Lexis
Journal in English Lexicology

19 | 2022
Phraseology and Paremiology in English

Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach

Stefan Th. Gries

Electronic version
URL: https://journals.openedition.org/lexis/6231
ISSN: 1951-6215

Publisher
Université Jean Moulin - Lyon 3

Electronic reference
Stefan Th. Gries, “Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach”, *Lexis* [Online], 19 | 2022, Online since 26 March 2022, connection on 26 March 2022. URL: http://journals.openedition.org/lexis/6231

Multi-word units (and tokenization more generally): a multi-dimensional and largely information-theoretic approach

Stefan Th. Gries

> Computational linguists have, to date, been unable to develop algorithms that reliably identify onomasiological units in English (compounds, verb-particle combinations, or idioms) written with intervening space characters. (Baayen, Milin & Ramscar [2016: 1176])

1. Introduction

1.1. General introduction

1 (Extended) Lexical units, or elements of the mental lexicon or constructicon, have long been of interest to a variety of fields including, but not limited to, various areas in applied linguistics (e.g. language teaching, register studies, and others), psycholinguistics, cognitive linguistics, and others. If one explores such work, one will quickly realize that many studies restrict their analyses to units (or *n*-grams or lexical bundles) with pre-specified lengths; for example, for reasons that, to me, seem entirely unrelated to any cognitive or linguistic motivation, many studies chose to study 4-grams, see, e.g., Cortes [2004] or Breeze [2013]. This is in spite of the seemingly obvious fact that such units can come in various sizes and levels of complexity, as is shown by even just cursory introspection:

- 2-element units such as *according to* or *because of*;
- 3-element units such as *in spite of* or *is that so?*;

- 4-element units such as *on the other hand* or *you gotta be kiddin'!*;
- 5-element units such as *be that as it may* or *as a matter of fact*;
- 6-element units such as *the fact of the matter is*, etc.

2 Thus, ultimately one would want to be able to 'let the data decide' on what the *n* in *n*-grams is. We need cut-off points that are (more) theoretically motivated and/or data-driven rather than a single arbitrary cut-off point forced on a corpus, and this question is what the present paper is all about: How do we identify MWUs that look like they consist of multiple words (in their orthography in corpora) and that might have varying lengths, but that, in speakers' uses, are actually not assembled on the fly but most likely rather stored as single units?

3 Defining such units even in theory is not uncontroversial – not even what to call them is (Wray [2002] alone identifies 60 different terms for formulaic, conventionalized word sequences) – because by their very nature such units can straddle the boundary of lexis, phraseology, and syntax. Many criteria have been used in the past to define these *n*-grams or multi-word units (MWUs) or phraseologisms, which include (as per Gries's [2008a] discussion of phraseology):

- the **nature of the elements** in a unit: just words or also other units (as in colligations);
- the **number of the elements** in a unit: see above;
- the **minimally required frequency** (or other) threshold;
- the **permissible distance** between elements of a unit;
- the degree of **lexical/syntactic flexibility** of the elements involved in the unit;
- semantic **(non-)compositionality**/(non)-predictability.

4 These kinds of MWUs are important for a variety of applications that span a range of research areas. With regard to speech production, Bell *et al.* [2003] discuss how words are shorter to produce when they are part of a more frequent 2-gram or 3-gram; with regard to speech comprehension, Underwood *et al.* [2004] show that subjects need fewer eye fixations to read formulaic sequences that are up to six words long; with regard to language acquisition, Bannard & Matthews [2008] show that children as young as two and three years old are faster and more accurate at repeating high-frequency phrases compared to lower-frequency phrases even when part frequency is controlled for; with regard to register studies, lexical *n*-grams have been shown to be useful for multidimensional register classification (Crossley & Louwerse [2007]) or the study of academic English (Biber, Conrad & Cortes [2004] or Simpson-Vlach & Ellis [2010]); in lexicography, MWUs are crucial for creating multi-word dictionary entries (Sinclair [1987]), etc.

5 Given the wide range of applications, the literature on the topic is vast and cannot possibly be done justice here. I will restrict my overview to several corpus-linguistic studies that have interesting properties that set the stage for the algorithm to be introduced further below; other interesting data-driven and more computational approaches include Wermter & Hahn [2004, 2005], Bestgen [2018], Nelson [2018], or Jeaco [2019]. One approach was proposed by Daudaravičius & Murcinkevičienė [2004] and this approach was extremely interesting for two reasons. First, they may well have been the first to develop an AM (called lexical gravity *G*) that does not just include the observed token frequencies in their approach – i.e. how often an element *E* (e.g., *according*) is observed with and without another element *F* (e.g., *to*) in the corpus – but also type frequencies, i.e. the number of different collocates of *E* (of which *F* is one). More specifically and for the above example, they include the co-occurrence frequency

of *E* and *F*, the total frequencies of each *E* and F, and the number of different collocate types of each E and F. Second, they extend the use of G to the identification of MWUs, which are defined as chains of words whose connecting G-values never fall below 5.5 (a somewhat arbitrary threshold, but not more arbitrary than, say, using *MI*-scores of 3 as a threshold).

6 I think Daudaravičius & Murcinkevičienė [2004] were very much ahead of their time and their approach has been underexplored (e.g., to validate it) and underutilized. It is not without potential shortcomings, though. One is that they themselves qualify their measure by stating that "reliable results can be obtained only for the words which form word pairs with a common sum of frequencies higher than 10 in the corpus". Another is that some of the results they present seem less than ideal. For instance, their most strongly attracted collocations are *the of*, *of the*, *in the*, *to the*, *a of*, etc. – while it is possible that these would be psycholinguistically relevant multi-word units (in the sense that speakers access them as one unit), they are still of a kind that most corpus linguists would not be interested in that much (what other linguistic or other application would want these ranked highest?) and they seem to be awfully stopword-heavy and frequency-determined. Similarly, the collocational chains they discuss in their Figure 5 for the sentence *He will work for a new free trade area embracing North America and Europe, an idea President Clinton is interested in* include some that seem too long and specific to be of any more general interest; these are the chains they stipulate:

- chain 1: He will work for a new
- chain 2: free trade area
- chain 3: North America and Europe, an idea
- chain 4: President Clinton is interested in

7 Arguably (and admittedly off the top of my head), I would expect as good chains maybe something like this instead (with the missing words not being part of chains):

- chain 1: will work for (a?)
- chain 2: (new?) free trade area
- chain 3: North America (and Europe?)
- chain 4: President Clinton
- chain 5: is interested in

8 However, this comment is of course a bit subjective and Gries [2010] found that, if one applies hierarchical cluster analysis on the basis of the *G*-values for each bigram type to the 4 registers and 19 subregisters of the BNC, the resulting clusters recognize the registers perfectly (and better than the more widely-used t-score).

9 One of the few studies taking up the idea of collocational chains based on *G* that I am aware of is Gries & Mukherjee [2010]. They extend the above ideas and extract *n*-grams from a variety of corpora covering different varieties of Asian Englishes. However, they modify Daudaravičius & Murcinkevičienė's approach to identifying chains: rather than just using 5.5 as a threshold value of *G* for joining 2-grams, they use an iterative process and compute the mean *G*-score for each *n*-gram and then, "for each *n*-gram *N* of length *l* and (mean) $G \geq 5.5$ [they] then tested whether there is another *n*-gram that contains the first *n*-gram *N*; has a length $l+1$; [and] has a higher $mean_G$-value" (Gries & Mukherjee [2010: 534]), which yields a variety of *n*-grams with different lengths (as merited by the average *G*-scores). They find that the *n*-grams arrived at in this way reliably separate speech from writing (but do not replicate the hypothesized evolutionary differences between the three Asian varieties and British English).[1]

10 Another interesting approach is that of O'Donnell's [2011] adjusted frequency list (AFL), which is conceptually similar to Jelinek [1990] or Kita *et al.* [1994]. The AFL approach, first, identifies all *n*-grams up to a certain length and with a user-defined minimum frequency in the corpus to which it is applied (in O'Donnell [2011], that frequency threshold was 3). Then, for each *n*-gram type, the two component *n*-minus-1-grams are derived. Lastly, the number of tokens in the frequency list of each *n*-minus-1-gram is decreased by the number of *n*-grams in which it is a component. This 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. This AFL approach is interesting because of its stepwise nature and, with the exception of a minimally required frequency, it requires little user intervention/ tweaking; in other words, it has very few researcher degrees of freedom. On the other hand, a notable downside of the AFL is that it only uses frequency information but does not even consider the degree of association of the *n*-grams it works with, a rather surprising choice given many decades of corpus-linguistic research on AMs and the identification of MWUs.

