Démonstrateur en-ligne du projet ANR PARSEME-FR sur les expressions polylexicales

Marine Schmitt, Elise Moreau, Mathieu Constant, Agata Savary

Démonstrateur en-ligne du projet ANR PARSEME-FR sur les expressions polylexicales

Marine Schmitt[1] Élise Moreau[2] Mathieu Constant[1] Agata Savary[3]
(1) Université de Lorraine, CNRS, ATILF, France, (2) Vivoka, France (3) Université de Tours, LIFAT, France
Marine.Schmitt@atilf.fr, elise.moreau@vivoka.com,
Mathieu.Constant@univ-lorraine.fr, agata.savary@univ-tours.fr

RÉSUMÉ

Nous présentons le démonstrateur en-ligne du projet ANR PARSEME-FR dédié aux expressions polylexicales. Il inclut différents outils d'identification de telles expressions et un outil d'exploration des ressources linguistiques de ce projet.

ABSTRACT

On-line demonstrator of the PARSEME-FR project on multiword expressions.

We present an on-line demonstrator of PARSEME-FR project on multiword expressions. It includes several multiword expression identification tools, and a browser of the linguistic resources built during this project.

MOTS-CLÉS : Expressions polylexicales, identification, corpus annoté, lexique.

KEYWORDS: Multiword expressions, identification, annotated corpus, lexicon.

1 Présentation générale

Les expressions polylexicales (EPs) sont des séquences d'éléments lexicaux montrant des irrégularités de composition à différents niveaux linguistiques. Leur identification est un composant essentiel du traitement automatique des langues, mais fait face à de nombreuses difficultés : ex. discontinuité, non-compositionnalité, variabilité, ... (Constant *et al.*, 2017). Le projet ANR PARSEME-FR [1] est dédié à ce type d'expressions et vise à développer de nouvelles méthodes de traitement en combinaison avec l'analyse syntaxique et sémantique. Il a conduit à la construction de nouvelles ressources logicielles et linguistiques, distribuées sous licences libres. Dans cet article, nous présentons un démonstrateur en-ligne public [2] qui permet de tester différents outils d'identification développés par les chercheurs du projet et de parcourir un corpus annoté en EPs et une ressource lexicale qui ont été automatiquement alignés. Il est dédié au traitement de la langue française, mais il est prévu une version multilingue.

A notre connaissance, aucun outil comparable n'existe dans la communauté francophone du TAL et ceux proposés par la communauté internationale sont peu nombreux. La version 1.0 du corpus PARSEME en 15 langues (Savary *et al.*, 2018), y compris le français, peut notamment être interrogée

1. http://parsemefr.lis-lab.fr

2. Le démonstrateur (https://mwedemonstrator.atilf.fr) a été développé avec le langage Python, à l'aide du framework Django pour l'intégration web. La base de données ayant servi à inclure le lexique et le corpus a été mise en place avec SQLite. La partie interface a été construite en HTML/CSS à l'aide du framework Bootstrap, ainsi qu'en Javascript/jQuery. Enfin, nous avons utilisé un container Docker pour faciliter la mise en place de l'environnement de développement.

en ligne via les systèmes de requêtage KonText [3] et NoSke [4] (Klyueva *et al.*, 2018). Au moins deux autres systèmes d'interrogation de corpus arborés, accessibles via l'infrastructure CLARIN, permettent la recherche des EP, à condition que celles-ci y soient explicitement annotées : PML Tree Query [5] et INESS [6]. Cependant, aucune ressource lexicale d'EP ne semble ni alignée ni interrogeable via les mêmes interfaces.

