Towards a Variability Measure for Multiword Expressions

Caroline Pasquer, Agata Savary, Jean-Yves Antoine, Carlos Ramisch

Towards a Variability Measure for Multiword Expressions

Caroline Pasquer and **Agata Savary** and **Jean-Yves Antoine**
University of Tours, France
first.last@univ-tours.fr

Carlos Ramisch
Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
carlos.ramisch@lis-lab.fr

Abstract

One of the outstanding properties of multiword expressions (MWEs), especially verbal ones (VMWEs), important both in theoretical models and applications, is their idiosyncratic variability. Some MWEs are always continuous, while some others admit certain types of insertions. Components of some MWEs are rarely or never modified, while some others admit either specific or unrestricted modification. This unpredictable variability profile of MWEs hinders modeling and processing them as "words-with-spaces" on the one hand, and as regular syntactic structures on the other hand. Since variability of MWEs is a matter of scale rather than a binary property, we propose a 2-dimensional language-independent measure of variability dedicated to verbal MWEs based on syntactic and discontinuity-related clues. We assess its relevance with respect to a linguistic benchmark and its utility for the tasks of VMWE classification and variant identification on a French corpus.

1 Introduction

Multiword expressions (MWEs), in particular verbal ones (VMWEs), are groups of words whose meaning does not derive from the meaning of their components and from their syntactic structure in a regular way (Gross, 1982), like ***pay** a **visit*** and ***take the cake*** 'be the most remarkable of its kind'. MWEs exhibit some degree of variability. On the one hand, they allow internal inflection (***paid** many **visits***), insertions (***pay** annual **visits***) and syntactic transformations (***visits paid** last month*). On the other hand, they can block variation that is usual/typical for ordinary expressions with the same syntactic structure, such as inflection (*#take a turn*[1] vs. ***take turns***), diathesis alternation (*#he cast the die* vs. ***the die is cast*** 'the point of no retreat is passed'), or adjunction of modifiers (*#take the sweet cake*). This leads to variation schemes which are specific to subclasses of MWE, that is, MWE variability is idiosyncratic.

Variability, also known as flexibility, has been considered a key property of MWEs in linguistic studies (Gross, 1988; Tutin, 2016; Nunberg et al., 1994; Sheinfux et al., 2017). It was also highlighted as a major challenge in NLP models and applications (Constant et al., 2017). Variants are pervasive (Jacquemin, 2001) and hinder straightforward search of MWE citation forms in a corpus (Nissim and Zaninello, 2013). They introduce discontinuities which challenge sequence labeling approaches. Even when employing parsers to cope with discontinuities, MWE recognizers can still fail to capture some syntactic transformations such as complex determiners, which can break a direct link between a verb and a noun in a dependency tree (***pay** a series of **visits***). These facts have important implications for downstream tasks and applications, e.g. parsers can heavily suffer from incorrectly identified MWEs (Baldwin et al., 2004).

The restricted variability of MWEs as compared to their regular counterparts can also be seen as an advantage in their automatic discovery (Weller and Heid, 2010; Tsvetkov and Wintner, 2014; Buljan and Šnajder, 2017). Substitution-based MWE discovery techniques based on lexico-semantic variability have been largely explored (Pearce, 2001; Farahmand and Henderson, 2016). Morphological and syntactic variability, however, have rarely been studied for MWE discovery (Ramisch et al., 2008) and even less so for in-context identification (Fazly et al., 2009).

Given the importance of MWE variability (Constant et al., 2017) as well as its gradual nature, especially for VMWEs, we suggest that this phenomenon should be subject to measurement. This paper presents measures of VMWE variability based on variant-to-variant similarity, taking syn-

[1] We use # to signal a loss of idiomatic meaning.

tactic variability and linear discontinuity into account (Sec. 2–3).[2] Our proposal is evaluated on a French corpus (Sec. 4). We assess the relevance of our measure with respect to a linguistic benchmark (Sec.5), and we study its usability for VMWE classification (Sec. 6) and variant identification (Sec. 7). Then, we conclude and sketch perspectives to extend our proposal to other languages and to an unsupervised framework (Sec. 8).