11 Wible *et al.* [2006] is another interesting study with a slightly different focus. Their method does not generate a (ranked) list of all MWUs contained in a corpus – rather, it aims at recursively finding all of the MWUs involving a given node word (much like a concordancing tool would). Once a node word has been defined, the algorithm generates continuous and discontinuous 2-grams within a specified window size around each token of the node word in the corpus and computes an *MI*-score for them. Then, all the 2-grams whose *MI*-score exceeds a specified threshold are "merged" into a single representation. The algorithm then considers new (continuous and discontinuous bigrams) containing the newly merged MWU and one other word, scores them for association, and merges the highest-scoring ones. This progress iterates until no new *n*-grams containing the node word and exceeding the threshold are found.

12 Finally, there is the MERGE algorithm (*M*ulti-word *E*xpressions from the *R*ecursive *G*rouping of *E*lements) by Wahl & Gries [2018, 2020]. The approach combines several aspects from the previous methods. The first step of the algorithm is to extract and count all 1- and 2-grams from a corpus and compute the corpus size. In the second step, for each 2-gram, one computes a measure of association strength, which in their study is the log-likelihood score G^2 (no relation to gravity *G*!). G^2 is computed from the usual 2×2 co-occurrence table shown in Table 1.[2]

Table 1. Schematic co-occurrence table for two linguistic elements *E* and *F*

	F	**other**	**Sum**
E	*a*	*b*	*a+b*
other	*c*	*d*	*c+d*
Sum	*a+c*	*b+d*	*a+b+c+d*

13 Specifically, the formula for G^2 is shown in (1); note that G^2 combines two dimensions of information: frequency and association.

$$(1)\ G^2 = 2 \cdot \sum_{a}^{d} observed \cdot ln \frac{observed}{expected}$$

14 As the third step, one finds the *n*-gram with the strongest attraction – i.e. highest G^2 for when observed *a* > expected *a*, merges this *n*-gram into a single new MWU, updates all the frequencies and the corpus size, and iterates. This process is repeated for a certain user-defined number of iterations (e.g., 10^3 or 10^4 or ...) or until a certain user-defined G^2-threshold is not exceeded anymore by any *n*-gram. Thus, MERGE is iterative/recursive like several of the methods discussed above, it involves frequency (like AFL) but also association (like Daudaravičius & Murcinkevičienė or Wible *et al.*) by virtue of using an AM reacting to both, and it returns all *n*-grams (like the AFL) rather than all for a node word (like Wible *et al.*).

15 To validate their approach and compare it to the closest conceptual predecessor, the AFL, Wahl & Gries [2018, 2020] discuss four case studies. The first validation consisted of having subjects rate MWUs that MERGE returned as good or bad MWUs for their being a "common reusable chunk"; a linear mixed-effects model modeling the rating as a function of the supposed MWU quality and their length (as a control) shows that the MWUs rated as good by MERGE scored significantly higher ratings than the MWUs rated as bad (R^2_c=0.64; p<0.001). The second validation consisted of having subjects rate MWUs again, but this time MWUs that were rated as good by either the MERGE algorithm or the AFL; a linear mixed-effects model modeling the rating as a function of which algorithm "recommended" the MWU and their length (as a control) shows that the MWUs rated as good by MERGE scored minimally, but significantly higher ratings than the MWUs rated as good by the AFL(R^2_c=0.03; p<0.003). The third study compared MERGE and the AFL with regard to which of them was better at identifying the MWUs that are annotated as such in the BNC; a 1-tailed exact binomial test shows that MERGE did better than the AFL ($p_{\text{MERGE>AFL}}$=0.015 and $p_{\text{AFL<MERGE}}$=0.018). Finally, the fourth study showed a high correlation (R^2>0.78) between the percentage of MWUs children learned over a period of time and the G^2-values MERGE returned for MWUs.

1.2. Motivation and overview of the present paper

16 While previous studies on MWU identification – those discussed or just mentioned above, but also scores of other studies – have yielded interesting results, there are multiple ways in which many of them could probably be improved, most of which are actually very general suggestions that would benefit most corpus-linguistic work on AMs (see Gries [2019]). First and as mentioned above, a shortcoming of specifically the AFL approach or Kita's cost criterion is that they **only use (token) frequency information and not association**, a really very surprising choice, given that it amounts to ignoring decades of work on association and collocations/phraseologisms.

17 Second, the studies that do include association usually involve **bidirectional AMs** such as *MI*, *t*, or G^2,[3] meaning they cannot distinguish *n*-grams where

- the first element attracts the second but not vice versa (such as *according to* or *upside down*);
- the second element attracts the first but not vice versa (such as *of course* or *for instance*);
- both elements really attract each other (such as *Sinn Fein* or *bona fide*).

18 Third and relatedly, many AMs – in particular *t* and *G*^{2} – **conflate token frequency and association**. That *can* be good, but it certainly loses a lot of information: a certain *t*- or *G*^{2}- or $p_{\text{Fisher-Yates}}$-value does not indicate whether it reflects high frequency and mediocre association or low to intermediate frequency but high association. For example, Gries (2022) shows that a fairly decent *G*^{2}-score of 81.66 can represent the former (a high frequency of 1965 but a mediocre odds ratio of 1.6) as well as the latter (a low frequency of 26 but a very high odds ratio of 201.5)!

19 Fourth, most approaches do not take the **numbers of co-occurring types** into consideration, and even lexical gravity *G* does not take into consideration the **distribution of the co-occurring types**. By that I mean, computing *G* would take into consideration the fact that, for instance, in some corpus, four different word types occur after word *w*, but it would not distinguish between (i) a scenario where the four types are all similarly likely after *w* (with, say, relative frequencies of 0.27, 0.26, 0.25, and 0.22) and (ii) a scenario where one of the four types accounts for nearly all uses of *w* (with, say, relative frequencies of 0.85, 0.06, 0.05, and 0.04).

20 Fifth, as far as I can tell, very little work takes the **dispersion of the candidate MWUs** across the corpus into consideration (but see Nelson 2018 for an interesting exception) although Stefanowitsch & Gries [2003] and Gries [2008b, 2019, 2020] have shown that underdispersion can undermine *all* frequencies and AMs from corpus data.

21 This paper is a first attempt to (i) describe an algorithm (called MERGE$_{\text{multidim}}$) that is based on the general workflow of MERGE by Wahl & Gries [2018, 2020] but designed to improve it with regard to every single one of the above issues. We want our improved approach

- to not just be based on frequency but, minimally, also on association;
- to be able to identify MWUs regardless of their direction of association, but also be able to utilize one or both direction(s) of association if required/desired;
- to consider whatever dimensions of information are used separately, i.e. avoid the kind of uncontrollable statistical conflation that characterizes many measures (like *G*^{2} or *t*) (at least initially; for lack of space, this first application here will also have to entertain a certain kind of heuristic strategy);
- to consider the numbers of word types before and after a potential source word unit (SWU) and their distributions where an SWU could be a single word (such as *in*) or something that has already been merged with something else (such as *in spite*);
- to consider how evenly an MWU candidate is dispersed throughout a corpus.

22 In other words, in this exploration I will consider an ideal MWU an expression that is frequent and evenly dispersed in a corpus and where at least one part attracts the other strongly (but where ideally both parts attract each other and/or are highly predictive of each other). In what follows, I discuss the algorithm and the dimensions of information that it considers and how (Section 2); then I turn to the results of an initial case study of a small, but well-known and straightforwardly manageable corpus (the Brown corpus of written American English from the 1960s) (Section 3), before I conclude and offer some suggestions for future exploration (Section 4).

