



This is a repository copy of *Discovering multiword expressions*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/153553/>

Version: Accepted Version

Article:

Villavicencio, A. orcid.org/0000-0002-3731-9168 and Idiart, M. (2019) Discovering multiword expressions. *Natural Language Engineering*, 25 (6). pp. 715-733. ISSN 1351-3249

<https://doi.org/10.1017/S1351324919000494>

This article has been published in a revised form in *Natural Language Engineering* <https://doi.org/10.1017/S1351324919000494>. This version is free to view and download for private research and study only. Not for re-distribution, re-sale or use in derivative works.
© Cambridge University Press.

Reuse

This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

Discovering Multiword Expressions

Aline Villavicencio and Marco Idiart
Federal University of Rio Grande do Sul, Brazil
University of Essex, UK

(Received July 2019)

Abstract

In this paper we provide an overview of research on Multiword Expressions, from a natural language processing perspective. We examine methods developed for modelling MWEs that capture some of their linguistic properties, discussing their use for MWE discovery and for idiomaticity detection. We concentrate on their collocational and contextual preferences, along with their fixedness in terms of canonical forms and their lack of word-for-word translatability. We also discuss a sample of the MWE resources that have been used in intrinsic evaluation setups for these methods.

1 Introduction

Multiword Expressions (MWEs) have already been described as *a pain in the neck* (Sag et al.2002a) and *hard going* (Rayson et al.2010) for natural language processing (NLP), but also considered to be *much ado about nothing* (de Marneffe et al.2009) and perhaps *plain sailing* (Rayson et al.2010) through the years. Despite any controversies, with a growing community and various events dedicated to them, interest in MWEs shows no indication of slowing down, as they can be viewed as providing not only challenges but also opportunities for designing new solutions for more accurate language processing (Constant et al.2017).

After almost two decades and thousands of citations since the publication of the *Pain in the Neck* paper by Sag et al. (2002a) what is it that makes them still an object of interest? First of all, MWEs come in all shapes, sizes and forms, from a (long) idiom like *keep your breath to cool your porridge* (as *keeping to your own affairs*) to a (short) collocation like *fish and chips*, and models designed for one category of MWE may not be adequate to other categories. Secondly, they may also display various degrees of idiosyncrasy, including lexical, syntactic, semantic and statistical (Baldwin and Kim2010), which may interact in complex ways. For instance, a *dark horse*, in addition to describing the colouring of an animal, may also be used to refer to *an unknown candidate who unexpectedly succeeds* and this second meaning cannot be fully inferred from the component words. As a consequence, their accurate detection and understanding may require knowledge that goes beyond the individual words and how they can be combined together (Fillmore1979).

In this paper we review some of the methods that have been adopted for computationally modelling MWEs, concentrating on their discovery from corpora. The paper is structured as follows: we start with a brief description of MWEs in section 2. Methods for MWE discovery are reviewed in section 3, with focus on discovering information from their collocational and contextual profiles (sections 4 and 5), as well as from the degree of rigidity of form and translatability (sections 6 and 7). We also discuss some of the MWE resources available (section 8). We finish with some conclusions and discussion of future possibilities.

2 What is in a Word/Multiword?

Multiword expressions are all around. According to estimates about four MWEs are produced per minute of discourse (Glucksberg1989). They feature prominently in the mental lexicon of native speakers (Jackendoff1997) in all languages and domains, in informal and in technical contexts (Biber et al.1999). They can be found in songs (*Joshua Tree* by U2, *Knocking on Heaven’s Door* by Guns ‘N’ Roses), in books (*Much ado about nothing*, *All is well that ends well* by Shakespeare), in newspaper headlines (*Spilling the beans about coffee’s true cost*¹), and in scientific texts (*dentate gyrus*, *long-term memory*, *word sense disambiguation*). Moreover, these expressions have also been found to have faster processing times compared to non-MWEs (compositional novel sequences) (Cacciari and Tabossi1988; Arnon and Snider2010; Siyanova-Chanturia2013). But what are they and how can we recognise them?

Different definitions have been proposed for them that describe them as recurrent or typical combinations of words that are formulaic (Wray2002) or that need to be treated as a unit at some level of description (Calzolari et al.2002; Sag et al.2002a). In fact, there may not even be a unified phenomenon but instead a set of features that interact in non-trivial ways and that fall in a continuum from idiomatic to compositional combinations (Moon1998).

As some of these definitions refer to words and the crossing of word boundaries (Sag et al.2002a), it is also important to adopt a clear definition of what a word is, either in terms of meaning, syntax, or whitespaces (Church2013; Ramisch2015). For example, the PARSEME guidelines (Ramisch et al.2018) define a word as a “linguistically (notably semantically) motivated unit”² and MWEs as containing at least two words even if they are represented as a single token (e.g. *snowman*). Here for the sake of simplicity we assume that words are separated by whitespaces in texts.³ Adopting clear and precise definitions for these target concepts provides the

¹ From the Guardian <https://www.theguardian.com/xero-digital-connectivity/2018/dec/11/spilling-the-beans-about-coffees-true-cost>

² https://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/?page=010_Definitions_and_scope/010_Words_and_tokens

³ Although simple to implement, this definition will not work for languages whose writing system does not use spaces like Chinese and Japanese, or for agglutinative languages in which a single word can in fact be an MWE (e.g. single-token compounds in Germanic languages) (Ramisch and Villavicencio2018).

basis for estimating their occurrence in human language and consequently for determining adequate vocabulary sizes, since the performance of many tasks seems to be linked to vocabulary size (Church2013). They are also important for designing clear evaluation setups for comparing different MWE processing methods. Discussions of alternative definitions for these and related concepts (e.g. phraseological units, phrasal lexemes, and collocations) along with the implications of the combinations they include can be found in (Moon1998; Seretan2011; Ramisch2015) and (Constant et al.2017).

Some of the core properties that have been used to describe MWEs include (Calzolari et al.2002):

- **high degree of lexicalisation**, with some component words not being used in isolation (e.g. *ad* from *ad hoc* and *sandboy* from *happy as a sandboy*),
- **breach of general syntactic rules** with reduced syntactic flexibility and limited variation (e.g. *by and large*/**short*/**largest*). Although it may be possible to find a canonical form for an MWE, it is not always easy to determine which elements form its obligatory core parts and which elements can be varied (if any), as they may allow discontinuity and some degree of modification (e.g. *throw NP to the hungry lions/wolves* as *sacrificing someone*),
- **idiomaticity or reduced semantic compositionality**, possibly involving figuration like metaphors, with the meaning of some expressions not being entirely predictable from their component words⁴. MWEs fall into a continuum of idiomaticity, from compositional expressions like *olive oil* (meaning an *oil made of olive*) to idiomatic expressions like *to trip the light fantastic* (meaning *to dance*),
- **high degree of conventionality and statistical markedness** reflecting a preference for some specific forms, or collocations, over plausible but low frequency variations, or anti-collocations (Pearce2001), (e.g. *strong tea* and *fish and chips* vs. the less common *powerful tea* and *chips and fish*).

Each of these characteristics may occur in varying degrees in a given expression, and Sag et al. (2002a) proposed a classification of MWEs in terms of how much variability they display. Fixed expressions do not display any morphological inflection or lexical variation (e.g. *in addition*/**additions* and *ad infinitum*). Semi-fixed expressions have fixed word order but display some morphological inflection (*coffee machine/machines*). Syntactically flexible expressions exhibit a large range of morphological and syntactic variation (*rock the political/proverbial/family/Olympic boat*).