2 Variant-to-variant similarity

To capture the variability of a VMWE, we rely on pairwise comparison of its occurrences. Fig. 1 shows the dependency trees of sentences containing two variants, henceforth V_1 and V_2, of ***prendre** une **décision*** 'to take a decision'. V_1 and V_2 exhibit some common and some divergent syntactic and linear properties. For instance, the noun *decision* governs a determiner (*det*) and an adjectival modifier (*amod*) both in V_1 and in V_2, and a relative clause (*acl:relcl*) in V_2. The verb *take* governs a nominal subject (*nsubj*), an object (*obj*) and adverbial modifiers (*adv*) in both V_1 and V_2, and an auxiliary (*aux*) in V_2. External elements are inserted between the lexicalized ones in both variants. Their POS are *adv* (twice), *det* and *adj* in V_1, and *pron*, *propn*, *aux* and *adv* in V_2, i.e. one POS (*adv*) is shared.[3]

In order to measure both these common characteristics and discrepancies, we define the similarity of two VMWE variants on the basis of the similarity of their components and of the external inserted elements. A *lexicalized component*, or simply a *component*, of a VMWE E is the one which is realized by the same lexeme in any variant of E.[4] All variants of E necessarily have the same number of lexicalized components, which are lemmatized and lexicographically sorted, yielding a canonical form of $E = (C_1, C_2, \ldots, C_n)$ which uniquely represents it.[5] By C_i^j we denote the form that component i takes in variant j. For instance, in Fig. 1 $C_1 =$ décision, $C_2 =$ prendre, $E =$ (décision, prendre), $C_1^1 =$ décision, $C_1^2 =$ décisions, $C_2^1 =$ prennent and $C_2^2 =$ prises. Similarity of objects (components or VMWEs) is measured by the Sørensen–Dice coefficient, which is defined as $S(O_1, O_2) = 2 \times |P(O_1) \cap P(O_2)|/(|P(O_1)| + |P(O_2)|)$, where $P(O_1)$ and $P(O_2)$ denote the sets of (relevant) properties exhibited by objects O_1 and O_2. We now define two variant-to-variant similarity measures: syntactic – focusing on the outgoing dependencies – and linear – based on insertions.

2.1 Syntactic similarity

Syntactic similarity S^S is based on the dependencies between a VMWE and its external elements. It allows us to account for long-distance arguments and modifiers not necessarily included between the lexicalized components. The similarity of each pair of lexicalized components is calculated first, and then averaged for the whole VMWE. For each component, the set of outgoing dependencies is considered and relations of the same type are counted once. In the two sentences given in Fig. 1, the syntactic similarity of the noun C_1 and the verb C_2 is:

$$S^S(C_1^1, C_1^2) = \frac{2 \times |\{\text{amod,det}\}|}{|\{\text{acl:relcl,amod,det}\}| + |\{\text{amod,det}\}|} = \frac{4}{5}$$

$$S^S(C_2^1, C_2^2) = \frac{2 \times |\{\text{adv,nsubj,obj}\}|}{|\{\text{adv,nsubj,obj}\}| + |\{\text{adv,aux,nsubj,obj}\}|} = \frac{6}{7}$$

Variant-to-variant syntactic similarity is the weighted average of the per-component scores:

$$S^S(V_1, V_2) = \sum_{i=1}^{n} w_i \times S^S(C_i^1, C_i^2)$$

where weights $w_1, \ldots, w_n$ sum up to 1. For instance, with uniform weights $w_1 = w_2 = \frac{1}{2}$:

$$S^S(V_1, V_2) = \frac{1}{2} \times \frac{4}{5} + \frac{1}{2} \times \frac{6}{7} = \frac{29}{35}$$

2.2 Linear similarity

Linear similarity S^L is defined for two VMWE variants in terms of the POS of the elements inserted between the lexicalized components. The number of insertions for the same POS is disregarded. In this way we focus on the quality of admitting an insertion of a certain POS, rather than on their count. For example, the two *adv* insertions in V_1 (*vraiment* '*really*' and *pas* '*NEG*') are only counted once:

[2] Morphological variability is disregarded in this paper, as it did not prove influential in the experiments described here.

