Without lexicons, multiword expression identification will never fly: A position statement

Agata Savary, Silvio Ricardo Cordeiro, Carlos Ramisch

Without lexicons, multiword expression identification will never fly: A position statement

Agata Savary
University of Tours, France
first.last@univ-tours.fr

Silvio Ricardo Cordeiro
Paris-Diderot University
France
firstmiddlec@gmail.com

Carlos Ramisch
Aix Marseille University,
Université de Toulon, CNRS
LIS, Marseille, France
first.last@lis-lab.fr

Abstract

Because most multiword expressions (MWEs), especially verbal ones, are semantically non-compositional, their automatic identification in running text is a prerequisite for semantically-oriented downstream applications. However, recent developments, driven notably by the PARSEME shared task on automatic identification of verbal MWEs, show that this task is harder than related tasks, despite recent contributions both in multilingual corpus annotation and in computational models. In this paper, we analyse possible reasons for this state of affairs. They lie in the nature of the MWE phenomenon, as well as in its distributional properties. We also offer a comparative analysis of the state-of-the-art systems, which exhibit particularly strong sensitivity to unseen data. On this basis, we claim that, in order to make strong headway in MWE identification, the community should bend its mind into coupling identification of MWEs with their discovery, via syntactic MWE lexicons. Such lexicons need not necessarily achieve a linguistically complete modelling of MWEs' behavior, but they should provide minimal morphosyntactic information to cover some potential uses, so as to complement existing MWE-annotated corpora. We define requirements for such a minimal NLP-oriented lexicon, and we propose a roadmap for the MWE community driven by these requirements.

1 Introduction

Multiword expression (MWE) is a generic term which encompasses a large variety of linguistic objects: compounds (***to and fro***, ***crystal clear***, *a* ***slam dunk*** 'an easily achieved victory')[1], verbal idioms (*to* ***take pains*** 'to try hard'), light-verb constructions (*to* ***pay*** *a* ***visit***), verb-particle constructions (*to* ***take off***), institutionalized phrases (***traffic light***), multiword terms (***neural network***) and multiword named entities (***Federal Bureau of Investigation***). They all share the characteristic of exhibiting lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasies (Baldwin and Kim, 2010). Most notably, they usually display non-compositional semantics, i.e. their meaning cannot be deduced from the meanings of their components and from their syntactic structure in a way deemed regular for the given language. Computational methods are, conversely, mostly compositional, therefore they often fail to model and process MWEs appropriately. Special, MWE-dedicated, treatment can be envisaged, provided that we know which parts of the text are concerned, i.e. we should be able to perform MWE identification.

MWE identification (MWEI) consists in automatically annotating MWEs occurrences in running text (Constant et al., 2017). In other words, we need to be able to distinguish MWEs (e.g. ***take pains***) from regular word combinations (e.g. *take gloves*) in context. This task proves very challenging for some categories of MWEs, as evidenced by two recent PARSEME shared tasks on automatic identification of verbal MWEs (Savary et al., 2017; Ramisch et al., 2018). We claim that the difficulty of this task lies in the nature of idiosyncrasies that various categories of MWEs exhibit with respect to regular word combinations. Namely, whereas many constructions (e.g. named entities) have a good generalisation potential for machine learning NLP methods, other MWEs, e.g. verbal ones, are mostly regular at the level of tokens, so the generalisation power of mainstream machine learning is relatively weak for them. However, they are idiosyncratic at the level of types (sets of surface realizations of the same ex-

[1] Henceforth, we highlight in bold the lexicalized components of MWEs, i.e. those always realized by the same lexemes.

Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), pages 79–91
Florence, Italy, August 2, 2019.

pression), therefore type-specific information, exploited by MWE discovery methods and encoded in lexicons, should be very helpful for MWEI.

This paper is a position statement based on an analysis of the state of the art in MWEI. We claim that, in order to make strong headway in MWEI, the community should bend its mind into coupling this task with MWE discovery via syntactic MWE lexicons. Such lexicons need not necessarily achieve a linguistically complete modelling of MWEs' behavior, but they should provide minimal morphosyntactic information to cover some potential uses, so as to complement existing MWE-annotated corpora. This also implies that, in building such lexicons, we can take advantage of the rich body of works dedicated to MWE discovery methods (Evert, 2005; Pecina, 2008; Seretan, 2011; Ramisch, 2015), provided that they are extended, so as to: (i) cover most syntactic types of MWEs, (ii) produce not only lists of newly discovered MWE entries but also their type-specific morphosyntactic properties.

The remainder of this paper is organized as follows. We discuss some linguistic properties of MWEs (Sec. 2) and state-of-the-art results (Sec. 3) relevant to our claims. We propose a scenario for coupling MWEI with MWE discovery via syntactic MWE lexicons (Sec. 6). Finally, we conclude by proposing a roadmap for the future efforts of the MWE community (Sec. 7).

2 The nature of MWEs

We propose to divide MWE categories roughly into two meta-categories, depending on the nature of the processes which provoke their lexicalization, that is, the assignment of conventional, fixed, non-compositional meanings. On the one hand, there are multiword named entities (NEs) and multiword terms, henceforth called *sublanguage MWEs* (SL-MWEs), whose form-meaning association is usually determined by sublanguage experts. Because such expert groups are more or less restricted and have dedicated nomenclature instruments (scientific publications, naming committees, etc.), and because technological domains and real-world entities to name develop rapidly, multiword terms and NEs strongly proliferate. On the other hand, *general language MWEs* (GL-MWEs)[2] are coined by much larger communities of speakers via informal processes, and take longer to be established in a language. This *proliferation speed* property (henceforth referred to as P_{prolif}) is the first SL-MWE vs. GL-MWE discrepancy we are interested in.

