Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles
Avignon, France, 27 juin au 1er juillet 2022

Identification des Expressions Polylexicales dans les Tweets

Nicolas Zampieri[1], Carlos Ramisch[2], Irina Illina[1], Dominique Fohr[1]
(1) Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
(2) Université d'Aix-Marseille, Université de Toulon, CNRS, LIS, Marseille, France
nicolas.zampieri@inria.fr, {illina, fohr}@loria.fr,
carlos.ramisch@lis-lab.fr

RÉSUMÉ

L'identification des expressions polylexicales (EP) dans les tweets est une tâche difficile en raison de la nature linguistique complexe des EP combinée à l'utilisation d'un langage non standard. Dans cet article, nous présentons cette tâche d'identification sur des données anglaises de Twitter. Nous comparons les performances de deux systèmes : un utilisant un dictionnaire et un autre des réseaux de neurones. Nous évaluons expérimentalement sept configurations d'un système état de l'art fondé sur des réseaux neuronaux récurrents utilisant des embeddings contextuels générés par BERT. Le système fondé sur les réseaux neuronaux surpasse l'approche dictionnaire, collecté automatiquement à partir des EP dans des corpus, grâce à son pouvoir de généralisation supérieur.

ABSTRACT

Identification of Multiword Expressions in Tweets

Multiword expression (MWE) identification in tweets is a complex task due to the complex linguistic nature of MWEs combined with the non-standard language use on social networks. In this article, we present this related task on English Twitter data. We compare the performance of two systems : lexicon-based and deep neural networks-based (DNN). We experimentally evaluate seven configurations of a state-of-the-art DNN system based on recurrent networks using pre-trained contextual embeddings from BERT. The DNN-based system outperforms the lexicon-based one thanks to its superior generalization power.

MOTS-CLÉS : expression polylexicales, identification, réseau social.

KEYWORDS: multiword expressions, identification, social media.

1 Introduction

Une expression polylexicale (EP) est une combinaison d'au moins deux unités lexicales (mots), par exemple *stand for* ou *icing on the cake*, qui présente une certaine forme d'idiosyncrasie au niveau morphologique, syntaxique ou sémantique (Baldwin & Kim, 2010). L'identification d'EP est définie comme l'annotation automatique des occurrences d'EP dans un corpus. L'identification d'EP se distingue de la découverte d'EP, qui consiste à extraire une liste d'EP à partir d'un corpus (Constant *et al.*, 2017). L'identification automatique d'EP est une tâche difficile du traitement automatique des langues, puisque les EP peuvent présenter des particularités : par exemple des discontinuités ou des imbrications. Les EP à forme compositionnelle sont ambiguës, comme par exemple l'EP *tomber sur la tête* qui peut être au sens propre ou polylexicale. De plus, seuls quelques corpus annotés en EP

Volume 1 : conférence principale, pages 365–373.

sont disponibles.

Les EP peuvent être utiles pour d'autres tâches du traitement automatique des langues, comme pour la détection automatique de la parole haineuse dans les réseaux sociaux. Zampieri *et al.* (2021) ont montré qu'intégrer l'information des mots appartenant au EP améliore la distinction entre les tweets haineux et normaux. Les travaux de recherche dans le présent article sont issus de l'article de Zampieri *et al.* (2022). Le but du papier de Zampieri *et al.* (2022) est d'étudier la robustesse de différents systèmes de l'identification des EP sur les tweets, afin d'utiliser les meilleurs systèmes pour étiqueter automatiquement des corpus de tweets annotés pour la détection de la parole haineuse.

Ici, nous nous intéressons à l'étude de l'efficacité de systèmes état de l'art d'identification d'EP dans des textes non standard, à savoir des tweets. En effet, les tweets utilisent souvent une syntaxe non standard et contiennent des fautes d'orthographe, des abréviations, etc. Nous émettons l'hypothèse que, dans ces conditions, la tâche d'identification d'EP devient encore plus difficile. Afin de mener à bien notre étude, nous proposons de confronter deux approches : la première utilise un dictionnaire et la seconde est fondée sur des réseaux neuronaux. Concernant cette dernière, nous évaluons différentes configurations.