[3] POS, morphological features and dependencies from UD: http://universaldependencies.org.

[4] Lexicalized components are highlighted in bold.

[5] We neglect rare cases of VMWEs sharing a canonical form, e.g. ***fermer les yeux*** '*close the eyes*'⇒'pretend not to see' vs. ***fermer l'oeil*** '*close the eye*'⇒'have a nap'.

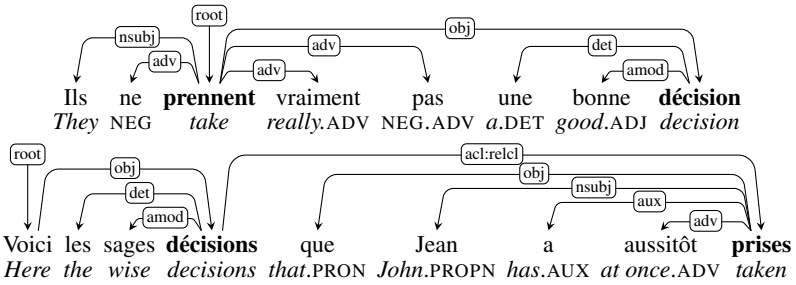

Figure 1: Two POS-tagged and dependency-parsed occurrences of ***prendre*** *une* ***décision*** 'take a decision'.

$$S^L(V_1, V_2) = \frac{2 \times |\{\text{adv}\}|}{|\{\text{adj,adv,det}\}| + |\{\text{adv,aux,pron,propn}\}|} = \frac{2}{7}$$

3 VMWE variability

Given the two similarity measures S^S and S^L between variants V_1 and V_2 of a VMWE E, the *rigidity* scores of E are the averages of all pairs of E's variants. For example, if *take decision* occurs 6 times, we average the scores S^S and S^L of $\binom{6}{2} = 15$ pairs:

$$R^Y(E) = \frac{1}{\binom{m}{2}} \times \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} S^Y(V_i(E), V_j(E))$$

where $Y \in \{\text{S}, \text{L}\}$, m is the number of E's variants in the corpus, and $V_i(E)$ is the i'th variant.

Note that the rigidity measures defined above range from 0 to 1. The *variability* of E can, thus, be defined as the complement of rigidity: $V_X^Y(E) = 1 - R_X^Y(E)$. Experiments were performed in order to estimate the relevance and utility of these measures. Parameter values were chosen empirically and are presented in Appendix A. In the long run, these parameters should be estimated experimentally, possibly in an application-specific manner.

4 Corpus

We use the French part of the PARSEME corpus[6] manually annotated for VMWEs in 18 languages (Savary et al., 2017). Among its 4 VMWE categories two are particularly relevant:

- *Light-verb constructions* (LVCs): combinations of the type Verb-(Adp)-(Det)-Noun where the verb is semantically void and the noun bears the meaning, e.g. ***faire*** *un* ***voeu*** 'make a wish'.
- *Idioms* (IDs): verbal phrases of various syntactic structures, often with non-compositional meaning and admitting both literal and idiomatic reading e.g. ***perdre pied*** '*lose foot*'⇒'lose self-confidence'

The VMWEs annotations in the corpus are accompanied by morphological and a syntactic layers, as shown in Fig. 1. In the morphological layer, lemmas, POS and morphological features are assigned to each token. The syntactic layer represents syntactic dependencies between tokens. Both result from manual annotation and use UD tagsets. The corpus is divided into a training corpus (TrC) and a test corpus (TeC). TrC contains 17,880 sentences, 450,221 tokens, and 4,462 VMWE occurrences, including 1,786 occurrences of 502 unique IDs and 1,362 occurrences of 672 unique LVCs. On average, each ID has 3.6 variants and each LVC has 2 variants. The frequency of individual VMWEs varies greatly (from 1 to 172) and so does the reliability of the variability estimation of each MWE. Hence, only the most frequent VMWEs are considered in Sec. 6.