2. Methods

23 To introduce and then evaluate the way MERGE$_{\text{multidim}}$ works, it is necessary to first discuss the dimensions of information that it considers in order to improve on existing suggestions/procedures. This in turn is best done on the basis of a tiny example corpus, for which we imagine we want to retrieve potential MWUs from it; this 'corpus' is represented in Figure 1 with some highlighting for letters/words/combinations that will feature prominently in the exposition below:

- the letter **b** is highlighted in bold type;
- the sequence c b is underlined;
- the sequence *b d* is italicized:

Figure 1: A fictitious corpus of three parts (with letters representing words)

Part/file	**Content**
1	a d c **b** e **b** f g h c **b** i j k a y z **b** n o a c c **b** p q r q a x r z n a
2	y i **b** c p x e j d g n k q r **b** x x c ***b*** *d* y z f o p q ***b*** *d* j e z ***b*** *d*
3	g g i o r j j **b** c d g j k r e j g f h k h f d h k o a c **b** r d g k **b**

24 The overall design of MERGE$_{\text{multidim}}$ is simple. Like MERGE, it is an iterative algorithm that

- starts from an 'untreated' corpus of SWUs (i.e. at the beginning each letter in Figure 1: is a SWU and a candidate for merging with something around it);
- generates all possible candidate MWUs by merging adjacent units (i.e., here, *a d*, *d c*, *c b*, *b e*, *e b*, ...);[4]
- computes "some information/score" for each of these candidate MWUs;
- picks the candidate MWU that scores the highest in step iii and merges its constituent SWUs into a new MWU;
- updates the corpus to reflect that merger (such that the newly-formed MWU is now also available as an SWU for future mergers (into longer MWUs)) and iterates.

25 Where MERGE$_{\text{multidim}}$ differs considerably from MERGE and all other approaches is the third step, which is what aims to address all shortcomings listed above. In what follows, I (i) outline the dimensions of information MERGE$_{\text{multidim}}$ should consider, (ii) clarify how they are operationalized (i.e. computed from the corpus data), and (iii) show how they contribute to a quantification of each candidate MWUs likelihood of being merged.

2.1. Dimensions of information

2.1.1. Token frequency

26 Dimension 1 is the **token frequency** of a candidate MWU in the corpus. This dimension is oriented such that, ignoring, or holding constant, all other dimensions (!), higher frequency of a candidate MWU increases the probability of something being merged into a MWU. This is simply because, again disregarding all other dimensions, we are

(likely) more interested in high(er)-frequency expressions (rather than hapaxes): low-frequency combinations of words – of which there will be extremely many, given the Zipfian distribution of word frequencies – are also extremely unlikely to be MWUs. In many even just moderately sized corpora, hapaxes account for half of all word types, but only a tiny number of them are likely to be part of a useful MWU. In its current implementation (in R, with some computations outsourced into C++ functions), the algorithm permits the user to define a minimum frequency threshold that a candidate MWU needs to exceed to be eligible for merging. The operationalization of this dimension is straightforward: I will collect the logged frequencies of each candidate MWU; in the above toy corpus, the candidate MWU *c b*, for instance, has a token frequency of 5 while the candidate MWU *b d* has a token frequency of 3.

2.1.2. Dispersion

27 Dimension 2 is the **dispersion** of a candidate MWU in the corpus. This dimension is oriented such that, disregarding all other dimensions, a more even dispersion of a candidate MWU in a corpus increases the probability of something being returned as a MWU, which reflects the fact that we are less interested in getting something returned as a MWU if it is only ever attested in one (potentially very small) part of a corpus. This should promote prototypical MWUs such as *according to*, which are going to be relatively widespread, or widely-used proper names (e.g., *New York* or *Los Angeles*), and this should demote, for instance, proper names that are not widespread (e.g., names of characters that occur in one particular novel but nowhere else but would still be returned by measures such as *MI*).

28 Dispersion will be operationalized here using a normalized version of the KL-divergence. The KL-divergence is a unidirectional measure quantifying how much one probability distribution *P* (here, how much in percent of a word's total occurrences is in each corpus part?) diverges from another probability distribution *Q* (here, what are the corpus part sizes in percent?); see Gries [2020] for the maybe first use of this measure for dispersion. For a corpus with *n* parts (here, *n*=3), the KL-divergence is computed as shown in (2); the normalization, which forces the values into the interval [0, 1] is shown in (3):

$$(2)\ KLD = \sum_{i=1}^{n} P_i \times log_2 \frac{P_i}{Q_i}$$

$$(2)\ KLD = \sum_{i=1}^{n} P_i \times log_2 \frac{P_i}{Q_i}$$

$$(3)\ KLD_{norm} = 1 - e^{-KLD}$$

$$(3)\ KLD_{norm} = 1 - e^{-KLD}$$

29 For instance, the candidate MWU *c b* is distributed across the corpus as follows:

- $^3/_5$=0.6 of its instances are in corpus part 1, which makes up $^{34}/_{101}$=0.3366 of the corpus;
- $^1/_5$=0.2 of its instances are in corpus part 2, which makes up $^{33}/_{101}$=0.3267 of the corpus;
- $^1/_5$=0.2 of its instances are in corpus part 3, which makes up $^{34}/_{101}$=0.3366 of the corpus.

30 Thus, we can compute the KLD as in equation (4):

$$(4)\ KLD = 0.6 \times log_2 \frac{0.6}{0.3366} + 0.2 \times log_2 \frac{0.2}{0.3267} + 0.2 \times log_2 \frac{0.2}{0.3366} = 0.2084118$$

$$(4)\ KLD = 0.6 \times log_2 \frac{0.6}{0.3366} + 0.2 \times log_2 \frac{0.2}{0.3267} + 0.2 \times log_2 \frac{0.2}{0.3366} = 0.2084118$$

and then normalize as in equation (5):

$$(5)\ KLD_{norm} = 1 - e^{-0.2084118} = 0.1881274$$

$$(5)\ KLD_{norm} = 1 - e^{-0.2084118} = 0.1881274$$

31 The higher that number (within the [0, 1] interval), the more clumpily distributed a candidate MWU is. In addition, the user can again define a certain minimum range threshold that a candidate MWU needs to exceed to be eligible for merging.

2.1.3. Type frequencies

32 Dimensions 3 and 4 are (i) the logged **type frequency** after each SWU and (ii) the logged type frequency before each SWU. This dimension is oriented such that, holding all other dimensions constant, a higher type frequency after a first SWU or before a second SWU decreases the probability of something being an MWU because one might assume that, if a word is attested with fewer collocate types, the connections to those attested ones is stronger than if a word can be attested with pretty much anything; cf. *hermetically* vs. *the*.[5] These two dimensions are straightforwardly also operationalizable with counts:

• for each first part of a candidate MWU, we count the number of different word types occurring after it:	
	• for the candidate MWU *c b*, the number of different word types after *c* is 4 (*b*, *c*, *d*, and *p*);
	• for the candidate MWU *b d*, the number of different word types after *b* is 9 (*c*, *d*, *e*, *f*, *i*, *n*, *p*, *r*, and *x*);
• for each second part of a candidate MWU, we count the number of different word types occurring before it:	
	• for *c b*, the number of different word types before *b* is 8 (*c*, *e*, *i*, *j*, *k*, *q*, *r*, and *z*);
	• for *b d*, the number of different word types before *d* is 6 (*a*, *b*, *c*, *f*, *j*, and *r*);

2.1.4. Normalized entropy

33 Dimensions 5 and 6 complement the type frequencies from the previous section with the **(normalized) entropy** (i) after each SWU and (ii) before each SWU. This is useful because the type frequencies only state the number of different words after/before some other word, but not also how much uncertainty that distribution contains. For example, we just saw that the number of different word types after *c* is 4 (*b*, *c*, *d*, and *p*), but that does not also utilize the frequencies of these four types after *c* (5, 1, 1, and 1 respectively). In other words, type frequencies cannot see that *c* is highly predictive of what follows it because *b* is five times more likely after *c* than each of the other three words.

34 The normalized entropy H_{norm} of a vector of probabilities (such as $^5/_8$, $^1/_8$, $^1/_8$, $^1/_8$) is computed as in Gries [2021: Section 3.1.1.2] and shown here in (6); for the four types after *c* this returns a value of 0.7743975.

$$(6)\ H_{norm} = -\frac{\sum_{i=1}^{n} p_i \times log_2 p_i}{log_2 n}\ (H_{norm}\ falls\ into\ the\ interval\ [0,1])$$

35 However, the entropy in the slot after/before each SWU *per se* is maybe not all that relevant – what seems more relevant is the effect that the current type has on the entropy of the slot. What does that mean? Consider *c b* again. We know the type frequency of SWUs after *c* is 4 (*b*, *c*, *d*, *p*) and that these letters/words occur with frequencies of 5, 1, 1, 1 respectively, for a H_{norm} of 0.7743975. That means, if it wasn't for the potential MWU *c b*, the frequencies of letters after *c* would be 1 (for *c* again), 1 (for *d*), and 1 (for *p*), with a maximal H_{norm}/uncertainty of 1. Put differently, without *c b*, *c* is not predictive of what comes next at all, and we can express that as the difference of H_{norm} after *c* with type *b* removed (1) minus H_{norm} after *c* (0.7743975), which yields 0.2256025.