2 État de l'art

L'identification automatique des EP a été abordée dans le passé à l'aide de modèles statistiques d'étiquetage de séquences, par exemple les champs aléatoires conditionnels *conditional random field* (CRF) (Constant *et al.*, 2012) ou un perceptron (Schneider *et al.*, 2014). Des analyseurs syntaxiques en dépendances par transitions (Constant & Nivre, 2016) et basés sur des grammaires de substitutions d'arbre (Green *et al.*, 2013) ont aussi été utilisés. L'identification des EP a également été réalisée à l'aide de dictionnaires et de systèmes basés sur des règles (Cordeiro *et al.*, 2016).

Les systèmes soumis à de récentes campagnes d'évaluation ont conduit à des avancées dans l'état de l'art (Schneider *et al.*, 2016; Savary *et al.*, 2017; Ramisch *et al.*, 2018, 2020). Le meilleur système de la campagne PARSEME 2017, nommé Transition, a été adapté à partir des travaux de Constant & Nivre (2016) en utilisant un analyseur syntaxique basé sur des transitions. En 2018, le meilleur système, TRAVERSAL, utilisait des arbres discriminants (Waszczuk, 2018). Certains modèles basés sur des réseaux neuronaux ont obtenus de bonnes performances, comme par exemple TRAPACC (Stodden *et al.*, 2018). L'édition 2020 a bénéficié d'avancées des modèles de langages pré-entraînés, comme en témoigne le meilleur système, *MTLB-struct*, basé sur le modèle BERT et utilisant un entraînement multi-tâches (Taslimipoor *et al.*, 2020). L'analyseur lexical de Liu *et al.* (2021) est un système basé sur un modèle BERT qui prédit les EP et les étiquettes de super-sens simultanément. Ce système a obtenu des résultats impressionnants sur le corpus Streusle (Schneider & Smith, 2015) et a également été évalué sur le corpus PARSEME (Ramisch *et al.*, 2018) et sur DimSum (Schneider *et al.*, 2016).

Certains articles ont analysé les performances de l'identification d'EP. Maldonado & QasemiZadeh (2018) ont montré que les performances des systèmes d'identification d'EP sont étroitement liées au taux d'EP non vues dans les corpus d'apprentissage. Savary *et al.* (2019) soutiennent que les dictionnaires sont nécessaires pour obtenir une meilleure généralisation de l'identification d'EP, où la généralisation est plus difficile que dans des tâches similaires telles que la reconnaissance d'entités nommées.

3 Methodologie

L'objectif de la tâche d'identification automatique des EP est d'étiqueter les mots qui composent les EP. Nous proposons d'analyser la robustesse de deux systèmes d'identification d'EP dans des tweets : un **système basé sur un dictionnaire par collecte *ad hoc*** avec la boîte à outils *mwetoolkit* (Cordeiro *et al.*, 2016), et un **système de reconnaissance lexicale** (LSR) utilisant des réseaux de neurones (Liu *et al.*, 2021).

Pour le système à base de dictionnaire par collecte *ad hoc*, nous extrayons une liste d'EP à partir de plusieurs corpus annotés manuellement. Chaque mot extrait des EP est lemmatisé et les formes canoniques des EP sont placées dans le dictionnaire d'EP. Le dictionnaire contient à la fois des EP comprenant des mots contigus et non contigus dans les corpus annotés. Pour ces derniers, seuls les mots composant l'EP sont conservés, sans tenir compte des mots intermédiaires.

Le système LSR est basé sur des réseaux neuronaux et a une plus grande capacité de généralisation par rapport au système basé sur le lexique. L'architecture LSR se compose du modèle BERT (Devlin *et al.*, 2019), suivi de deux couches bidirectionnelles *long-short term memory* (LSTM) et d'une couche de CRF. Nous utilisons ce système dans nos expérimentations étant donné qu'il est récent et qu'il a obtenu de bons résultats lors d'évaluations. Nous nous intéressons à l'étude de différentes configurations d'apprentissage du système LSR : variation de la quantité et de la nature des corpus d'apprentissage et utilisation de différents schémas d'étiquettes "BIO" (*beginning, inside, outside*) [1] (voir Figure 1).