To sum up, MWEs can be characterised as possibly discontinuous word combinations that display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasies (Baldwin and Kim2010). These properties can be distributed in different ways in MWE categories such as:

⁴ This property is also related to semantic decomposability (Nunberg et al.1994): by considering non-standard meanings for the components of an expression, its meaning can be compositionally constructed (e.g. *spill beans* as *reveal secrets* with *spill* as *reveal* and *beans* as *secrets*).

- **proper names:** *Manchester United*,
- **collocations:** *emotional baggage, heavy rain*,
- **compounds:** *pinch of salt, friendly fire*,
- **idioms:** *keep NP in NP's toes, throw NP to the lions/wolves*,
- **support verbs:** *wind blows, make a decision, go crazy*,
- **prepositional verbs:** *look for, talk NP into*,
- **verb-particle constructions:** *take off, clear up*,
- **lexical bundles:** *I don't know whether*.

More detailed inventories of categories are discussed by Sag et al. (2002a), Constant et al. (2017) and Ramisch et al. (2018). For instance, the PARSEME annotation guidelines (Ramisch et al.2018) focus on verbal MWEs in over 25 languages including Bulgarian, French, Portuguese and Turkish.

3 Can we detect them automatically?

There has been considerable work on describing MWEs and cataloguing their properties, and some popular resources are discussed in section 8. As their manual construction is time consuming and requires expert knowledge, much effort has been devoted to automatically extracting MWEs from corpora. This task, known as MWE discovery⁵ aims to determine if a given sequence of words forms a genuine MWE or if it can be treated as standard combination of words (e.g. *small boy*). For MWE discovery the hope is that some form of salience is present such that MWEs stand out and can be automatically detected. In this context, methods based on statistical markedness have been particularly popular since they rely on association and entropic measures calculated from corpus counts (Manning and Schütze1999; Pecina2010) and are inexpensive and independent of language and MWE category. They have been used to detect preferences of various types, including:

- **Collocational preferences.** Given that the “collocations of a given word are statements of the habitual or customary places of that word” (Firth1957a), these methods search for word sequences that are particularly recurrent in corpora and can form MWEs.
- **Contextual Preference.** Assuming the Distributional Hypothesis that you shall know a (multi)word by the company it keeps (Firth1957a) these methods have been used to detect discrepancies between the meaning of an MWE and those of its parts, as an indication of idiomaticity.
- **Canonical Form Preferences.** As MWEs may display different types of inflexibility, evidence of marked preferences for very few of the expected morphological, lexical, and syntactic variants can be use as indications of an MWE.

⁵ A related task, known as MWE identification, focuses on finding (and labelling) occurrences of a particular MWE in a text, usually with the help of previously compiled MWE resources (Constant et al.2017). In this paper we concentrate on the task of discovery, in particular in methods for finding MWEs and determining how idiomatic they can be.

- **Multilingual Preferences.** These methods are often based on detecting unexpected asymmetries in translations.

In the next sections we present a general overview of these methods.

4 Collocational preferences

Assuming that words that like to co-occur more frequently than by chance are indicative of MWEs (Manning and Schütze1999; Pecina and Schlesinger2006), this statistical markedness can be detected by measures of association strength. In a typical scenario, a list of candidate MWEs is generated, for example from n-grams (Manning and Schütze1999) or from relevant syntactic patterns for the target MWE categories (Justeson and Katz1995). The list of candidates is then ranked according to the score of association strength, and those with stronger associations are expected to be genuine MWEs.

Formally, we consider a candidate MWE as a generic n-gram with n word tokens w_1 through w_n . Its frequency in a corpus C of size N and lexicon L is denoted by $f(w_1 \dots w_n)$. From the corpus frequencies it is possible to estimate probabilities using maximum likelihood estimation, for instance the unigram probability ($p(w_1)$) and the n-gram probability ($p(w_1 \dots w_n)$):

$$p(w_1) = \frac{f(w_1)}{N} \quad , \quad p(w_1 \dots w_n) = \frac{f(w_1 \dots w_n)}{f(* \dots *)}$$

or the probability that the word w_1 occurs in the left of a bigram

$$p(w_i*) = \frac{f(w_i*)}{f(* *)}$$

or even the probability that two words appears separated by a certain number of words

$$p(w_i * * w_j) = \frac{f(w_i * * w_j)}{f(* * * *)}$$

Here $*$ represents the sum over all possible words in L in that position.

A central question of the collocation problem is if the observed frequency of a given combination of words is higher than what would be expected from pure chance. Of course language is far from a random distribution of words, yet a notable discrepancy certainly represents something special in language. To access that we have to measure the strength of the association between words, that formally demands us to come up with a clear expression for the predicted frequency in the case of pure chance, a baseline sometimes referred to as a null hypothesis. The usual choice is to consider statistical independence, or that the frequency of a sequence corresponds to the product of the unigram probabilities⁶ of its members scaled by

⁶ Rigorously the quantities of interest should be marginal probabilities such as $p(*w_i*)$ for words occurring in the inner part of the MWE candidate, and $p(w_1*)$ and $p(*w_n)$ for words occurring at the extremes, where the symbol $*$ represents any word in the corpus. In a very large corpus, however, it is expected that the marginal probabilities are not significantly different from the unigram probabilities.

the size of the corpus,

$$H_0 : f_{\emptyset}(w_1 \dots w_n) = Np(w_1) \dots p(w_n)$$

Therefore the association measure has to be a function that gauges some kind of distance between the observed data and the prediction. This can be formulated both in terms of frequencies

$$A(w_1 \dots w_n) = D[f(w_1 \dots w_n), f_{\emptyset}(w_1 \dots w_n)]$$

or in terms of probabilities

$$A(w_1 \dots w_n) = D'[p(w_1 \dots w_n), p(w_1) \dots p(w_n)]$$

However we must have in mind that the true probabilities are not known, just the maximum likelihood estimates that we can obtain from a finite sample (in this case a corpus). This fact raises an important issue of statistical significance of the association itself in the case of low frequencies. In order to circumvent this problem there are many association measures that are deduced from known statistical tests. This results in more generalised versions of association measures that not only depend of unigram frequencies but also of other possible combinations involving n-grams of lower orders than the target. In the next sections we discuss some of these measures.

4.1 Pointwise Mutual Information

By far the most widely used association measure is the Pointwise Mutual Information (PMI) (Church and Hanks1990) and its variations. It is derived for bigrams directly from the mutual information between two random variables, using the log-ratio between the observed co-occurrences of the sequence and of the individual words.

$$PMI = \log \frac{p(w_1 w_2)}{p(w_1) p(w_2)} = \log \frac{f(w_1 w_2)}{f_{\emptyset}(w_1 w_2)}$$

PMI values can be positive, denoting affinity between the words, 0 denoting independence between them, or negative, denoting lack of affinity. Moreover, the closer the counts for the sequence are to the word counts, the stronger the association between the words and the more exclusively they like to co-occur.

One well known issue with PMI is its bias towards infrequent events. Its upper bound, corresponding to the case of perfect association ($f(w_i) = f(w_i w_j) = f(w_i w_j)$), is $-\log(f(w_i w_j)/N)$. Therefore a moderately associated low frequency bigram could, in principle, have a better score than a highly associated high frequency bigram (Bouma2009). To correct this, alternative statistical measures based on suitable normalisation of PMI have been proposed (Bouma2009). One popular variant is the lexicographer's mutual information (LMI), or salience score (Kilgarriff et al.2004), which adjusts a PMI value by multiplying it by the frequency, reintroducing the importance of meaningful recurrence.