36 Before a reader thinks this is an artificial example, consider *according to* in the Brown corpus. There are 5 different words after *according*, 4 of which occur once and one of which (*to*) occurs 136 times. That means the entropy after *according* in general is quite low because it predicts *to* so well: H_{norm}=0.1052225. But if the observed collocate of *to* is left aside, *according* is not predictive of what's next anymore at all because every other type occurs equally often after it, H_{norm}=1. Thus, what we compute here is the entropy difference for a slot: how much does a word in a slot after/before a SWU reduce the entropy after/before that SWU? The fact that this value – 1-0.1052225 =0.8947775 – is so high for *according*→*to* indicates that this might be a good MWU. At the same time, the reverse does not hold: *to* is not that predictive of *according* because there are so many other words preceding *to* and *according* does not reduce the entropy before *to* much.

2.1.5. Association

37 Finally, dimensions 7 and 8 involve **association**: (i) the degree to which the first SWU of the candidate MWU attracts the second one and, separately, (ii) the degree to which the second SWU of the candidate MWU attracts the first; of course, higher association should promote MWU status whereas low(er) association or repulsion should demote MWU status. Each of these two dimensions is also operationalized with the *KLD*. Consider Table 2, which shows the frequencies of *c* and *b* and their combination in the corpus in the usual 2×2 table format.

Table 2. Observed co-occurrence frequencies for *c b* (with relevant row % in parentheses)

	b	**other**	**Sum**
c	*5* ($^{5}/_{8}$=0.625)	*3* ($^{3}/_{8}$=0.375)	*8*
other	*8*	*85*	*93*
Sum	*13* ($^{13}/_{101}$=0.1287)	*88* ($^{88}/_{101}$=0.8713)	*101*

38 The degree to which *c* attracts a following *b* is computed as the normalized *KLD* of the probabilities of *b* when *c* is present ($^5/_8$ vs. $^3/_8$) from the probabilities of *b* in general ($^{13}/_{101}$ vs. $^{88}/_{101}$):

$$(7)\ KLD = 0.625 \times log_2 \frac{0.625}{0.1287} + 0.375 \times log_2 \frac{0.375}{0.8713} = 0.9687$$

$$(7)\ KLD = 0.625 \times log_2 \frac{0.625}{0.1287} + 0.375 \times log_2 \frac{0.375}{0.8713} = 0.9687$$

$$(8)\ KLD_{norm} = 1 - e^{-0.9687} = 0.6204$$

$$(8)\ KLD_{norm} = 1 - e^{-0.9687} = 0.6204$$

39 By analogy, the degree to which *b* attracts *c* before it is computed as the normalized KL-divergence of the probabilities of *c* when *b* is present ($^5/_{13}$ vs. $^8/_{13}$) from the probabilities of *c* in general ($^8/_{101}$ vs. $^{93}/_{101}$), yielding 0.4049.

2.2. Picking a candidate MWU and an example

40 We have now defined 8 dimensions that are computed for every single candidate MWU (as per steps i to iii above), as represented here in Table 3 for the two candidate MWUs referred to above.

Table 3. Observed dimension values for *b d* and *c b*

	unlogged token freq of cand.	**1-disp. of cand. MWU**	**unlogged type freq slot 1**	**unlogged type freq in slot 2**	**entropy difference in slot 1**	**entropy difference in slot 2**	**association SWU_1 → SWU_2**	**association SWU_2 → SWU_1**
cand.: *b d*	3	0.199126	6	9	0.069372	0.029215	0.14481	0.243436
cand.: *c b*	5	0.8118726	8	4	0.095055	0.225603	0.62043	0.4049
range / interval	[1, 5]	[0.199126,	[2, 8]	[2, 9]	[-0.918296,	[-0.918296,	[0.001248,	[0.00063,
		0.8118726]			0.1677511]	0.2256025]	0.889927]	0.83905]

41 However, how does one use this for deciding whether, in this small example space, the MWU to be recognized next is *b d* or *c b*? Put differently, how do we do step iv from above, picking a candidate MWU (especially when this will need to be done for millions of candidate MWUs in a real corpus)? One solution would be to define an *n*-dimensional space – one for each dimension considered – or hypercube where (i) by default at least, each dimension of the hypercube is of the same length (namely 1) and (ii) each dimension is oriented such that high values on it increase a candidate's probability to be merged into an MWU so that (iii) we can locate each candidate MWU's position in

this space and use its distance from the origin of this hypercube as its “MWU-ness” index. For this, some of the dimensions collected above need to be prepared:

- for dimension$_1$ (token frequency), we theoretically have positive integers and actually values from 1 to 5 so, to make this work, the values will be logged to the base of 10 and [0, 1]-transformed/min-max scaled, which can be easily done with an R function call such as the following: x.transformed <- (y<-x-min(x))/max(y);
- for dimension$_2$ (dispersion), we already have a measure that theoretically ranges from 0 to 1 but it has the wrong orientation – high values meaning uneven distribution, which is expected to be correlated with low MWU status – so we use 1-dispersion as our value in the hypercube;
- for dimension$_3$ and dimension$_4$ (type frequencies), we theoretically have positive integers and actually values from 2 to 8 and 2 to 9 so these values will be logged to the base of 10, [0, 1]-transformed, and subtracted from 1;
- for dimension$_5$ and dimension$_6$ (entropy differences), we have a measure that theoretically ranges from -1 to +1 so we use [0, 1]-transform/min-max scale again;
- for dimension$_7$ and dimension$_8$ (associations), we already have values in the interval [0,1] with the right orientation so we can use those directly.

42 If we perform these steps here and then compare the two candidates’ distances from the origin of the hypercube, the result is not surprising, given that we can see already in Table 3 that *c b* has scores that are more associated with MWU status: in terms of their Euclidean distances, *b d* and *c b* score the values of 1.51095 and 2.059107 respectively, which would lead to the decision that, of those two at least, *c b* should be merged into a new MWU. Thus, the algorithm would now merge each of the five sequences of *c* and *b* in the corpus into a new word/SWU *c_b*, store all the dimensions of information that led to the formation of the new unit for potential follow-up analysis, and iterate.

43 In this example, *c b* is actually one of the highest-scoring candidates of all. For expository reasons, however, it is instructive to quickly look at the lowest scoring candidate, *a c*, whose dimensional values are represented in Table 4:

Table 4. Observed dimension values for *q a*

	unlogged token freq of cand.	**1-disp. of cand. MWU**	**unlogged type freq slot 1**	**unlogged type freq in slot 2**	**entropy difference in slot 1**	**entropy difference in slot 2**	**association SWU$_1$ → SWU$_2$**	**association SWU$_2$ → SWU$_1$**
cand.: *a c*	2	0.5651022	6	4	0.002592	-0.0127533	0.3164643	0.2391363

44 It’s no wonder that *a c* is not a good candidate for a MWU: it occurs only twice, which also means it is a bit underdispersed, there are decent numbers of other words around *a* and *c* (‘undermining’ their mutual associations via the type frequencies), and the actual associations between both words are small; for instance, *a* is much more attracted to a following *y* (association *a*→*y*=0.55) and *y* is actually also much more attracted to a preceding *a* (association *a*←*y*=0.84).

45 Before we apply this algorithm to a real data set, one potential objection needs to be anticipated. One of the shortcomings of some previous studies I mentioned above was that they used AMs that conflated token frequency and association (e.g. G^2 and t) but then MERGE$_{\text{multidim}}$ seems to do the same because it collects 8 dimensions of information and conflates them into one Euclidean distance. While that is correct, there is also one very important difference: The conflation of dimensions is unavoidably explicit and tweakable for a particular objective. In other words, the problem with G^2 is not merely that it conflates dimensions of information – the problem is that it does so in a way that is not under the researcher's control. The degree to which token frequency and association raise G^2 mathematically depends on the corpus size and the frequencies of the co-occurring parts (see again Gries 2022) and cannot be tweaked by the researcher. Put differently, the researcher cannot say "for the current application, I am computing G^2-values but I will compute them in a way that prioritizes association over frequency". In MERGE$_{\text{multidim}}$, however, the researcher is in charge and can, if so desired, say "I will take the 8 values for my candidate MWUs but when the 8-dimensional hypercube is constructed, then the dimensions "association SWU1→SWU2" and "association SWU1←SWU2" will not have length 1, but length 2", which means their impact on the computation of the Euclidean distances for all candidate MWUs will be exactly twice as high as all others. Thus, the present approach does not leave the researcher at the mercy of peculiarities of the data (frequencies of SWUs, corpus size) or certain formulae, with them hoping that whatever way G^2 or some other statistic conflates dimensions will be useful – instead, MERGE$_{\text{multidim}}$ forces them, or permits them, to be explicit about what dimension of information is supposed to be (more) important (than others), which I consider a distinct advantage.

