50-something years of work on collocations

What is or should be next …*

Stefan Th. Gries
University of California, Santa Barbara

This paper explores ways in which research into collocation should be improved. After a discussion of the parameters underlying the notion of 'collocation', the paper has three main parts. First, I argue that corpus linguistics would benefit from taking more seriously the understudied fact that collocations are not necessarily symmetric, as most association measures imply. Also, I introduce an association measure from the associative learning literature that can identify asymmetric collocations and show that it can also distinguish collocations with high and low association strengths well. Second, I summarize some advantages of this measure and brainstorm about ways in which it can help re-examine previous studies as well as support further applications. Finally, I adopt a broader perspective and discuss a variety of ways in which *all* association measures – directional or not – in corpus linguistics should be improved in order for us to obtain better and more reliable results.

Keywords: collocation, directionality, association measure, ΔP (delta P), dispersion

1. Introduction

1.1 Definitional features of phraseologism and collocation

Perhaps the most famous quote in corpus linguistics is Firth's (1957: 179) "You shall know a word by the company it keeps". Thus, the notion of collocation, or more generally co-occurrence, has now been at the centre of much corpus-linguistic work for decades. As is so often the case, however, this does not mean that we as a field have arrived at a fairly unanimous understanding of what collocations are (in general), how they are best retrieved/extracted, how their strength or other characteristics are best measured/quantified, etc. It is therefore not

International Journal of Corpus Linguistics 18:1 (2013), 137–165. DOI 10.1075/ijcl.18.1.09gri
ISSN 1384–6655 / E-ISSN 1569–9811 © John Benjamins Publishing Company

surprising that the notion of 'collocation' is probably best characterized as a radial category whose different senses are related to each other and grouped around one or more somewhat central senses, but whose senses can also be related to each other only rather indirectly.

This definitional situation regarding 'collocation' is somewhat similar to that of 'phraseologism', another notion for which every scholar seems to have their own definition. In a previous publication, Gries (2008a) attempted to tease apart a variety of dimensions that researchers of phraseologisms/phraseology should always take a stand on when they use 'phraseologism'. These dimensions are not new – in fact, they are implicit in pretty much all uses of 'phraseologism' – but they are not always made as explicit as comprehensibility, comparability, and replicability would demand. For 'phraseologism', this is the list of dimensions proposed, to which a possible separation of lexical flexibility and syntactic flexibility (or commutability / substitutability) could be added:

i. the nature of the elements involved in a phraseologism;
ii. the number of elements involved in a phraseologism;
iii. the number of times an expression must be observed before it counts as a phraseologism;
iv. the permissible distance between the elements involved in a phraseologism;
v. the degree of lexical and syntactic flexibility of the elements involved;
vi. the role that semantic unity and semantic non-compositionality / non-predictability play in the definition.

It is a useful starting point to consider the dimensions that underlie most of the work using collocations, and given the at least general similarity of 'phraseologism' and 'collocation' (cf. e.g. Evert's (2009: 1213) statement that "[t]here is considerable overlap between the phraseological notion of collocation and the more general [Firthian] empirical notion"), several characteristics are similar, too:

i. the nature of the elements observed; for collocations at least, these elements are words; once more general categories such as parts of speech or others are considered, researchers typically use 'colligation' or 'collostruction' for such cases;
ii. the number of collocates *l* that make up the collocation; the most frequent value here is "two" but others are possible and lead to the territory of notions such as multi-word units, n-grams, lexical bundles, etc.;
iii. the number of times *n* an expression must be observed before it counts as a collocation; often, *n* is defined as "occurring more frequently than expected by chance" but other thresholds and many statistics other than raw frequencies of co-occurrence are used, too;

iv. the distance and/or (un)interruptability of the collocates; the most frequent values here are "directly adjacent", "syntactically/phrasally related but not necessarily adjacent" (as in the V *into* V-*ing* construction), or "within a window of *x* words" or "within a unit (e.g. a sentence)";
v. the degree of lexical and syntactic flexibility of the collocates involved; typically, the word 'collocation' is used with word forms, but studies using lemmas are also common;
vi. the role that semantic unity and semantic non-compositionality / non-predictability play in the definition; often, it is assumed that the *l* words exhibit something unpredictable in terms of form and/or function.

On the one hand, these are, I think, useful criteria – just like with phraseologisms, studies can only benefit from making clear what their definition of 'collocation' implies on each of the above dimensions. On the other hand, it is also plain to see that one's definition of collocation may have to vary from application to application – compare a computational system designed to identify proper names to an applied-linguistics context trying to identify useful expressions for foreign learners – and that these can easily conflict with each other. For instance, the potential collocation *in the* consists of two specific and adjacent lexical elements, the collocation is very frequent ($n > 500{,}000$ in the BNC) and more frequent than expected by chance ($MI > 2$) – but at the same time *in the* has virtually nothing unpredictable or interesting about it in terms of both form and function, and many researchers would prefer assigning collocation status to something more functionally useful (even if rarer) such as *because of* or *according to* (cf. again Evert 2009 for useful discussion and exemplification).

1.2 Association measures to quantify collocation strength and the present study

In attempts to come to a potentially much more generally applicable definition of 'collocation', and to cope with increasingly larger corpora and, thus, larger numbers of candidates for collocation status, the last fifty years or so have resulted in many studies on the second characteristic in the above list, namely on how to best extract, identify, and measure collocations given their frequencies of co-occurrence. Also, computing time has become exponentially cheaper over the last few decades so the possibility that, once we throw enough data and computing power at the right measure or algorithm, we get a good list of collocations, has become increasingly attractive. As a result, many of the studies during the last fifty years have been devoted to developing, surveying, and comparing measures of collocational attraction/repulsion, i.e. association measures that quantify the strength and/or the reliability of a collocation. Good recent overviews include, for example, Evert (2005), Wiechmann (2008), and Pecina (2009), who discuss and review association measures in the domains of both lexical co-occurrence and lexico-grammatical co-occurrence: Evert (2005) focuses, among other things, on the statistical properties of association measures and their geometric interpretation, Wiechmann (2008) compares altogether 47 different association measures with regard to how well they match up with psycholinguistic reading-time data, and Pecina (2009) compares more than 80 measures for collocation extraction.

While these numbers of association measures just mentioned are quite large, nearly all of the ones that are used with any frequency worth mentioning are at least in some way based on a co-occurrence table of observed frequencies as exemplified in Table 1 and a comparison of (parts of) this table with (parts of) the table of frequencies expected by chance.

Table 1. Schematic co-occurrence table underlying most collocational statistics

	word$_2$: present	**word$_2$: absent**	**Totals**
word$_1$: present	*a*	*b*	*a*+*b*
word$_1$: absent	*c*	*d*	*c*+*d*
Totals	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*

Thus, while the number of measures that have been proposed is staggering, this high quantity of measures has not also lead to a corresponding increase in diversity and/or new ideas as well as quality, and in fact most of the measures that did make it into mainstream corpus linguistics are lacking (in a variety of different ways, many of which are routinely discussed when a new measure has been proposed). In this paper, I will – as admittedly many before me – try to breathe some new life into the domain of collocation studies, but hope I will do so with some viewpoints that are underrepresented in collocation studies.