The second property (henceforth, P_{discr}) is the *nature of discrepancies* which statistically distinguish MWEs from regular word combinations. SL-MWEs exhibit peculiarities at the level of tokens (individual occurrences). For instance multiword NEs are usually capitalized and often contain, follow or precede trigger words (*Bureau*, *river*, *Mr.*). Multiword terms often contain words which are less likely in general than in technical language (*neural*). GL-MWEs, conversely, are mostly regular at the level of tokens (e.g. they use no capitalization, are rarely signaled by triggers, and contain common frequent words) but idiosyncratic at the level of types (sets of surface realizations of the same expression). For instance, *to* ***take pains*** 'to try hard' does not admit noun inflection (i.e. *to take the pain* cannot be interpreted idiomatically), while similar regular word combinations like *to take gloves* and *to relieve pains* have very similar meaning to their morphosyntactic variants *to take the glove* or *to relieve the pain*.

The third relevant property (P_{sim}) is the *component similarity* among MWEs. A strong similarity, whether at the level of surface forms or at the level of semantics, often occurs between components of different SL-MWEs. For instance, new multiword terms are often created by modification or specialization of previously existing ones (***neural network***, ***neural net***, ***recurrent neural network***, ***neural network pushdown automata***, etc.). Also, many types of NEs come in series in which some components are identical and some others vary within a given semantic class, e.g. ***American/Brazilian/French/Ethiopian Red Cross***, ***Nigerian Red Cross Society***, ***Iranian/Iraki Red Crescent Society***, ***Saudi Red Crescent Authority***. In GL-MWEs, the degree of P_{sim} depends on the category. It is stronger in light-verb constructions, i.e. verb-noun combinations in which the verb is semantically void or bleached, and the noun is predicative[3], as in *to* ***make*** *a* ***decision*** and *to* ***pay*** *a* ***visit***. Many light-verb constructions are similar to each other be-

[2] The border between SL-MWEs and GL-MWEs is fuzzy, but this characterization is useful for our argumentation.

[3] A noun is predicative if it has at least one semantic argument, according to the PARSEME guidelines (http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1).

cause of the predicative nature of the nouns but also because they contain one of the few very frequent light verbs like *make*, *take*, etc. (Savary et al., 2018). Note, however, that these verbs, are also highly frequent in regular constructions, i.e. P_{sim} is moderate but P_{discr} is still restricted to the level of types. Component similarities are weaker among inherently reflexive verbs, like (PL) ***znaleźć się*** 'find oneself'. On the one hand, inherently reflexive verbs always contain a (mostly uninflected) reflexive clitic (here: *się*) governed by a verb. On the other hand, semantically similar verbs do not systematically form inherently reflexive verbs, e.g. (PL) *wyszukać* 'find' is a synonym of *znaleźć* 'find' but **wyszukać się* 'find oneself' is ungrammatical. Finally, verbal idioms, which cover diverse syntactic structures, are largely dissimilar to each other but similar to regular constructions, e.g. *to* ***take pains*** 'to try hard' is a MWE but *to take aches* is not.

The fourth property (P_{ambig}) is the very *low ambiguity* of word combinations appearing in MWEs. These combinations are ambiguous because they can occur both with idiomatic and with literal readings, as in examples (1) vs. (2) below. Ambiguity is considered one of the major challenges posed by MWEs in NLP (Constant et al., 2017). However, recent work (Savary et al., 2019) shows that, although most combinations of MWEs' components could potentially be used literally, they are rarely used so in corpora. Namely, in 5 languages from different language genera, the idiomaticity rate of verbal MWEs, i.e. the proportion of idiomatic occurrences with respect to the total number of idiomatic and literal occurrences, ranges from 0.96 to 0.98. This means that, whenever the morphosyntactic conditions for an idiomatic reading are fulfilled, this reading occurs almost always. A similarly high idiomaticity rate (0.95) was also observed for Polish on other, non-verbal categories of MWEs: nominal, adjectival, and adverbial GL-MWEs, as well as multiword NEs (Waszczuk et al., 2016). This property might be related to the fact that ambiguity is reduced with the addition of words to the context, a hypothesis that has been employed in word-sense disambiguation for many years (Yarowsky, 1993).

(1) We often **took pains** not to harm them.
'We often tried hard not to harm them.'

(2) I could not take the pain any longer.

Finally, the fifth property (P_{zipf}) we are interested in is the *Zipfian distribution* of MWEs. As most language phenomena, few MWE types occur frequently in texts, and there is a long tail of MWEs occurring rarely (Ha et al., 2002; Ryland Williams et al., 2015). The success of machine learning generalization relies on dealing with rare or unseen events, based on their similarity with frequent ones. Such similarity is hard to define for the heterogeneous phenomena included under the MWE denomination.

3 State of the art in MWE identification

In this section we offer a comparative analysis of state-of-the-art results with respect to two axes: SL-MWEs vs. GL-MWEs and seen vs. unseen data. All results are indicated in terms of the F1-measure, with the exact-match metric. In other words, a prediction for a text fragment is considered correct only when the identified unit corresponds to exactly the same words as in the gold standard.[4] For most SL-MWE results, the F1-measure additionally accounts for categorisation, i.e. a correctly identified span of words must also be assigned the correct NE category.

3.1 Identification of sublanguage MWEs

For SL-MWEs, identification methods have been developed for decades, but most often fuse multiword objects with single-word ones. Two typical examples are NE recognition and term identification. In these two domains, state-of-the-art results have been encouraging or good already in early systems and evaluation campaigns.

In the CoNLL 2002 and 2003 shared tasks on NE recognition (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003), dedicated mainly to person, organization and location names, the top-3 systems obtained F1-measures of 0.71, 0.74, 0.77, and 0.86, with datasets of 20,000, 13,000, 18,000 and 35,000 annotated NEs, for German, Dutch, Spanish and English, respectively. All of these systems used machine learning techniques such as hidden Markov models, decision trees, MaxEnt classifiers, conditional random fields, support-vector machines, recurrent neural networks, with features that often included external entity list lookup.[5] Yadav and Bethard (2018) provide more recent state-of-the-art results for NE

[4] The same metric is called *MWE-based*, as opposed to *token-based*, in the PARSEME shared task campaigns.