So far we have discussed association between two words. One option for handling larger candidates is the generalisation of the mutual information to account for

many variables. However, as this generalisation is not unique (Van de Cruys2011), various proposals have been made for calculating the equivalent of PMI for n-grams. One of these is the specific total correlation (STC), which is the direct extension of the formula above for $w_1 \dots w_n$ and it is based on the so called total correlation proposed by Watanabe (1960). Similarly, if the Interaction Information (McGill1954) is used as a basis, we have the variant called Specific Interaction Information (SII) (Van de Cruys2011). One important difference of between STC and SII is that former is zero only if all words are independent while the latter is zero if at least one is not associated to the others. Table 1 displays these two measures for trigrams.

Another alternative for n-grams is to maintain the original PMI formulation with two variables (w_1 and w_2) but to allow each variable to contain nested expressions as one word (e.g. $w_1=first_class$ and $w_2=lounge$, and $w_1=recurrent$ and $w_2=neural_network$) (Seretan2011).

Name	Formula
Association Measures	
1. Pointwise Mutual Information (PMI)	$\log \frac{p(w_1 w_2)}{p(w_1)p(w_2)} = \log \frac{f(w_1 w_2)}{f_0(w_1 w_2)}$
2. Specific Total Correlation (STC)	$\log \frac{p(w_1 w_2 w_3)}{p(w_1)p(w_2)p(w_3)} = \log \frac{f(w_1 w_2 w_3)}{f_0(w_1 w_2 w_3)}$
3. Specific Information Interaction (SII)	$\log \frac{p(w_1 w_2)p(w_2 w_3)p(w_1 w_3)}{p(w_1)p(w_2)p(w_3)p(w_1 w_2 w_3)}$
4. Students-t based association (t)	$\frac{f(w_1 \dots w_n) - f_0(w_1 \dots w_n)}{\sqrt{f(w_1 \dots w_n)}}$
5. Dice	$\frac{n f(w_1 \dots w_n)}{f(w_1) + \dots + f(w_n)}$
6. χ^2 based association	$\sum_{\substack{v \in (w_1, \bar{w}_1) \\ u \in (w_2, \bar{w}_2)}} \frac{(f(vu) - f_0(vu))^2}{f(vu)}$

Table 1. Table of association measures.

4.2 Other measures

In addition to PMI, another score that has been used for MWE discovery is the probability of occurrence of the n-gram itself. When used on its own, it cannot detect if the high score is merely by chance because the words are frequent, or if the co-occurrence is meaningful. However, it has been used in conjunction with other measures like PMI, generating LMI.

Two other popular measures are the Student t-test based measure and the Dice coefficient (Table 1), which in common with PMI, also take into consideration the expected counts to detect meaningful occurrences between words. For instance, Student's *t*-test is based on hypothesis testing, assuming that if the words are independent, their observed and expected counts are identical. The Dice coefficient,

also known as Normalised Expectation (Pecina and Schlesinger2006), differs from both of these measures by having an upperbound of 1 for perfect correlation.

There are also measures based on contingency tables that record not only the marginal frequencies of the words, but also the probability of “non-occurrence”, and using all possible combinations of w_i and \bar{w}_i (all words but w_i), for all i . These measures, which include Pearson χ^2 (Table 1) and the more robust log-likelihood ratio (Dunning1993), compare the co-occurrence of two words with all other combinations in which they occur.

Over the years, many other association measures have been defined for MWE discovery, and Pecina and Schlesinger (2006) compiled as many as 82 measures for bigram collocation discovery found in the literature. They show that these measures capture different aspects of MWEs, and as a consequence, when combined together they can generate better results in terms of MWE discovery than if used in isolation. In fact, in comparative evaluations, no single measure has been found to be the optimal best for extracting MWEs of any category or in any language, confirming that an empirical exploration of these measures is needed for a particular category and language combination (Evert and Krenn2005; Pearce2002; Villavicencio et al.2007). Likewise, as these measures can be used to produce ranked lists of MWE candidates, as discussed before, defining a threshold that separates genuine MWEs from non-MWEs, also seems to depend on the particular target MWEs, and on whether the task benefits more from recovering more MWEs at the expense of allowing more noise, or not. Evaluation of how closely a given measure captures the MWEs of a particular domain and language are usually done by means of gold standard resources or manual validation by expert judges.

5 Contextual preferences

When deriving the meaning of a combination of words, one widely adopted strategy is to build it from the meanings of the parts, following the Principle of Compositionality⁷. This principle allows a meaning to be assigned to larger units and sentences, even if they contain unseen combinations of words. However, it is not adequate for handling idiomatic MWEs since it may lead to an unrelated meaning being derived (e.g. for *trip the light fantastic*). Considerable effort has been employed in methods for detecting idiomaticity, both at the level of MWE types, discovering the degree of idiomaticity that an MWE usually displays, and at the level of MWE tokens, deciding for a specific occurrence if it is idiomatic or not. For example, the first task would be used to identify that the meaning of *access road* can, in general, be inferred from its parts (*a road for giving access to a place*) while the second task would be to decide if in a sentence like *the exam was a piece of cake* the occurrence of *piece of cake* should be interpreted literally as *a slice of a baked good*, or idiomatically as *something easy*. For both tasks information about the contexts in which an MWE occurs has been found to be a good indicator of idiomaticity and we now discuss some of the measures that have been proposed for these tasks.

⁷ Attributed to Frege (1892 1960)

5.1 Type idiomatcity

If a word can be characterised by “the company it keeps” (Firth1957a) and given that words that occur in similar contexts have similar meanings (Turney and Pantel2010) we can approximate the meaning of an MWE by aggregating its affinities with its contexts. We can also find words and MWEs with similar meanings measuring how similar their affinities are. These affinities can be determined from distributional semantic models (or vector space models) which have been used to represent word meaning (and possibly subword and phrase meaning) as numerical multidimensional vectors in a putative semantic space (Lin1998; Pennington et al.2014; Mikolov et al.2013). These models are capable of reaching high levels of agreement with human judgements about word similarity (Baroni et al.2014; Camacho-Collados et al.2015; Lapesa and Evert2017). They vary according to factors like the following:⁸

- **Type of model:** count-based and predictive models (Baroni et al.2014). Count-based models generate vectors derived from co-occurrence counts between words and their contexts (Lin1998; Pennington et al.2014). Predictive models represent words as real-valued vectors projected onto low-dimensional space whose distances are adjusted as part of learning to predict words from contexts (or vice-versa) (Mikolov et al.2013; Baroni et al.2014).
- **Type of pre-processing** applied to the input corpus: such as lemmatisation and part-of-speech tagging. While state-of-the-art modes for English have been constructed without any pre-processing, for morphologically richer languages like French and Portuguese pre-processing the corpus can lead to better models (Cordeiro et al.2019).
- **Type of context:** in bag-of-words (BOW) models (Mikolov et al.2013), the contexts of a target word are represented as an unordered set of words that does not differentiate between their positions or relations to the target. In models based on syntactic dependencies (Lin1998; Levy and Goldberg2014) contexts are further distinguished in terms of their syntactic relations to the target (e.g. *dog* as subject vs. as object of the target).
- **Window size:** defines the number of words around the target that are included as contexts (Lapesa and Evert2014). These windows can be symmetric or asymmetric in relation to the target, and may incorporate a decay factor for prioritising words that are closer to the target.
- **Number of vector dimensions** used for representing words. These range from sparse vectors with as many dimensions as words in the vocabulary to denser and more compact representations. Reductions in the number of dimensions can be obtained using explicit context filtering, such as using only the n more frequent or salient contexts (Padró et al.2014; Salehi et al.2014), or adopting techniques like singular value decomposition.
- **Measures of association strength** between a target word and its contexts. These measures help to detect more salient co-occurrences that are not just

⁸ A detailed discussion of these models can be found in (Clark2015).

due to chance, and some of them were discussed in the previous section such as χ^2 , t-score, PMI and Positive PMI (PPMI) (Curran and Moens2002; Padró et al.2014).