Le schéma "BIObio" est similaire au schéma d'étiquette BIO original avec des catégories d'EP et des étiquettes de super-sens Liu *et al.* (2021). Les étiquettes "b", "i" et "o" ont la même signification que les étiquettes "B", "I" et "O", mais l'EP étiquetée ainsi est imbriquée dans une autre EP. Les catégories des EP sont concaténées aux balises "B" et "b". Les catégories lexicales et les catégories d'EP (par exemple, VID pour les verbales idiomatiques, VPC pour les constructions verbe-particule) sont concaténées avec les balises initiales "B" et "b". Le schéma d'étiquettes "BIOo-cat" concatène les catégories d'EP sur les étiquettes "B" et "I". Le jeu d'étiquettes "BIOo" n'utilise pas les catégories d'EP. Contrairement à Liu *et al.* (2021), ces deux schémas ne contiennent pas les étiquettes de super-sens. Le système LSR peut prédire une séquence d'étiquettes invalide : par exemple, dans une phrase un mot étiqueté avec un "I" apparaît sans qu'un mot n'ai été précédemment étiqueté avec un "B". Pour corriger ces séquences d'étiquettes invalides, nous appliquons un filtrage sur les sorties du système LSR.

4 Configuration des expériences

Dans cette section, nous décrivons les corpus et les configurations de nos systèmes. La Table 1 présente les statistiques de trois corpus utilisés.

— **Streusle** est un corpus de critiques en ligne annoté en termes de super-sens et d'EP *faibles* (par exemple, *narrow escape*, *do not be surprised*) et *fortes* (par exemple, *go out of my way*, *close call*) (Schneider & Smith, 2015). Le corpus est annoté en 20 catégories d'EP. Le corpus est divisé en trois parties : apprentissage, développement et test. Nous utilisons la version 4.3

1. Chaque unité lexicale séparée par un espace (appelé token) est marqué par l'étiquette "B" s'il se trouve au début d'une EP, "I" s'il est à l'intérieur d'une EP et "O" s'il n'appartient pas à une EP.

BIOo	O	B	o	o	I	O	O	B	I	O
BIOo-cat	O	B-V.LVC.full	o	o	I-V.LVC.full	O	O	B-N	I-N	O
BIObio	O-PRON	B-V.LVC.full.v.social	o-DET	o-ADJ	I_	O-P-p.purpose	O-DET	B-N-n.body	I_	O-PUNCT
Sentence	I	**had**	a	routine	**surgery**	for	an	**ingrown**	**toenail**	.

FIGURE 1 – Exemple d'étiquetage "BIO" pour le modèle LSR. Cet exemple possède deux EP : *had surgery* et *ingrown toenail*. L'étiquette du premier mot composant une EP est "B" (*begin*), les mots suivants de l'EP sont étiquettés "I" (*inside*). Pour le "BIOo-cat" et le "BIObio", les catégories des EP sont ajoutées aux étiquettes : catégorie lexicale (par exemple, "V" pour verbale, "N" pour nominale), et catégorie des EP (par exemple, "LVC.full" pour les constructions à verbe support pleines, de l'anglais *full light-verb construction*). Pour la configuration "BIObio", des identifiants de supersenses (par exemple "n.body" pour *body parts*) sont ajoutés aux étiquettes. Des catégories lexicales sont également ajoutées aux balises "O" (*outside*) mais pas aux étiquettes "I", comme dans l'article de Liu *et al.* (2021).

Corpus		#phrases	#token	#EP
Streusle	App	2 724	44 822	2 425
	Dev	554	5 394	283
	Test	535	5 381	281
PARSEME	App	3 471	53 201	331
	Test	3 965	71 002	501
DimSum partie tweet	App	987	18 247	1 112
	Test	500	6 627	362

TABLE 1 – Nombre de phrases, de token et d'occurrences d'EP des partitions standard dans les corpus d'apprentissage, de développement et de test pour les corpus Streusle, PARSEME et DimSum.

du corpus Streusle.

— Le corpus **PARSEME** (Ramisch *et al.*, 2018) contient, entre autres, des news, des extraits de site internet et de Wikipédia. Il est annoté uniquement en termes d' EP verbales fortes. Six catégories d'EP verbales sont considérées. Le corpus PARSEME n'est disponible qu'en version 1.1, et il est divisé en corpus d'apprentissage et de test, sans corpus de développement.

— Le corpus **DimSum** (Schneider *et al.*, 2016) contient des critiques en ligne, des transcriptions de discussions (TED talk) et des tweets. Dans notre travail, nous n'utilisons que la partie tweet de ce corpus car nous concentrons nos expériences sur les tweets. Le corpus est annoté en termes d'EP fortes sans étiquettes de catégories. Nous exploitons la partie test de ce corpus pour évaluer nos systèmes d'identification.