5 Linguistic relevance

It order to estimate the relevance of our measures, we refer to an existing corpus study by Tutin (2016). There, 30 French VMWEs of the form Verb-(Det)-Noun are studied with respect to 5 morpho-syntactic variation types. This yields 6 *variability levels* depending on how many of the 5 variability types a VMWE exhibits. This is illustrated in Tab. 1 with three VMWEs which stand at distinct levels of the variability spectrum.

Tutin's variability types are defined in terms of complex linguistic phenomena, such as admitting passivization and relative constructions, which have to be validated manually. We, conversely, are in need of fully automatic procedures. Therefore we capture the VMWE variability in distinct ways. It is interesting to see how far both approaches agree on their conclusions.

[6] http://hdl.handle.net/11372/LRT-2282

Variability type	Examples
Noun's number inflection	***prendre*** *la/les* ***décision(s)*** *'take the decision(s)'* ***fermer*** *la/les* ***porte(s)*** *'close the door(s)'* ***donner lieu/#*** *lieux 'give place(s)'*
Passivization	***décision prise*** *'decision taken'* ***porte fermée*** *'door closed'* *#lieu donné 'place given'*
Noun's determiner variation	***prendre*** *la/ma/cette* ***décision*** *'take the/my/this decision'* *#fermer une/ma/cette porte 'close the/my/this door'* *#donner un/mon/ce lieu 'give a/my/this place'*
Relative construction	*la* ***décision*** *qu'il* ***prend*** *'the decision which he takes'* *#la porte qu'il ferme 'the door which he closes'* *#lieu qu'il donne 'place which it gives'*
Adjunction of noun modifiers	***prendre*** *une grande* ***décision*** *'take a great decision'* *#fermer la grande porte 'close the great door'* *#donner un grand lieu 'give a great place'*

Table 1: Tutin's variability types for ***prendre*** *une* ***décision*** '*take a decision*'⇒'make a decision' (level 5), ***fermer la porte*** '*close the door*'⇒'hinder' (level 2), ***donner lieu*** '*give place*'⇒'lead to' (level 0).

Tutin's variability level	0	1	2	3	4	5	All
# TrC VMWEs	6	3	2	3	1	3	18
# TrC occurrences	69	114	8	18	7	54	270
Aggregated level	S_{0-1}		S_{2-4}			S_5	S

Table 2: Distribution of the VMWEs extracted from the PARSEME training corpus into Tutin's classes.

To this aim, we extract from TrC all occurrences of the 30 VMWEs covered by Tutin and retain those with at least 2 occurrences (measuring similarity requires two variants at least). Tab. 2 shows the distribution of the resulting set S of 18 VMWEs into Tutin's levels. While their corpus frequency is relatively high at levels 0, 1 and 5, it is low at levels 2, 3 and 4. Therefore we aggregate neighbor levels into 3 subsets: S_{0-1}, S_{2-4} and S_5. For each VMWE in S we calculate V^L and V^S with weight $w_i = 1$ for the noun and 0 for the verb and the determiner (if any). As shown by the corresponding boxplots in Fig. 2 (a–b), V^L tends to increase with Tutin's level. That is to say, the more variable VMWEs are (as judged by a linguist expert on the basis of a manual corpus study), the higher is their automatically calculated linear variability value. Tutin's extreme levels 0–1 and 5 are particularly well discriminated by V^L.[7] No interesting tendency could be observed for the syntactic variability of the noun. We hypothesize that different outgoing dependencies have different roles in modeling syntactic variability. For instance in ***aller dans*** *le bon* ***sens*** '*go to the right direction*'⇒'evolve positively', the dependency between the noun and the modifier *bon* '*good*' probably tells us more about the rigidity of this VWME than its case-marking preposition *dans* '*in*' or its determiner *le* '*the*'. In future work, we would like to address experimental estimation of weights for different dependency relations in S^S.