[5] Results of the same systems without external entity list lookup are not provided.

recognition based on neural networks on the same datasets. There, the best results mostly exceed 0.78 for German, 0.85 for Dutch and Spanish, and 0.9 for English, even without external dictionary lookup. In Slavic languages, where NE recognition is substantially hardened by the rich declension of nouns and adjectives, stable benchmarking data are still missing.[6] Sample results can be cited in Polish, where relatively rich NE-annotated corpora and lexicons are available. Reference tools achieve the F1-measure of 0.71 (Marcińczuk et al., 2017) and 0.77 (Waszczuk et al., 2013) with methods based on conditional random fields.

As for term identification, several domain-specific benchmarking datasets allowed for system development and comparison. For instance, the best systems for biomedical term identification obtain F1-measure of about 0.81, 0.85 and 0.88 on disorder, chemical and gene/protein names, respectively (Campos et al., 2012).

While single-word and multiword NEs and terms are fused in the above results, good hints exist that the results on multiword NEs and terms are comparable or better than results on single-word items. Firstly, the majority of NEs and terms in corpora consist of several words. For instance, in the 110,000-token English Wiki50 corpus (Vincze et al., 2011), around 65% of annotated NEs and terms consist of at least 2 words. Also in the JNLPBA and i2b2 shared tasks on biomedical and medical NE recognition, 55% and 58%, respectively, of all terms are multiword terms (Campos et al., 2012). Secondly, some NE recognition efforts were explicitly dedicated to boosting performance for multiword NEs and terms. For instance, Downey et al. (2007) achieve F1=0.74 on the recognition of multiword named entities in a web corpus with a very simple system based on n-gram statistics. A baseline system using bidirectional recurrent neural networks (BiLSTM) by Campos et al. (2012) achieves the F1-measure of 0.74 and 0.81 on bigrams, which are the most frequent multiword terms in the i2b2 and JNLPBA corpora.

3.2 Identification of general-language MWEs

Within GL-MWEs, multilingual benchmarking data are available mainly for verbal MWEs via editions 1.0 and 1.1 of the PARSEME shared tasks (Savary et al., 2017; Ramisch et al., 2018). In edition 1.1, the scores (across 19 languages) for the top-3 systems range from 0.5 to 0.58. The per-language scores vary greatly due to corpus size variety and typological differences between languages. Table 1 shows the corpus sizes and the best system F1-measure for the 6 languages whose corpora contain at least 5,000 annotated verbal MWEs.[7] The results of the best systems, with and without neural networks, never exceed 0.68, with the exception of Romanian, which has a low percentage of unseen data in the test corpus.

	BG	FR	PL	PT	RO	TR
#verbal MWEs	6.7K	5.7K	5.2K	5.5K	5.9K	7.1K
unseen ratio	.33	.50	.28	.28	.05	.75
Best non-NN F1	.63	.56	.67	.62	.83	.45
Best NN F1	.66	.61	.64	.68	.87	.59

Table 1: Sizes of the corpora (in thousands of annotated verbal MWEs), the ratio of unseen verbal MWEs in the test corpora and the best system performance, without (non-NN) and with neural networks (NN), in the PARSEME shared task 1.1 for 6 languages with the largest corpora.

These results are not directly comparable to those from Sec. 3.1 because evaluation measures partly differ (e.g., NE recognition includes categorisation), the sets of languages hardly overlap, and corpus sizes are largely below those of the CoNLL corpora.[8] Still, it is clear that MWEI is a particularly hard problem and it is important to understand the vulnerabilities (if any) of current approaches.

3.3 Challenges of unseen data

The PARSEME shared task 1.1 introduced phenomenon-specific evaluation measures which

[6] In the first shared task on NE recognition in Balto-Slavic languages (Piskorski et al., 2017), only test data but no annotated training data were published.

[7] Hungarian is left out because its corpus consists of specialized law texts. Language codes in the tables are: Bulgarian (BG), French (FR), Polish (PL), Portuguese (PT), Romanian (RO), Turkish (TR).

[8] The PARSEME shared task 1.0 results for Czech, with 12,000 annotated verbal MWEs, come up to $F1 = 0.72$ with a non-neural system. This might be comparable to the CoNLL-2002 results for Dutch, with 13,000 annotated NEs and the top F1-measure of 0.74 for a non-neural system. However, as many as 69% of the annotated verbal MWEs in the Czech corpus are inherently reflexive verbs (IRVs), such as ***se bavit*** 'amuse oneself'⇒'play', which are relatively easy to predict due to the moderate strength of P_{sim}. The Czech corpus was not annotated from scratch but converted from a previously annotated resource, and inherently reflexive verbs are probably over-represented there. The rate of inherently reflexive verbs in other Slavic languages in the PARSEME corpora range from 0.3 to 0.48.

focus on known challenges posed by MWEs. Thus, results were reported separately for continuous vs. discontinuous, multi-token vs. single-token, seen vs. unseen, and identical-to-train vs. variant-of-train verbal MWEs.[9] The most dramatic performance differences appear in the seen vs. unseen opposition. A verbal MWE from the corpus is considered seen if another verbal MWE with the same multiset of lemmas is annotated at least once in the training corpus. For instance, given the occurrence of ***has*** *a new* ***look*** in the training corpus, the following verbal MWEs from the test corpus would be considered:

- seen: ***has*** *a new* ***look***, ***had*** *an appealing* ***look***, ***has*** *a* ***look*** *of innocence*, *the* ***look*** *that he* ***had***
- unseen: ***has a look*** *at this report*, ***gave*** *a* ***look*** *to the book*, ***walk*** *that he* ***had***, ***took part***, etc.

Tab. 2 shows the PARSEME shared task 1.1 results achieved on seen and unseen data for 3 of the 6 previously analysed languages. French and Turkish were left out since no lemmas are provided for 20-30% of their test data. Romanian is skipped because only 5% of its test corpus corresponds to unseen data. We focus on the overall best systems in the closed and open track[10]: TRAVERSAL (Waszczuk, 2018) and SHOMA (Taslimipoor and Rohanian, 2018). The former applies sequential conditional random fields extended to tree structures, while the latter feeds word embeddings to convolutional and recurrent neural networks, which are given to a decision layer based on conditional random fields. On unseen data in the 3 languages under study, TRAVERSAL's score never exceeds 0.20, and the performance is 3.9 (for Portuguese) to 6.1 (for Bulgarian) times worse than on seen data. SHOMA's generalization power is greater: it achieves a score of 0.18 (for Polish) to 0.31 (for Bulgarian and Portuguese) on unseen data, which is still 2.5 (for Portuguese) to 4.6 (for Polish) times worse than for seen expressions.