Comme les corpus PARSEME et DimSum sont annotés en termes d'EP fortes, nous ne prenons en compte que les annotations d'EP fortes du corpus Streusle, les EP faibles ne sont pas prises en compte (sauf pour la configuration LSR_1). Pour tous les corpus, à l'exception de DimSum test, on filtre les EP : dans une phrase donnée, lorsqu'un mot est commun à deux EP (chevauchement d'EP) ou que deux EP sont imbriquées, on supprime la deuxième EP [2]. Ce phénomène est peu fréquent et se produit dans moins de 5% des phrases.

Pour la configuration du système à base de dictionnaire, nous extrayons les EP de tous les corpus

2. l'EP la plus courte si elles commencent à la même position.

présentés ci-dessus, à l'exception du corpus de test de DimSum : corpus d'apprentissage de développement et de test de Streusle, corpus d'apprentissage et de test de PARSEME, et corpus d'apprentissage de DimSum. Le dictionnaire obtenu contient 3255 EP. Nous utilisons le corpus d'apprentissage de DimSum pour régler les paramètres du système basé sur le dictionnaire. Nous avons évalué l'utilisation de l'étiquetage morpho-syntaxique (*part-of-speech*) avec les lemmes des mots composants une EP. L'étiquetage morpho-syntaxique n'a pas montré d'amélioration dans l'identification d'EP sur le corpus de développement. Donc, nous utilisons uniquement les lemmes. Nous avons également expérimenté plusieurs valeurs pour ajuster l'écart maximal entre les mots composant les EP discontinues. La valeur optimale, 3, est utilisée dans les expériences ci-dessous.

Pour le modèle LSR, nous entraînons sept configurations. Pour chaque configuration, nous effectuons cinq apprentissages avec initialisation aléatoire. Nous rappelons que les configurations LSR proposées diffèrent en ce qui concerne les données d'apprentissage et la granularité des étiquettes.

— La configuration $\mathbf{LSR}_1$ correspond au système proposé dans Liu *et al.* (2021). Dans cette configuration, nous entraînons le modèle LSR sur le corpus d'apprentissage de Streusle utilisant le schéma d'étiquettes "BIObio" comme dans Liu *et al.* (2021) avec les EP faibles et fortes. Il s'agit d'un système d'étiquetage complexe qui compte plus de 600 étiquettes.
— La configuration $\mathbf{LSR}_2$ est également entraînée sur le corpus d'apprentissage de Streusle. Nous utilisons les étiquettes "BIOo-cat". Les étiquettes de super-sens ne sont pas utilisées. Le nombre final d'étiquettes est de 42.
— La configuration $\mathbf{LSR}_3$ est également entraînée sur le corpus d'apprentissage de Streusle. Nous utilisons le schéma d'étiquettes "BIOo" qui contient seulement 4 étiquettes. L'objectif des configurations LSR_1, LSR_2 et LSR_3 est d'étudier l'impact de différents schémas d'étiquetage sur l'identification d'EP.
— La configuration $\mathbf{LSR}_4$ utilise le corpus d'apprentissage DimSum pour entraîner le modèle LSR. Le corpus DimSum n'étant pas étiqueté en termes de catégories pour les EP, nous utilisons le schéma d'étiquetage "BIOo" avec 4 étiquettes.
— La configuration $\mathbf{LSR}_5$ est entraînée sur les corpus d'apprentissage de DimSum (tweets) et de Streusle (non-tweets). Nous utilisons les étiquettes "BIOo" avec 4 étiquettes, comme dans LSR_4.
— La configuration $\mathbf{LSR}_6$ est l'union des prédictions de deux modèles. Le premier est entraîné à l'aide des corpus d'apprentissage de PARSEME et de Streusle et utilise uniquement des EP verbales (14 étiquettes). Le second est entraîné sur le corpus d'apprentissage de Streusle pour prédire les EP non-verbales (30 étiquettes). Cette configuration utilise le jeu d'étiquettes "BIOo-cat". Si la prédiction finale possède une imbrication ou un chevauchement d'EP, nous choisissons de conserver l'EP qui apparaît en premier.
— La configuration $\mathbf{LSR}_7$ est identique à celle de LSR_6, à l'exception du jeu d'étiquettes. Dans cette configuration, nous adoptons le schéma d'étiquettes "BIOo" avec 4 étiquettes.