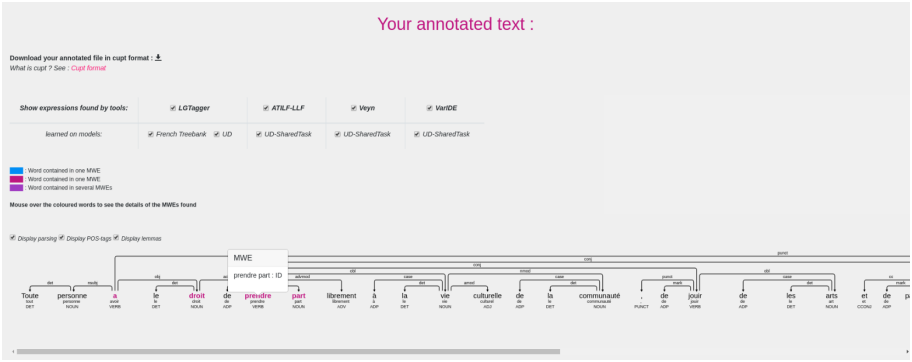

FIGURE 1 – Résultats de l'application d'outils d'identification

2 Tester des outils d'identification

La plateforme sert de vitrine aux différents outils développés lors du projet PARSEME-FR, qu'un utilisateur peut tester sur les textes de son choix. Un certain nombre d'outils sont dédiés à l'identification des expressions polylexicales verbales. Le système ATILF-LLF (Al Saied *et al.*, 2018) s'appuie sur un algorithme d'analyse par transitions et un modèle SVM de classification. Le système VarIDE (Pasquer *et al.*, 2018) se fonde sur les propriétés de variabilité des expressions à l'aide d'un modèle bayésien de classification. Le système Veyn (Zampieri *et al.*, 2018) réalise un étiquetage séquentiel s'appuyant sur des réseaux de neurones récurrents. Chacun de ces systèmes a participé à une des deux éditions de la compétition internationale PARSEME sur l'identification des expressions polylexicales verbales (Savary et al. 2017, Ramisch et al. 2018). Le système LGTagger (Constant & Sigogne, 2011) annote tous les types d'expressions non-verbales continues, via un étiquetage séquentiel se fondant sur les champs markoviens conditionnels (CRF) et exploitant des ressources lexicales.

La plateforme donne la possibilité de tester ces outils simultanément ou indépendamment sur le texte de son choix. Le texte peut être soit édité dans un champ texte, soit être téléversé dans un format brut ou au format CONLL-U [7]. Pour un texte brut, un prétraitement est appliqué au moyen de l'outil UDPipe (Straka & Straková, 2017) qui produit automatiquement la tokenisation, la lemmatisation, l'étiquetage morphosyntaxique et morphologique, ainsi que l'analyse syntaxique en dépendances dans le schéma *Universal Dependencies* (Nivre *et al.*, 2016). La plateforme affiche les résultats comme montré dans la figure 1. Il y a la possibilité de télécharger le résultat au format CUPT [8], qui est une extension du format CONLL-U intégrant une couche supplémentaire d'annotation pour les EPs.

3. http://lindat.mff.cuni.cz/services/kontext/corpora/corplist
4. http://corpora.phil.hhu.de/bonito/parseme.cgi/first_form
5. https://lindat.mff.cuni.cz/services/pmltq
6. http://clarino.uib.no/iness/page
7. https://universaldependencies.org/format.html
8. http://multiword.sourceforge.net/cupt-format

3 Explorer un corpus annoté et une ressource lexicale alignés

La plateforme permet d'explorer le corpus annoté en expressions verbales que l'équipe du projet PARSEME-FR a constitué pour les données françaises de l'édition 2017 de la compétition PARSEME (Candito *et al.*, 2017). Les expressions annotées ont été alignées automatiquement avec des entrées des tables du lexique-grammaire d'expressions figées et de noms prédicatifs (Tolone, 2012). L'exemple de la figure 2 montre une entrée du lexique extraite et structurée automatiquement par un script Python depuis les tables du lexique-grammaire. Cet alignement se fonde sur un algorithme simple faisant correspondre les éléments lexicaux figés des expressions du corpus et du lexique, avec plus ou moins de flexibilité. L'interface montre pour chaque entrée polylexicale verbale, sa tête ainsi que ses arguments, leurs catégories grammaticales et leurs réalisations syntaxiques, soit sous un format tabulaire soit sous un format graphique. L'utilisateur peut accéder pour chaque entrée aux tables du lexique-grammaire d'origine. L'interface montre également les instances annotées de cette entrée dans le corpus. A noter que s'il existe plusieurs entrées pour une expression, toutes les entrées sont alignées avec l'occurrence.

Il existe un outil de recherche dans le corpus et le lexique permettant de filtrer les expressions selon différents critères : par exemple, la valeur lexicale de sa tête verbale, sa catégorie (ex. expression idiomatique), sa longueur et autres critères avancés (ex. traits morphologiques). A chaque occurrence de l'expression, l'utilisateur a la possibilité de visualiser la ou les entrées du lexique alignées.