- **Measures of similarity, distance or divergence** between word vectors. These measures have been used to find word vectors that display similar affinities with their contexts, like cosine (explained below), Manhattan distance, Kullback-Leibler divergence, Jensen-Shannon, Dice and Jaccard.

A major advantage of vector space models is the possibility of using algebra to model complex interactions between words. Similarity or relatedness can be modelled as a comparison between word vectors, for instance, as the normalised inner product (the cosine similarity):

$$sim_{cos}(w_1, w_2) = \hat{\mathbf{v}}(w_1) \cdot \hat{\mathbf{v}}(w_2)$$

where $\hat{\mathbf{v}}(w)$ is the normalised⁹ word vector of the word w . Compositional meaning also can be modelled as a mathematical function that composes the vectors of the words in an MWE, but this time not to compare but to add information. The simplest of all is the additive model (Mitchell and Lapata2008) but there are alternative possibilities including other operations (Mitchell and Lapata2010; Reddy et al.2011; Mikolov et al.2013; Salehi et al.2015). For the additive model the vector for a two word compound ($\mathbf{v}_\beta(w_1, w_2)$) can be defined as

$$\mathbf{v}_\beta(w_1, w_2) = \beta \hat{\mathbf{v}}(w_{head}) + (1 - \beta) \hat{\mathbf{v}}(w_{mod}),$$

where w_{head} (or w_{mod}) indicates the semantic *head* (or *modifier*) of the compound, and $\beta \in [0, 1]$ is an adjustable parameter (usually set to 1/2) that might control the relative importance of the head to the compound semantics (Reddy et al.2011). For example in *flea market*, it is the head (*market*) that has a larger contribution to the overall meaning, and β may be used to reflect this.

The degree of compositionality can be calculated between the corpus-derived vector of the MWE, $\mathbf{v}(w_1w_2)$ (e.g. for *rocket_science*)¹⁰, and the compositionally constructed vector containing the combination of the component words, $\mathbf{v}_\beta(w_1, w_2)$ (e.g. *rocket* and *science*):

$$comp(w_1w_2) = \cos(\mathbf{v}(w_1w_2), \mathbf{v}_\beta(w_1, w_2)).$$

MWEs that presented low values of *comp* are candidates to be idiomatic MWEs (Cordeiro et al.2019).

This score can be used both to validate a given candidate MWE and also to assign a degree of idiomaticity to it, since MWEs fall on a continuum of idiomaticity (McCarthy et al.2003; Reddy et al.2011; Salehi et al.2018). The success of this score hangs on how linguistically accurate the compositional models and similarity measures used are. The good news is that recent work has demonstrated that additive

⁹ $\hat{\mathbf{v}}(w) = \mathbf{v}(w)/\|\mathbf{v}(w)\|$ and $\|\cdot\|$ is the Euclidean norm.

¹⁰ This is usually done during pre-processing by connecting the words of the MWE using underscores so it corresponds to a unit (for instance *rocket science* becomes *rocket_science*).

compositional models associated with cosine similarity are suitable for detecting idiomaticity of noun compounds (Cordeiro et al.2019), and have outperformed other variants in similar tasks (Reddy et al.2011; Salehi et al.2015), including in predicting intra-compound semantics (Hartung et al.2017).

5.1.1 Additional Semantic Measures

Alternative measures for approximating idiomaticity have included comparing the distributional neighbourhood of an MWE with those of the component words, that is, the words that are closest to each of them in vector space. Assuming that compositional MWEs share more distributional neighbours with their component words, the overlap between their neighbours has been used as an indication of the degree of compositionality (McCarthy et al.2003). Additionally, the rank position of these neighbours can also be considered.

Semantic information about MWEs and their possible senses can also be obtained from resources like dictionaries and thesaurus, including synonyms, antonyms, definitions and examples. Some resources, like WordNet (Fellbaum1998) also include similarity measures like Wu-Palmer (Wu and Palmer1994) and Leacock-Chodorow (Leacock and Chodorow1998). However, their coverage for MWEs may be limited, and they may not be available for a given domain or language, restricting their applicability for idiomaticity detection.

5.2 Token idiomaticity

So far we discussed methods for discovering MWEs and deciding how idiomatic they can be, and these could be useful for building resources. However, when faced with a particular sequence of words, a speaker (as well as an automatic system) must decide whether in that sentence they can be treated as simple isolated words or if they are components of a unit, an MWE. Sometimes the syntactic context may help to disambiguate them, as in the sentence *Does the bus stop here?* where *bus stop* could be flagged as a possible MWE occurrence except that *stop* is a verb and the MWE *bus stop* is formed by two nouns. However, there are cases where both idiomatic and literal readings are possible with exactly the same syntactic configuration. For instance for *kick the bucket* more information is needed to disambiguate if a kicking event took place with a literal interpretation of the words, or a dying event with idiomatic interpretation. Although for some MWEs one of the meanings will be predominant, ambiguity is not the exception: an analysis of idiomatic verb-noun combinations (VNCs) revealed that many of them were also used with their literal senses in corpus (Fazly et al.2009). Therefore, for a given MWE occurrence in a sentence, we need to determine if it is used in a literal or an idiomatic meaning.

Token idiomaticity detection can be seen as a word sense disambiguation task, where information from the surrounding words in the sentential context can be used to help disambiguate the MWE sense. Returning to the case of *kick the bucket*, although both the literal and the idiomatic senses are possible, sentences in which the idiomatic sense occurs will include words that may not be compatible with

the literal sense (e.g. *illnesses, hospitals* and *funerals*). In previous work on token idiomaticity detection this sentential context has been modelled in terms of lexical chains, assuming that a literal sense displays strong cohesive ties with the context, which are absent for the idiomatic sense (Sporleder and Li2009).

To solve this ambiguity something akin to compositionality prediction, described in the previous section, has to take place. But this time, instead of comparing the compositional vector of the MWE formed by the combination of the parts with the corpus generated vector for the MWE, we must compare the vectors for the literal (e.g. *hitting the bucket*) and idiomatic (e.g. *dying*) senses with the vectors containing a representation of the sentential context in which the MWE occurs. In this case the sentential context can be represented using sentence-level distributional models such as Skip-Thought Vectors (Kiros et al.2015), or it can be compositionally constructed from the vector representations of the words in the sentence using an operation like vector addition. For token idiomaticity detection in VNC, King and Cook (2018) compared the use of different distributional models for representing the target sentences in which the VNCs occur, from word-level (Mikolov et al.2013) to sentence-level models (Kiros et al.2015). They found that representing a sentential context using the additive model obtained the best results. Alternatives to the additive model include concatenating word vectors of specific parts of the sentential context (Taslimipoor et al.2017).

