# Chapter 5
# Multi-word Expressions: A Novel Computational Approach to Their Bottom-Up Statistical Extraction


**Alexander Wahl and Stefan Th. Gries**

**Abstract** In this paper, we introduce and validate a new bottom-up approach to the identification/extraction of multi-word expressions in corpora. This approach, called Multi-word Expressions from the Recursive Grouping of Elements (MERGE), is based on the successive combination of bigrams to form word sequences of various lengths. The selection of bigrams to be "merged" is based on the use of a lexical association measure, log likelihood (Dunning, *Computational Linguistics* 19:61–74, 1993). We apply the algorithm to two corpora and test its performance both on its own merits and against a competing algorithm from the literature, the adjusted frequency list (O'Donnell, *ICAME Journal* 35:135–169, 2011). Performance of the algorithms is evaluated via human ratings of the multi-word expression candidates that they generate. Ultimately, MERGE is shown to offer a very competitive approach to MWE extraction.

## 1 Introduction

Consider the following word sequences:

(1) a. Kick the bucket (idiom)
b. Apple pie (compound)
c. Strong coffee (habitual collocation, cf. *powerful coffee* is less correct)
d. To put up with (multi-word verbs)
e. You know what I mean? (speech formula)

A. Wahl
Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

S. T. Gries (✉)
University of California, Santa Barbara, Santa Barbara, CA

Justus Liebig University Giessen, Giessen, Germany
e-mail: stgries@linguistics.ucsb.edu



P. Cantos-Gómez, M. Almela-Sánchez (eds.), *Lexical Collocation Analysis*, Quantitative Methods in the Humanities and Social Sciences,
https://doi.org/10.1007/978-3-319-92582-0_5

f. A penny saved is a penny earned (proverb)
g. Barack Obama (proper name)[1]

While these sequences represent a variety of syntactic structures and lexical phenomena, they all have something in common: they are conventionalized combinations, taken up and reproduced by speakers who have used them – or heard them used by others – before. In other words, they do not represent novel creations of individual language users, assembled from scratch on the basis of regular rules of grammar and semantics that operate on individual words. In this article, we will use the term *multi-word expressions* (MWEs) to collectively refer to these various kinds of sequences.[2]

MWEs have generated a great amount of interest in linguistics over the past few decades, spurred largely by researchers who realized that earlier linguistic approaches were generally ill-equipped to handle such sequences. While these earlier approaches did acknowledge that highly salient MWEs with unpredictable meanings (viz., idioms) must be stored, such non-compositionality was considered a rather marginal linguistic feature – indeed, rule-based regularity was thought to be the dominant motif of language. However, in what has become a foundational paper in MWE research, Pawley and Syder (1983) point to the subtlety with which conventionalization among sequences of words may appear. What they term "native-like selection" describes production choices that L1 speakers make but which L2 speakers struggle with. Specifically, native speakers do not just choose words on the basis of word-level semantics and syntax, whereby two synonyms would be equally valid productions in a phrasal formulation. For example, while *strong* and *powerful* are both adjectives that share at least one sense, L1 speakers produce *strong coffee* but not *powerful coffee*. That is, although both formulations ostensibly communicate the same meaning, only the *strong coffee* sequence "feels" native-like. It must be the case, then, that L1 speakers store representations across usage events that describe the specific combination of the word type *strong* with the word type *coffee*. And, crucially, note that *strong coffee* appears decomposable into individual semantic units and thus does not seem to be an idiom expected to be stored in memory.

Works such as Pawley and Syder's have helped to shift linguists' thinking that what is stored versus assembled may actually be a much larger proportion of discourse than originally thought. Indeed, a number of studies have now set out to count the density of MWEs in discourse (e.g., Erman and Warren 2000; Foster 2001; Biber et al. 2004). And while results vary considerably based on how they operationalize and count sequence formulaicity, most studies find that between one third and a half of sequences appear to instantiate dependencies between specific lexical types. Moreover, the types of MWEs that have been shown to make up

[1]The list of types of MWEs above is by no means exhaustive or clear-cut; however, this list is inspired by a useful taxonomy in Siyanova-Chanturia, Conklin, and Schmitt (2011).

[2]Numerous terms, with partially overlapping definitions, have been broadly used to refer to the same general collection of phenomena (terms including fixed expressions, formulaic expressions, n-grams, phraseologisms, and others).

discourse are not dominated by any one kind, ranging from the subtle collocational preferences of native speakers to well-known lexical compounds.

With this emergent appreciation for the extent of between-word formulaicity, various subfields have shifted attention to MWEs. These include the use of MWEs as the basis for the differentiation between varieties of the same language (Gries and Mukherjee 2010) and between genres within a single language (Biber et al. 2004); creating multi-word dictionary entries in lexicographic work (Sinclair 1987); development of native-like abilities in second language acquisition (e.g., Sinclair 1987; Simpson-Vlach and Ellis 2010); exploration of the role of MWEs in child acquisition (Bannard and Matthews 2008) and adult language processing (Bod 2009); and creating native-like speech in natural language generation (Lareau et al. 2011), among many others.

The increasing research foci on MWEs have been accompanied by the ongoing development of methods for the identification of such sequences in discourse. Unsurprisingly, the traditional method for such identification is through hand annotation. However, this method is slow, expensive, not necessarily objective, or replicable across raters, and it does not scale up well to large corpora. One important way of addressing these limitations is through automated computational approaches for the extraction of MWEs from corpora. These approaches typically generate a list of candidate multi-word structures from a corpus and then score and rank them according to some statistical metric of co-occurrence strength. Those items ranked highest represent the algorithm's best hypotheses for true MWEs, and those ranked lowest represent the algorithm's best hypotheses for what are not MWEs. Ultimately, these items must be hand curated to more or less of a degree, with the removal of erroneous results.

These algorithms vary along a number of dimensions relating to how MWEs are defined, counted, and extracted (issues that we return to in the next sections); thus, they will yield different lists of MWEs that they hypothesize in a given text. At the same time, they all rely on the premise that MWEs ought to be discoverable through word co-occurrence counts. This is because, over diachronic time, linguistic structures that are recurrently used become increasingly conventionalized in meaning and form; thus, conventionalization/formulaicity tends to correlate with usage frequency.

The current article presents an implemented algorithm that we have developed for the extraction of MWEs, entitled MERGE (Multi-word Expressions from the Recurrent Grouping of Elements)[3]. As we will discuss below, this algorithm differs from many traditional approaches to MWE extraction in that it identifies sequences of various sizes that may or may not include "gaps" in them. In this way, it is designed to be sensitive to the many different structural formats that MWEs can take in language, from sequences that are adjacent (e.g., *apple pie*) to discontinuous (e.g., *as ... as*), from those that are shorter to longer (e.g., *that's what she said*).

[3]Specifically, the algorithm was first developed in the first author's Ph.D. dissertation, which was co-supervised by the second author.

MERGE accomplishes this through a recurrent mechanism that builds on existing lexical association measures from the corpus linguistic literature on the extraction of MWEs. Furthermore, as we will demonstrate below, it offers a potentially superior method over other existing approaches that identify MWEs of different sizes, an issue we return to later.

In the next section, we return to the issue of defining MWEs, discussing terminological and definitional variation in the literature, and explaining how MWEs are operationalized in the present article; also, we discuss algorithmic approaches to MWE extraction, covering the role that lexical association measures have played in this research as well as how they are adapted to the current algorithm. In Sect. 3, we report two empirical studies to validate the performance of the algorithm using human participant ratings of model output. The first study in Sect. 3.1 compares human ratings of items extracted early by the algorithm to those extracted at later iterations, under the premise that, if MERGE is finding MWEs effectively, early-item ratings ought to be higher. The second study in Sect. 3.2 compares ratings assigned to output from MERGE to ratings assigned to output from another algorithm from the literature that identifies MWEs, in order to demonstrate that MERGE does offer competitive performance to an existing approach. Finally, in Sect. 4, we offer conclusions and directions for future research.