The first main part of this paper, Section 2, is devoted to (i) introducing and arguing in favour of the notion of 'directional collocation' and to (ii) proposing as well as exemplifying a simple directional association measure derived from the domain of associative learning, Δ*P*, and (iii) exploring its results in general and in reference to more established measures.

The second main part of this paper, Section 3, is more speculative and brainstorming in nature. After a very brief interim recap, I summarily highlight what I believe to be the main advantages of Δ*P* and continue with referring to ways in which Δ*P* can maybe shed new light on results from previous studies. Furthermore, I also briefly speculate on Δ*P*'s extendability to multi-word units. The final part, Section 4, is concerned with at least three ways in which probably *nearly all* association measures – bidirectional or directional – must be improved upon; all of these have to do with different ways of increasing the resolution of how we study the kind of data represented schematically in Table 1.

2. Towards exploring a directional association measure

2.1 Directional approaches to the association of collocations

As mentioned above, all association measures currently in wider use are based on co-occurrence tables of the sort exemplified in Table 1. Nearly all such measures reflect the mutual association of word$_1$ and word$_2$ to each other and this type of approach has dominated corpus-linguistic thinking about collocations for the last fifty years. However, what all these measures do *not* distinguish is whether word$_1$ is more predictive of word$_2$ or the other way round. This holds even for those measures that are supported most theoretically and supported strongly empirically such as $p_{\text{Fisher-Yates exact test}}$ or G^2 (a.k.a. *LL*, the log-likelihood measure). In other words, nearly all measures that have been used are bidirectional, or symmetric. However, as Ellis (2007) and Ellis & Ferreira-Junior (2009: 198) point out correctly, "associations are not necessarily reciprocal in strength". More technically, bidirectional/symmetric association measures conflate two probabilities that are in fact very different: $p(\text{word}_1|\text{word}_2)$ is not the same as $p\,(\text{word}_2|\text{word}_1)$, just compare $p\,(\textit{of}\,|\textit{in spite})$ to $p\,(\textit{in spite}|\textit{of})$.

While it is difficult to not recognize this difference in probabilities and its potential impact, just like the notion of dispersion this issue has not been explored very much. One measure that addresses this in part is Minimum Sensitivity *MS* (cf. Pedersen 1998), which is defined in (1).

(1) $$MS = minimum\,(\frac{a}{a+b}, \frac{a}{a+c})$$

In Wiechmann's (2008) comparative study, *MS* is the measure that is most strongly correlated with psycholinguistic reading time data, followed by the (insignificantly worse) $p_{\text{Fisher-Yates}}$. However, in spite of its good performance, I think that *MS* is somewhat dangerous as an association measure for the simple reason that any one *MS*-value does not reveal what it actually means. More specifically, if one obtains $MS = 0.2$, then this value *per se* does not even reveal whether that 0.2 is $^{a}/_{a+b}$ or $^{a}/_{a+c}$, or $p\,(\textit{of}\,|\textit{because})$ to $p\,(\textit{because}|\textit{of})$!

A second measure that has been studied and that is actually implied by *MS* is simple conditional probability as exemplified in (2).

(2) a. $p(word_2 | word_1) = \frac{a}{a+b}$

b. $p(word_1 | word_2) = \frac{a}{a+c}$

This measure has been used with at least some success in some studies on how predictability affects reduction in pronunciation (cf. Bell et al. 2009 and Raymond & Brown 2012 for recent applications). However, there is so far hardly any work which explored its use as a measure of collocational strength. Two exceptions are Michelbacher et al. (2007, 2011). Michelbacher et al. (2007) compute conditional probabilities based on adjective/noun collocates in a window of 10 words around node words in the BNC and correlate them with the University of South Florida Association Norms. They find that conditional probabilities are fairly good at identifying asymmetric associations in the norming data but perform much less successfully in identifying symmetric associations; in addition, a classifying task based on conditional probabilities did better than chance, but still resulted in a high error rate of 39%.

The final measure, also proposed by Michelbacher et al. (2007), is based on the differences of ranks of association measures (such as chi-square values). For such rank measures, a collocation *x y* is explored by (i) computing all chi-square tests for collocations with *x*, ranking them, and noting the rank for *x y*, and by (ii) computing all chi-square tests for collocations with *y*, ranking them, and noting the rank for *x y*, and (iii) comparing the difference in ranks. In tests analogous to those of conditional probabilities, this rank measure does not perform well with asymmetric associations but a little better with symmetric ones; in the additional classification task, the rank measure came with an even higher error rate than conditional probabilities (41%). In their (2011) study, additional rank measures are also based on raw co-occurrence frequencies, G^2, and *t*, and the corpus-based data are compared to the results of a free association task undertaken specifically for that study. The results of the rank measures in that study are much more compatible with the subjects' reactions in the experiment both qualitatively ("[f]or about 80% of the [61] pairs [...] the statistical measures indicate the correct direction of association", Michelbacher et al. 2011: 266) and quantitatively; of the rank measures, G^2 performs best but, in spite of the huge computational effort involved in the thousands of ranked G^2-values, not better than conditional probability (Michelbacher et al. 2011: 270).

2.2 A measure from associative learning: ΔP

While the vast majority of quantitative work on collocations has utilized symmetric measures, we have seen that at least two studies are available that take directionality of collocations more seriously. The first study of Michelbacher et al. provided rather mixed results, but the second provided support for both conditional probability and their rank measures. However, I think there may be room for improvement. First, it may be a problem of conditional probabilities that the probability distribution of, say, word_2 given word_1 is not normalized against that of not-word_2 given word_1.

Second, the computational effort that goes into the computation of the rank measures is huge: since the computation of a directional association score of even a single word pair can require the computations of tens or hundreds of thousands of, say, G^2 or t-tests, which seems less than optimal given that, in the quantitative analysis of Michelbacher et al. (2011), conditional probabilities did just as well as G^2.

Third, Michelbacher et al. (2011) is a very laudable study in how they try to combine corpus-linguistic data and psycholinguistic evidence. However, one cannot help but notice that the corpus-based statistics they use do not (necessarily) correspond to anything cognitive or psycholinguistic: to the best of my knowledge, there are, for instance, no cognitive, psychological, or psycholinguistic theories that involve something like ranks of G^2-values.