It is also interesting to see which unseen verbal MWEs categories have been correctly identified by both systems. Tab. 2 reveals that generalization is the strongest for inherently reflexive verbs and light-verb constructions, likely due to the moderate inter-MWE component similarity (P_{sim}) discussed in Sec. 2. Still, it is far below the generalization power in SL-MWEs (see below), probably because P_{discr} is related to types but not tokens.

As far as SL-MWE identification is concerned, we are aware of only one study explicitly dedicated to the impact of unseen data. Namely, Augenstein et al. (2017) compare the performance of 3 state-of-the-art named-entity recognition tools on 19 NE-annotated datasets in English. For the CoNLL corpora cited in Sec. 3.1, the scores achieved on unseen data range from 0.81 to 0.94. The scores for out-of-domain unseen data are significantly lower but still exceed 0.61 for the 2 best systems. Unseen NEs are defined in this study as those with surface forms present only in the test, but not in the training data, which differs from the PARSEME shared task 1.1 definition (where data with different surface forms are considered seen if they have seen multisets of lemmas). Still, morphosyntactic variability in English NEs should be relatively low, therefore we may safely deduce that MWEI on unseen data performs significantly better on SL-MWEs in a morphologically-poor language than on GL-MWEs in morphologically-rich languages. We believe that this is more related to the SL-MWE vs. GL-MWEs distinction than to typological differences between languages.[11]

To conclude, the challenges posed by unseen data to MWEI seem significantly harder for GL-MWEs than for SL-MWEs. We attribute this fact to the different nature of the two phenomena. SL-MWEs differ from regular word combinations at the level of tokens (P_{discr}) and exhibit strong similarities among components (P_{sim}). These properties can be leveraged by machine learning tools, whether supervised (e.g. using character-level features or word embeddings, to account for surface and semantic similarity of NEs components, respectively) or unsupervised (e.g. based on contrastive measures for terms), notably to generalize over unseen data. Conversely, GL-MWEs are mostly idiosyncratic at the level of types but not tokens (P_{discr}) and show moderate or weak component similarities (P_{sim}). These charateristics are hard to tackle by systems which model MWEI as a tagging problem, except if features based on type-

[9] http://multiword.sourceforge.net/sharedtaskresults2018

[10] In the closed track, systems are only allowed to use the provided training/development data. In the open track, they can additionally use external resources (lexicons, word embeddings, language models trained on external data, etc.).

[11] PARSEME shared task 1.1 results for identical-to-train vs. variant-of-train items, presented in the next section, corroborate this intuition: TRAVERSAL and SHOMA handle morphosyntactic variability much better than lexical novelty.

		BG				PL				PT			
		IRV	LVC	VID	All	IRV	LVC	VID	All	IRV	LVC	VID	All
TRAVERSAL	seen	.89	.63	.55	.76	.92	.76	.57	.85	.89	.77	.69	.78
	unseen	.26	.06	.07	.13	.26	.20	.04	.17	.12	.25	.07	.20
SHOMA	seen	.92	.65	.58	.78	.90	.69	.58	.82	.86	.88	.84	.87
	unseen	.59	.21	.10	.31	.24	.19	.04	.18	.42	.35	.08	.31

Table 2: PARSEME shared task 1.1 identification scores on seen and unseen data for TRAVERSAL and SHOMA. Verbal MWE categories are inherently reflexive verbs (IRVs), light-verb constructions (LVCs) and verbal idioms (VIDs).

specific idiosyncrasies are used. The few token-specific hints (if any) which may help such systems generalize over unseen data are mostly limited to the presence of particular light verbs or function words. Their role resembles the one of trigger words and nested entities in NE recognition (Sec. 2), but, differently from the latter, they are also highly frequent in regular constructions, which hinders their discriminative power for GL-MWEs.

3.4 Progress potential in seen data

Since unseen GS-MWEs prove drastically hard to identify, it is interesting to understand how much progress might be achieved on seen data. We believe that this potential of improvement is relatively high due to several factors.

Firstly, the low effective ambiguity of MWEs (P_{ambig}) means that identifying morphosyntactically well-formed combinations of previously seen MWE components constitutes a strong baseline for MWEI. For instance, Pasquer et al. (2018b) propose a very simple baseline for verb-noun MWE identification in which previously seen verb-noun pairs are tagged as MWEs as soon as they have the same lemmas as a seen MWE and maintain a direct dependency relation, whatever the label and direction of this dependency. This very simple method achieves F1=0.88 on French. A comparable result was observed in the 2016 DiMSUM shared task (Schneider et al., 2014), in which a rule-based baseline was ranked second. This system extracted MWEs from the training corpus and then annotated them in the test corpus based on lemma/part-of-speech matching and heuristics such as allowing a limited number of intervening words (Cordeiro et al., 2016).

Secondly, there is a large gap to bridge for seen data whose surface form is not identical to the ones seen in train. Tab. 3 shows that, indeed, the difference between identical-to-train and variant-of-train scores ranges from 0.12 (in Polish for TRAVERSAL and Portuguese for SHOMA) to 0.37 (in Bulgarian for SHOMA). At the same time, Pasquer et al. (2018a) show that morphosyntactic variability, relatively high in verbal MWEs, can be neutralized with dedicated methods. Namely, co-occurrences of previously seen MWE components can be effectively recognized by a Naive Bayes classifier, with features leveraging type-specific idiosyncrasies (P_{discr}). This method scored the best in the PARSEME shared task 1.1 for Bulgarian, even if it was restricted to the seen data only.

		BG	PL	PT
TRAVERSAL	identical to train	.85	.92	.87
	variants of train	.55	.80	.72
SHOMA	identical to train	.89	.95	.93
	variants of train	.52	.71	.81

Table 3: PARSEME shared task 1.1 identification scores on identical-to-train and variant-of-train data for TRAVERSAL and SHOMA.