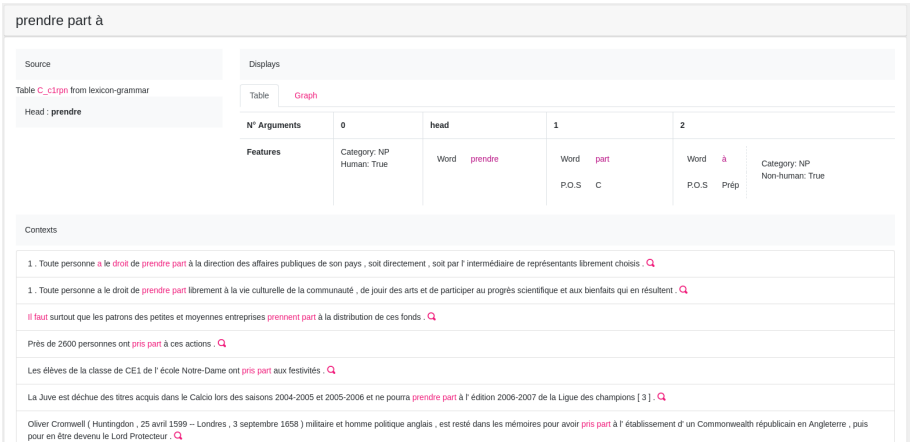

FIGURE 2 – L'entrée du lexique *prendre part*, avec les exemples annotés de cette expression dans le corpus.

Remerciements

Ce travail a bénéficié du financement de l'Agence nationale de la recherche via le projet PARSEME-FR (ANR-14-CERA-0001) et partiellement de l'action COST IC1207 PARSEME[9]. Les auteurs remercient Marie Candito, Yannick Parmentier, Carlos Ramisch, Éric Laporte et Takuya Nakamura pour leurs retours précieux.

9. http://www.parseme.eu

Références

AL SAIED H., CANDITO M. & CONSTANT M. (2018). A transition-based verbal multiword expression analyzer. In *Multiword expressions at length and in depth : Extended papers from the MWE 2017 workshop*, volume 2, p. 209 : Language Science Press.

CANDITO M., CONSTANT M., RAMISCH C., SAVARY A., PARMENTIER Y., PASQUER C. & ANTOINE J.-Y. (2017). Annotation d'expressions polylexicales verbales en français. In J.-Y. A. IRIS ESHKOL, Ed., *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, Actes de TALN, volume 2 : articles courts, p. 1–9, Orléans, France.

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.

CONSTANT M. & SIGOGNE A. (2011). MWU-aware part-of-speech tagging with a CRF model and lexical resources. In *Proc. of the ACL 2011 Workshop on MWEs*, p. 49–56, Portland, OR, USA.

KLYUEVA N., VERNEROVA A. & QASEMIZADEH B. (2018). Querying multi-word expressions annotation with CQL. In J. HAJIC, Ed., *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, TLT 2018, Prague, Czech Republic, January 23-24, 2018*, p. 73–79.

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 N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France : European Language Resources Association (ELRA).

PASQUER C., RAMISCH C., SAVARY A. & ANTOINE J.-Y. (2018). Varide at parseme shared task 2018 : Are variants really as alike as two peas in a pod ? In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 283–289, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

SAVARY A., CANDITO M., MITITELU V. B., BEJČEK E., CAP F., VOMÍR ČÉPLÖ S., CORDEIRO S. R., ERYIĞIT G., GIOULI V., VAN GOMPEL M., HACOHEN-KERNER Y., KOVALEVSKAITĖ J., KREK S., BES KIND C. L., MONTI J., ESCARTÍN C. P., VAN DER PLAS L., QASEMIZADEH B., RAMISCH C., DERICO SANGATI F., STOYANOVA I. & VINCZE V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In S. MARKANTONATOU, C. RAMISCH, A. SAVARY & V. VINCZE, Eds., *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*, p. 87–147. Berlin : Language Science Press.

STRAKA M. & STRAKOVÁ J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada : Association for Computational Linguistics.

TOLONE E. (2012). *Analyse syntaxique à l'aide des tables du Lexique-Grammaire français*. Sarrebruck, Allemagne : Éditions Universitaires Européenes. ISBN 978-3-8381-8194-3 (352 pp.).

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.