In this paper, I would therefore like to propose to use a different measure, a measure first discussed by Ellis (2007) and then used in the above-cited work by Ellis & Ferreira-Junior (2009). This measure is called ΔP and is defined in (3) and below:

(3) $\Delta P = p\,(\mathit{outcome}\,|\,\mathit{cue}=\mathit{present}) - p\,(\mathit{outcome}\,|\,\mathit{cue}=\mathit{absent})$

> ΔP is the probability of the outcome given the cue (P(O|C)) minus the probability of the outcome in the absence of the cue (P(O|-C)). When these are the same, when the outcome is just as likely when the cue is present as when it is not, there is no covariation between the two events and $\Delta P = 0$. ΔP approaches 1.0 as the presence of the cue increases the likelihood of the outcome and approaches −1.0 as the cue decreases the chance of the outcome – a negative association.
> (Ellis 2007: 11; cf. that paper also for experimental validation of ΔP in the domain of associative learning theory)

Thus, ΔP addresses all three above shortcomings of the directional measures explored so far: it normalizes conditional probabilities, it is computationally extremely easy to obtain, and it arose out of associative learning theory and can thus lay more claim to being a psychologically/psycholinguistically realistic measure. If this logic is applied to Table 1, two perspectives can be distinguished, depending on whether the outcome is the choice of word$_2$ and the cue is the presence or absence of word$_1$ (in the rows, cf. (4)) or whether the outcome is the choice of word$_1$ and the cue is the presence or absence of word$_2$ (in the columns, cf. (5)):

(4) $\Delta P_{2|1} = p(word_2 \mid word_1 = present) - p(word_2 \mid word_1 = absent) = \frac{a}{a+b} - \frac{c}{c+d}$

(5) $\Delta P_{1|2} = p(word_1 \mid word_2 = present) - p(word_1 \mid word_2 = absent) = \frac{a}{a+c} - \frac{b}{b+d}$ Note: typo in published ms.

More concretely, if we apply (4) and (5) to the data shown in Table 2 (in (6) and (7) respectively), the difference is striking: *of* is not a good cue for *course*, but *course* is quite a strong cue for *of*.[1]

Table 2. Co-occurrence table for *of* and *course* in the spoken component of the BNC

	course: present	*course*: absent	Totals
of: present	5610	168,938	174,548
of: absent	2257	10,233,063	10,235,320
Totals	7867	10,402,001	10,409,898

(6) $\Delta P_{2|1} = p(course \mid word_2 = of) - p(course \mid word_2 \neq of) = \frac{5610}{174548} - \frac{2257}{10235320} \approx 0.032$

(7) $\Delta P_{1|2} = p(of \mid word_2 = course) - p(of \mid word_2 \neq course) = \frac{5610}{7867} - \frac{168938}{10402001} \approx 0.697$

On the one hand, this may seem only too obvious – *of* occurs with very many different types and a large number of tokens, but *course*'s distribution is much more restricted and, thus, *course* is a better cue to *of* than vice versa. On the other hand, it is just as obvious that all standard measures do not differentiate this, as is evident from computing some standard collocational statistics for Table 2. As is shown in Table 3, many of these are very high (MI, t, G^2, $p_{\text{Fisher-Yates}}$), but since they conflate two potential directions of association, they do not reveal that the association is in fact only high in one direction. Note also that, as argued above, the *MS*-value as such, here $^{5610}/_{174,548}$, does not reveal whether the association of *of* to *course* is very similar to that of *course* to *of*: all it says is that the value of the weaker direction is 0.032 – whether the other direction has a sensitivity of 0.033 (i.e. a bit larger) or 0.66 (i.e. much larger) is not clear. Note finally that Michelbacher et al.'s (2007) rank measure can also not identify *of course*'s asymmetry since *of course* scores rank 1 in *both* chi-square rankings!

Table 3. Collocational statistics for of and course in the spoken component of the BN

2-word unit	*MI*	*t*	Dice	G^2	$p_{\text{Fisher-Yates}}$	*MS*
of course	5.41	476.97	0.062	36,693.85	$< 10^{-320}$	0.032

Inequalities of the above kind, where one Δ*P* is very different from the other Δ*P* for the same word pair, are by no means restricted to *of course* – rather, they are frighteningly frequent, which in turn casts a serious shadow of doubt on very many previous studies involving the standard collocational measures; cf. Michelbacher et al. (2007: Section 3.4) for a similar finding. Given the frequency of such asymmetries, it seems useful to have a measure that can handle them well, but it is necessary first to explore the discriminatory power of Δ*P* as well as its correlation with some of the currently standard measures, which is what I will do in the following three sub-sections.

2.3 Validation 1: Bigrams with a high mutual association

2.3.1 *The overall behaviour of* ΔP *and other association measures*

To explore whether Δ*P* can identify strong collocations and asymmetries in them in more than just *of course*, I computed all the above collocational statistics and the two different Δ*P*-values for 262 two-word units annotated as such in the spoken component of the BNC. These units should exhibit a high degree of attraction, which motivated their consideration as multi-word units in the first place.

Two kinds of observations can be made. First, the means of all measures – the symmetric ones and both Δ*P*-values – suggest that these two-word units are strongly associated with each other, as represented in Table 4. If we explore the central 95% of all measures, however, we find that some of the traditional measures return highly negative values – indicating repulsion of the bigrams' components, which is surprising given the bigrams' status of multi-word units – whereas the Δ*P*-values and MS venture into "repulsion territory" only ever so slightly.

Table 4. Some collocational statistics for bigrams in the spoken component of the BNC

	MI	*t*	G^2	$\Delta P_{1\|2}$	$\Delta P_{2\|1}$	*MS*
mean	7.65	466.4	1064.11	0.28	0.2	0.1
0.025 quantile	−3.68	−13.52	−287.05	−0.01	−0.01	0
0.975 quantile	22.79	3226.43	12909.76	1	1	1

Second, the directional measures indicate a sizable proportion of collocations that are asymmetric. For more than a quarter of all bigrams (25 + 43 = 68 out of 262), there is a large difference between the two Δ*P*-values (≥0.5 or ≤–0.5), which is represented in the left panel of Figure 1, but by definition not revealed by any of the standard bidirectional collocational statistics. The right panel of Figure 1 shows the difference even more precisely: Δ*P* $\text{word}_2|\text{word}_1$ is represented on the *x*-axis, Δ*P* $\text{word}_1|\text{word}_2$ is represented on the *y*-axis, every circle represents a bigram, with the size of the circle being proportional to its frequency, and overplotting is represented in shades of grey; in both plots, the "x" marks *of course*. It is clear that, in a way that is not obviously related to frequency of co-occurrence, the two-word units in question are very different in how one word may attract the other much more/less than the other.

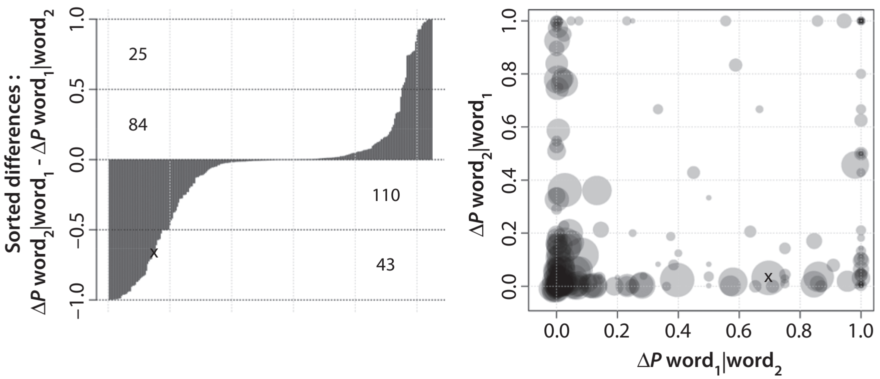


Figure 1. Pairwise differences between Δ*P*-values for 262 two-word units in the spoken part of the BNC: sorted differences (left panel); Δ*P*-values in both directions of association (right panel)

2.3.2 *The correlation of* ΔP *with other association measures*

Let us now also assess the way(s) in which the results of the two Δ*P*-measures differs from four much more established but less precise standard measures, eight scatterplots representing the relevant correlations were created. In these plots in Figure 2, Δ*P*-measures are on the *x*-axis (Δ*P* $\text{word}_2|\text{word}_1$ in the top panels and Δ*P* $\text{word}_1|\text{word}_2$ in the bottom panels), the bidirectional measures are on the *y*-axis (from left to right: *MI*, $\log_{10}$ G^2/LL, $\log_{10}$ *t*, and Dice), and each of the 262 two-word units is indicated by a point (where overplotting leads to darker points). In addition, dashed lines median-dichotomize the coordinate system, a locally-weighted smoother summarizes the correlation, and the point representing *of course* is circled.