46 Now that MERGE$_{\text{multidim}}$ has been explicated on the basis of a small example, where one could easily compute/look up everything by hand, let us now apply it to a more realistic example, the Brown corpus. The version used here was set to lower case (and tagged, which treated punctuation marks as 'words') but I am not using the POS tags here. Everything else will be done as described above, i.e., all dimensions will be computed as before and equally-weighted; in this first application of MERGE$_{\text{multidim}}$, the number of iterations was set to 500 and each candidate MWU to be considered needed to have a minimum token frequency of three and to occur in more than one of the 500 files.

3. Results

47 After the R script implementing MERGE$_{\text{multidim}}$ completed 500 iterations, I identified all MWUs it had returned, checked which of them were part of larger MWUs (i.e. which of them were only 'stepping stones' to something longer that the algorithm also considered a MWU), and categorized them in terms of what kind of MWU, if any, they might constitute. The current categorization is not watertight, fully formalized, and mutually exclusive; no theoretical significance should be given to what are merely intended as convenient and flexible cover terms, but the groups will give nevertheless a decent impression of this first-ever performance of the algorithm.

48 Let us begin with the first category of **not-so-successful MWUs**, which comprises 154 instances falling into several categories; since some MWUs involve punctuation marks (esp. commas), in what follows, I show each MWU between underscores:

- instances that seem completely useless: _*, dostoevsky*_, _*, patting*_, _*, reserving*_, _*(ap*_, _** ya_, _** yf_, _*1.1 billion*_, etc.;
- many instances that begin with a determiner: _*a mammoth*_, _*a quart*_, _*a sportswriter*_, _*a truism*_, _*the advent*_, _*the advisability*_, _*the churchyard*_, _*the drizzle*_, _*the foreseeable*_, _*the hulks*_, _*the migrant*_, _*the riverbank*_, _*the safest*_, etc.;
- instances of *to* plus (an infrequent) verb: _*to cultivate*_, _*to deprive*_, _*to emulate*_, _*to envision*_, _*to implement*_, _*to inquire*_, _*to re-enter*_, _*to rebut*_, etc.

49 In all fairness, one has to admit that, while most of these are not useful, some are, namely for instance (i) those that during a later iteration get amalgamated into a longer MWU that seems more useful (*the foreseeable→the foreseeable future*) or (ii) those that could be seen as part of a place name (*the treasury*); I will return to the evaluation of these items briefly below.

50 A tiny bit more useful were some MWUs consisting of a **comma and a month or a state name**, as is customary in the US: _*, ala*_, _*, calif*_, _*, mich*_, _*, ore*_, etc.

51 However, much more interesting were many instances of expressions that I (again, informally) categorized as **compounds** (including some hyphenated expressions); in order to provide a good representation of the range of the results, I quote most of them here: _*aluminum foil*_, _*armed forces*_, _*atmospheric tests*_, _*ballistic missile(s)*_, _*barbed wire*_, _*bathing suits*_, _*bermuda shorts*_, _*booby traps*_, _*cellulose acetate*_, _*collective bargaining*_, _*communist bloc*_, _*coronary artery*_, _*dairy cows*_, _*december 31*_, _*differential equations*_, _*drainage ditch*_, _*dressing gown*_, _*expandable styrene*_, _*fairy tales*_, _*forked tongue*_, _*golden calf*_, _*household chores*_, _*human beings*_, _*hydrogen atoms*_, _*interior designers*_, _*interstate commerce*_, _*joint chiefs*_, _*juvenile delinquency*_, _*left ventricle*_, _*livery stable*_, _*manned bombers*_, _*megaton bomb(s)*_, _*molecular weights*_, _*monroe doctrine*_, _*nobel prize*_, _*nuclear weapons*_, _*orange juice*_, _*peace corps*_, _*peaceful coexistence*_, _*polaris missiles*_, _*pulitzer prize*_, _*racial discrimination*_, _*raw sewage*_, _*ray diffraction*_, _*real estate*_, _*respiratory infections*_, _*roman catholic(s)*_, _*sand dunes*_, _*shirt sleeves*_, _*siamese cats*_, _*sidewalk cafe*_, _*southeast asian*_, _*spiritual beings*_, _*sugar bowl*_, _*swivel chair*_, _*theatre guild*_, _*thermal conductivity*_, _*thyroid gland*_, _*vending machines*_, _*vocational rehabilitation*_, _*wave lengths*_, _*absent-mindedly*_, _*ante-bellum*_, _*far-reaching*_, _*flat-bottomed*_, _*good-natured*_, _*post-world*_, _*self-reliant*_, _*x-ray diffraction*_, etc. To me at least, these seem like excellent MWUs in the sense of, for instance, clearly being useful to learners of English.

52 Then, there was a wider range of MWUs I labeled **phrase**, by which I mean that many of them are part of recognizable, maybe decently frequent, and sometimes quite useful phrases/expressions or collocations (like *conspicuously absent* or *formally entrenched*), but they do not quite reach the level of "compound": _*desperate urgency*_, _*difficult to envision*_, _*he's hurting*_, _*i am*_, _*large supermarkets*_, _*microscopic examination*_, _*miscellaneous receipts*_, _*n't budge*_, _*nineteenth century*_, _*of nomenclature*_, _*of the loveliest*_, _*scottish rite*_, _*skilled manpower*_, _*the midst*_, _*to secede*_, _*v. united states*_, _*years ago*_, etc.

53 Then, there is a range of MWUs I informally called **fixed expressions**, which are the following and which also seem extremely useful (note that several of them are 3- or 4-

grams): *_at least_*, *_in abeyance_*, *_in accordance_*, *_in addition_*, *_in conjunction_*, *_in retrospect_*, *_in spite_*, *_in unison_*, *_inversely proportional_*, *_mutually exclusive_*, *_firmly entrenched_*, *_conspicuously absent_*, *_of yesteryear_*, *_the founding fathers_*, *_the remainder_*, *_to the hilt_*, *_truth or falsity_*, *_under the auspices of_*, *_up for grabs_*, *_vice versa_*.

54 The algorithm was also very successful at discovering a variety of **foreign-language expressions** (some of them even in their preferred syntactic uses, e.g. at the end of enumerations as evidenced by their combination with punctuation marks): *_, et cetera_*, *_, etc._*, *_ad hoc_*, *_bel canto_*, *_de chambre_*, *_de facto_*, *_dolce vita_*, *_el dorado_*, *_en route_*, *_esprit de corps_*, *_et al_*, *_in lieu_*, *_in vitro_*, *_in vivo_*, *_laissez-faire_*, *_pas de deux_*, *_per annum_*, *_per capita_*, *_per cent_*, *_per centum_*, *_piano concerto_*, *_status quo_*.

55 A slightly related group are the following, which I called **parentheticals** because they highlight the connection of some words to certain syntactic positions; the list will make clear what I mean: *_, alternatively_*, *_, jr_*, *_, ltd_*, *_, ma' am_*, *_. furthermore_*, *_. interestingly_*, *_. miraculously_*, *_. to summarize_*.