Pour chaque configuration LSR décrite ci-dessus, nous utilisons le corpus de développement de Streusle pour ajuster les paramètres. Concernant le filtrage des sorties LSR, nous avons évalué quelques heuristiques et adopté les suivantes : nous supprimons les EP à token unique, les étiquettes "I" qui ne sont pas précédées d'un "B", et les EP contenant des tokens spéciaux (@USER, URL, et hashtags). La longueur maximale de l'écart des EP non contiguës a également été ajustée et définie à 2, nous supprimons donc tous les EP contenant des écarts strictement supérieurs à 2.

Pour donner un ordre d'idée de l'impact sur l'environnement de chacune des méthodes développées dans cet article, nous donnons quelques indices concernant le temps de calcul. Le temps d'exécution de la méthode par dictionnaire est de l'ordre de quelques secondes (sur CPU) et pour le système LSR

de l'ordre d'une minute. Le temps d'apprentissage d'un système LSR est de l'ordre de 30 minutes avec un GPU.

Nous évaluons les systèmes d'identification d'EP à l'aide de mesures standard, qui ont été appliquées pour les campagnes d'évaluation PARSEME (Savary *et al.*, 2017) et DimSum (Schneider *et al.*, 2016). La mesure *MWE-based* est le score F1 pour les EP entièrement prédites. La mesure *token-based* est la mesure F1 pour les tokens appartenant à une EP, évaluant les correspondances partielles. La mesure *MWE-link-based* est le score F1 basé sur la correspondance des paires de mots adjacents, et tient compte des EP partiellement correctes.

5 Résultats

Dans cette partie, nous présentons les résultats obtenus pour l'identification d'EP sur le corpus de test de DimSum.

La Table 2 montre que le système à base de dictionnaire atteint 28,7% pour le score F1 *MWE-based*. Cette performance est assez faible car 78% des EP présentes dans le corpus de test de DimSum ne sont pas présentes dans le dictionnaire d'EP et donc ne peuvent pas être trouvées par le système. Toutes les configurations LSR surpassent l'approche à base de dictionnaire par collecte *ad hoc*. Nous observons que les configurations LSR_{2-7} améliorent le rappel et la précision (en termes de *MWE-based*) par rapport à l'approche à base de dictionnaire. En effet, les configurations LSR ont une meilleure capacité de généralisation et peuvent détecter des EP qui ne sont pas présentes dans le corpus d'apprentissage. Le LSR_5 obtient les meilleurs résultats en termes de *MWE-based*, *token-based* et *MWE-link-based*. En comparant les scores *token-based* obtenus par les systèmes LSR_5 et dictionnaire (56,8% contre 28,5%), nous observons que le système LSR_5 prédit mieux les EP partielles.

Afin d'étudier l'impact des schémas d'étiquettes, nous comparons les trois configurations LSR_1, LSR_2 et LSR_3 entraînées sur le même corpus avec des schémas différents : "BIObio", "BIOo-cat" et "BIOo". A partir de la Table 2, nous observons qu'un jeu d'étiquette complexe ("BIObio") diminue les scores F1. En effet, le système LSR_1 obtient 36,1% de *MWE-based* contre 43,3% atteint par les systèmes LSR_2 ou LSR_3. Nous observons la même performance pour les configurations utilisant le schéma "BIOo-cat" ou "BIOo" indiquant que les catégories des EP n'aident pas le système. Ce constat est confirmé avec les résultats obtenus par les systèmes LSR_6 et LSR_7.

Maintenant, nous nous concentrons sur les configurations exploitant le même schéma d'étiquettes mais différents corpus d'apprentissage : LSR_3, LSR_4, LSR_5 et LSR_7. Nous observons que LSR_4 a les scores F1 les plus bas, atteignant 41,2% de *MWE-based*. Cela peut être dû au fait qu'il est entraîné avec le plus petit corpus d'apprentissage, contenant 987 phrases, par rapport aux corpus des systèmes LSR_3, LSR_5 et LSR_7 qui contiennent plus de 2724 phrases. Le système LSR_7, qui est entraîné sur les corpus d'apprentissage de Streusle et de PARSEME, n'améliore pas les résultats, en termes de score F1, par rapport au système LSR_3, qui utilise uniquement Streusle. Cela peut être dû au fait que le système LSR_7 utilise deux modèles de réseaux neuronaux entraînés indépendamment. Le système LSR_5, qui est entraîné sur les corpus d'apprentissage de Streusle et de DimSum, obtient les meilleurs scores F1 avec respectivement 46,5%, 56,8% et 54,0% de *MWE-based*, de *token-based* et de *MWE-link-based*. Cela peut être dû au fait que le système LSR_5 est entraîné sur des données de tweets et de non-tweets.