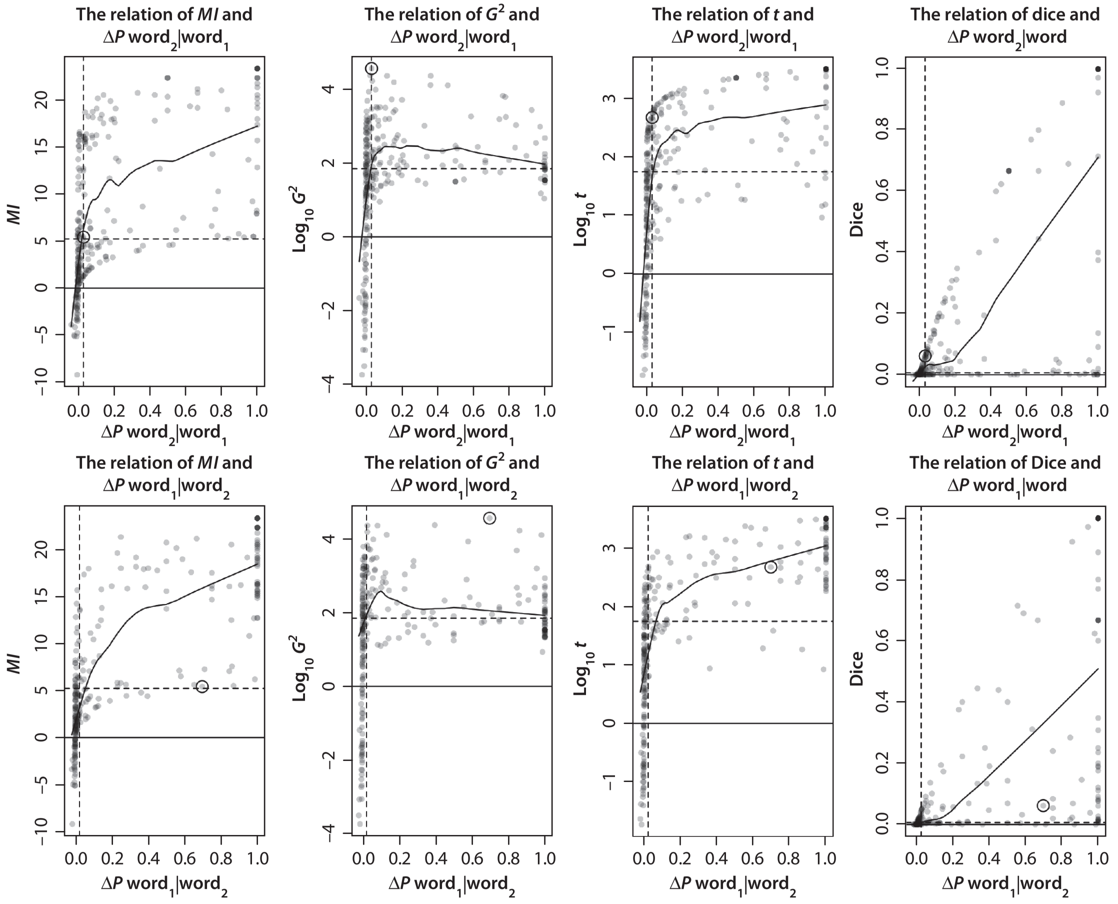


Figure 2. Pairwise correlations of both ΔP-measures and bidirectional collocational measures for 262 two-word units in the spoken part of the BNC

On the whole, the results paint a mixed picture at best. In fact, the results seem not so bad at first sight because the most widely-used measures – *MI*, G^2, and *t* – exhibit the larger than expected correlations one wants to see for elements that are two-word units, and the two Δ*P*-measures are often positively correlated with the bidirectional measures. However, the G^2 measure, probably the best or second-best measure on mathematical grounds as it is the best approximation to $p_{\text{Fisher-Yates}}$ (cf. also Evert 2009: 1235), is hardly correlated with either Δ*P*, and even the measures that do exhibit some sort of correlation – *MI* and *t* – exhibit a vast range of variability across the range of Δ*P*-values. It is this fact and the distribution of words to be discussed below with regard to Figure 3 that indicate clearly that bidirectional measures conflate two pieces of information that should probably not be conflated: $p(\text{word}_2|\text{word}_1)$ and $p(\text{word}_1|\text{word}_2)$. However, the extent of the problem becomes even more obvious when we explore the collocations for which the most extreme differences in the collocational statistics are observed.

2.3.3 *Bigrams with high* ΔP *differences*

While the previous section has shown that there are pronounced differences between the bidirectional measures and Δ*P*, it is only in a second step that one can really appreciate the exact nature of the differences, namely when one compares units for which particularly large differences are observed. Consider Figure 3.

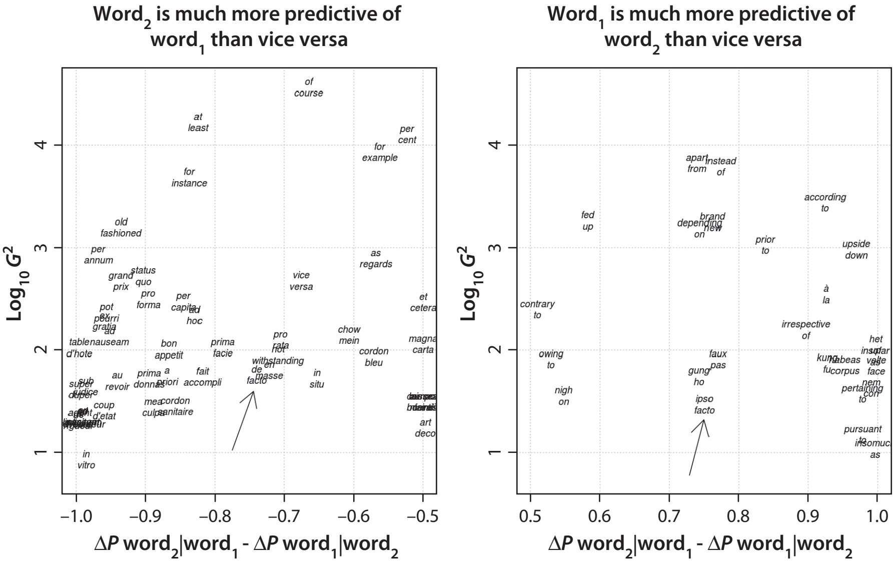


Figure 3. The distribution of two-word units according to G^2 (on the *y*-axis, logged) against the two extremes of $\Delta P\ \text{word}_2|\text{word}_1 - \Delta P\ \text{word}_1|\text{word}_2$[2]