## 2 Multi-word Expressions: Their Definition and Extraction

### *2.1 The Definition of Multi-word Expressions*

Numerous terminologies have been used in the literature to refer to formulaic, conventionalized word sequences: Wray (2002) identifies 60 terms, and her count is not exhaustive. Crucially, not all of these terms have been used to refer to exactly the same phenomena, and often the same term may be used in different works to refer to somewhat different phenomena. Despite variability in definitions, Gries (2008) identifies several different criteria that commonly appear across many definitions of formulaic language. He argues that the more researchers are consistent in defining their terms via a common set of criteria such as the ones he proposes, the easier it will be to compare studies. Thus, we define here our use of the term *multi-word expression* with reference to these criteria in an attempt to be explicit about the kinds of sequences that MERGE learns. In this discussion, we also note how the sequences that MERGE is tasked with identifying differ from (and are often more realistic/complete than) the kinds of sequences that more conventional extraction approaches are designed to identify.

Of the ways in which definitions of MWEs vary that Gries (2008) mentions, perhaps that which is most oft-cited in MWE research is the *role of semantic (non-)compositionality*. For some researcher, semantic non-compositionality (e.g., *kick the bucket* has nothing to with kicking or buckets) is a prerequisite for

formulaicity. For others, whether or not a word sequence is compositional is a basis for categorizing word sequences into different types (e.g., idiomatic versus non-idiomatic formulaic language; see Conklin and Schmitt 2012). And still in other approaches, there may be no direct accounting for semantics at all; instead, frequency-based metrics may be the sole means for identifying MWEs. Since most corpora are not annotated with the kind of semantic information that would distinguish non-compositional from compositional sequences, it is this last approach that we adopt.

Gries (2008) also notes that definitions of formulaic language vary in terms of the *types of units that can make up a co-occurrence* and the *lexical and syntactic flexibility* among these units. The most prototypical type of MWE comprises two or more words that do not admit any variation or only admit variation at the level of differing inflections (though often researchers may work with lemmatized corpora to avoid such inflectional variation). Exceptions include, for example, Gries' (2008) definition of phraseologism, which includes co-occurrences between words and paradigmatic slots that accept any number of word types representing a lexical class (e.g., *as tall as* versus *as red as*, *he spilled the beans* versus *she spilled the beans*, etc.).

Sag et al. (2002) taxonomize such lexico-syntactic flexibility, distinguishing between fixed expressions, semifixed expressions, and flexible expressions. Fixed expressions include sequences such as *by and large*, ad hoc, and *Palo Alto*, and often exhibit lexico-syntactic irregularities. Semifixed expressions allow some inflectional variations and include many non-decomposable idioms, compound nominals, and proper names. Finally, syntactically flexible MWEs include verb-particle constructions, decomposable idioms, and light verb constructions. Admittedly, the theoretical inclusion of flexible slots offers a more complete picture of MWEs as elements that interact with and are embedded within larger syntactic phrasal and clausal structures. However, computationally accounting for paradigmatic flexibility within MWEs quickly becomes a much more complex grammar induction problem, which is beyond the scope of most collocation studies. Accordingly, the MWEs that MERGE is tasked with identifying for now comprise strict co-occurrences of word forms.

The remaining three criteria that Gries (2008) identifies are where extraction algorithms tend to vary the most. Two of these are the *number of units in the MWE* and the *syntagmatic distance between units*. Regarding the first of these, often corpus linguists just focus on bigrams, as they are easy to extract computationally and handle statistically. Regarding the second criterion, researchers tend to focus on sequences whose elements are strictly adjacent. However, real MWEs may in principle be of any length, and they may involve discontinuous sequences, and thus an ideal algorithm ought to be able to extract such variable-length, possibly discontinuous MWEs. Indeed, some existing research has developed techniques for extracting adjacent MWEs of variable lengths (e.g., Nagao and Mori 1994; Daudaravičius and Murcinkevičienė 2004; Gries and Mukherjee 2010; O'Donnell 2011), as well as MWEs of variable lengths containing gaps (e.g., Ikehara et al. 1996; Da Silva et al. 1999; Wible et al. 2006). Similarly, the MERGE algorithm is designed to extract variable-length sequences that are both continuous and

discontinuous, and it is designed to do so in a way that improves upon existing approaches. It is important to note that MERGE's ability to include gaps in MWEs allows for spaces in which different lexical items of particular paradigms might be located, as discussed with regard to the *lexical and syntactic flexibility* criterion. However, MERGE does not directly learn anything about these paradigms.

Gries' (2008) final criterion is the *role of unit co-occurrence frequency* in defining a particular notion of formulaic language. Again, this is one of the criteria for which there is great variation among automated extraction techniques. As mentioned above, usage frequency is correlated with formulaicity. As such, direct corpus counts of sequence frequency may serve as a measure of MWE status (e.g., Biber et al. 2004), and some automatic extraction approaches are based on frequency counts (e.g., O'Donnell 2011). However, not all MWEs can be captured via frequency: idioms, for example, are typically low frequency yet clearly memorized; for example, an expression such as *blithering idiot(s)* occurs approximately once per 50 m words (in the Corpus of Contemporary American English) and yet is known to most native speakers of American English.

## 2.2 The Extraction of Multi-word Expressions

The identification of MWEs of different sizes and the use of lexical association measures present a paradox. On the one hand, most lexical association measures are designed for bigrams and do not scale to larger co-occurrences in obvious or uncontroversial manners. For this reason, the work that draws on these measures has tended to focus on such bigrams, neglecting interesting larger co-occurrences. One possibility of circumventing this problem is to use a simpler measure such as frequency, which is counted in the same way regardless of sequence length. However and as mentioned above, frequency counts alone may miss interesting co-occurrences that are low-frequency yet high-saliency, such as idioms. Still, assuming a particular algorithm were to manage a solution to this contradiction and could assign strength values to MWEs of different sizes, there is still the quandary of how to identify the correct size of a particular MWE. In other words, a high-scoring bigram such as *in spite* may simply be a part of a larger "true" MWE such as *in spite of*. Or, two adjacent high-scoring trigrams such as *be that as* and *as it may* may exhibit a one-word overlap such that the true MWE is the five-gram that spans them both. Simply extracting all 2- through *n*-grams and then scoring and ranking them will result in a list of many such cases. Thus, it would be desirable to develop an extraction approach whose ultimate output does not include such fragmentary cases.

In the next subsection, we provide a brief discussion of lexical association measures, given the central role they have played in MWE research in general and the role one measure plays in MERGE in particular. Then, in Sect. 2.2.1, we turn to the description of recent extraction techniques that address the issues that we have just raised in different ways.

**Table 5.1** Schematic 2 × 2 table for co-occurrence statistics/association measures

| | $Word_2$ = present | $Word_2$ = absent | Totals |
|---|---|---|---|
| $Word_1$ = present | obs: $a$<br>exp: $(a + b) \times (a + c)/n$ | obs: $a$<br>exp: $(a + b) \times (b + d)/n$ | $a + b$ |
| $Word_1$ = absent | obs: $a$<br>exp.: $(c + d) \times (a + c)/n$ | obs: $a$<br>obs: $a$ | $c + d$ |
| Totals | $a + c$ | $b + d$ | $a + b + c + d = n$ |

### 2.2.1 Traditional Lexical Association Measures

Numerous lexical association measures have been developed by corpus linguists to quantify the amount of statistical attraction between words in bigram relationships (Pecina (2009) reviews 80 separate measures). Most of these measures are based on *contingency tables*, such as the one in Table 5.1, which represents schematically the observed and expected frequencies of occurrence of the constituents of a bigram (or any bipartite collocation, for that matter) and their co-occurrence.