Thirdly, significant progress can also be achieved if another important challenge is explicitly addressed: discontinuity of verbal MWEs. For instance, Rohanian et al. (2019) employ neural methods combining convolution and self-attention mechanisms and obtain impressive improvements over the best PARSEME shared task systems.

Finally, not only annotated training corpora but also MWE lexicons can provide information about seen data. The two next sections describe the state of the art in lexical description of MWEs, and integration of MWE lexicons in NLP methods.

4 Lexicons of MWEs

Describing MWEs in dictionaries dedicated to human users has a long-standing lexicographic tradition, but its synergies with NLP have not been straightforward (Gantar et al., 2018). More formal linguistic modeling of MWEs has also been carried out for decades, notably in the frameworks

of the Lexicon Grammar (Gross, 1986) and of the Explanatory Combinatorial Dictionary (Mel'čuk et al., 1988; Pausé, 2018). These approaches assume that units of meaning are located at the level of elementary sentences (predicates with their arguments) rather than of words, and MWEs, especially verbal, are special instances of predicates in which some arguments are lexicalized. Those works paved the way towards systematic syntactic description of MWEs, but suffered from insufficient formalization and required substantial accommodation to be applicable to NLP (Constant and Tolone, 2010; Lareau et al., 2012).

With the growing understanding of the challenges which MWEs pose to NLP, a large number of (fully or partly) NLP-dedicated lexicons have been created for many languages (Losnegaard et al., 2016). These resources can be classified notably along 3 axes, according to (i) the account of the morpho-syntactic structure of a MWE and its variants, (ii) lexicon-corpus coupling, (iii) number of entries.

Along axis (i), there is a gradation in the complexity of the related formalisms. The simplest are raw lists of MWEs, sometimes accompanied with selected morphosyntactic variants, collected from large corpora or automatically generated (Steinberger et al., 2011).

More elaborate are approaches based on finite-state-related formalisms. They usually indicate the morphological categories and features of individual MWE components, and offer rule-based combinatorial description of their variability patterns (Karttunen et al., 1992; Breidt et al., 1996; Oflazer et al., 2004; Silberztein, 2005; Krstev et al., 2010; Al-Haj et al., 2014; Lobzhanidze, 2017; Czerepowicka and Savary, 2018). They mostly cover continuous (e.g. nominal) MWEs in which morphosyntactic phenomena remain local (Savary, 2008). Therefore, additionally to the intentional format, i.e. rules describing the analysis and production of MWE instances, they often come with an extensional format, which stores the MWE instances (inflected forms) themselves. Plain-text extensional lists can be straightforwardly matched against a text. Such finite-state frameworks do not account for deep syntax and for interactions of MWE lexicalized components with external elements. Therefore, they are not well adapted to verbal MWEs.

Finally, there exist syntactic lexicons in which MWEs are most often covered jointly with single words. On the one hand, there are approaches meant to be theory-neutral (Grégoire, 2010; Przepiórkowski et al., 2017; McShane et al., 2015), i.e. they implicitly assume the existence of regular grammar rules, and explicitly describe only those MWE properties which do not conform to these rules. Although these lexicons suffer from insufficient formalization (Lichte et al., 2019), they could be successfully applied to parsing after ad hoc conversion to particular grammar formalisms. On the other hand, some approaches accommodate some types of MWEs directly in the lexicons of computational grammars within particular grammatical frameworks: head-driven phrase structure grammar (Sag et al., 2002; Copestake et al., 2002; Villavicencio et al., 2004; Bond et al., 2015; Herzig Sheinfux et al., 2015), lexical functional grammar (Attia, 2006; Dyvik et al., 2019), tree-adjoining grammar (Abeillé and Schabes, 1989, 1996; Vaidya et al., 2014; Lichte and Kallmeyer, 2016), and dependency grammar (Diaconescu, 2004).

Along axis (ii), most recent approaches are usually coupled with corpora, but to a different degree. PDT-Vallex (Urešová, 2012) is a Czech valency dictionary fully aligned with the Prague Dependency Treebank, i.e. new frames were added as they were encountered during manual annotation of the corpus. These frames are also linked to their corpus instances. Similarly, SemLex (Bejček and Straňák, 2010), is a MWE lexicon bootstrapped from pre-existing dictionaries (not necessary corpus-based) and further developed hand-in-hand with the PDT annotation. It contains syntactic structures of MWE entries to which corpus occurrences are linked. In Walenty (Przepiórkowski et al., 2014), a Polish valency dictionary, the initial set of entries stems from pre-existing single-word e-dictionaries, which were then extended to MWEs and described as exhaustively as possible as to their valency frames. All frames are documented with attested examples, preferably but not necessarily from the National Corpus of Polish. In DUELME (Grégoire, 2010), a Dutch MWE lexicon, all MWE were automatically acquired from a large raw corpus on the basis of a short list of morpho-syntactic patterns. Lexicon entries contain example sentences illustrating the use of MWEs. Finally, when MWEs were directly accommodated in implemented formal grammars,

the choice of MWEs to model is rarely documented but was probably motivated by a possibly high syntactic and semantic variety of constructions rather than by corpus frequencies, even if attested examples support the grammar engineering.

Along axis (iii), the sizes of the existing MWE lexical resources vary greatly, from several dozen to several tens of thousands of MWE entries. This coverage is often inversely correlated with the richness and precision of the linguistic description.

5 MWE lexicons in MWE identification

Handcrafted MWE lexicons, as those addressed in the previous section, can significantly enhance MWEI. In sequence tagging MWEI methods, such resources can be used as sources of lexical features (Schneider et al., 2014). In parsing-based approaches they may serve as a basis for word-lattice representation of an input sentence, in which the compositional vs. MWE interpretation of a word sequence is represented jointly (Constant et al., 2013). The impact of lexical resources on MWEI is explicitly addressed by Riedl and Biemann (2016). Using a CRF-based MWEI system, they show that the addition of an automatically discovered lexicon of MWEs can benefit MWEI quality.