56 The second-to-last group of MWUs are 'just' **names** of people, places, institutions, etc. as well as **titles** (with or without names); I provide a selection here, which again include several 3-grams or even longer MWUs: *_general motors_*, *_herald tribune_*, *_international harvester_*, *_interstate commerce commission_*, *_johns hopkins_*, *_ku klux klan_*, *_manchester guardian_*, *_new englander_*, *_new yorker_*, *_rca victor_*, *_the iliad_*, *_the manchester guardian_*, *_the milwaukee braves_*, *_the sheraton-biltmore_*, *_united nations_*, *_albert schweitzer_*, *_babe ruth_*, *_benjamin franklin_*, *_casey stengel_*, *_dag hammarskjold_*, *_de gaulle_*, *_don quixote_*, *_f.d. r_*, *_fidel castro_*, *_jesus christ_*, *_joseph mccarthy_*, *_julius caesar_*, *_lou gehrig_*, *_mary jane_*, *_moise tshombe_*, *_noel coward_*, *_patrice lumumba_*, *_sam rayburn_*, *_sam spade_*, *_sargent shriver_*, *_theodore roosevelt_*, *_willie mays_*, *_baton rouge_*, *_beverly hills_*, *_cape cod_*, *_ciudad trujillo_*, *_coney island_*, *_el paso_*, *_hong kong_*, *_lake champlain_*, *_las vegas_*, *_lewisohn stadium_*, *_los angeles_*, *_n. y._*, *_new orleans_*, *_new york_*, *_new zealand_*, *_notre dame_*, *_pacific northwest_*, *_prairie du chien_*, *_puerto rico_*, *_san diego_*, *_san francisco_*, *_san juan_*, *_santa barbara_*, *_staten island_*, *_the bronx_*, *_the dominican republic_*, *_the kirov_*, *_the kremlin_*, *_the netherlands_*, *_the parthenon_*, *_the philippines_*, *_the union of soviet socialist republics_*, *_the united states_*, *_the ussr_*, *_u. s._*, _u. s. s. r._, *_virgin islands_*, *_gov. vandiver_*, *_italian consul_*, *_lt. gov._*, *_premier khrushchev_*, *_prime minister_*, *_prince souvanna phouma_*, *_sen. wayne_*, *_senator joseph mccarthy_*, etc.

57 The final group here is a bunch of MWUs where MERGE$_{multidim}$ successfully recognized that the pre-treatment of the corpus (tagging with udpipe) resulted in splitting things up that actually should have been/remained together; I am referring to these as **(re-)united-into-ones** but list them here in the way the algorithm returned them: *_- fro nt_*, *_a nalyses_*, *_a nalysis_*, *_bur nt_*, *_ca n't_*, *_co -operative_*, *_confro nt_*, *_dumo nt_*, *_gon na_*, *_i nches_*, *_i vy_*, *_in fant_*, *_over -all_*, *_teen - agers_*, *_thei r_*, *_vermo nt_*, *_vien na_*, *_viet nam_*, *_waterfro nt_*, *_wo n't_*.

4. Discussion and concluding remarks

4.1. Evaluation and discussion

58 Clearly, the results are not perfect: There is a sizable number of expressions returned by MERGE$_{\text{multidim}}$ that do not seem overly useful; post-hoc exploration of especially the larger number of 2-grams consisting of a determiner (esp. *the*) plus an adjective or a noun indicates that they have low token frequencies but are nonetheless returned by MERGE$_{\text{multidim}}$ due to high values on (some of) the other dimensions; for instance, we get several *to*+verb units, all of which involve relatively low-frequency verbs (e.g., *deprive*, *implement*, *incultivate*, ...).

59 At the same time, MERGE$_{\text{multidim}}$ also returns many very useful expressions, and one interesting aspect of those is that the algorithm is able to recognize different groups of them (because of its ability to integrate multiple dimensions of information). For example, some of the MWUs MERGE$_{\text{multidim}}$ returned are ones that many might superficially also expect from an implementationally much simpler *MI*-based approach (because they are, e.g., relatively infrequent 2-grams consisting of a first and a last name). However, on the whole, that impression would be mistaken and would not do MERGE$_{\text{multidim}}$ justice. That is because, if one computes *MI*-scores on all candidate MWUs from the Brown corpus, all top-scoring 100 resulting MWUs would be hapaxes! And even if one adds the additional criterion of a minimum frequency (of 3, like above), the results look a bit better than they do without the frequency threshold, but still not nearly as convincing as those of MERGE$_{\text{multidim}}$: The top results of such an *MI*+frequency threshold are *systemic linkage*, *sultan ahmet*, *sancho panza*, *ku klux*, *klux klan*, *amici curiae*, *hwang pah*, *grands crus*, ..., which is not terrible, but those of MERGE$_{\text{multidim}}$ are *los angeles*, *hong kong*, *dolce vita*, *klux klan*, *ku klux klan*, *puerto rico*, ... Biased as I might well be, I submit that it is hard to imagine a regular MWU application that would prefer the former list over the latter, and Evert [2009: 1239] cites similar bad results for *MI*. (Note also that MERGE$_{\text{multidim}}$ does indeed find left-to-right MWUs like *upside down* and right-to-left MWUs such as *in accordance*.)

60 As mentioned above, this superior performance over *MI* or *MI*+frequency is of course because MERGE$_{\text{multidim}}$ can recognize both the kinds of infrequent examples (that *MI* might prioritize) and the kinds of frequent examples (that more frequency-focused measures like G^2 or *t* might prioritize): It finds *Las Vegas*, *in vitro* and *in vivo*, *Julius Caesar*, *vending machines*, and *Ku Klux Klan* (with frequencies < 5), but also *United States*, *New York*, and *Rhode Islands* (with frequencies > 100). More broadly, since MERGE$_{\text{multidim}}$ by definition handles multiple dimensions, it prioritizes MWUs that do not only have high *MI*-values (like *systemic linkage* or *grands crus*) and/or high *t*-values, but ones that on top of high association also exhibit high frequencies, even dispersions, predictive power, etc. Of course, the same logic applies to any other dimension as well. Just as MERGE$_{\text{multidim}}$ does not return MWUs only because they are highly associated (but rare and underdispersed), so also it does not return MWUs only because they are highly frequent and evenly dispersed (but not strongly attracted). Note in particular the absence of the usual frequent and well-dispersed suspects (such as *of the* or *in the*, etc.) in the MWUs returned here, which is one reason why, to my mind, MERGE$_{\text{multidim}}$ outperforms Wahl & Gries's MERGE and some other approaches, which return many such very-high-frequency, but otherwise useless *n*-grams. Thus, while, again, the

results are not perfect, the way in which multiple dimensions of corpus-linguistic information are considered simultaneously has appealing characteristics and leads to many appealing results.

4.2. Where to go from here

61 What are the next steps? And how can the algorithm be improved? In a sense, the next steps follow quite naturally from everything discussed above, and there are plenty. First and most obviously, the number of iterations should be increased to see what else the algorithm returns when it goes beyond the top 500 MWUs. For instance, will *according to* emerge (it does not (yet)), will *in accordance with* (it does), will *in spite of* (it does), will we find things like *on the other hand* (not yet), etc.? Similarly obviously, one would want to apply MERGE$_{\text{multidim}}$ to other corpora and validate its output in ways that Wahl & Gries did for MERGE. A direct comparison with Wahl & Gries's results is not straightforward, but a quick glance at their results along the lines just alluded to shows that MERGE$_{\text{multidim}}$ seems quite a bit better at least in the way that it avoids many MWUs that their G^2-based approach returned: *in the*, *if you*, *of the*, *and I*, *on the*, *to get*, *to the*, etc. – G^2 just reflects too much of frequency (Gries 2022). While MERGE$_{\text{multidim}}$ has not yet been applied to the kinds of spoken data Wahl & Gries considered, it seems unlikely that, given its architecture, MERGE$_{\text{multidim}}$ would return less useful MWUs and that many of them. A straightforward follow-up would be to apply MERGE$_{\text{multidim}}$ to, say, the spoken part of the ICE-GB, which is work currently ongoing.

62 The next major kind of follow-up would be to explore the effect that different minimal settings have on MERGE$_{\text{multidim}}$'s output. Two kinds of settings are available: (i) the low-hanging fruit of different threshold values for token frequency (3 above) and range ($^2/_{500}$ above) and (ii) the much more ambitious and resource-intensive exploration of different weightings for the 8 dimensions. For example, does the too-high number of relatively rare 2-grams beginning with a determiner or with *to* above get addressed (while not making other aspects of the results worse) when the token frequency dimension gets assigned a weight > 1? Or do those get addressed when the relative importances of type frequency or directional attraction from the second word to *to* or the determiner are downgraded a bit? Or, would it be useful to not use all 8 dimensions but base the computation of the Euclidean distance on the 6 or 7 highest values so that, in a sense, one low-scoring dimension does not automatically downgrade a candidate too much?