[7]Wilcoxon-Mann-Whitney (WMW) test confirms that S_5 differs from S_{0-1} with significance at $\alpha = 0.05$.

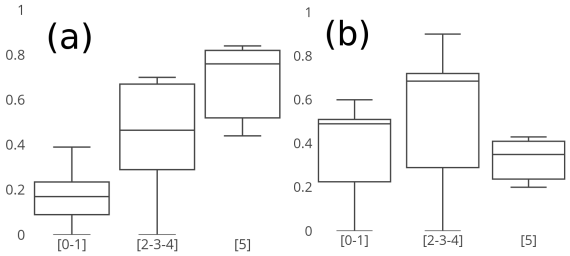

Figure 2: Tukey boxplots of V^L and V^S (y-axis) as a function of Tutin's levels (x-axis).

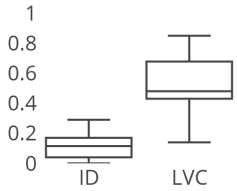

Figure 3: Tukey boxplots of V^L (y-axis) as a function of VMWE categories (x-axis).

6 VMWE classification

LVCs are known to have a relatively regular morphosyntactic behavior as compared to IDs, which tend to be more rigid. We expect our variability measures to help discriminate these categories. We selected those VMWEs whose frequency in TrC was higher than 9, i.e. 12 IDs and 17 LVCs.[8] We then calculated V^S and V^L for each selected VMWE. As shown in Fig. 3, a strong ID vs. LVC discriminative power can be attributed especially to V^L, given that the variability of IDs never exceeds 0.3, while it reaches 0.94 for LVCs.[9]

7 Identification of VMWE variants

As shown by Fazly et al. (2009), English MWEs exhibit lower variability than non-MWEs. Thus, variability measures can help identify MWEs in running text. We test this hypothesis for French using S^L and S^S, which model variant similarity differently from this seminal work. To this aim, we adapted the method proposed by Savary and

[8]This threshold is a trade-off between keeping enough variant pairs to be compared to capture the variability profile of a VMWE, and enough VMWEs to evaluate V^S and V^L. Increasing this value e.g. to 19 would yield at least 190 comparisons per VMWE (vs. 45 here) but keep only 8 VMWEs.

[9]These results are statistically significant at $\alpha = 0.01$ according to the WMW test.

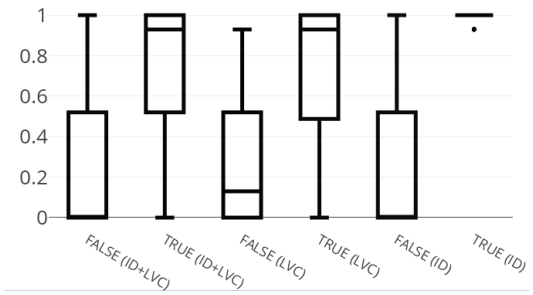

Figure 4: VMWE identification with S^L : Tukey boxplots of False vs. True positives

Cordeiro (2018) to consider all VMWEs of the form Verb-(Det)-Noun annotated in TrC and extract their candidate occurrences in TeC. For instance, if TrC contains the expression e ***perdre pied*** *'lose foot'*⇒'lose self-confidence', then the extracted TeC candidates, noted $Cand(e)$, contain true variants of e (e.g. *ces obstacles me font* ***perdre pied*** 'these obstacles make me lose my self-confidence'), literal readings of e (e.g. *il a perdu le pied gauche 'he lost his left foot'*), and coincidental occurrences of e's components (e.g. *traces des pieds de l'enfant perdu 'traces of the lost child's feet'*). Our hypothesis is that S^S and S^L should be able to distinguish true VMWEs from literal and accidental occurrences, thus being useful for supervised VMWE identification. More precisely, we hypothesise that the more a candidate resembles a known VMWE occurrence, the more chances it has to be a VMWE.