The results clearly show how traditional measures often return high bidirectional associations for two words (reflected by high G^2-values in both panels) regardless of whether word$_1$ "selects" word$_2$ or word$_2$ "selects" word$_1$: high G^2-values are obtained across the whole range of ΔP-values. In the left panel, word$_2$ "selects" word$_1$ much more strongly than vice versa, and for many two-word units there such as *for instance*, *for example*, *old fashioned*, and *status quo*, to name but a few, that is quite obvious (esp. if word$_1$ is a frequent function word). Also, there are cases where the two-word unit is an expression in a language other than English, but its first component is also an English word, as in *pot pourri*, *coup d'etat*, or *grand prix*, which are nicely brought out as such by ΔP. Finally, for others in that panel, their being in the left panel may be less obvious because their non-English origin may make them appear as a single lexical item even more and may cloud an intuitive and immediate recognition of the direction of stronger attraction; to my first impression, *fait accompli* or *mea culpa* were cases in point.

By contrast, in the right panel, word$_1$ "selects" word$_2$ much more strongly than vice versa. Again, there are many obvious cases (esp. with function words as word$_2$), such as *instead of*, *according to*, *owing to*, *pertaining to*, etc.[3] Then, there is *volte face*, whose position is of course due to *face*'s promiscuity after many words, and finally some expressions that, again, I would have not expected in such an extreme position *a priori*, such as *kung fu*, *gung ho*, and *faux pas*.

A particularly nice illustration of the difference between directional and bidirectional association is the pair of *de facto* and *ipso facto*, which are highlighted by small arrows and whose statistics are summarily presented in Table 5. As is obvious, according to all bidirectional statistics – *MI*, *t*, G^2, and $p_{\text{Fisher-Yates}}$, the two words are highly attracted to each other, but what all of them fail to reveal is the directionality: given *ipso*, *facto* is a near certainty, but given *facto*, the association is less unambiguous, since words like *de*, *post*, and others are competing for the slot before it.

Table 5. Collocational statistics for *de facto* and *ipso facto* in the spoken component of the BNC

2-word unit	*MI*	*t*	Dice	G^2	$p_{\text{Fisher-Yates}}$	*MS*	$\Delta P_{2\mid 1}$	$\Delta P_{1\mid 2}$
de facto	14.65	278.02	0.02	58.2	$< 10^{-12}$	0.001	0.01	0.75
ipso facto	21.31	1613.22	0.4	29.82	$< 10^{-6}$	0.25	1	0.25

2.4 Validation 2: Randomly chosen bigrams

The above sections should have indicated the degree to which directional association measures can provide a wealth of information not available from the current standard of the discipline. However, it needs to be borne in mind that, so far, I only discussed a sample of two-word units for which strong effects were to be expected, given that the sample was based on expressions identified as multi-word units in the BNC. It is therefore necessary to demonstrate that the results obtained above are not invalidated by findings for words from a broader range of collocations, especially when collocations are included that, according to the standard set of measures, should exhibit a much wider range of associations, from rather low to maybe medium-high. This section is devoted to this issue.

In order to put the above results to the test, I generated a frequency list of all sentence-internal bigrams in the spoken component of the BNC. From that list, I sampled randomly 237 two-word collocations from eight logarithmically-defined frequency bins and computed the above standard bidirectional association measures as well as the two Δ*P*-measures.[4] As a first step, we can again compute the means of the association measures and the central 95% around the means. We find that, just like most traditional association measures, Δ*P* recognizes that these bigrams have been chosen randomly. While that of course does not mean that all their values are 0 or really small – after all, even completely unrelated words do not occur purely randomly in a corpus – the mean Δ*P*-values and the range of the central 95% are much smaller than before (and again include 0), as they should if Δ*P* is not overly sensitive.

Table 6. Some collocational statistics for a random sample of two-word collocations in the spoken component of the BNC

	MI	*t*	G^2	$\Delta P_{1\|2}$	$\Delta P_{2\|1}$	*MS*
mean	2.28	126.18	14687.63	0.08	0.05	0.03
0.025 quantile	−2.23	−15.08	0.24	−0.01	−0.01	0
0.975 quantile	6.31	802.39	147225.3	0.4	0.52	0.12

However, the last row of Table 6 also reveals that there are some high Δ*P*-values, which is why one should explore what the bigrams are for which high differences between Δ*P*-values are obtained. Again, in keeping with the fact that these bigrams were selected randomly, there are only very few instances – 8 out of 237 instances, i.e. 3.3% –where differences ≥0.5 between Δ*P*-measures are observed. And, the two-word collocations that do exhibit these large differences are in fact examples underscoring Δ*P*'s utility remarkably because a random-sampling of bigrams may of course sample some bigrams that are interesting, and this is what happened here: two of the five highly negative differences are discourse markers (*I mean* and *I think*), *I'm* is probably a single unit in speakers' linguistic systems, and *the faintest* is most likely in this group given the tendency of *faintest* to occur in *the faintest idea*. For the highly positive differences, the hedge *sort of* is *de facto* a multi-word unit, *can't* is of the same type as *I'm*, and *lack of* is arguably also intuitively plausible as a collocation that association measures should identify.

Table 7. Maximal differences ΔP word$_2$|word$_1$ minus ΔP word$_1$|word$_2$ for a random sample of two-word collocations in the spoken component of the BNC

the faintest	*I'm*	*I mean*	*the biggest*	*I think*	*sort of*	*ca n't*	*lack of*
−0.961	−0.889	−0.775	−0.697	−0.518	0.836	0.858	0.873

3. A directional association measure: Advantages and further applications

This study is of course not the first to "discover" that collocations exhibit directional effects. For example, Stubbs (2001: 29) discusses this using the examples of

- *bonsai*, which predicts *tree* to the right much more strongly than *tree* predicts *bonsai* to the left: while both Δ*P*s are relatively close to 0, their difference is two orders of magnitude in the BNC;
- *cushy*, which predicts *job* to the right much more strongly than *job* predicts *cushy* to the left: while both Δ*P*s are relatively close to 0, their difference is close to three orders of magnitude in the BNC.