Generally, lexical association measures are based on various mathematical formulae that compare observed frequency cell value(s) to expected frequency cell value(s). Using an association measure's formula, one can calculate an association score for each bigram type; these scores may then be used to rank the bigrams in a corpus by strength. While each measure's scores represent different units, often a positive value will indicate statistical association between two words: that is, that the two words co-occur more often than might be expected by chance. Conversely, a negative value will indicate statistical repulsion, or that two words occur less frequently than might be expected by chance.

Of the measures that have been developed, some have emerged as more popular than others. For example, mutual information (*MI*) is among the most well-known association measure. However, *MI* and transitional probability – which is not usually considered a lexical association measure but nonetheless measures sequence strength – exhibit a similar problem. They rank very low-frequency, high-contingency bigrams too highly (e.g., a bigram in which both component words are hapaxes; see Daudaravičius and Murcinkevičienė 2004); alternatives such as $MI^k$ fare somewhat better in this respect (see McEnery 2006, Evert 2009:1225). Another, and maybe the most popular, lexical association measure that has yielded quite good results (e.g., Wahl 2015) and does not appear oversensitive to very low frequencies is log likelihood (Dunning 1993), whose formula is given in (2).

(2) $\text{log likelihood} = 2\sum_{i=a}^{d} \text{obs} \times \log \frac{\text{obs}}{\text{exp}}$

Unlike other measures, log likelihood takes into account observed and expected values from all four frequency cells ($a$, $b$, $c$, and $d$) of the contingency table. It also provides a close approximation to Fisher's exact test (Evert 2009:1235), considered on mathematical grounds to be the best method for quantifying statistical association (yet its computational cost to implement makes it prohibitive for iterative applications like MERGE). Due to these strong credentials, log likelihood is the

measure we use in the present implementation of MERGE[4]. One final point that should be made is that (2) will always result in positive values. Thus, in order for log likelihood scores to correspond to the convention in which positive values denote statistical attraction between words and negative values repulsion, the product of eq. 1 must be multiplied by $-1$ when the observed frequency of a bigram is less than the expected (following Evert 2009:1227).

### 2.2.2 Some Newer Developments

In this section, we discuss some newer developments in MWE extraction research. First, we discuss two studies that use a so-called lexical gravity approach; then, we turn to O'Donnell's (2011) adjusted frequency list; finally, we discuss work on discontinuous MWEs, focusing in particular on the recursive bigram approach by Wible et al. (2006).

Daudaravičius and Murcinkevičiené (2004) develop a new lexical association measure known as lexical gravity (*LG*). The distinctive feature of this measure is that, unlike all other measures used with at least some frequency, it takes the type frequency of the token frequencies (in particular in cell *b*) into account; see Gries (2012) for detailed exemplification. At its heart, *LG* is based on the sum of the forward and backward transitional probabilities (TPs) of a two-way co-occurrence. However, each TP is weighted by the type frequency (i.e., the number of different word types) that can occupy its outcome slot, given its cue. Thus, for a given (forward or backward) TP, there is a reward for promiscuity in possible outcomes and a punishment for faithfulness (this is because a high TP is more impressive when it occurs in the context of many possible outcomes).

While *LG*, like other association measures, is principally a two-way co-occurrence metric, Daudaravičius and Murcinkevičiené 2004 develop a technique for extending it to the identification of $n + 2$-grams. Their algorithm moves through the corpus incrementally and considers any uninterrupted sequence of bigrams with *LG* values exceeding 5.5 as constituting an MWE or *collocational chain* in their terminology (they do not motivate their choice of 5.5 as their threshold value, but at $\text{df} = 1$ this corresponds to a *p*-value of approximately 0.02). In a later paper, Gries and Mukherjee (2010) refine this technique by basing the collocational chain criterion on *mean LG*. Specifically, they extract *n*-grams of various lengths and score them on the basis of the mean *LG* of their component bigrams, discarding those *n*-grams with mean *LG*s below 5.5. Then, they proceed through the list, discarding *n*-grams that are contained by one or more $n + 1$-grams with a higher mean *LG* score. The resulting list constitutes their algorithm's hypothesis of the MWEs in the corpus.

[4]Note that while log likelihood is developed in Dunning (1993) as a lexical association measure, it is in fact a multiple of another measure known as the Kullback-Leibler (K-L) divergence from the field of information theory (Evert 2005). K-L divergence was not developed to quantify word co-occurrences, but rather to measure the difference between two discrete probability distributions that share the same domain.

Rather than adapting lexical association measures to co-occurrences beyond the bigram, another set of approaches circumvent this problem by employing frequency counts as a metric of MWE strength. One of the seminal works on MWE extraction, by Nagao and Mori 1994, takes this approach, as does the more recent adjusted frequency list (AFL) by O'Donnell (2011). This latter algorithm works by first identifying all *n*-grams up to some size threshold in a corpus. Next, only *n*-grams exceeding some frequency threshold are retained in the AFL along with their frequency (in his paper, the author set this frequency threshold to three). Then, for each *n*-gram, starting with those of threshold length and descending by order of length, the two component *n*-minus-1-grams are derived. Finally, the number of tokens in the frequency list of each *n*-minus-1-gram is decremented by the number of *n*-grams in which it is a component. Like the lexical gravity approaches, this procedure prevents the kinds of overlaps and redundancies that would result from a brute-force approach of simply extracting all *n*-grams of various sizes and then ranking them based on frequency. However, in using the AFL, there is the possibility that low-frequency, high-contingency MWEs would be ignored.

One drawback of these approaches is that, as implemented, they do not allow for discontinuous MWEs. Most corpus linguistic work has shied away from the challenges of the combinatorial explosion entailed by extracting MWEs with discontinuities. Notable exceptions include an early approach by Ikehara et al. (1996) (itself based on the work by Nagao and Mori), Da Silva et al.'s (1999) LocalMax algorithm, and an algorithm by Wible et al. (2006), all of which are capable of identifying both continuous and discontinuous MWEs. We will focus on this last approach, which also crucially differs from other approaches in that it does not generate a list of ranked MWEs hypotheses contained in a corpus. Instead, it is designed to find all of the MWEs that a given node word participates in (in this way, it is more akin to a concordancer). The algorithm represents what we will call a recursive bigram approach. Upon selection of a node word to be searched, the algorithm generates continuous and discontinuous bigrams within a specified window size around each token of the node word in the corpus; these bigrams consist of all those that have the node word as one of their elements. Next, the algorithm scores these bigrams on the basis of a lexical association measure (they use *MI*), and all those bigrams whose score exceeds a specified threshold are "merged" into a single representation. The algorithm then considers new continuous and discontinuous bigrams, in which one of the elements is one of the new, merged representations, and the other element is a single word within the window. The new bigrams are scored, and winners are chosen and merged. This progress iterates until no more bigrams exceeding the threshold are found. Ultimately, the algorithm generates a list of MWEs of various sizes that contain the original node word. Importantly, the model never has to calculate association strengths for co-occurrences larger than two elements, since one element will always be a word, and, after the first iteration, the other element will always be a word sequence containing the node word.