The systems competing in PARSEME shared tasks used lexical resources to a much lesser degree. In both editions only one, rule-based, system applied a MWE lexicon, for French in edition 1.0 and for English, French, German and Greek in edition 1.1 (Nerima et al., 2017). Other systems, even those from the open track, employed only one type of external resources, namely word embeddings, but no MWE lexicons. This is probably due mainly to the fact that the competition was meant to promote cross-lingual methods, but few or no MWE lexical frameworks offer large MWEs lexicons for many languages. The resources covered by the (Losnegaard et al., 2016) survey are numerous and cover at least 19 languages, but their formats are not uniform so MWE identifiers cannot easily integrate them. Another reason might be that the complex constraints imposed by MWEs, especially verbal ones, call for complex formalisms, whose expressive power is hard to accommodate with mainstream machine learning methods. Still, current MWEI identifiers are able to benefit from rich joint syntactic and MWE annotation, notably to neutralize variability (cf. Sec. 3).

6 Towards syntactic lexicons for MWE identification

As discussed in Sec. 2, MWEs exhibit a Zipfian distribution (P_{zipf}), which means that the power to generalize over unseen data is crucial for high-quality MWEI. However, as seen in Sec. 3, current MWEI methods badly fail on unseen data. At the same time, performance on seen items can be very high if morphosyntactic variability is appropriately accounted for.

The straightforward idea is then to maximize the quantity of the seen data. This proposal is of course trivial with respect to most learning problems in NLP. But we believe that its applicability is particularly relevant in the domain of GL-MWE identification for at least four reasons. Firstly, there is a particularly acute discrepancy between the performance on seen vs. unseen data, as discussed in Sec. 3, so the potential of the gain in this respect is huge. Secondly, unsupervised discovery of (previously unseen) MWEs has a rich bibliography and proves particularly effective when type-specific idiosyncrasies are exploited (P_{discr}), for instance, in verb-noun idiom discovery (Fazly et al., 2009). Thirdly, the low effective ambiguity of word combinations occurring in MWEs (P_{ambig}) implies scarcity of naturally occurring negative examples. Therefore, the Zipfian distribution (P_{zipf}) can be partly balanced, with minor bias, by complementing a (small) annotated corpus with several minimal positive occurrence examples for lower-frequency MWEs discovered in very large corpora by unsupervised methods. Fourthly, the relatively low proliferation speed (P_{prolif}) of GL-MWEs makes them good candidates for large-coverage lexical encoding. Thus, it should be possible to produce relatively stable and high-quality lexical resources via manual validation of unsupervised discovery methods.

The conclusions from Sec. 4 and 5 also speak in favor of the use of lexical MWE resources in MWEI, especially if they are offered in a unified format for many languages, and if they carry information similar to what can be found in treebanks.

These observations lead us to propose the following scenario for future development in MWEI.

- Automatic identification of GL-MWEs should be systematically coupled with MWE discovery via syntactic lexicons.
- In such lexicons, for each MWE type, one

should be able to retrieve at least: (i) the lemmas and parts of speech of its lexicalized components, (ii) its syntactically least marked dependency structure preserving the idiomatic reading (Savary et al., 2019),[12] (iii) the description of some of its morphosyntactic variants[13] preserving the idiomatic reading, e.g. those judged most frequent or most discriminating.

- If the lexicon is stored in an intentional format, it should be distributed with its extensional equivalent. The simplest form of an extensional format is a set of corpus examples for each MWE entry, with syntactic and MWE annotation.
- The extensional format should be compatible with standard corpus formats,[14] so as to require minimal effort from corpus-based tools in completing the existing corpora with the lexicon examples.
- The lexicon should encode with high priority those MWEs which occur rarely or never in the reference corpora, i.e. the corpora annotated for MWEs and used for training MWE identifiers. This is in sharp contrast to the existing NLP-oriented MWE lexicons more or less strongly coupled with reference corpora (see Sec. 4).

Note that exhaustiveness of this description, and notably of the morphosyntactic variation, is not required. This feature should make the lexical encoding adventure relatively feasible, with the help of fully and/or semi-automatic methods.

7 Roadmap

To complement the proposal of MWE discovery/identification interface from the previous section, we suggest that the MWE community should more thoroughly address the challenges posed to MWEI by unseen data. In the short run, future shared tasks on MWEI might, for instance, propose subtasks dedicated specifically to unseen data. New MWEI tools may leverage the type-specific idiosyncrasy of MWEs (P_{discr}), so as to achieve better generalization over unseen data.

The community should also put more effort into the development of large-coverage syntactic MWE lexicons. To this end, the MWE discovery task should be redefined so that not only bare lists of MWE candidates but also their syntactic structures for at least some morphosyntactic variants are extracted (Weller and Heid, 2010). Many existing discovery methods are dedicated to selected MWE categories, syntactic patterns and languages. New methods should, conversely, be more generic so as to cover the large variety of MWE categories and adapt to many languages. In order to incrementally achieve high quality for such resources (e.g. via manual validation), MWE discovery should not be performed from scratch, but should take as input and enrich existing MWE lexicons. MWE discovery evaluation measures should explicitly account for this enrichment aspect.

Steps should also be taken towards defining MWE lexicon formats which would be compatible with the recommendations from Sec. 6. To this end, a shared task on lexicon format definitions and/or lexicon construction methods could be organized. A mid-long-term objective of the community would then be to produce unified multilingual reference datasets which would consist both of MWE-annotated corpora (extended to new, non-verbal MWE categories) and of NLP-oriented MWE lexicons. We believe that these steps are necessary to bridge the performance gap between MWEI and other NLP tasks, so that MWEI becomes a regular component of traditional NLP text analysis pipelines.

Acknowledgments

This work was funded by the French PARSEME-FR project (ANR-14-CERA-0001).[15] We are grateful to Jakub Waszczuk and Kilian Evang for their valuable feedback at an early stage of our proposal. We also thank the anonymous reviewers for their useful comments.

[12] A form with a finite verb is less marked than one with an infinitive or a participle, a non-negated form is less marked than a negated one, the active voice is less marked than the passive, a form with an extraction is more marked than without, etc.

[13] Following (Savary et al., 2019), we understand a variant of a given MWE as a set of all its occurrences sharing the same *coarse syntactic structure*, i.e. the same lexicalized lemmas, POS and dependency relations.