6 Canonical Form Preferences

Methods for MWE discovery have also used information about the fixedness displayed by some MWEs in comparison with ordinary word combinations (Sag et al.2002b).¹¹ Characteristics like limited lexical and syntactic flexibility (Sag et al.2002b), have been used as indicators in tasks such as MWE discovery and idiomaticity detection. For instance, the expression *to make ends meet* cannot undergo changes in determiners (**to make some/these/many ends meet*), pronominalization (**make them meet*), modification (**to make month ends meet*), and so on.

One common strategy to detect fixedness is to generate all variants that would be expected for a given combination of words and verify which of them occurs in a very large corpus. The assumption is that absence (or very limited presence) of expected variants is an indication of idiomaticity (Ramisch et al.2008a; Fazly et al.2009). These variants can be of two types: lexical and syntactic variants.

Lexical variants can be generated by lexical substitution of the component words using synonyms from resources like WordNet (Pearce2001; Ramisch et al.2008a) and inventories of semantic classes (Villavicencio2005) or using similar words from distributional semantic models. For instance, for *nut case* variants would include *hazelnut case, cashew case, nut briefcase* and *nut luggage*. A possible measure of

¹¹ For expert annotation, the PARSEME annotation guidelines use inflexibility for various MWE detection tests. For instance, if *a regular morphological change that would normally be allowed by general grammar rules lead to ungrammaticality or to an unexpected change in meaning* this is an indication of a (morphologically inflexible) MWE (from the PARSEME annotation guidelines (Ramisch et al.2018)).

lexical fixedness (LF) proposed by (Fazly et al.2009), compares how the PMI of a target MWE deviates from the average PMI of possible variants of this target

$$\text{LF}(w_1\dots w_n) = \frac{\text{PMI}(w_1\dots w_n) - \overline{\text{PMI}}}{\sigma_{\text{PMI}}}$$

where $\overline{\text{PMI}}$ is the average on the variants and σ_{PMI} is the standard deviation. LF was defined in the context of detecting idiomaticity in VNCs and the variants were obtained from a certain number of close synonyms of the verb and the noun, but it can be adapted to larger n-grams using generalisations of PMI as discussed in section 4. The reasoning behind using PMI is to avoid the possible confound caused by high frequency lexical substitutes.

Syntactic variants can be generated according to regular syntactic rules that apply to a given MWE category, such as passivisation, pluralisation, change of determiners or adverbial modification for verbal MWEs (e.g. *?the bucket was kicked/?kick a bucket/?kick the buckets*). Due to the fact that syntactic variants may present different number of words it is no longer suitable to compare PMIs. Instead (Fazly et al.2009) defined a syntactic fixedness (SF) measure based on the probability of occurrence in the corpus of a given syntactic pattern, among a set of m syntactic patterns used to generate the syntactic variants. The proposed fixedness measure is the Kullback-Leibler divergence between the probability distribution for the typical syntactic behaviour $p(pt)$ and the distribution of occurrences of syntactic patterns given that the target n-gram is involved $p(pt|w_1\dots w_n)$.

$$\text{SF}(w_1\dots w_n) = \sum_{pt=1}^m p(pt|w_1\dots w_n) \log \frac{p(pt|w_1\dots w_n)}{p(pt)}$$

Large values of SF indicate that the target n-gram presents syntactic pattern frequencies that are very different from the typical frequency distribution expected for that kind of n-gram and this is interpreted as higher degree of syntactic fixedness (Fazly et al.2009). If the syntactic patterns are approximately uniformly distributed SF is related to the Entropy of Permutation and Insertion (EPI) proposed by Ramisch et al. (2008b),

$$\text{EPI}(w_1\dots w_n) = - \sum_{pt=0}^m p(pt|w_1\dots w_n) \log(p(pt|w_1\dots w_n))$$

Nonetheless, EPI can be used in more general contexts. Low values of EPI indicate some degree of fixedness.

Similarly, for some types of MWEs, fixedness can be captured by entropic measures of word order as the Permutation Entropy (Zhang et al.2006) defined as

$$\text{PE}(w_1\dots w_n) = - \sum_k p_k(w_1\dots w_n) \log(p_k(w_1\dots w_n))$$

where $p_k(w_1\dots w_n)$ is the probability of occurrence in the corpus of the k^{th} permutation of the n-gram $w_1w_2\dots w_n$. PE is also indirectly related to the association strength of the components of a candidate, since if there is no special association between words, the probability of them appearing in multiple orders should be similar, lead-

ing to high PE values (Villavicencio et al.2007). One of the advantages of using PE as an association measure is that it can be applied to MWEs of arbitrarily large sizes, without the need to be redefined.

If a MWE candidate passes a criteria for fixedness (a rigid adherence to a canonical form) based in the measures described in this section it is very likely an idiomatic MWE. Therefore fixedness is an informative score for MWE discovery.

Fixedness has also been incorporated in methods for detecting token idiomaticity, such as those discussed in section 5.2. The assumption is that when the idiomatic sense is used it tends to occur in the canonical form of the MWE, while the literal sense is less rigid and may occur in more patterns (Fazly et al.2009). Fazly et al. (2009) propose a method based on canonical forms learned automatically from corpora, where distributional vectors for canonical and non-canonical forms are learned and then an MWE token is classified as idiomatic if it is closer to the canonical form vectors. Methods that incorporate both information about the canonical form of an MWE and distributional information about its sentential contexts (section 5.2) have found them to be complementary and outperform models that use only one of them (Fazly et al.2009; King and Cook2018).

7 Multilingual Preferences

Idiomatic MWEs resist word-for-word translation, often generating unnatural, non-sensical or incorrect translations (e.g. *o fim da picada* in Portuguese, lit. *the end of the bridle path* meaning *something unacceptable*). When parallel resources are available this lack of direct translatability can be measured using information such as asymmetries in word alignments between source and target languages (Melamed1997; Caseli et al.2010; Attia et al.2010; Tsvetkov and Wintner2012).

The degree of idiomaticity of an MWE has also been calculated from the overlap between the translation of an MWE and the translations of its component words. Moreover, the translations for the MWE and for each of its component words can also be compared using string distance metrics that can help to account for any inflectional differences between them and determine whether the translations share a substring (Salehi et al.2014). For instance, the translation for *public* into Persian is contained in the translation for *public service*. These string similarity measures have been found to lead to better results for MWE idiomaticity detection when combined with information from distributional similarity models of the source and target language (Salehi et al.2018).

8 MWE Resources

Evaluation of MWE discovery methods can be performed intrinsically or extrinsically. In intrinsic evaluation, the results produced by a model are compared to a gold standard, usually a dictionary, electronic resource or dataset where MWEs have been manually curated using expert annotations from linguists or lexicographers, or collected via crowdsourcing. While the former provides high quality and robust annotations, it is usually costly and time-consuming to obtain. The latter provides

a faster way of gathering judgements from usually large groups of non-experts to reduce the impact of subjectivity on the scores. In extrinsic evaluation the results produced are incorporated in an NLP application such as machine translation or text simplification, with the expectation that the quality of the MWE resource will be reflected in the performance of the task. However, the results may be influenced by the the particular integration of the information into the application. In this section we focus on some of the resources that have been used for intrinsic evaluation of MWE tasks and further discussion about extrinsic evaluations can be found in (Constant et al.2017). In particular we discuss some of the main corpora that have been annotated with MWEs, as well as datasets containing human judgments about MWE properties.