### 2.2.3 Co-occurrence Versus Grammar-Based MWE Extraction

The methods for MWE extraction discussed thus far are based on recurrent co-occurrences between word forms or, sometimes, lemmas. Furthermore, they are unsupervised: while gold standard lists of MWEs may be used a posteriori to evaluate algorithms' performance, there are not parameters of the algorithm trained on labels prior to evaluation. In contrast to this paradigm, a parallel line of research for the identification of MWEs has been pursued in the field of computational linguistics. While methods vary, these researchers prototypically use supervised approaches whereby sequence labelers and/or parsers are trained on a partition of a corpus that is enriched with additional features besides just the boundaries between word forms or lemmas (see, e.g., Spence et al. 2013, Constant et al. 2017 for an up-to-date survey). For example, these features may include parts of speech labels, syntactic dependencies, MWE tags, and morphological and frequency/statistical association information. Once training has converged, the algorithm is tested on another partition of the corpus in order to see how it can match the MWE tags (and possibly other features).

Research has suggested that these labeler- and parser-based supervised approaches achieve a higher level of precision and recall than *n*-gram-based approaches. That said, unsupervised co-occurrence-based approaches present a different domain of application. To the extent that they do not rely on a corpus already enriched with MWE and POS labels, syntactic dependencies, and other features, they may be applied in a much broader set of contexts – for example, for the case of smaller languages with few corpus resources or with texts from specialized domains. In many of these circumstances, while the set of POS and syntactic category types (if not tokens) may be exhaustively known, it is not necessarily the case that the set of MWE types are known. Thus, unsupervised co-occurrence-based approaches allow for the exploratory, bottom-up investigation of what MWEs might exist within a particular domain.

### 2.2.4 MERGE: A New Recursive Bigram Approach

Similar to the algorithm developed by Wible et al. (2006), the MERGE algorithm embodies a recursive bigram approach. But unlike this earlier work, our algorithm is designed to extract all MWEs in a corpus (not just those that contain a particular node word). It begins by extracting all bigram tokens in a corpus. These include adjacent bigrams, as well as bigrams with one or more words intervening, up to some user-defined discontinuity parameter (similar to Wible et al.'s use of a window). The tokens for each bigram type are counted, as are the tokens for each individual word type, and the total corpus size (in words) is tallied. Next, these values are used to calculate log likelihood scores. The highest-scoring bigram is selected as the winner, and it is merged into a single representation; that is, it is assigned a data structure representation equivalent to the representations of individual words (this differs from Wible and colleagues' approach, wherein

multiple winners were chosen at an iteration on the basis of a threshold association value). We call these representations *lexemes*. At the next stage, all tokens of co-occurring word lexemes in the corpus that instantiate the winning bigram are replaced by instances of the new, merged representation. More specifically, if the winning bigram type is the combination of the lexeme "in" followed by a one-word gap and followed by the lexeme "of," the newly created lexeme would be "in _ of." Furthermore, at each point in the corpus where this co-occurrence is attested, the leftmost word position is populated with the new lexeme ("in" becomes "in _ of") and the other word positions in the co-occurrence (i.e., "of") are populated with placeholder objects that point to the leftmost word position of the co-occurrence.

Frequency information and bigram statistics must then be updated. New candidate bigrams are created through the co-occurrence in the corpus of individual word lexemes with tokens of the new merged lexeme. For example, the lexeme *in _ of* can now co-occur with *spite*, which occurs in the gap between *in* and *of*. Furthermore, certain existing candidate bigrams may have lost tokens. That is, some of these tokens may have partially overlapped with tokens of the winning bigram (i.e., they shared a particular word token). Since these word tokens in effect no longer exist, these candidates' frequency counts must be adjusted downward. For example, some or all of the occurrences of the individual word *in* followed by *spite* have ceased to exist, since many/all of the relevant tokens of *in* were swallowed up by the merge that created *in _ of*. And because of this, the frequency of the individual word types found in the winner must be reduced by the number of winning bigram tokens. Finally, the corpus frequency has decreased, since individual words have been consumed by two-word sequences. After these adjustments in frequency information have been made, new bigram strengths can be calculated.

The cycle then iteratively repeats from the point at which a winning bigram is chosen above, and this iteration continues until the lexical association strength of the winning bigram reaches some minimum cutoff threshold. After cycle cutoff, the output of the algorithm is a corpus, parsed in terms of MWEs, and a list of lexemes, from individual words to MWEs of different sizes, with and without gaps.

Because the input to candidate bigrams at later iterations may be output from previous iterations, MERGE can grow MWEs unrestricted in size, which is similar to the Wible et al. (2006) algorithm. Another key difference, however, is that one element of their candidate bigrams must always be a single word and the other a word sequence (at least after the first iteration, where both elements are single words). In contrast, at later iterations, MERGE can choose a winning bigram that comprises two single words, a single word and a word sequence, or two word sequences. Moreover, assuming a sufficiently sized gap parameter, one element may in principal occur inside the gap of another element. Even more unusual scenarios are possible: *as _ matter* and *a _ of fact* could be interleaved to form *as a matter of fact*. Thus, there are many possible paths of successive merges that result in a particular MWEs, provided that the distance between the leftmost words of the two elements of a bigram never exceeds the discontinuity parameter.

Thus, MERGE sits at the vanguard in terms of MWE extraction research in that it identifies MWEs that are co-occurrences of (dis)continuous words of various lengths, on the basis of statistical measures of lexical association.

## 3 Empirical Evaluation of the Algorithm

It is necessary to determine whether MERGE does in fact do a reasonable job of identifying MWEs. In this section, we report two different empirical studies. In Sect. 3.1, we discuss a study in which human participants rated sequences extracted by the algorithm for how well these sequences reflect "true" MWEs. Specifically, we are testing the hypothesis that the point in time when MERGE labels an expression a MWE can distinguish MWEs that are highly formulaic from MWEs that are not. After that, in Sect. 3.2, we discuss another such rating study; this time, however, the output of MERGE is compared to the output of a different automated MWE extraction approach from the literature, the AFL, to test the hypothesis that MWEs returned by MERGE will score higher in formulaicity than MWEs returned by the AFL approach.

### *3.1 Rating Study 1: "Good" vs. "Bad" MWEs*

In this study, we explore how human participants rate MWEs that differ along two crucial dimensions. The first of these dimensions is captured in a binary variable BINRANK, *early* vs. *late*, which states when during MERGE's application a MWE is identified: early (which, if MERGE is successful, should be MWEs that are rated as highly formulaic) or late (which should be MWEs that should not be rated as highly formulaic).

The second dimension is captured in a numeric variable SIZE which could take on values from 2 to 5 and just provides the number of lexical constituents of the MWE. In Sect. 3.1.1, we discuss how the MWEs we used in the experiment were obtained; in Sect. 3.1.2, we describe how the experiment was designed and undertaken; in Sect. 3.1.3, we discuss how the results were analyzed statistically; in Sect. 3.1.4, we present the results of the statistical analysis, and in Sect. 3.1.5, we provide an interim summary and discussion of this first case study.

#### 3.1.1 Materials

The input data for the algorithm comprised two corpora: the Santa Barbara Corpus of Spoken American English (SBC; Du Bois, Chafe, Meyer, and Thompson 2000; Du Bois, Chafe, Meyer, Thompson, and Martey 2003; Du Bois and Englebretson 2004; 2005) and the spoken component of the Canadian subcorpus of the Interna-

tional Corpus of English (ICE-Canada Spoken; Newman and Columbus 2010). SBC includes about 250,000 words, while ICE-Canada Spoken includes about 450,000, for a combined total of 700,000 words.