[14] PARSEME corpora for verbal MWEs use an extension of the CoNNL-U format (https://universaldependencies.org/format.html) called cupt (http://multiword.sourceforge.net/cupt-format)

[15] http://parsemefr.lif.univ-mrs.fr/

References

Anne Abeillé and Yves Schabes. 1989. Parsing idioms in lexicalized TAGs. In *Proceedings of the 4th Conference of the European Chapter of the ACL, EACL'89, Manchester*, pages 1–9.

Anne Abeillé and Yves Schabes. 1996. Non-compositional discontinuous constituents in Tree Adjoining Grammar. In Harry Bunt and Arthur van Horck, editors, *Discontinuous Constituency*, pages 279–306. Mouton de Gruyter, Berlin, Germany.

Hassan Al-Haj, Alon Itai, and Shuly Wintner. 2014. Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27(2):130–170.

Mohammed A. Attia. 2006. Accommodating multiword expressions in an Arabic LFG grammar. In *Proceedings of the 5th international conference on Advances in Natural Language Processing*, FinTAL'06, pages 87–98, Berlin. Springer.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition. *Comput. Speech Lang.*, 44(C):61–83.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA.

Eduard Bejček and Pavel Straňák. 2010. Annotation of multiword expressions in the Prague dependency treebank. *Language Resources and Evaluation*, 44(1–2):7–21.

Francis Bond, Jia Qian Ho, and Dan Flickinger. 2015. Feeling our way to an analysis of English possessed idioms. In *Proceedings of the 22nd International Conference on Head- Driven Phrase Structure Grammar*, pages 61–74, Stanford, CA. CSLI Publications.

Elisabeth Breidt, Frédérique Segond, and Guiseppe Valetto. 1996. Formal Description of Multi-Word Lexemes with the Finite-State Formalism IDAREX. In *Proceedings of COLING-96, Copenhagen*, pages 1036–1040.

David Campos, Sérgio Matos, and José Luís Oliveira. 2012. Biomedical named entity recognition: A survey of machine-learning tools. In Shigeaki Sakurai, editor, *Theory and Applications for Advanced Text Mining*, chapter 8. IntechOpen, Rijeka.

Matthieu 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.

Matthieu Constant, Joseph Le Roux, and Anthony Sigogne. 2013. Combining compound recognition and PCFG-LA parsing with word lattices and conditional random fields. *TSLP Special Issue on MWEs: from theory to practice and use, part 2 (TSLP)*, 10(3).

Matthieu Constant and Elsa Tolone. 2010. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In Michele De Gioia, editor, *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008). Seconde partie*, volume 1 of *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, pages 79–93. Aracne. ISBN 978-88-548-3166-7.

Ann Copestake, Fabre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin, Ivan A. Sag, and Dan Flickinger. 2002. Multiword expressions: linguistic precision and reusability. In *Proceedings of LREC 2002*.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. 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)*, pages 910–917, San Diego, California, USA. Association for Computational Linguistics.

Monika Czerepowicka and Agata Savary. 2018. *SEJF - A Grammatical Lexicon of Polish Multiword Expressions*, volume 10930 of *Lecture Notes in Computer Science*. Springer Cham.

Stefan Diaconescu. 2004. Multiword expression translation using generative dependency grammar. In *Advances in Natural Language Processing. EsTAL 2004*, volume 3230 of *Lecture Notes in Computer Science*, pages 243–254, Berlin, Heidelberg. Springer.

Doug Downey, Matthew Broadhead, and Oren Etzioni. 2007. Locating complex named entities in web text. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2733–2739, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Helge Dyvik, Gyri Smørdal Losnegaard, and Victoria Rosén. 2019. Multiword expressions in an LFG grammar for Norwegian. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and Parsing of Multiword Expressions*, pages 41–72. Language Science Press, Berlin.

Stefan Evert. 2005. *The statistics of word cooccurrences: Word pairs and collocations*. Ph.D. thesis, Univ. of Stuttgart, Stuttgart, Germany.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Polona Gantar, Lut Colman, Carla Parra Escartín, and Héctor Martínez Alonso. 2018. Multiword Expressions: Between Lexicography and NLP. *International Journal of Lexicography*.

Nicole Grégoire. 2010. DuELME: a Dutch electronic lexicon of multiword expressions. *Language Resources and Evaluation*, 44(1-2).

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

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F. J. Smith. 2002. Extension of Zipf's law to words and phrases. In *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, pages 1–6, Stroudsburg, PA, USA. Association for Computational Linguistics.

Livnat Herzig Sheinfux, Tali Arad Greshler, Nurit Melnik, and Shuly Wintner. 2015. Hebrew verbal multiword expressions. In *Proceedings of the 22nd International Conference on Head-Driven Phrase Structure Grammar, Nanyang Technological University (NTU), Singapore*, pages 122–135, Stanford, CA. CSLI Publications.

Lauri Karttunen, Ronald M. Kaplan, and Annie Zaenen. 1992. Two-Level Morphology with Composition. In *Proceedings of COLING-92, Nantes*, pages 141–148.

Cvetana Krstev, Ranka Stanković, Ivan Obradović, Duško Vitaš, and Milos Utvic. 2010. Automatic Construction of a Morphological Dictionary of Multi-Word Units. *LNAI*, 6233:226–237.

François Lareau, Mark Dras, Benjamin Boerschinger, and Myfany Turpin. 2012. Implementing lexical functions in xle.

Timm Lichte and Laura Kallmeyer. 2016. Same syntax, different semantics: A compositional approach to idiomaticity in multi-word expressions. In *Empirical Issues in Syntax and Semantics 11*, pages 111–140, Paris. CSSP.

Timm Lichte, Simon Petitjean, Agata Savary, and Jakub Waszczuk. 2019. Lexical encoding formats for multi-word expressions: The challenge of "irregular" regularities. In Yannick Parmentier and Jakub Waszczuk, editors, *Representation and Parsing of Multiword Expressions*, pages 41–72. Language Science Press, Berlin.