63 Third, one might also consider simplifying the algorithm a bit. One possibility (that I did not report on) is concerned with the "bidirectionality" of some of the dimensions: type frequency, entropy difference, and association. Maybe we do not need to distinguish both directions of association but can just pick the larger of the two values. This might be justifiable on the basis of the argument that, maybe, it does not matter for the probability of something being a mentally-represented MWU whether the direction is $\text{SWU}_1 \rightarrow \text{SWU}_2$ (as in *upside down*) or $\text{SWU}_1 \leftarrow \text{SWU}_2$ (as in *in accordance*). In other words, maybe 5 dimensions are sufficient (token frequency, dispersion, some combined type frequency value, some combined entropy-difference value, and some combined association value).[6]

64 Lastly and relatedly, are there combinations of dimension weightings we want to explore for different applied purposes? In this first programmatic paper, I used a

unifying Euclidean distance to find the best MWU (and we just said the number of dimensions to be used is up for debate), but maybe other measures would do better (e.g., the harmonic mean of the dimension values; I thank Simon Todd for that idea) or the algorithm could use all 8 dimensions but without the subsequent conflation? Would it be interesting, for certain kinds of applications, to focus on MWUs with the best dispersion values (neglecting or downgrading their other dimensions for a moment) and then focus on MWUs with the highest association values (neglecting or downgrading their other dimensions for a moment)?

65 There is also a very general issue requiring some consideration, namely the best way of evaluating the output. Many returned MWUs are straightforwardly good (most of the things labeled as compounds, fixed expressions, and names above), but other MWUs are more problematic. What does one do with *the brink*? This does not look that useful and in the first 500, in fact the first 1000 iterations, nothing is done with this expression, but it is conceivable that at some later point, this will become part of *at/on/to the brink*, which is certainly a useful MWU. In other words, the question becomes, how does one deal with cases that are not useful (yet), but where a qualified human analyst sees potential? One would not want to give the algorithm full credit for them already since they were not returned as proper MWUs (yet); at the same time, counting them as false positive already also seems wrong, namely overly conservative...

66 If all of the above seems like MERGE$_{\text{multidim}}$ raises more questions than it answers, this is probably correct. However, I don't think this is because of obvious shortcomings of MERGE$_{\text{multidim}}$ but because (i) the notion of MWU is a fuzzy, radial category that is relevant to a huge number of applications with potentially conflicting demands and because (ii), to be honest, I think that much previous work has not taken seriously all the dimensions that are actually relevant: As far as I can tell, there is very little or no work that went beyond token frequency and association and maybe a simplistic range threshold although we have decades of cognitive-linguistic and psycholinguistic evidence for the relevance of all other dimensions discussed here. Thus, I don't think that MERGE$_{\text{multidim}}$ raises so many questions because it is such a lacking approach – I think it raises so many questions because it is among the first approaches to consider all these dimensions of corpus-linguistic information, which we know to be important from hundreds of other applications, and it considers them simultaneously. And I think, without exaggeration, this kind of work is absolutely fundamental to corpus linguistics because what this paper is really about at the most fundamental level is tokenization/segmentation (recall again Colson [2018]) and, in a sense, what could be more important than that? Without proper tokenization, we do not get proper collocation results (because our counts to compute AMs will be off and certain collocations would not be recognized), without proper tokenization, we do not get proper keywords results (because our counts to compute keyness will be off and preferences of certain MWUs to some target corpus would not be recognized), without proper tokenization, we get lexical bundle analyses that do not respect what really are linguistic units, etc., etc. For example, Leech & Fallon's [1992] paper doing essentially a cultural keywords analysis of Brown and LOB is by virtue of its (tokenization) design unable to see which of the MWUs identified in this paper are cultural keywords for American English: their "tokenization" did not permit them to see *Ku Klux Klan*, *vending machines*, and even *to secede* as tokens for which their difference coefficient/keyness might return that they are key for American data (relative to British data), but if one

processes one's corpora with something like MERGE$_{\text{multidim}}$ first, such MWUs could be recognized and entered into the analysis properly.

67 Thus, the ideal outcome of this paper would therefore be if it stimulated many follow-up studies (e.g. in the form of dissertations that have the space to play around with multiple settings) to different data to determine good/better settings for the application of MERGE$_{\text{multidim}}$ in various linguistic settings (and to speed up my current implementation) because MERGE$_{\text{multidim}}$'s output, a more powerful tokenization, is a precondition for virtually everything that follows.

BIBLIOGRAPHY

BAAYEN Harald, MILIN Petar & RAMSCAR Michael, 2016, "Frequency in lexical processing", *Aphasiology*, 30(11), 1174-1220.

BANNARD Colin & MATTHEWS Danielle, 2008, "Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations", *Psychological Science* 19(3), 241-248.

BELL Alan, JURAFSKY Daniel, FOSLER-LUSSIER Eric, GIRAND Cynthia, GREGORY Michelle & GILDEA Daniel, 2003, "Effects of disfluencies, predictability, and utterance position on word form variation in English conversation", *The Journal of the Acoustical Society of America* 113(2), 1001-1024.

BESTGEN Yves, 2018, "Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test", *Corpora* 13(2), 205-228.

BIBER Douglas, CONRAD Susan & CORTES Viviana, 2004, "If you look at ...: Lexical bundles in university teaching and textbooks", *Applied Linguistics* 25(3), 371-405.

BREEZE Ruth, 2013, "Lexical bundles across four legal genres", *International Journal of Corpus Linguistics* 18(2), 229-253.

COLSON Jean-Pierre, 2018, "From Chinese word segmentation to extraction of constructions: two sides of the same algorithmic coin", *in Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, NM, 41-50.

CORTES Viviana, 2004, "Lexical bundles in published and student disciplinary writing: Examples from history and biology", *English for Specific Purposes* 23(4), 397-423.

CROSSLEY Scott A. & LOUWERSE Max, 2007, "Multi-dimensional register classification using bigrams", *International Journal of Corpus Linguistics* 12(4), 453-478.

DAUDARAVIČIUS Vidas & MARCINKEVIČIENĖ Rūta, 2004, "Gravity counts for the boundaries of collocations", *International Journal of Corpus Linguistics* 9(2), 321-348.

DUNN Jonathan, 2018, "Multi-unit association measures", *International Journal of Corpus Linguistics* 23(2), 183-215.

EVERT Stefan, 2009, "Corpora and collocations", *in* LÜDELING Anke & KYTÖ Merja (Eds.), *Corpus linguistics: An international handbook, Vol. 2*, Berlin & New York: Mouton de Gruyter, 1212-1248.

GRIES Stefan Th., 2008a, “Phraseology and linguistic theory: a brief survey”, *in* GRANGER Sylviane & MEUNIER Fanny (Eds.), *Phraseology: an interdisciplinary perspective*, Amsterdam & Philadelphia: John Benjamins, 3-25.

GRIES Stefan Th., 2008b, “Dispersions and adjusted frequencies in corpora”, *International Journal of Corpus Linguistics* 13(4), 403-437.

GRIES Stefan Th., 2010, “Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora”, *in Proceedings of Corpus Linguistics 2009*, Liverpool: University of Liverpool.

GRIES Stefan Th., 2019, “15 years of collostructions: some long overdue additions/corrections (to/ of actually all sorts of corpus-linguistics measures)”, *International Journal of Corpus Linguistics* 24(3), 385-412.

GRIES Stefan Th., 2020, “Analyzing dispersion”, *in* PAQUOT Magali & GRIES Stefan Th. (Eds.), *A practical handbook of corpus linguistics*, Berlin & New York: Springer, 99-118.

GRIES, Stefan Th., 2021, *Statistics for Linguistics with R.*, 3rd rev. & ext. ed., Boston & Berlin: Mouton de Gruyter.

GRIES Stefan Th., 2022, “What do (some of) our association measures measure (most)? Association?”, *Journal of Second Language Studies* 5(1), 1-33.

GRIES Stefan Th. & MUKHERJEE Joybrato, 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.

JEACO Stephen, 2019, “Exploring collocations with The Prime Machine”, *International Journal of Computer-Assisted Language Learning and Teaching* 9(3), 29-49.

JELINEK Frederick, 1990, “Self-organized language modeling for speech recognition”, *in* WAIBEL Alex & LEE Kai-Fu (Eds.), *Readings in speech recognition*, San Mateo, CA: Morgan Kaufmann, 450-506.

KITA Kenji, KATO Yasuhiko, OMOTO Takashi & YANO Yoneo, 1994, “Automatically extracting collocations from corpora for language learning”, *Journal of Natural Language Processing* 1(1), 21-33.

LEECH Geoffrey & FALLON Roger, 1992, “Computer corpora: What do they tell us about culture?”, *ICAME Journal* 16, 29-50.