To maximize the likelihood that study participants would be familiar with the MWEs that appear, it was decided to use corpora that comprise recent North American English, since the participants are young college students in the USA. Furthermore, it was decided to use spoken language data that span a variety of discourse genres (the files of the corpora include face-to-face and telephone conversations, academic lectures, religious sermons, political debates, business meetings, radio programs, and many others). The greater formality of written language means that it is more likely to contain low-frequency, unfamiliar word combinations.[5]

These criteria greatly limited the candidate corpora, so we decided to combine two smaller corpora to generate as large a data set as possible. Note that, although more than half of the words in the combined corpus are from Canadian speech, while the study participants are from the USA, the differences between these two varieties are relatively minute compared to the differences between, say, US and British varieties (the reason why the ten million word spoken component of the British National Corpus was not used).

The formatting of both corpora was then standardized. All tags and transcription characters that were not part of the lexical representation of the words themselves were removed, including markers of overlap in talk, laughter, breathing, incomprehensible syllables, pauses, and other non-lexical vocalizations, among other features.

Following corpus preprocessing, the MERGE algorithm was run on the data set. The maximum gap size threshold was set to one – that is, the algorithm could acquire MWEs with one or more gaps within them, provided that these gaps were no longer than one word long. The algorithm was run for 20,000 iterations. Bigrams that span a boundary between turns-at-talk were not permitted.

Next, output MWEs were selected for use as experimental stimuli. These included the first 40 and last 40 merged items for each size of MWE in terms of the number of words that they contained, from MWEs of two words to MWEs of five words. While the model did extract sequences of six or more words, these were relatively few in number, so a maximum size of five words was chosen. Thus, 320 different MWE types were selected, with half belonging to an *early* bin and half to a *late* bin.

[5]This assumption that written language exhibits a greater lexical diversity than the often more repetitive, simpler topics kind of language you find in spoken data (especially in conversation) is widely held and supported, for instance, by a quick computation of lexical diversity statistics in the ICE-GB. Guiraud's measure of lexical diversity returns a value of approximately 64.3 for all the written data in ICE-GB, which is completely different than the mean of 500 random samples of the same number of words from the spoken data without replacement, mean = 41.2, IQR = 0.14.

### 3.1.2 Experimental Design

Four different versions of the rating survey were then created, each containing 80 MWEs. Each version included 10 two-word MWEs from the early bin, 10 two-word MWEs from the late bin, 10 three-word MWEs from the early bin, and so forth. Each group of 10 words was selected at random, without replacement, from all the MWEs that exhibited the same bin identity and were of the same size. Five copies of each version of the survey were then created (with stimuli ordered randomized within each copy), for a total of 20 surveys. Each stimulus item was also accompanied by an example utterance sourced from the corpus that contained the item, so that study participants had a sense of the use of the candidate MWE in context.

Next, the survey instructions were prepared. As discussed in Sect. 2, there are various criteria involved in defining/identifying MWEs, which differ from study to study. However, as we have mentioned, a common thread among different definitions and types of MWEs is that they are maintained in and reused from memory across usage events, rather than constructed on line from regular rules. In order to tap into nonspecialist intuitions about this notion, the instructions asked participants to rate sequences, on a seven-point Likert scale, for how well they represented *common, reusable chunks* (with seven indicating strong agreement). The instructions were supplemented with both good and bad examples of common, reusable chunks, based on the opinion of the researcher. These examples were sourced from the MERGE output and were not included as stimulus items.

Finally, 20 participants were recruited from introductory linguistics courses at the University of California, Santa Barbara. Each participant was placed in a quiet room by themselves and given as much time as they needed to complete the survey.

### 3.1.3 Statistical Analysis

The judgment data were analyzed with what is currently the state of the art for psycholinguistic data with dependent numeric (or potentially ordinal) variables, a linear mixed-effects model; we used the software language and environment R (R Core Team 2016) with the packages lmer (Bates et al. 2015) for the overall model selection process, lmerTest (Kuznetsova et al. 2016) to obtain $p$-values (based on Satterthwate's approximations), as well as MuMIn (Barton 2015) to obtain $R^2$ values for our regression models (Nakagawa and Schielzeth 2010, Johnson 2014). The dependent variable in our regression model was RATING, i.e., those ratings provided by the subjects. As independent variables, we entered the above-mentioned predictors SIZE (as an orthogonal polynomial to the second degree) and BINRANK as well as their interaction. The random-effects structure we used was the maximal random-effects structure that converged without warnings (following Barr et al. 2013): varying intercepts for every $n$-gram and every experimental subject as well as slopes for SIZE and BINRANK for every $n$-gram and every subject.

Note that this approach to evaluation differs from many of the approaches adopted in the literature on supervised MWE identification. There, algorithm

**Table 5.2** Results for the fixed-effects part of the regression model (REML)

| Predictor | coef | se | df | $t$ | $p_{2\text{-tailed}}$ |
|---|---|---|---|---|---|
| Intercept | 5.69 | 0.16 | 29.6 | 34.86 | $<10^{-15}$ |
| SIZE (polynomial 1) | −26.26 | 4.13 | 129.6 | −6.36 | $<10^{-8}$ |
| SIZE (polynomial 2) | −13.04 | 2.85 | 162.5 | −4.57 | $<10^{-5}$ |
| BINRANK: *early* → *late* | −3.87 | 0.2 | 31 | −19.17 | $<10^{-15}$ |
| SIZE (polynomial 1): BINRANK | 15.88 | 4.93 | 178.6 | 3.22 | 0.0015 |
| SIZE (polynomial 2): BINRANK | 11.66 | 3.92 | 322.2 | 2.98 | 0.0031 |

performance is compared against MWE labels/decisions as to whether a particular sequence is or is not an MWE provided by human subjects, which are considered to be to be the gold standard. Here, we make no such Boolean either-or claims but use scalar information instead. Because of this methodological choice, the conventional Boolean-based evaluation metrics of "precision" and "recall" are not available, and instead we use regression to assess the degree of correlation between human ratings and algorithm performance.

### 3.1.4 Results

The results of the linear mixed-effects model indicated a significant correlation (LR chi-squared 87.08, df = 5, $p < 10^{-15}$, from a ML-comparison to a model without fixed effects) with a high/strong overall effect: $R^2$marginal, the $R^2$-value that quantifies the amount of variance explained by the fixed effects, is 0.643, and all fixed effects entered into the model reached standard levels of significance; see Table 5.2 for the corresponding results.

Compared to the above-mentioned fixed-effects, the random-effects structure, while having some effect, did less in terms of variance explanation: $R^2$conditional, the $R^2$-value that quantifies the amount of variance explained by both fixed and random effects, is 0.84, and the main random-effects contributions were made by both varying intercepts and by the different GRAM slopes for BINRANK; the product-moment correlation between the observed ratings and the one predicted by our model is $r = 0.93$.

Figure 5.1 is a visual effects-plot representation of both our fixed- and random-effects results. On the $x$-axis, we show the predictor SIZE, on the $y$-axis the predicted judgments by the experimental participants (averaged across MWEs). Each thin blue and red line represents a single participant's regression line for the BINRANK, *early*, and BINRANK, *late* data, respectively (highlighting the individual variation quantified by the random-effects structure), whereas the red and blue confidence bands indicate the impact the interaction of the two fixed effects has on the predicted judgments.