Irina Lobzhanidze. 2017. Computational Model of Modern Georgian Language and Searching Patterns for On-line Dictionary of Idioms. In *Twelfth International Tbilisi Symposium on Language, Logic and Computation 18-22 September, 2017, Lagodekhi, Georgia*.

Gyri Smørdal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on mwe resources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Michał Marcińczuk, Jan Kocoń, and Marcin Oleksy. 2017. Liner2 — a generic framework for named entity recognition. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 86–91, Valencia, Spain. Association for Computational Linguistics.

Marjorie McShane, Sergei Nirenburg, and Stephen Beale. 2015. The Ontological Semantic treatment of multiword expressions. *Lingvisticæ Investigationes*, 38(1):73–110.

Igor Mel'čuk, Nadia Arbatchewsky-Jumarie, Louise Dagenais, Léo Elnitsky, Lidija Iordanskaja, Marie-Noëlle Lefebvre, and Suzanne Mantha. 1988. *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques*, volume II of *Recherches lexico-sémantiques*. Presses de l'Univ. de Montréal.

Luka Nerima, Vasiliki Foufi, and Eric Wehrli. 2017. Parsing and MWE detection: Fips at the PARSEME shared task. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 54–59, Valencia, Spain. Association for Computational Linguistics.

Kemal Oflazer, Özlem Çetonoğlu, and Bilge Say. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. In *Second ACL Workshop on Multiword Expressions, July 2004*, pages 64–71.

Caroline Pasquer, Carlos Ramisch, Agata Savary, and Jean-Yves Antoine. 2018a. 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)*, pages 283–289. Association for Computational Linguistics.

Caroline Pasquer, Agata Savary, Carlos Ramisch, and Jean-Yves Antoine. 2018b. If you've seen some, you've seen them all: Identifying variants of multiword expressions. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. The COLING 2018 Organizing Committee.

Marie-Sophie Pausé. 2018. Modelling french idioms in a lexical network. *Studi e Saggi Linguistici*, 55(2):137–155.

Pavel Pecina. 2008. *Lexical association measures: Collocation extraction*. Ph.D. thesis, Faculty of

Mathematics and Physics, Charles Univ. in Prague, Prague, Czech Republic.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

Adam Przepiórkowski, Jan Hajič, Elżbieta Hajnicz, and Zdeňka Urešová. 2017. Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, 30(1):1–38.

Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk, and Marcin Woliński. 2014. Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)*, pages 83–91, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Carlos Ramisch. 2015. *Multiword expressions acquisition: A generic and open framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer. https://doi.org/10.1007/978-3-319-09207-2.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 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)*, pages 222–240. Association for Computational Linguistics.

Martin Riedl and Chris Biemann. 2016. Impact of MWE resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions, (MWE 2016)*, Berlin, Germany.

Omid Rohanian, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the gap: Attending to discontinuity in identification of multiword expressions. *CoRR*, abs/1902.10667.

Jake Ryland Williams, Paul R. Lessard, Suma Desu, Eric M. Clark, James P. Bagrow, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific Reports*, 5.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expresions: A Pain in the Neck for NLP. In *Proceedings of CICLING'02*. Springer.

Agata Savary. 2008. Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2):1–53.

Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Sla vomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Lie bes kind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Fe derico Sangati, Ivelina Stoyanova, and Veronika Vincze. 2018. PARSEME multilingual corpus of verbal multiword expressions. In Stella Markantonatou, Carlos Ramisch, Agata Savary, and Veronika Vincze, editors, *Multiword expressions at length and in depth. Extended papers from the MWE 2017 workshop*, pages 87–147. Language Science Press, Berlin.

Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa I nurrieta, and Voula Giouli. 2019. Literal occurrences of multiword expressions: Rare birds that cause a stir. *The Prague Bulletin of Mathematical Linguistics*, 112:5–54.

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)*, pages 31–47, Valencia, Spain. Association for Computational Linguistics.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: running the MWE gamut. *Transactions of the ACL*, 2:193–206.

Violeta Seretan. 2011. *Syntax-based collocation extraction*. Text, Speech and Language Technology. Springer.

Max Silberztein. 2005. NooJ's dictionaries. In *Proceedings of LTC'05, Poznań*, pages 291–295. Wydawnictwo Poznańskie.

Ralf Steinberger, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. 2011. JRC-NAMES: A freely available, highly multilingual named entity resource. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 104–110, Hissar, Bulgaria. Association for Computational Linguistics.

Shiva Taslimipoor and Omid Rohanian. 2018. SHOMA at PARSEME Shared Task on automatic identification of vmwes: Neural multiword expression tagging with high generalisation. *CoRR*, abs/1809.03056.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *Proceedings of the 6th Conference on Natural Language Learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Zdeňka Urešová. 2012. Building the PDT-Vallex valency lexicon. In *On-line Proceedings of the fifth Corpus Linguistics Conference*, University of Liverpool.

Ashwini Vaidya, Owen Rambow, and Martha Palmer. 2014. Light verb constructions with 'do' and 'be' in Hindi: A TAG analysis. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing*, pages 127–136.

Aline Villavicencio, Ann Copestake, Benjamin Waldron, and Fabre Lambeau. 2004. Lexical Encoding of MWEs. In *ACL Workshop on Multiword Expressions: Integrating Processing, July 2004*, pages 80–87.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011. Multiword expressions and named entities in the wiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 289–295, Hissar, Bulgaria. Association for Computational Linguistics.

Jakub Waszczuk. 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)*, pages 275–282. Association for Computational Linguistics.

Jakub Waszczuk, Katarzyna Glowinska, Agata Savary, Adam Przepiórkowski, and Michal Lenart. 2013. Annotation tools for syntax and named entities in the National Corpus of Polish. *IJDMMM*, 5(2):103–122.

Jakub Waszczuk, Agata Savary, and Yannick Parmentier. 2016. Promoting multiword expressions in A* TAG parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 429–439, Osaka, Japan. The COLING 2016 Organizing Committee.

Marion Weller and Ulrich Heid. 2010. Extraction of German Multiword Expressions from Parsed Corpora Using Context Features. In *LREC*.

Vikas Yadav and Steven Bethard. 2018. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Yarowsky. 1993. One sense per collocation. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.