Similar observations have been made by others, too: Kjellmer (1991) distinguishes left-predictive and right-predictive collocations (examples for right-predictive collocations he mentions are *Achilles heel* or *moot point*); Smadja (1993) approaches collocation extraction using a completely underused mean/variance-approach to positions of collocates around node words, which can also reveal the directionality of collocations; Bartsch (2004), Handl (2008), and Evert (2009: 1245) mention directionality of collocations, etc. However, up until even the most recent overview and testing studies of collocation/association measures (e.g. Pecina 2009), the issue of directionality has not received the attention that one of the most important notions in our field deserves; Michelbacher et al. (2007, 2011) seem to be the only dedicated studies. The remainder of this section is devoted to the potential and implications of a stronger emphasis on directional collocations. But first let me recap very briefly the main advantages that Δ*P* has to offer, before I venture off into increasingly more speculative and brainstorming-type of suggestions and, in Section 4, general desiderata for the study of collocations:

- obviously and as mentioned above, ΔP is more sensitive than all traditional measures because, unlike them, it can tease apart which collocates in a collocation exhibit the strongest or weakest amounts of attraction or repulsion to the other collocate(s);
- ΔP is very easy to compute: unlike many traditional measures it makes no distributional assumptions (normality, variance homogeneity, etc.), it involves neither complicated formulae nor computationally intensive exact tests, and, unlike Michelbacher et al.'s rank measures, it does not require (tens of) thousands of collocate tests to get a mere two rank scores for a single collocation;
- unlike many other statistics, ΔP is easy to understand: contrary to, say, the formula for G^2 or $p_{\text{Fisher-Yates}}$, it involves nothing but a mere difference of percentages, something even the least mathematically-inclined person will find easy to comprehend, but at the same time, it is not as utterly arbitrary as, say, Kilgarriff's (2009) add-*n* approach;
- ΔP is not a significance test and, thus, avoids all sorts of arguments often levelled (and many times erroneously so) against the significance testing paradigm; however, this does not mean that ΔP could not benefit from being combined with other information, most notably dispersion and/or frequency (as exemplified above in the right panel of Figure 1 where the frequency of a collocation was represented by the size of a bubble);
- ΔP has received quite some experimental support in studies on the psychology of learning (cf. Ellis 2007: 11f. for examples).

Given ΔP's results and above characteristics/advantages, one potentially interesting area could be lexicography. The fact that ΔP provides directionality information could make it an interesting and objective tool to use – together with other/existing heuristics, of course – when it comes to considering the directionality of dictionary entries, i.e. which part of a complex expression to choose as the headword under which an expression will be found. In cases where such a decision is not obvious – which it often will be, but maybe not always – one would make the part the headword that, as a cue, leads to a higher ΔP.

Note, however, that the first point, ΔP's increased sensitivity, has much more to offer than just different results to explore. The inbuilt directionality of ΔP should be especially important to the growing number of corpus linguists who view linguistics as a part of cognitive science and approach corpus data with a cognitive and/or psycholinguistically informed perspective. This is because, from such a perspective, it is very obvious that similarity in general is not necessarily symmetric – just because many widely-used statistical measures are symmetric (e.g. Dice, Jaccard) does not mean that adopting an asymmetric view on similarity would not be more cognitively realistic and (corpus-)linguistically revealing (cf. Tversky

1977 on asymmetric similarity and Shanks 1995 on asymmetric measures in the context of associative learning). As an example of applying this perspective to something at the very centre of corpus linguistics, consider two studies by Gries (2001, 2003). These studies applied this perspective to near-synonymous *-ic* and *-ical* adjectives such as *economic*/*economical*, *symmetric*/*symmetrical*, *alphabetic*/ *alphabetical*, etc. and showed that one can explore the similarity of the meanings of the two adjectives on the basis of how many collocates of one adjective are also collocates of the other.

The biggest contribution that I see such directional measures as making may be to help us improve the fit, or understand the lack of it, between corpus-linguistic and psycholinguistic data on, say, the association of words. Again, Michelbacher et al.'s work was groundbreaking in how they tested their measures against psycholinguistic reference data, and it is one of the goals of this study to inspire similar follow-up studies. For example, Mollin (2009), a study Michelbacher et al. (2011) also mention, demonstrates discrepancies and non-correlations between co-occurrence data from the BNC and association data from the Edinburgh Associative Thesaurus. However, Mollin's five kinds of co-occurrence data – raw co-occurrence frequency, *MI*, *z*, *MI*3, and G^2 – were all bidirectional, whereas association data are essentially directional, since one word is provided as a stimulus and the other is the response. It may therefore be the case that part of the unexplained variance in her correlations is simply due to the fact that the directionality of the word-association task had no proper counterpart in her corpus statistics; a reanalysis of these data *may* be interesting.[5] Similar comments apply more generally: psycholinguists, who have been more eager to problematize and test our corpus-linguistic lexical association measures than we have ourselves, have produced an array of results that are not always easy to reconcile: sometimes, bidirectional measures such as co-occurrence frequency or *MI* predict subjects' or speakers' patterning well, but sometimes unidirectional transitional probabilities (e.g. $^{a}/_{a+b}$) fare better. Again, it might be useful to shift the focus on to directional measures, in particular directional measures that relate a mere transitional probability $^{a}/_{a+b}$ to its counterpart $^{c}/_{c+d}$.

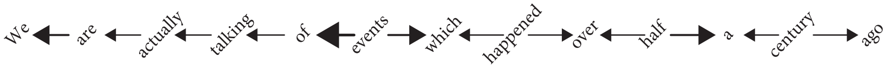
Just to state the obvious: this logic does of course not only apply to collocational studies nor is it restricted to the works of others. For example, Gries et al. (2005) tested which of several bidirectional corpus-based statistics – raw frequency of co-occurrence, p(verb in construction|verb), or a dichotomized collostruction strength as measured with $p_{\text{Fisher-Yates}}$ – predicted subjects' sentence completions best. In that experiment as well as a follow-up based on self-paced reading-time data, collostruction strength turned out to be by far strongest predictor. However, a re-analysis of these data that added ΔP construction|verb and its interactions to the predictors in a logistic regression – ΔP construction|verb because the subjects were given a sentence fragment ending in a verb and asked to complete the sentence with whatever construction they wanted – shows that the ΔP-measure is a significant predictor of subjects' sentence completions both on its own and in a significant interaction. Given all of the above, this may now not be surprising anymore, but the virtually complete absence of serious research on directional association measures at the time – in fact, till very recently – did not help us see this possible connection, and re-analyses of both collocational and colligational/collostructional studies may result in similar findings.

However, even for those corpus linguists who are not (yet) willing to endorse cognitive- or psycholinguistic perspectives on their data, ΔP provides much potential and food for thought. One application that may be interesting to explore is the extension of ΔP to the study of multi-word units involving more than two lexical items. Most studies that attempt to use association measures for multi-word units are based on *MI*, but use ready-made software tools that compute quick but dirty versions of the candidate expressions' *MI*-values. In these, the expected frequency is computed on the assumption of complete independence. For natural language, this assumption is of course wrong and, correspondingly, so are these estimates. However, as early as 1990, Jelinek proposed an iterative approach that is not only more useful to obtain better *MI*-values, but can also immediately be applied with ΔP-values.[6]

1. V_0 is the vocabulary of single words, $i = 0$, and λ is a user-defined threshold value;
2. find all pairs x, y in V_i for which $MI(x, y) > \lambda$;
3. let $\mathrm{V}_{\mathrm{i}+1}$ be V_i augmented with high *MI* pairs;
4. increment *i* and go back to step 2.

That is, in an iterative process, collocations bound together by high ΔP-values would be successively amalgamated into larger multi-word units, until no further amalgamation appears worthwhile. One step in such an approach as yet to be fleshed out is exemplified in Figure 4, which, for realistic testing, of course needs to be applied to large corpora.

Whatever the exact implementation will look like, I hope it is obvious that there are many ways in which thinking about directional association measures can inject new ideas into the study of co-occurrence and association measures.[7] Alas, in the next and final section, I want to outline a few ways in which *all* association measures would ideally be improved upon.