The main effect of BINRANK, *early* vs. *late,* is the most crucial finding in this experiment: the (blue) early MWEs, the ones hypothesized to be highly formulaic, do indeed have highly significantly higher overall ratings than the (red) late MWEs,

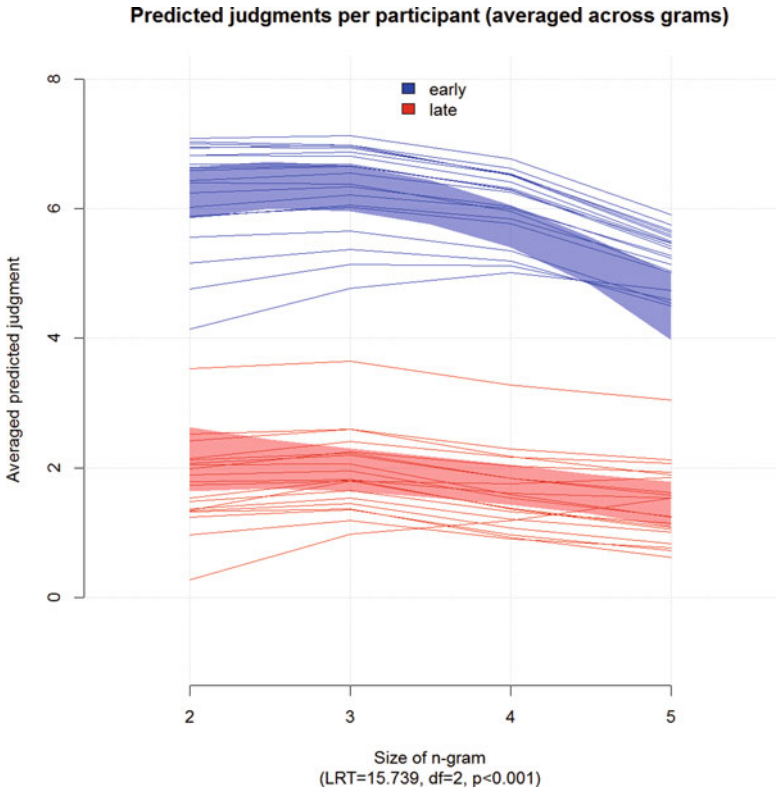


**Fig. 5.1** The interaction of poly(SIZE, 2): BINRANK

which confirms the main hypothesis formulated above. The main effect of SIZE, on the other hand, consists of the expected weak negative correlation such that the longer the MWE, the lower its ratings. This is to some extent a reflection of the fact that the longer an expression, the less likely it is to indeed be a stored unit in the subjects' mental lexicons rather than "creatively" assembled on the spot and the less likely subjects were to recognize it as an expression they would give a high rating. This finding is compatible with the frequencies of lengths of MWEs in corpora: the spoken component of the British National Corpus contains >65 K MWEs of length 2, ≈10.5 K MWEs of length 3, 675 MWEs of length 4, and 10 of length 5.

While the main effects just discussed are relatively straightforward to interpret, they also participate in a highly significant interaction. Crucially for the purposes of the present paper, the interaction is of such a nature that it does not negate (any part of) the effect of BINRANK. Instead, it reflects the fact that MWEs returned late by MERGE do not decrease much in formulaicity as they become longer: we believe that, in some sense, this is little more than a floor effect, and in general, there's a negative effect of SIZE such that longer MWEs are less formulaic than shorter ones. Since MWEs with BINRANK (*late*) are already also much less formulaic than those

with BINRANK (*early*), there is just not that much "judgment space" to decrease to, as is evidenced by the fact that the fixed-effects confidence interval for the red regression line is not only compatible with a straight and completely horizontal regression line but when SIZE = 5 is very close to the minimally possible judgment value of 1.

### 3.1.5 Interim Summary

The main finding of our first experiment is that the MERGE algorithm does indeed seem successful in identifying highly formulaic MWEs at an early stage of its application and returns less formulaic ones at a later stage (when association strengths decrease). This finding is compatible with our above hypothesis and, thus, constitutes a first piece of encouraging evidence in favor of MERGE. However, more evidence is needed to begin to make a solid case for MERGE, and we will provide more evidence in the next section. Specifically, in Sect. 3.2, we contrast the MWEs returned by MERGE with those of a competing proposal, namely, O'Donnell's AFL discussed above in Sect. 2.2.2.

## *3.2 Rating Study 2: AFL vs. MERGE*

One of the major dimensions along which algorithms vary, as discussed in Sect. 2, is how they quantify the statistical strength of MWEs in order to rank MWEs from "better" to "worse." Many approaches, such as MERGE, use lexical association measures, which take into account various pieces of frequency information relevant to a target word co-occurrence. The drawback of such measures is that they have typically been limited to two-way co-occurrences and are thus not viable for comprehensively finding longer MWEs in a corpus (such as *it goes without saying*); this is because of the facts that just about all measures are based on co-occurrence tables of the type shown in Table 5.1 and that it is not obvious how to compute the expected frequencies of more than two words (since complete conditional independence is ridiculously anticonservative, see Gries 2010:275). The collocational chain approaches in Daudaravičius and Murcinkevičienė (2004) and Gries and Mukherjee (2010) and the recursive bigram approaches of Wible et al. (2006) and MERGE are innovative in their abilities to overcome this limitation. An alternative, however, to dealing with this would be to use a measure that was not limited to two-way co-occurrences, such as simple frequency counts.

This is precisely what another algorithm from the literature, the adjusted frequency list (AFL), does (O'Donnell 2011). Under this approach, candidate MWEs are ranked based simply on how often they occur. But remember that certain word sequences may represent true MWEs yet be low frequency. Idioms are a prototypical example of such sequences. We would thus anticipate a frequency-based approach such as the AFL to fail to identify many good MWEs that follow this pattern.

Conversely, lexical association measures are designed to be able to find such low-frequency yet high-contingency sequences, so an approach like MERGE that has adapted such a measure to sequences beyond bigrams ought to be able to not only find low-frequency MWEs but ones of various sizes. In this section, we thus compare MERGE and the AFL in another rating experiment in order to test the hypothesis that an approach such as MERGE that scales lexical association up to co-occurrences greater than 2 is superior to an approach that obviates this by using frequency, which is not inherently restricted to bigrams.

A final note should be made regarding discontinuities in MWEs. Remember that MERGE is designed to be able to find them; the AFL is not. Already, then, it can be claimed that MERGE offers something beyond the AFL in that it identifies an additional format of possible MWE. The present study will therefore be limited to comparing the performances of the algorithms in their ability to find MWEs with purely adjacent words. To this end, MERGE's max gap size parameter will be set here to zero.

### 3.2.1 Materials

The same corpora used in experiment 1 were also used here, with the same preprocessing procedures. Next, the algorithms were run and the top 1000-ranked items from the output of each were selected for further consideration. In the case of MERGE, this involved simply running the algorithm for 1000 iterations. In the case of the AFL, the minimum frequency threshold was set to 5 and the 1000 items with highest frequencies were selected. We then decided to focus on the MWEs that the two algorithms did not agree on rather than the MWEs that they had in common. Thus, two groups of items were created: the first group comprised those items found in the AFL output but not in the MERGE output; the second group comprised those items found in the MERGE output but not in the AFL output; this means the two lists do not share any items (and the overlap of the lists is not relevant since we are comparing the algorithms on the basis of an external "gold standard," the subjects' ratings). This allowed a highly tractable examination of how the respective performances of the two algorithms contrasted, as stimulus items fell into one of two categories.[6] The two groups of disjunctive output contained 180 items each. An even distribution of sampling from across the range of items was

[6]Note that there would have been difficulties in comparing the performance of the algorithms on the basis of the output that they had in common (i.e., by seeing which algorithm's ranking of output best correlated with participant-assigned ratings of this output). Since the strength metrics used to rank output were different for each model, the algorithm-assigned strength values would have to have been rank-ordered to make them comparable across algorithms. But the fact that the AFL is based on integer frequency means that there are numerous ties, whereas the log likelihood decimal values used by MERGE make for virtually no ties (at least at higher scores). Thus, the rank order distributions of the two model outputs were intractably different.