We extracted 195 candidates $c \in Cand(e)$ from TeC. For each candidate c, we calculated the minimum similarities $S^L(e,c)$, $S^S(e,c)$ and the average of both $S^{L-S}(e,c)$ over all occurrences of e in TrC.[10] Interesting results were obtained mainly with S^L. Fig. 4 shows pairwise comparison of the minimal value of $S^L(e,c)$ when IDs and LVCs are considered jointly (boxplots 1–2), or separately (boxplots 3–6). In each case S^L clearly delimits false from true positives.[11]

8 Conclusions and future work

We defined syntactic and linear measures of VMWE variability. They use pairwise similarity based on expert linguistic knowledge. We showed their statistically significant correlation with a linguistic benchmark. We also discovered that linear similarity proves useful in VMWE classification and identification, which is particularly interesting in comparison to the seminal work by Fazly et al. (2009), who do not consider this kind of similarity.

These definitions and estimations should be further improved to deal with other MWE categories, not only verb-noun combinations. Our similarity measures rely on language-independent assumptions: they can be applied to any MWE-annotated corpus containing POS tags and dependency trees. If these morphosyntactic annotations use the unified UD tagsets, cross-language MWE variability studies can be carried out. Therefore, our experiments will be extended to all languages accounted for in the PARSEME corpus. Task-specific parameter tuning should show which parameters are shared by all/many languages and/or tasks, and which have to be language- and task-specific. Morphological variability, including both inflection and derivation (as in ***refaire appel*** *'re-make appeal'*⇒'to call on again'), temporarily abandoned for French, could be examined in a multilingual context. Finally, the measures should be adapted to an unsupervised context, to scale them up to larger VMWE vocabularies and languages with no MWE-annotated corpora. For instance, MWE variant candidates could be extracted from automatically parsed text, using lists of known MWE lemmas (Savary and Cordeiro, 2018).

We believe that with these extensions our variability measures will offer a unified framework for describing variability profiles of MWEs, which should be useful both in theoretical and applied research. They could help: (i) disambiguate literal vs. idiomatic readings of VMWEs, (ii) conflate variants of the same MWE to reduce information variation in text, (iii) measure the sensitivity of NLP tools to variability, (iv) define variability-specific evaluation measures in MWE identification to boost the efficient recognition of variants.

Acknowledgments

This work was funded by the French PARSEME-FR grant (ANR-14-CERA-0001). We are grateful to Silvio Cordeiro for sharing the script for VMWE candidate extraction and for his helpful assistance. We also thank the anonymous reviewers for their useful comments.

[10] In this section, all similarities S are estimated as the average of the four coefficients presented in App. B.

[11] This is confirmed by the WMW test with significancy at $\alpha = 0.01$ for both IDs and LVCs.

References

Timothy Baldwin, Emily M. Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2004. Road-testing the English Resource Grammar over the British National Corpus. In *Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

Maja Buljan and Jan Šnajder. 2017. Combining Linguistic Features for the Detection of Croatian Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, Valencia, Spain.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics* 43(4):837–892. https://doi.org/10.1162/COLI_a_00302.

Meghdad Farahmand and James Henderson. 2016. Modeling the non-substitutability of multiword expressions with distributional semantics and a log-linear model. In *Proc. of the ACL 2016 Workshop on MWEs*. Berlin, pages 61–66. http://anthology.aclweb.org/W16-1809.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics* 35(1):61–103. https://doi.org/10.1162/coli.08-010-R1-07-048.