Prediction directionality of words in a sentence in the spoken component of the BNC

Figure 4. Δ*P*-values in a randomly-chosen sentence from D8Y.xml from the BNC. Arrows connect the word at the arrow's origin to the neighbouring word for which it is more predictive; the line width of arrows is proportional to Δ*P*

4. What's next and a more general exploration of what's wrong with our collocations

While an approach to association measures that can take the direction of association into consideration to a greater extent is useful, it is still not the end of the story. Put differently, if this paper does no more than stimulate some research on this, it has done one of its main jobs. However, much remains to be done in much more comprehensive ways, and several strategies to improve Δ*P* or similar directional measures suggest themselves.

The probably most obvious candidate is weighting the Δ*P*-values by, say, observed frequencies. The idea is that, all other things being equal, a Δ*P*-value carries more weight if the percentage $p(\mathrm{O}|\mathrm{C})$ – or the difference of the percentages – is based on more observations than on fewer ones, which is compatible with what we know about learning in general. It would be particularly useful to explore how the extraction of collocations or the prediction of, or correlation with, psycholinguistic reference data can be improved by such a weighting scheme.

A second candidate is just as obvious – especially since it applies to any corpus-linguistic frequency or association measure – but unfortunately is also very much under-researched: any frequency and any collocation measure – bidirectional or directional – should be weighted by the dispersion of the observations across corpora or corpus parts. Gries (2008b) shows that both frequencies as well as co-occurrence information (e.g. lexical collocations or lexico-grammatical colligations or collostructions) can hugely overestimate relevance as well as association strengths when dispersion is neglected. In a similar vein, Gries (2010a) demonstrates that some dispersion measures, or frequencies adjusted for dispersion, are more strongly correlated with psycholinguistic reaction times than simple observed frequencies. Thus, all other things being equal, a Δ*P*-value should carry more weight if the percentage $p(\mathrm{O}|\mathrm{C})$ or the difference of the percentages is based on observations that are more equally and widely dispersed in the corpus under consideration than on observations that are clumped together in only a small corpus part.

Unfortunately, the picture is even more complex than that – given the complexity of the data, it is actually amazing how often our often crude approximations work, an issue to which I will return below. To approach the issue at hand, consider Table 8, which exemplifies fictitious data that are well-behaved and studied in more detail than is often found: the co-occurrence of some word$_1$ and some word$_2$ is studied, and to account for dispersion, co-occurrence data are evaluated in three different corpus parts. Ideally, one would of course have many more corpus parts, but for the current purposes three parts will suffice; the crucial point to notice is that, in this data set, the collocation word$_1$ word$_2$ is quite similar in all three corpus parts A, B, and C (according to odds ratios and ΔP-values); the bold numbers will be explained below.

However, all standard association measures used in corpus linguistics are based only on the above types of 2×2 co-occurrence data: not only do they neither include directionality nor dispersion – they also do not include type frequencies of lexical items or constructions/patterns. Consider Table 9, which provides a more fine-grained resolution of the data in Table 8. Note that, while the bold figures in Table 8 and Table 9 are identical, implications and interpretations change, which is because the variation in cells *b*, *c*, and *d*, which remains neglected in Table 8, is now taken into consideration. In fact, the changes are remarkable. While the bold numbers that most analysts only consider have not changed, the *implications* of the data are as different as can be, especially when one tries to also understand how speakers store all sorts of linguistic and extra-linguistic probabilistic information in their minds.

How are these implications different? This can be demonstrated on the basis of a polemic thought experiment in three parts. First, how do most traditional corpus linguists studying word$_1$ proceed these days? They look at the table that arises from combining the three subtables of Table 8 ($a=77$, $b=174$, $c=270$, $d=29150$), notice the somewhat large *MI*-value (4.71) and maybe even the large odds ratio (47.78) and go on to interpret word$_1$'s semantics on the basis of the strong and significant association with word$_2$.

The second perspective is already much better and, unfortunately, also much rarer. Given the above, a more progressive corpus linguist studying word$_1$ would not be satisfied with the first approach and (i) consider also the association of word$_1$ with word$_2$ in the three subcorpora separately (dispersion) and (ii) maybe even also use an additional directional association measure. This corpus linguist would also interpret word$_1$'s semantics on the basis of the strong and significant association with word$_2$, and maybe so with more confidence because the findings are robustly attested in three different corpus parts.

Table 8. Fictitious co-occurrence data of words word$_1$ and word$_2$ in three corpus parts

corpus part A	word$_2$	other words	Totals
word$_1$	**30**	60	**90**
other words	100	9800	9900
Totals	**130**	9860	**9990**

corpus part B	word$_2$	other words	Totals
word$_1$	**22**	62	**84**
other words	70	9600	9670
Totals	**92**	9662	**9754**

corpus part C	word$_2$	other words	Totals
word$_1$	**25**	52	77
other words	100	9750	9850
Totals	**125**	9802	**9927**

Table 9. Fictitious co-occurrence data of words word$_{1–30}$ in three corpus parts

corpus part A	word$_2$	word$_3$	word$_4$	word$_5$	word$_6$	20 words$_{7–26}$	Totals
word$_1$	**30**	60	0	0	0	0	**90**
word$_{27}$	28	1	1250	1180	250	0	2709
word$_{28}$	32	0	770	600	1900	979	4281
word$_{29}$	21	0	280	557	0	1200	2058
word$_{30}$	19	0	200	163	350	120	852
Totals	**130**	61	2500	2500	2500	2299	**9990**

corpus part B	word$_2$	word$_3$	word$_4$	word$_5$	word$_6$	20 words$_{7–26}$	Totals
word$_1$	**22**	**30**	**30**	**1**	**1**	**0**	**84**
word$_{27}$	20	1	0	2	1	3022	3046
word$_{28}$	30	0	1	3	0	4280	4314
word$_{29}$	14	0	1	0	4	1850	1869
word$_{30}$	6	1	0	0	1	433	441
Totals	**92**	32	32	6	7	9585	**9754**

corpus part C	word$_2$	word$_3$	word$_4$	word$_5$	word$_6$	20 words$_{7–26}$	Totals
word$_1$	**25**	2	2	1	1	46	77
word$_{27}$	30	0	0	0	0	2980	3010
word$_{28}$	20	0	0	1	0	2100	2121
word$_{29}$	10	0	1	0	0	2468	2479
word$_{30}$	40	1	0	0	0	2199	2240
Totals	**125**	3	3	2	1	9793	**9927**

Only the third and virtually non-existent approach towards studying word$_1$ is the really fruitful one, however. It is this corpus linguist that would notice that, to study word$_1$,

- in corpus part A, one should talk much more about word$_3$ than about word$_2$: word$_1$ occurs only with two different types at all (word$_2$ and word$_3$) and the entropy of this distribution is very low because word$_3$ is much more strongly associated with word$_1$ than word$_2$ (as can be seen, for example, from the Pearson residuals for this table);[8]
- in corpus part B, one should talk much more about word$_3$ and word$_4$ than about word$_2$: word$_1$ occurs with four different types (word$_{2–5}$) and the entropy is somewhat higher because word$_3$ and word$_4$ both exhibit a high association to word$_1$ than word$_2$;
- in corpus part C, one should focus on word$_2$: word$_1$ occurs with many different types (note the rightmost column) and word$_2$'s association to word$_1$ is by far the strongest.