Annotated corpora

- The largest initiative in terms of language diversity is the PARSEME project (Savary et al.2015), which resulted in the creation of corpora for over 25 languages (Ramisch et al.2018) containing annotations of verbal MWEs.¹²
- The Supersense-Tagged Repository of English with a Unified Semantics for Lexical Expressions (STREUSLE) (Schneider and Smith2015) provides comprehensive manual annotations of MWEs and of noun and verb semantic supersenses in a corpus of online reviews in English.¹³
- Detecting Minimal Semantic Units and their Meanings shared task data (DIMSUM) extended the STREUSLE corpus with additional domains and resulted in a comprehensive annotation of MWEs in running text for English (Schneider et al.2016). The corpus contains over 90,000 words and 5,000 MWEs.¹⁴
- The VNC-Tokens dataset (Cook et al.2008) contains 2,984 sentences from the British National Corpus that contain VNCs, marked according to whether their sense is idiomatic, literal or unclear, with up to 100 sentences for each of 53 different combinations.¹⁵
- For detecting compositionality in context, Korkontzelos et al. (2013) produced annotations for the occurrences in context of target phrases, like *old school*, with a figurative or literal meaning in 4,350 sentences from WaCky corpus.¹⁶

Datasets

- The English Compound Noun Compositionality Dataset (ECNC) (Reddy et al.2011) contains crowdsourced judgments about the degree of compositionality for a set of 90 English noun–noun (e.g. *zebra crossing*) and adjective–noun (e.g. *sacred cow*) compounds. For each compound an average of 30 judgments were collected for three numerical scores: the degree to which the first word contributes to the meaning of the compound (e.g. *zebra* to *zebra crossing*),

¹² <https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-2842>

¹³ <https://github.com/nert-nlp/streusle>

¹⁴ <https://github.com/dimsum16/dimsum-data>

¹⁵ http://cs.unb.ca/~ccook1/English_VNC_Cook.zip

¹⁶ <https://www.cs.york.ac.uk/semEval-2013/task5/index.php%3Fid=full.html>

the same for the second word (e.g. *crossing* to *zebra crossing*) and the degree to which the compound can be compositionally constructed from its parts. A Likert scale from 0 (most idiomatic) to 5 (most compositional) was used.¹⁷

- The Noun Compositionality Dataset (Ramisch et al.2016; Cordeiro et al.2019) uses the same protocol as Reddy et al. (2011) and extends the ECNC with judgments collected from native speakers for 190 new compounds for English, and 180 compounds for two additional languages, French and Portuguese. Additionally, for Portuguese the annotations were extended to include lexical substitution candidates for each of the compounds, resulting in the Lexical Substitution of Nominal Compounds Dataset (LexSubNC) (Wilkins et al.2017).¹⁸
- The Dataset of English Noun Compounds Annotated with Judgments on Non-Compositionality and Conventionalization (Farahmand et al.2015; Yazdani et al.2015) provides judgments for 1,042 English noun-noun compounds. Each compound contains two binary judgments by 4 expert annotators, both native and non-native speakers: one for its compositionality and one for its conventionalization.¹⁹
- The Norwegian Blue Parrot Dataset (Kruszewski and Baroni2014) has judgments for modifier-head phrases in English. These include annotations about the phrase being an instance of the concept denoted by the head (e.g., *dead parrot* and *parrot*) or a member of the more general concept that includes the head (e.g., *dead parrot* and *pet*), along with typicality ratings.²⁰
- The German Noun-Noun Compound Dataset (Roller et al.2013) contains judgments for a set of 244 German compounds using a compositionality scale from 1 to 7. Each compound has an average of around 30 judgments obtained through crowdsourcing. This resource has also been enriched with feature norms (Roller and Schulte im Walde2014).²¹
- A Representative Gold Standard of German Noun-Noun Compounds (Ghost-NN) (Schulte im Walde et al.2016) includes human judgments for 868 German noun-noun compounds about their compositionality, corpus frequency, productivity and ambiguity. The annotations were performed by the authors, linguists, and through crowdsourcing. A subset of 180 compounds has been selected for balancing these variables and for these the annotations were done only by experts.²²

Other collections containing MWEs include the SemEval datasets for keyphrase extraction (Kim et al.2010) and for noun compound interpretation (Nakov2008;

¹⁷ http://sivareddy.in/papers/files/ijcnlp_compositionality_data.tgz

¹⁸ <http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds>

¹⁹ https://github.com/meghdadFar/en_ncs_noncompositional_conventionalized

²⁰ <http://marcobaroni.org/PublicData/NBP.zip>

²¹ <https://www.ims.uni-stuttgart.de/forschung/ressourcen/experiment-daten/feature-norms.en.html>

²² <https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/ghost-nn.html>

Hendrickx et al.2013; Butnariu et al.2010), MWE-aware treebanks (Rosén et al.2015), MWE lists²³ as well as lexical resources (Losnegaard et al.2016).

9 Conclusions

MWEs are complicated, unruly, unpredictable and difficult. They are the telltale sign of non-native speakers and are one big stumbling block for many applications to achieve a more natural and precise handling of human language. Whole decades of research have been devoted to them, and their behaviour still defies attempts to fully capture them. However, they are also a frequent informal and very efficient communicative device to transmit whole complex concepts in a conventional manner, and in the words of Fillmore, Kay and Connor (1988) *the realm of idiomaticity in a language includes a great deal that is productive, highly structured and worthy of serious grammatical investigation*. In this paper we provided an overview of research on computational modelling of MWEs, revisiting some representative methods for MWE discovery. We concentrated in particular on methods for the detection of word combinations that qualify as MWEs, and that identify some of their characteristics, like their degree of fixedness and idiomaticity.

However, this paper only scratches the surface of MWE research, and additional discussions can be found in (Constant et al.2017; Ramisch and Villavicencio2018; Pastor and Colson2019). Moreover, progress in related areas is paving the way for a better understanding of how people learn, store and process MWEs, and for the development of computational approaches for dealing with them. For instance, advances in word representations have brought new possibilities for MWE research. In particular, crosslingual word embeddings (Søgaard et al.2019) provide fertile grounds for the exploration of multilingual asymmetries linked to idiomaticity, while richer contextually-aware word representation models like ELMo (Peters et al.2018) can be incorporated in methods for token idiomaticity detection.

One possible source of clues of how to improve MWE processing comes from studies of how the brain performs the task. Experimental studies dedicated to investigate how humans process language is growing in number and involve a series of increasingly sophisticated techniques for measuring brain activity. The focus is to understand with increasing accuracy what are the brain regions used in language processing and how their interactions varies temporally and spatially with linguistic complexity. These studies can provide clues about how MWEs are stored and processed by the human brain. The use of eye-tracking information has already brought benefits for tasks like part-of-speech tagging (Barrett et al.2016; Barrett et al.2018). MWEs have been found to have faster processing times compared to non-MWEs (compositional novel sequences) and these effects have been found in both research using eye-tracking and EEG (Siyanova-Chanturia2013). Investigations of the use of gaze features from the GECO corpus (Cop et al.2017) produced promising results in tasks like discovery (Rohanian et al.2017), and further advances are expected

²³ <http://multiword.sourceforge.net/>

with increasing availability of larger collections of eye-tracking data. There is still a large gap that has to be overcome to connect the algorithms we develop for NLP and the algorithm actually used by the brain. The hope is that the gap will close soon. MWEs are here to stay and for the foreseeable future will still be on the limelight of research.