Configurations (corpus d'apprentissage)	BIO	MWE-based			Token-based	MWE-link-based
		Précision	Rappel	score F1	score F1	score F1
Lexicon-based	-	45,5	21,0	28,7	28,5	25,9
LSR_1 (ST)	BIObio	$45,5 \pm 3,4$	$29,9 \pm 2,0$	$36,1 \pm 2,4$	$47,6 \pm 1,3$	$43,8 \pm 1,4$
LSR_2 (ST)	BIOo-cat	53,7 $\pm 1,1$	36,4 $\pm 2,6$	43,3 $\pm 1,6$	53,5 $\pm 2,1$	51,2 $\pm 2,1$
LSR_3 (ST)	BIOo	49,0 $\pm 2,7$	39,2 $\pm 4,1$	43,3 $\pm 1,5$	54,7 $\pm 1,8$	52,0 $\pm 2,0$
LSR_4 (DSM)	BIOo	**61,1** $\pm 2,7$	31,2 $\pm 2,6$	41,2 $\pm 2,3$	51,8 $\pm 3,1$	48,5 $\pm 2,7$
LSR_5 (ST-DSM)	BIOo	60,4 $\pm 2,5$	37,9 $\pm 0,9$	**46,5** $\pm 0,3$	**56,8** $\pm 1,0$	**54,0** $\pm 1,5$
LSR_6 (ST-PSM)	BIOo-cat	53,2 $\pm 1,5$	37,1 $\pm 2,0$	43,6 $\pm 1,3$	54,1 $\pm 1,7$	50,9 $\pm 1,7$
LSR_7 (ST-PSM)	BIOo	50,0 $\pm 4,4$	**39,9** $\pm 3,3$	44,1 $\pm 0,9$	54,7 $\pm 2,1$	51,9 $\pm 2,4$

TABLE 2 – Résultats d'identification d'EP sur les tweets de test de DimSum. Pour chaque résultat, le score moyen et l'écart type de 5 apprentissages sont donnés (sauf pour la configuration à base de dictionnaire). "ST", "DSM" et "PSM" représentent respectivement Streusle, DimSum et PARSEME. La colonne "BIO" représente le schéma d'étiquettes pour les configurations LSR.

6 Conclusion

Dans ce travail, nous avons étudié les performances des systèmes d'identification d'EP, basés sur un dictionnaire ou des réseaux de neurones, en se concentrant sur des corpus de tweets. Nous avons proposé et évalué 7 configurations pour le système LSR. Nous avons constaté que les systèmes LSR surpassent le système à base de dictionnaire. La meilleure configuration du système LSR est celle qui est entraînée sur des tweets et des données non-tweets.

A cour terme, nous envisageons d'effectuer une analyse détaillée des erreurs des systèmes LSR_5 et LSR_7 qui ont des performances similaires. Dans des travaux futurs, nous aimerions combiner des approches basées sur des réseaux neuronaux et sur le dictionnaire pour augmenter la généralisation de l'identification d'EP.

Références

BALDWIN T. & KIM S. (2010). Multiword expressions. In *Handbook of Natural Language Processing*, p. 267–292 : CRC Press, Taylor and Francis Group, Boca Raton, 2nd edition.

CONSTANT M., ERYIĞIT G., MONTI J., VAN DER PLAS L., RAMISCH C., ROSNER M. & TODIRASCU A. (2017). Survey : Multiword expression processing : A Survey. *Computational Linguistics*, **43**(4), 837–892. DOI : 10.1162/COLI_a_00302.

CONSTANT M. & NIVRE J. (2016). A transition-based system for joint lexical and syntactic analysis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 161–171, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1016.

