolingual and Multilingual Perspectives, p. 60–67, Malaga, Spain.

VERLINDE S., BINON J. & SELVA T. (2006). The base lexicale du français (blf) : A multifunctional online database for learners of french. In C. O. ELISA CORINO, CARLA MARELLO, Ed., *Proceedings of the 12th EURALEX International Congress*, p. 471–481, Torino, Italy : Edizioni dell'Orso.

ZAMPIERI N., SCHOLIVET M., RAMISCH C. & FAVRE B. (2018). Veyn at parseme shared task 2018 : Recurrent neural networks for vmwe identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 290–296, Santa Fe, New Mexico, USA : Association for Computational Linguistics.