References

- Inbal Arnon and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, 62:67–82, 01.
- Mohammed Attia, Lamia Tounsi, Pavel Pecina, Josef van Genabith, and Antonio Toral. 2010. Automatic extraction of Arabic multiword expressions. In *Proc. of the COLING 2010 Workshop on MWEs*, pages 19–27, Beijing.
- Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkha and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition.
- Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don’t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. of the 52nd ACL (Volume 1: Long Papers)*, pages 238–247, Baltimore, MD, USA, Jun. ACL.
- Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.
- Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In Anna Korhonen and Ivan Titov, editors, *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 302–312. Association for Computational Linguistics.
- Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Pearson Education Ltd, Harlow, Essex, 1st edition. 1204 p.
- G. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen.
- Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2010. Semeval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In Erk and Strapparava (Erk and Strapparava2010), pages 39–44.
- Cristina Cacciari and Patrizia Tabossi. 1988. The comprehension of idioms. *Journal of Memory and Language*, 27:668–683, 12.
- Nicoleta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine Macleod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proc. of the Third LREC (LREC 2002) (con2002)*, pages 1934–1940.
- José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1–7, Beijing, China, July. Association for Computational Linguistics.

- Helena de Medeiros Caseli, Carlos Ramisch, Maria das Gracas Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.
- Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.
- Kenneth Church. 2013. How many multiword expressions do people know? *TSLP*, 10(2):4:1–4:13.
- Stephen Clark, 2015. *Vector Space Models of Lexical Meaning*, chapter 16, pages 493–522. John Wiley & Sons, Ltd.
2002. *Proc. of the Third LREC (LREC 2002)*, Las Palmas, Canary Islands, Spain, May. ELRA.
2008. *Proc. of the 46th ACL: HLT (ACL-08: HLT)*, Columbus, OH, USA, Jun. ACL.
- Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.
- Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. In Nicole Grégoire, Stefan Evert, and Brigitte Krenn, editors, *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 19–22, Marrakech, Morocco, Jun.
- Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2017. Presenting geco: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, 49(2):602–615, Apr.
- Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised compositionality prediction of nominal compounds. *Computational Linguistics*, 45(1):1–57.
- James Curran and Marc Moens. 2002. Scaling context space. In *Proc. of the 40th ACL (ACL 2002)*, pages 231–238, Philadelphia, PA, USA, Jul. ACL.
- Marie-Catherine de Marneffe, Sebastian Padó, and Christopher D. Manning. 2009. Multiword expressions in textual inference: Much ado about nothing? In *Proceedings of the 2009 Workshop on Applied Textual Inference*, pages 1–9, Suntec, Singapore, Aug. Association for Computational Linguistics.
- Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Comp. Ling.*, 19(1):61–74.
- Katrin Erk and Carlo Strapparava, editors. 2010. *Proc. of the 5th SemEval (SemEval 2010)*, Uppsala, Sweden, Jul. ACL.
- Stefan Evert and Brigitte Krenn. 2005. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, 19(4):450–466.
- Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, Denver, Colorado, June. Association for Computational Linguistics.
- Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MIT Press, May. 423 p.
- Charles J. Fillmore, Paul Kay, and Mary Catherine O’Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64:501–538, Sep.
- Charles J. Fillmore. 1979. Innocence: A second idealization for linguistics. *Annual Meeting of the Berkeley Linguistics Society*, 5, 07.
- John R. Firth. 1957a. *Papers in Linguistics 1934-1951*. Oxford UP, Oxford, UK. 233 p.
- John R Firth. 1957b. A synopsis of linguistic theory, 1930-1955.

- Gottlob Frege. 1892–1960. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50. Translated, as ‘On Sense and Reference’, by Max Black.
- Sam Glucksberg. 1989. Metaphors in conversation: How are they understood? why are they used? *Metaphor and Symbolic Activity*, 4(3):125–143.
- Zellig Harris. 1954. Distributional structure. *Word*, 10:146–162.
- Matthias Hartung, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano. 2017. Learning compositionality functions on word embeddings for modelling attribute meaning in adjective-noun phrases. In *Proceedings of the 15th Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*.
- Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. Semeval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of *SEM 2013, Volume 2 – SemEval*, pages 138–143. ACL, June.
- Ray Jackendoff. 1997. Twistin’ the night away. *Language*, 73:534–559.
2005. *Comp. Speech & Lang. Special issue on MWEs*, 19(4).
2010. *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing*, 44(1-2), Apr.
- John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(01):9–27.
- Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX*.
- Su Nam Kim, Olena Medelyan, Min-Yen Kan, and Timothy Baldwin. 2010. Semeval-2010 task 5 : Automatic keyphrase extraction from scientific articles. In Erk and Strapparava (Erk and Strapparava2010), pages 21–26.
- Milton King and Paul Cook. 2018. Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb-noun combinations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 345–350, Melbourne, Australia, Jul. Association for Computational Linguistics.
- Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.
- Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. Semeval-2013 task 5: Evaluating phrasal semantics. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013, Atlanta, Georgia, USA, June 14-15, 2013*, pages 39–47. The Association for Computer Linguistics.
- Germán Kruszewski and Marco Baroni. 2014. Dead parrots make bad pets: Exploring modifier effects in noun phrases. In Johan Bos, Anette Frank, and Roberto Navigli, editors, *Proceedings of the Third Joint Conference on Lexical and Computational Semantics, *SEM@COLING 2014, August 23-24, 2014, Dublin, Ireland.*, pages 171–181. The *SEM 2014 Organizing Committee.
- Gabriella Lapesa and Stefan Evert. 2014. A large scale evaluation of distributional semantic models: Parameters, interactions and model selection. *Transactions of the Association for Computational Linguistics*, 2:531–545.
- Gabriella Lapesa and Stefan Evert. 2017. Large-scale evaluation of dependency-based dsms: Are they worth the effort? *EACL 2017*, page 394.
- C. Leacock and M. Chodorow. 1998. Combining local context and wordnet similarity for word sense identification. In Christiane Fellbaum, editor, *MIT Press*, pages 265–283, Cambridge, Massachusetts.
- Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proc. of the 52nd ACL (Volume 2: Short Papers)*, pages 302–308, Baltimore, MD, USA, Jun. ACL.

- Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774. Association for Computational Linguistics.
- Gyri Smørðal Losnegaard, Federico Sangati, Carla Parra Escartín, Agata Savary, Sascha Bargmann, and Johanna Monti. 2016. Parseme survey on MWE resources. In *Proc. of LREC 2016*, Portoroz, Slovenia.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, USA. 620 p.
- Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In Francis Bond, Anna Korhonen, Diana McCarthy, and Aline Villavicencio, editors, *Proc. of the ACL Workshop on MWEs: Analysis, Acquisition and Treatment (MWE 2003)*, pages 73–80, Sapporo, Japan, Jul. ACL.
- William J. McGill. 1954. Multivariate information transmission. *Psychometrika*, 19(2):97–116.
- I. Dan Melamed. 1997. Automatic discovery of non-compositional compounds in parallel data. In *Proc. of the 2nd EMNLP (EMNLP-2)*, pages 97–108, Brown University, RI, USA, Aug. ACL.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proc. of the 46th ACL: HLT (ACL-08: HLT) (con2008)*, pages 236–244.
- Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.
- R Moon. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford Studies in Lexicography. Clarendon Press.
- Preslav Nakov. 2008. Paraphrasing verbs for noun compound interpretation. In *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 46–49.
- Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. In Stephen Everson, editor, *Language*, pages 491–538. Cambridge University Press.
- Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014. Nothing like good old frequency: Studying context filters for distributional thesauri. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014) - short papers*, Doha, Qatar, Oct.
- Gloria Corpas Pastor and Jean-Pierre Colson. 2019. *Computational and Corpus-based Phraseology*. John Benjamins.
- Darren Pearce. 2001. Synonymy in collocation extraction. In *WordNet and Other Lexical Resources: Applications, Extensions and Customizations (NAACL 2001 Workshop)*, pages 41–46, Jun.
- Darren Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Proc. of the Third LREC (LREC 2002) (con2002)*, pages 1530–1536.
- Pavel Pecina and Pavel Schlesinger. 2006. Combining association measures for collocation extraction. In James Curran, editor, *Proc. of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 651–658, Sidney, Australia, Jul. ACL.
- Pavel Pecina. 2010. Lexical association measures and collocation extraction. In *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing (jou2010)*, pages 137–158.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In

- Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, Jun. Association for Computational Linguistics.
- Carlos Ramisch and Aline Villavicencio. 2018. Computational treatment of multiword expressions 2nd edition, 04.
- Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008a. An evaluation of methods for the extraction of multiword expressions. In *Proc. of the LREC 2008 Workshop on MWEs*, pages 50–53, Marrakech.
- Carlos Ramisch, Aline Villavicencio, Leonardo Moura, and Marco Idiart. 2008b. Picking them up and figuring them out: Verb-particle constructions, noise and idiomaticity. In Alex Clark and Kristina Toutanova, editors, *Proc. of the Twelfth CoNLL (CoNLL 2008)*, pages 49–56, Manchester, UK, Aug. The Coling 2008 Organizing Committee.
- Carlos Ramisch, Silvio Cordeiro, Leonardo Zilio, Marco Idiart, Aline Villavicencio, and Rodrigo Wilkens. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 156.
- Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archana Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoia 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, Santa Fe, New Mexico, USA, Aug. Association for Computational Linguistics.
- Carlos Ramisch. 2015. *Multiword Expressions Acquisition: A Generic and Open Framework*, volume XIV of *Theory and Applications of Natural Language Processing*. Springer.
- Paul Rayson, Scott Piao, Serge Sharoff, Stefan Evert, and Begoña Villada Moirón. 2010. Multiword expressions: hard going or plain sailing? In *Lang. Res. & Eval. Special Issue on Multiword expression: hard going or plain sailing (jou2010)*, pages 1–5.
- Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)*, Chiang Mai, Thailand, November.
- Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha. 2017. Using gaze data to predict multiword expressions. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 601–609. INCOMA Ltd.
- Stephen Roller and Sabine Schulte im Walde. 2014. Feature norms of german noun compounds. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 104–108. ACL, April.
- Stephen Roller, Sabine Schulte im Walde, and Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In Valia Kordoni, Carlos Ramisch, and Aline Villavicencio, editors, *Proc. of the 9th Workshop on MWEs (MWE 2013)*, pages 32–41, Atlanta, GA, USA, Jun. ACL.
- Victoria Rosén, Gyri Smørdal Losnegaard, Koenraad De Smedt, Eduard Bejček, Agata Savary, Adam Przepiórkowski, Petya Osenova, and Verginica Mititelu. 2015. A survey of multiword expressions in treebanks. In *of TLT 2015*.
- Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002a. Multiword expressions: A pain in the neck for NLP. In *Proc. of the 3rd CICLing*

- (*CICLing-2002*), volume 2276/2010 of *LNCS*, pages 1–15, Mexico City, Mexico, Feb. Springer.
- Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002b. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.
- Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In Gosse Bouma and Yannick Parmentier, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 472–481, Gothenburg, Sweden, April. The Association for Computer Linguistics.
- Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proc. of NAACL/HLT 2015*, pages 977–983, Denver, CO.
- Bahar Salehi, Paul Cook, and Timothy Baldwin. 2018. Exploiting multilingual lexical resources to predict MWE compositionality. 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 343–373. Language Science Press., Berlin.
- Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, Nov.
- Nathan Schneider and Noah A. Smith. 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*, pages 1537–1547, Denver, Colorado, May–Jun. Association for Computational Linguistics.
- Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. SemEval 2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM). In *Proc. of SemEval*, San Diego, California, USA, Jun.
- Sabine Schulte im Walde, Anna HäTTY, Stefan Bott, and Nana Khvtisavrishvili. 2016. GhoSt-NN: A representative gold standard of german noun-noun compounds. In *LREC*.
- Violeta Seretan. 2011. *Syntax-Based Collocation Extraction*, volume 44 of *Text, Speech and Language Technology*. Springer, Dordrecht, Netherlands, 1st edition. 212 p.
- Anna Siyanova-Chanturia. 2013. Eye-tracking and erps in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon*, 8(2):245–268.
- Anders Søgaard, Ivan Vulic, Sebastian Ruder, and Manaal Faruqui. 2019. *Cross-Lingual Word Embeddings*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 754–762, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Shiva Taslimipoor, Omid Rohanian, Ruslan Mitkov, and Afsaneh Fazly. 2017. Investigating the opacity of verb-noun multiword expression usages in context. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 133–138, Valencia, Spain, Apr. Association for Computational Linguistics.
- Yulia Tsvetkov and Shuly Wintner. 2012. Extraction of multi-word expressions from small parallel corpora. *Natural Language Engineering*, 18(04):549–573.

- Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.
- Tim Van de Cruys. 2011. Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 16–20, Portland, Oregon, USA, Jun. Association for Computational Linguistics.
- Aline Villavicencio, Valia Kordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In Jason Eisner, editor, *Proc. of the 2007 Joint Conference on EMNLP and Computational NLL (EMNLP-CoNLL 2007)*, pages 1034–1043, Prague, Czech Republic, Jun. ACL.
- Aline Villavicencio. 2005. The availability of verb-particle constructions in lexical resources: How much is enough? In *Comp. Speech & Lang. Special issue on MWEs (jou2005)*, pages 415–432.
- Satosi Watanabe. 1960. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.*, 4(1):66–82, Jan.
- Rodrigo Wilkens, Leonardo Zilio, Silvio Ricardo Cordeiro, Felipe Paula, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2017. LexSubNC: A dataset of lexical substitution for nominal compounds. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017)*, Montpellier, France. <https://aclweb.org/anthology/W17-6941>.
- Alison Wray. 2002. *Formulaic Language and the Lexicon*. Cambridge UP, Cambridge, UK. 348 p.
- Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal, Sep. Association for Computational Linguistics.
- Yi Zhang, Valia Kordoni, Aline Villavicencio, and Marco Idiart. 2006. Automated multiword expression prediction for grammar engineering. In Begoña Villada Moirón, Aline Villavicencio, Diana McCarthy, Stefan Evert, and Suzanne Stevenson, editors, *Proc. of the COLING/ACL Workshop on MWEs: Identifying and Exploiting Underlying Properties (MWE 2006)*, pages 36–44, Sidney, Australia, Jul. ACL.