CONSTANT M., SIGOGNE A. & WATRIN P. (2012). Discriminative strategies to integrate multiword expression recognition and parsing. In *Proceedings of the 50th Annual Meeting of the Associa-*

tion for Computational Linguistics (Volume 1 : Long Papers), p. 204–212, Jeju Island, Korea : Association for Computational Linguistics.

CORDEIRO S., RAMISCH C. & VILLAVICENCIO A. (2016). UFRGS&LIF at SemEval-2016 task 10 : Rule-based MWE identification and predominant-supersense tagging. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 910–917, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1140.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186.

GREEN S., DE MARNEFFE M.-C. & MANNING C. D. (2013). Parsing models for identifying multiword expressions. *Computational Linguistics*, **39**(1), 195–227. DOI : 10.1162/COLI_a_00139.

LIU N. F., HERSHCOVICH D., KRANZLEIN M. & SCHNEIDER N. (2021). Lexical semantic recognition. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, p. 49–56, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.mwe-1.6.

MALDONADO A. & QASEMIZADEH B. (2018). Analysis and insights from the parseme shared task dataset. In *Multiword expressions at length and in depth*, p. 149–176, Berlin : Language Science Press.

RAMISCH C., CORDEIRO S. R., SAVARY A., VINCZE V., BARBU MITITELU V., BHATIA A., BULJAN M., CANDITO M., GANTAR P., GIOULI V., GÜNGÖR T., HAWWARI A., IÑURRIETA U., KOVALEVSKAITĖ J., KREK S., LICHTE T., LIEBESKIND C., MONTI J., PARRA ESCARTÍN C., QASEMIZADEH B., RAMISCH R., SCHNEIDER N., STOYANOVA I., VAIDYA A. & WALSH A. (2018). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 222–240, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

RAMISCH C., SAVARY A., GUILLAUME B., WASZCZUK J., CANDITO M., VAIDYA A., BARBU MITITELU V., BHATIA A., IÑURRIETA U., GIOULI V., GÜNGÖR T., JIANG M., LICHTE T., LIEBESKIND C., MONTI J., RAMISCH R., STYMNE S., WALSH A. & XU H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, p. 107–118, online : Association for Computational Linguistics.

SAVARY A., CORDEIRO S. & RAMISCH C. (2019). Without lexicons, multiword expression identification will never fly : A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, p. 79–91, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/W19-5110.

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. DOI : 10.18653/v1/W17-1704.

SCHNEIDER N., DANCHIK E., DYER C. & SMITH N. A. (2014). Discriminative lexical semantic segmentation with gaps : Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, **2**, 193–206. DOI : 10.1162/tacl_a_00176.

SCHNEIDER N., HOVY D., JOHANNSEN A. & CARPUAT M. (2016). SemEval-2016 task 10 : Detecting minimal semantic units and their meanings (DiMSUM). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, p. 546–559, San Diego, California : Association for Computational Linguistics. DOI : 10.18653/v1/S16-1084.

SCHNEIDER N. & SMITH N. A. (2015). A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 1537–1547, Denver, Colorado : Association for Computational Linguistics. DOI : 10.3115/v1/N15-1177.

STODDEN R., QASEMIZADEH B. & KALLMEYER L. (2018). TRAPACC and TRAPACCS at PARSEME shared task 2018 : Neural transition tagging of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 268–274, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

TASLIMIPOOR S., BAHAADINI S. & KOCHMAR E. (2020). MTLB-STRUCT @parseme 2020 : Capturing unseen multiword expressions using multi-task learning and pre-trained masked language models. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, p. 142–148, online : Association for Computational Linguistics.

WASZCZUK J. (2018). TRAVERSAL at PARSEME shared task 2018 : Identification of verbal multiword expressions using a discriminative tree-structured model. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 275–282, Santa Fe, New Mexico, USA : Association for Computational Linguistics.

ZAMPIERI N., ILLINA I. & FOHR D. (2021). Multiword expression features for automatic hate speech detection. In E. MÉTAIS, F. MEZIANE, H. HORACEK & E. KAPETANIOS, Éds., *Natural Language Processing and Information Systems*, p. 156–164, Cham : Springer International Publishing.

ZAMPIERI N., RAMISCH C., ILLINA I. & FOHR D. (2022). Identification of multiword expressions in tweets for hate speech detection. In *Proceedings of the 13th Language Resources and Evaluation Conference*, p. to appear, Marseille, France : European Language Resources Association.