NELSON Robert, 2018, “How ‘chunky’ is language? Some estimates based on Sinclair’s Idiom Principle”, *Corpora* 13(3), 431-460.

O’DONNELL Matthew Brook, 2011, “The adjusted frequency list: A method to produce cluster-sensitive frequency lists”, *ICAME Journal No. 35*, 135-169.

SIMPSON-VLACH Rita & ELLIS Nick C., 2010, “An Academic Formulas List: New methods in phraseology research”, *Applied Linguistics* 31(4), 487-512.

SINCLAIR John M., 1987, *Collins COBUILD English language dictionary*, Ann Arbor: Collins.

STEFANOWITSCH Anatol & GRIES Stefan Th., 2003, “Collostructions: investigating the interaction between words and constructions”, *International Journal of Corpus Linguistics* 8(2), 209-243.

UNDERWOOD Geoffrey, SCHMITT Norbert & GALPIN Adam, 2004, “The eyes have it: An eye-movement study into the processing of formulaic sequences”, *in* SCHMITT Norbert (Ed.), *Formulaic Sequences: Acquisition, Processing, and Use*, Amsterdam & Philadelphia: John Benjamins, 153-172.

WAHL Alexander & GRIES Stefan Th., 2018, "Multi-word expressions: A novel computational approach to their bottom-up statistical extraction", *in* CANTOS-GÓMEZ Pascual & ALMELA-SÁNCHEZ Moisés (Eds.), *Lexical collocation analysis: advances and applications*, Berlin & New York: Springer, 85-109.

WAHL Alexander & GRIES Stefan Th., 2020, "Computational extraction of formulaic sequences from corpora: Two case studies of a new extraction algorithm", *in* CORPAS PASTOR Gloria & COLSON Jean-Pierre (Eds.), *Computational phraseology*, Amsterdam & Philadelphia: John Benjamins, 84-110.

WERMTER Joachim & HAHN Udo, 2004, "Collocation extraction based on modifiability statistics", *Coling 2004: Proceedings of the 20th International Conference on Computational Linguistics*, Geneva: Association for Computational Linguistics, vol. 2, 980-986.

WERMTER Joachim & HAHN Udo, 2005. "Paradigmatic modifiability statistics for the extraction of complex multi-word terms", *in* MOONEY Raymond *et al.* (Eds.), *Proceedings of the 5th Human Language Technology Conference and 2005 Conference on Empirical Methods in Natural Language Processing*, Vancouver: Association for Computational Linguistics, 843-850.

WIBLE David, KUO Chin-Hwa, CHEN Meng-Chang, TSAO Nai-Lung, HUNG Tsung-Fu, 2006, "A Computational Approach to the Discovery and Representation of Lexical Chunks", *The 13th Conference on Natural Language Processing (TALN 2006). April 10-13, 2006. Leuven (Belgium), 2006*, Leuven, Belgium.

WRAY Alison, 2002, *Formulaic language and the lexicon*, Cambridge: Cambridge University Press.

XU Ying, GOEBEL Randy, RINGLSTETTER Christoph & KONDRAK Grzegorz, 2010, "Application of the Tightness Continuum Measure to Chinese information retrieval, *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, Beijing: Coling 2010 Organizing Committee, 54-62.

NOTES

[1] Based on the results from the *G*-values, they also propose a measure called "lexical stickiness", which quantifies how much words like to occur in multi-word units as opposed to on their own and which essentially relates the behavior of a word to the continuum from the open-choice to the idiom principle; while interesting, I won't discuss this here further since it does not involve a different means of MWU identification.

[2] Other AMs can of course be used as well.

[3] Dunn [2018] is a study that uses multiple measures of directional collocation strength (△*P*); see also Colson [2018] who uses directional transitional probabilities and, citing Xu *et al.* [2010], comments on the high degree of similarity between MWU identification and tokenization/word segmentation in Chinese.

[4] Candidate MWUs could be generated from non-adjacent SWUs, but for simplicity's sake, I will not consider this here.

[5] However, this is not a particularly strong expectation: Daudaravičius & Murcinkevičienė developed their measure *G* with the reverse orientation.

[6] An initial exploration of the top 500 MWUs suggests that the 8 dimensions can be conflated into 4 different principal components (explaining ≈89% of the variance of the original 8 dimensions).

ABSTRACTS

It has been argued that most of corpus linguistics involves one of four fundamental methods: frequency lists, dispersion, collocation, and concordancing. All these presuppose (if only implicitly) the definition of a unit: the element whose frequency in a corpus, in corpus parts, or around a search word are counted (or quantified in other ways). Usually and with most corpus-processing tools, a unit is an orthographic word. However, it is obvious that this is a simplifying assumption borne out of convenience: clearly, it seems more intuitive to consider *because of* or *in spite of* as one unit each rather than two or three. Some work in computational linguistics has developed multi-word unit (MWU) identification algorithms, which typically involve co-occurrence token frequencies and association measures (AMs), but these have not become widespread in corpus-linguistic practice despite the fact that recognizing MWUs like the above will have a profound impact on just about all corpus statistics that involve (simplistic notions of) words/units. In this programmatic proof-of-concept paper, I introduce and exemplify an algorithm to identify MWUs that goes beyond frequency and bidirectional association by also involving several well-known but underutilized dimensions of corpus-linguistic information: *frequency*: how often does a potential unit (like *in_spite_of*) occur?; *dispersion*: how widespread is the use of a potential unit?; *association*: how strongly attracted are the parts of a potential unit?; *entropy*: how variable is each slot in a potential unit?

The proposed algorithm can use all these dimensions and weight them differently. I will (i) present the algorithm in detail, (ii) exemplify its application to the Brown corpus, (iii) discuss its results on the basis of several kinds of MWUs it returns, and (iv) discuss next analytical steps.

On soutient généralement que la linguistique de corpus recourt à l'une des quatre méthodes de base suivantes : listes de fréquences, dispersion, collocation et concordance. Toutes ces méthodes présupposent (ne serait-ce qu'implicitement) la définition de ce qu'est une unité, à savoir l'élément dont la fréquence dans un corpus, dans des extraits de corpus, ou dans l'environnement textuel d'un mot étudié est calculée (ou quantifiée d'une quelconque manière). En règle générale et pour la majorité des outils de traitement de corpus, une unité est un mot orthographique. Cependant, il est évident qu'il s'agit là d'une hypothèse simplificatrice résultant d'un souci de facilité : il est évident qu'il semble plus intuitif de considérer que les mots *because of* ou *in spite of* constituent chacun une unité plutôt que deux ou trois. Certains travaux en linguistique computationnelle ont développé des algorithmes pour l'identification des unités multi-mots (*multi-word units* en anglais, MWU), qui reposent généralement sur des fréquences de cooccurrence de tokens et des mesures d'association (*association measures* en anglais, AM), mais ces algorithmes ne se sont pas généralisés en linguistique de corpus, malgré le fait que la reconnaissance des MWU, à l'image de celles susmentionnées, pourrait avoir un effet significatif sur la quasi-totalité des statistiques de corpus qui se fondent sur les notions (simplistes) de « mots/unités ». Dans cet article programmatique où je souhaite valider un concept, je présente et illustre un algorithme d'identification des MWU qui va bien au-delà de la fréquence et de l'association bidirectionnelle en intégrant également plusieurs dimensions bien connues – mais sous-utilisées – de l'information en linguistique de corpus : la *fréquence* : combien de fois une unité potentielle (comme *in_spite_of*) se rencontre-t-elle ?, la *dispersion* : à quel point l'utilisation d'une unité potentielle est-elle répandue ?, *l'association* : à quel point les constituants d'une potentielle unité s'attirent-ils plus ou moins fortement ?, *l'entropie* : à quel point chaque emplacement d'une potentielle unité est-il variable ?

L'algorithme proposé a recours à ces quatre dimensions et les pondère différemment. Je vais (i) présenter l'algorithme en détail, (ii) exemplifier son application au corpus Brown, (iii) discuter les résultats obtenus sur la base de plusieurs types de MWU qu'il renvoie, et (iv) envisager les prochaines étapes de l'analyse.

INDEX

Keywords: corpus linguistics, multi-word units, n-grams, frequency, dispersion, association, entropy
Mots-clés: linguistique de corpus, unités multi-mots, n-grams, fréquence, dispersion, association, entropie

AUTHOR

STEFAN TH. GRIES

UC Santa Barbara & JLU Giessen
stgries@linguistics.ucsb.edu