Gaston Gross. 1988. Degré de figement des noms composés. *Langages* 90:57–72.

Maurice Gross. 1982. Une classification des phrases "figées" du français. *Revue québécoise de linguistique* 11(2).

Christian Jacquemin. 2001. *Spotting and Discovering Terms through Natural Language Processing*, The MIT Press.

Malvina Nissim and Andrea Zaninello. 2013. Modelling the internal variability of multiword expressions through a pattern-based method. In *ACM Transactions on Speech and Language Processing , Special issue on Multiword Expressions*. volume 10.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language* 70:491–538.

Darren Pearce. 2001. Synonymy in collocation extraction. In *Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*. Pittsburgh, PA, pages 41–46.

Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An evaluation of methods for the extraction of multiword expressions. In *Proc. of the LREC 2008 Workshop on MWEs*. Marrakech, pages 50–53.

Agata Savary and Silvio Ricardo Cordeiro. 2018. Literal readings of multiword expressions: as scarce as hen's teeth. In *Proceedings of the 16th Workshop on Treebanks and Linguistic Theories (TLT 16)*. Prague, Czech Republic.

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, Valencia, Spain, pages 31–47. http://www.aclweb.org/anthology/W/W17/W17-1704.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2017. *Verbal MWEs: Idiomaticity and flexibility*, Language Science Press, to appear, pages 5–38.

Yulia Tsvetkov and Shuly Wintner. 2014. Identification of multiword expressions by combining multiple linguistic information sources. *Computational Linguistics* 40(2):449–468. https://doi.org/10.1162/COLI_a_00177.

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

Marion Weller and Ulrich Heid. 2010. Extraction of German multiword expressions from parsed corpora using context features. In *Proc. of LREC 2010*. Valletta, pages 3195–3201.

A Parameter weights

Weigths of lexicalized components : 'VERB': 0 'NOUN': 1 'DET': 0

Features for S^L : 'ADJ': 1 'ADV': 1 'INTJ': 1 'NOUN': 1 'CCONJ': 1 'NUM': 1 'PROPN': 1 'VERB': 1 'AUX': 1 'SCONJ': 1 'ADP': 1 'PRON': 1 'X': 1 'PART': 1 'SYM': 1 'DET': 1 '_': 0 'PUNCT': 0

Features for S^S : 'aux:pass': 1 'nmod:poss': 1 'nummod': 1 'det': 1 'nsubj:pass': 1 'acl:relcl': 1 'amod': 1 'acl': 1 'expl': 0 'xcomp': 0 'root': 0 'iobj': 0 'goeswith': 0 'advcl': 0 'appos': 0 'compound': 0 'fixed': 0 'obl': 0 'mark': 0 'parataxis': 0 'punct': 0 'csubj': 0 'nmod': 0 'flat:name': 0 'orphan': 0 'discourse': 0 '_': 0 'flat:foreign': 0 'dep': 0 'cop': 0 'aux': 0 'dislocated': 0 'obj': 0 'advmod': 0 'conj': 0 'vocative': 0 'reparandum': 0 'nsubj': 0 'case': 0 'cc': 0 'ccomp': 0

B Similarity coefficients used in the variant-to-variant similarity

Similarity between two datasets X and Y is given by the following formulae:

card(X $\cap$ Y) = a

card(X $\cup$ Y) = a + b + c

card(X) = a + b

card(Y) = a + c

Jaccard : $\frac{a}{a+b+c}$

Sørensen-Dice : $\frac{2a}{2a+b+c}$

Sneath-Sokal : $\frac{a}{a+2(b+c)}$

Cosinus : $\frac{a}{\sqrt{((a+b).(a+c))}}$

The variant-to-variant similarity defined in Sec. 7 uses the arithmetic mean of these four coefficients.