**Table 5.3** Random sampling of output from AFL and MERGE

| AFL | | MERGE | |
|---|---|---|---|
| *He is* | *Well it* | *Auto reverse* | *Good afternoon* |
| *And just* | *They all* | *In the middle of* | *Melissa Soligo* |
| *But if you* | *And how* | *We need* | *They weren't* |
| *Because the* | *To their* | *To make sure* | *Must have been* |
| *And this* | *Of it* | *Square root* | *Next week* |
| *And I think* | *A real* | *I want you* | *A good idea* |
| *It the* | *Says the* | *You think* | *I wanted to* |
| *Get a* | *With that* | *Kind of thing* | *We'll see* |
| *Before the* | *There and* | *Let us* | *Thanks very much* |
| *What kind of* | *So this* | *Major depression* | *A great* |

achieved by partitioning the two rank-ordered item groups into 10 bins and then randomly sampling 18 items from each bin. These items were then used in our experimental design.

### 3.2.2 Experimental Design

On the basis of the items sampled as described above, groups of stimuli for the surveys were created, with each group containing 45 items sampled randomly without replacement from each of the two groups of 180 items above. Thus, each survey contained 90 items – 45 generated by MERGE and 45 generated by the AFL. In Table 5.3, we provide a random sampling of 20 stimuli sourced from the 180 AFL items and 20 sourced from the 180 MERGE items.

One can immediately appreciate the qualitative difference between many of the items in these two lists. While the high-frequency sequences represented in the AFL output comprise many combinations of function words, the MERGE output comprises many sequences combining function and content words. The combinations include structures such as noun phrases (*a good idea*) or compound nouns (*square root*), compound prepositions (*in the middle of*), whole utterances (*thanks very much*), or phrasal verbs (*to make sure*), among others. Furthermore, while these combinations may be lower overall in frequency, their component words are mutually contingent. This type of relationship of mutual contingency is precisely the statistical pattern that lexical association measures like log likelihood are designed to capture.

At the next stage (and as in the first study), 20 surveys were created, including 5 of each version, each to be rated by a single participant. Again, the order of presentation of stimulus items for each survey was randomized, and each stimulus item was accompanied by an utterance sourced from the corpus containing that stimulus item, so that study participants had a sense of the use of the candidate MWEs in context. Pilot testing revealed that the ratings assigned across the two stimulus groups did not differ significantly, despite the apparent qualitative

difference in the stimulus patters seen in Table 5.3. It is possible that the instructions to identify *common, reusable chunks* are to blame for this result. While they yielded successful results in the first study, the instructions did not appear effective here; this may be because they failed to tap into intended intuitions about memorization. For example, the idea of *commonness* may trigger intuitions about frequency rather than memory, and *reusability* may trigger notions about utility. To try to more explicitly target intuitions about memorization, the instructions were altered. In the new version, study participants were asked to rate sequences based on whether, in their opinion, they represented a *complete unit of vocabulary*. The hope was that participants' understanding of the notion of vocabulary would be roughly analogous to the linguistic notion of a lexicon, since these US students would have grown up learning vocabulary lists in spelling classes, etc. Again as in the first study, 20 participants were recruited from an introductory linguistics course at the University of California, Santa Barbara. Each participant was placed in a quiet room by themselves and given as much time as they needed to complete the survey.

### 3.2.3 Statistical Analysis

The data were analyzed with a linear mixed-effects model as outlined above for experiment 1. In this case study, the dependent variable was again RATING, i.e., the numerical rating provided by subjects for MWEs; the independent variable was the binary variable ORIGIN, which specified from which list of MWEs – AFL vs. MERGE – the rated MWE was from (recall that we used items that were returned by only one algorithm). As above, the random-effects structure was maximal, including varying intercepts and slopes for both subjects and MWEs.

### 3.2.4 Results

The linear mixed-effects model we fitted resulted in a significant fit (LR chi-squared = 5, df = 1, $p = 0.0254$, from a ML-comparison to a model without fixed effects) but not a particularly strong correlation: $R^2$marginal = 0.02 and $R^2$conditional = 0.37; see Table 5.4 for the corresponding results.

As is obvious from the above statistics, the overall effect is weak – the product-moment correlation between the observed ratings and the one predicted by our model is $r = 0.68$ – and the random-effects structure explains more of the variance than the fixed effects. We visualize the findings in Fig. 5.2. On the *x*-axis, we

**Table 5.4** Results for the fixed-effects part of the regression model (REML)

| Predictor | coef | se | df | *t* | $p_{1\text{-tailed}}$ |
|---|---|---|---|---|---|
| Intercept | 3.93 | 0.27 | 19.7 | 14.6 | <10–11 |
| ORIGIN: *AFL* → *MERGE* | 0.59 | 0.25 | 22.8 | 2.31 | 0.0151 |

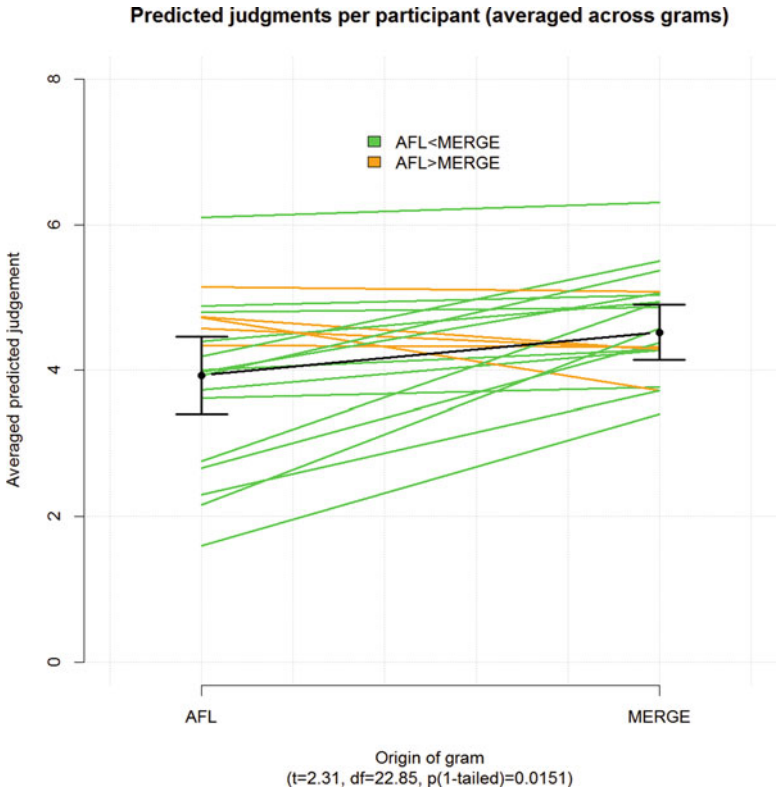


**Fig. 5.2** The main effect of ORIGIN

represent the two levels of ORIGIN, on the *y*-axis the predicted judgments by the experimental participants (averaged across MWEs). Each green and orange line represents a single participant's regression line; a green line represents a participant's predicted median ratings for MWEs from the MERGE list, which are higher than those for the AFL list; an orange line represents the opposite relation, and the black points/lines (with confidence intervals) indicate the overall effect of ORIGIN.

The main effect of ORIGIN provides support for the hypothesized usefulness of the MERGE algorithm. While the effect is not strong and variable across subjects and MWEs, there is a significant difference such that the randomly sampled MWEs from the MERGE algorithm score higher average formulaicity judgments than the randomly sampled MWEs from the AFL algorithm. Given the small effect size, the evidence is not conclusive but nonetheless compatible with our hope/expectation of MERGE outperforming the AFL approach. In the next section, we will present our conclusions.

## 4 Discussion and Conclusion

In this paper, we presented a new recursive algorithm to identify MWEs in corpora, which we called MERGE. We motivated its application and characteristics and, more importantly, attempted to validate it in two experimental ways. In a first experiment, we demonstrated that MWEs returned by MERGE early, as predicted by MERGE's design, indeed score higher in formulaicity than MWEs returned by MERGE late, a robust main effect that is largely unqualified by an interaction with the size of an MWE. In a second experiment, we demonstrated that MWEs returned by MERGE score higher in formulaicity than MWEs returned by the AFL algorithm. While both case studies are small and can only begin to set the stage for the large and comprehensive set of tests that will ultimately be necessary for any new corpus-based algorithm, we interpret these first two significant results as good initial support for MERGE.

In terms of methodological implications, MERGE's performance provides further evidence for the effectiveness of lexical association measures in identifying meaningful word co-occurrences, especially compared to the use of raw frequency counts, as in the AFL. While the AFL found many high-frequency, low-contingency strings which do not obviously represent stored, meaningful units, MERGE was much more effective in its ability to single out salient sequences (i.e., sequences that occur more often than may be expected based on their individual word frequencies), a hallmark of lexical association measures. Furthermore, MERGE's performance exemplifies one effective way of scaling up lexical association measures to co-occurrences beyond the bigram. While the current study speaks to the good performance of the log likelihood association measure in this implementation, further work is needed to determine whether other association measures, such as the widely used *MI*-score, or newer measures such as *LG* (which includes type frequencies) or $\Delta P$ (which is directional, see Gries 2013), likewise yield good results when implemented in MERGE.

The MERGE algorithm offers a relatively simple approach that harnesses the proven potency of lexical association measures, and adapts them to MWEs of various sizes, with and without gaps. But MWEs are not merely crystallized sequences of words – the "slots" within them, or at their edges, may allow some (limited) set of words (i.e., a part-of-speech category) to fill them. In the future, it would be desirable if MERGE could be adapted to not only learn where the gaps were, but also what word paradigms might fill them; specifically, what the set of types is as well as their frequency distribution and maybe entropy. Furthermore, since members of the same paradigm may comprise different numbers of words, it would also be desirable if MERGE could be adapted to recognize identical word sequences containing gaps of different sizes as instantiating the same MWE (e.g., *as _ as* in *as funny as* versus *as __ as* in *as truly hilarious as*).

Conventionalized, memorized, multi-word sequences represent an important component in modern language sciences research, both at the level of cognitive and grammatical theory as well as in the applied domain of computer technologies. Being able to identify them automatically, using the explosion of corpus resources

that are ever more available, is an increasingly important goal for researchers in various disciplines. The MWEs extracted by MERGE, which exhibit strong similarities to humanlike knowledge of formulaic language, indicate that this algorithm is a powerful tool for such work.

## References


Bannard, C., & Matthews, D. (2008). Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science, 19*, 241–248.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Barton, K. (2015). MuMin: Multi-model inference. *R package version, 1*(13), 4 http://cran.r-project.org/web/packages/MuMIn/index.html.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science, 33*(5), 752–793.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*(1), 45–61.

Constant, M., Eryigit, G., Monti, J., van der Plas, L., & Ramisch, C. (2017). Michael, and Amalia Todirascu. *Multiword Expression Processing: A Survey. Computational Linguistics., 43*(4), 837–892.

Da, S., Joaquin, F., Dias, G., Guilloré, S., & Pereira Lopes, J. G. (1999). Using LocalMax algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, 849–849.

Daudaravičius, V., & Murcinkevičiene, R. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics, 9*(2), 321–348.

Du Bois, J. W., & Englebretson, R. (2004). *Santa Barbara corpus of spoken American English, part 3*. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W., Chafe, W. L., Meyers, C., Thompson, S. A., & Martey, N. (2003). *Santa Barbara corpus of spoken American English, part 2*. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W., & Englebretson, R. (2005). *Santa Barbara corpus of spoken American English, part 4*. Philadelphia: Linguistic Data Consortium.

Du Bois, J. W., Chafe, W. L., Meyers, C., & Thompson, S. A. (2000). *Santa Barbara corpus of spoken American English, part 1*. Philadelphia: Linguistic Data Consortium.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text, 20*(1), 29–62.

Evert, S. (2005). The statistics of word co-occurrences: Word pairs and collocations. *Ph. D. Dissertation. Universität Stuttgart.*

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 2, pp. 1212–1248). Berlin & New York: Mouton de Gruyter.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 75–93). Harlow: Longman.

Green, S., de Marneffe, M.-C., Bauer, J., & Manning, C. D. (2013). Parsing models for identifying multiword expressions. *Computational Linguistics, 39*(1), 195–227.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–25). Amsterdam: John Benjamins.

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez & M. Almela (Eds.), *A mosaic of corpus linguistics: Selected approaches* (pp. 269–291). Peter Lang: Frankfurt am Main.

Gries, S. T. (2012). Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language, 36*(3), 477–510.

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next .... *International Journal of Corpus Linguistics, 18*(1), 137–165.

Gries, S. T., & Mukherjee, J. (2010). Lexical gravity across varieties of English: An ICE-based study of *n*-grams in Asian Englishes. *International Journal of Corpus Linguistics, 15*(4), 520–548.

Ikehara, S., Shirai, S., & Uchino, H. (1996). A statistical method for extracting uninterrupted and interrupted collocations from very large corpora. *Proceedings of the 16e Conference on Computational linguistics, 1*, 574–579.

Johnson, P. C. D. (2014). Extension of Nakagawa and Schielzeth's R2GLMM to random slopes models. *Methods in Ecology and Evolution, 5*(9), 944–946.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B.. (2016). lmerTest: Tests in linear mixed effects models. R package version 2.0–30. https://CRAN.R-project.org/package=lmerTest

Lareau, F., Dras, M., Börschinger, B., & Dale, R. (2011). Collocations in multilingual natural language generation: Lexical functions meet lexical functional grammar. In *Proceedings of ALTA'11* (pp. 95–104).

McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present. Abington*. New York: Routledge.

Nagao, M., & Mori, S. (1994). A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of Japanese. Proceedings of the. In *15th conference on computational linguistics* (pp. 611–615).

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews, 85*(4), 935–956.

Newman, J., & Columbus, G. (2010). *The international Corpus of English – Canada*. Edmonton, Alberta: University of Alberta.

O'Donnell, M. B. (2011). The adjusted frequency list: A method to produce cluster-sensitive frequency lists. *ICAME Journal, 35*, 135–169.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191–225). London: Longman.

Pecina, P. (2009). *Lexical association measures: Collocation extraction*. Prague: Charles University.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Sag, I. A., Baldwin, T., bond, F., Copestake, A., & Flickinger, D. (2002). *Multiword expressions: A pain in the neck for NLP. Proceedings of the third international conference on intelligent text processing and computational linguistics* (pp. 1–15). Mexico City.

Simpson-Vlach, R., & Ellis, N. (2010). An academic formulas list. *Applied Linguistics, 31*(4), 487–512.

Sinclair, J. (1987). *Collins COBUILD English language dictionary*. Ann Arbor: Collins.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*(2), 251–272.

Wahl, A. (2015). Intonation unit boundaries and the storage of bigrams: Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics, 13*(1), 191–219.

Wible, D., Kuo, C.-H., Chen, M.-C., Tsao, N.-L., & Hung, T.-F. (2006). *A computational approach to the discovery and representation of lexical chunks. Paper presented at TALN 2006*. Leuven.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.