In sum, association measures should be weighted by, or incorporate in some other way, the nature, or entropy, of the type-token distributions involved. Note that this is a more stringent requirement going beyond even the one and only association measure that includes type frequencies, lexical gravity *G*, as proposed by Daudaravičius & Marcinkevičienė (2004). Lexical gravity *G* is extremely interesting because its computation of the association between word$_1$ and word$_2$ takes the number of types with which word$_1$ and word$_2$ are attested into consideration. Initial comparative studies have shown very promising results: in Gries (2010b), bigram gravities allow to perfectly recover the sampling structure of corpora and outperform *t*; in Gries & Mukherjee (2010), it is demonstrated how *G* can be extended to multi-word utterances, and Ferraresi & Gries (2011) show how *G* allows to recover a type of collocations qualitatively different from more traditional measures.

Given *G*'s incorporation of type frequencies, why is this still not sufficient? Why do we still need a better measure than that, one that includes the entropy of the type-token distribution? The answer to this question is obvious from (the first rows of both panels in) Table 10.

The problem that will ultimately have to be addressed is that gravity *G* could treat these two distributions identically since it is only concerned with the number of types word$_1$ occurs with, and in this case, word$_1$ occurs with six other types in both cases. However, again for a linguist it is plain to see that the distributions are very different in how strongly word$_2$ and word$_3$ are associated with word$_1$. And intuitively we all know this: few corpus linguists would be surprised by the results for *de facto* and *ipso facto* discussed above (cf. Table 5). Recall that *facto*

Table 10. Fictitious co-occurrence data of words word$_{1-7}$ in two corpus parts

corpus part A	**word$_2$**	**word$_3$**	**word$_4$**	**word$_5$**	**word$_6$**	**word$_7$**	**Totals**
word$_1$	**30**	56	1	1	1	1	**90**
other words	**100**	3	3	3264	3265	3265	**9900**
Totals	**130**	59	4	3265	3266	3266	**9990**

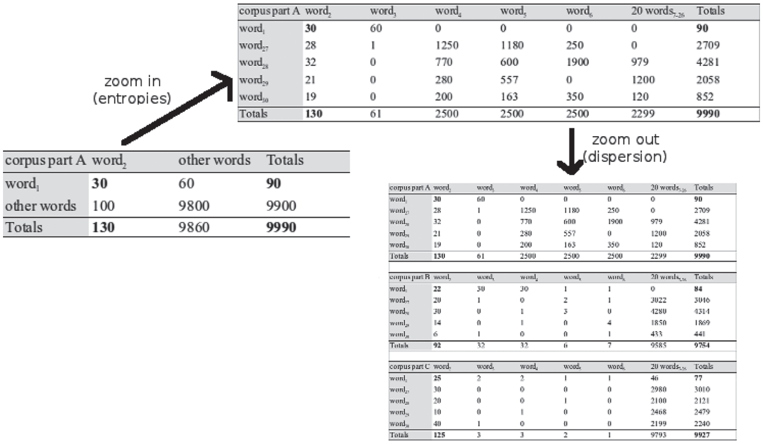
corpus part B	**word$_2$**	**word$_3$**	**word$_4$**	**word$_5$**	**word$_6$**	**word$_7$**	**Totals**
word$_1$	**30**	29	28	1	1	1	**90**
other words	**100**	3	3	3264	3265	3265	**9900**
Totals	**130**	32	31	3265	3266	3266	**9990**

was more predictive of *de* than vice versa while *ipso* was more predictive of *facto* than vice versa. Why is that? Type-token distributions of course: nothing much other than *facto* happens after *ipso*, which means the entropy of that distribution will be minimal. But there are a few words that can precede *facto*: *ipso*, *de*, *post*, and maybe more, which is why the entropy of *that* distribution will be higher and which is why one should not only consider type frequency, but also the token frequency distributions of types, as argued above.

To wrap up, collocation has been, and will remain, one of the most important concepts in corpus linguistics. However, after many decades of "more of the same", of proposing literally dozens of measures based on 2×2 tables that hide much of the interesting variability in the data, it is time to explore new ways of studying collocations:

– *directional measures*: what directionality effects do nearly all of our measures miss?
– *dispersion*: how homogeneous are associations really across corpora or corpus parts?
– *type-token distributions and/or their entropies*: what huge amounts of variability in the *b*, *c*, and *d* cells of our simplistic co-occurrence tables do we miss?
– *extendability to multi-word units*: how do we best approach association measures for multi-word units?

In other words, we need to be aware of the larger dimensionality of our data in its contexts: instead of focusing on tables such as the first panel in Table 8, we need to (i) zoom in to discover the variability in cells *b*, *c*, and *d* and quantify it in terms of entropy etc. but also (ii) zoom out to explore the variability of such panels in different corpus parts (cf. Figure 5). The current state of the art of the field is to remain at the simple and cushy left part when what we should be doing is exploring the variability both below and above where we currently are.



corpus part A	$word_2$	other words	Totals
$word_1$	**30**	60	**90**
other words	100	9800	9900
Totals	**130**	9860	**9990**

corpus part A	$word_2$	$word_3$	$word_4$	$word_5$	$word_6$	20 $words_{7\text{-}26}$	Totals
$word_1$	**30**	60	0	0	0	0	**90**
$word_{27}$	28	1	1250	1180	250	0	2709
$word_{28}$	32	0	770	600	1900	979	4281
$word_{29}$	21	0	280	557	0	1200	2058
$word_{30}$	19	0	200	163	350	120	852
Totals	**130**	61	2500	2500	2500	2299	**9990**

corpus part A	$word_2$	$word_3$	$word_4$	$word_5$	$word_6$	20 $words_{7\text{-}26}$	Totals
$word_1$	**30**	60	0	0	0	0	**90**
$word_{27}$	28	1	1250	1180	250	0	2709
$word_{28}$	32	0	770	600	1900	979	4281
$word_{29}$	21	0	280	557	0	1200	2058
$word_{30}$	19	0	200	163	350	120	852
Totals	**130**	61	2500	2500	2500	2299	**9990**

corpus part B	$word_2$	$word_3$	$word_4$	$word_5$	$word_6$	20 $words_{7\text{-}26}$	Totals
$word_1$	**22**	30	30	1	1	0	**84**
$word_{27}$	20	1	0	2	1	3022	3046
$word_{28}$	30	0	1	3	0	4280	4314
$word_{29}$	14	0	1	0	4	1850	1869
$word_{30}$	6	1	0	0	1	433	441
Totals	**92**	32	32	6	7	9585	**9754**

corpus part C	$word_2$	$word_3$	$word_4$	$word_5$	$word_6$	20 $words_{7\text{-}26}$	Totals
$word_1$	**25**	2	2	1	1	46	**77**
$word_{27}$	30	0	0	0	0	2980	3010
$word_{28}$	20	0	0	1	0	2100	2121
$word_{29}$	10	0	1	0	0	2468	2479
$word_{30}$	40	1	0	0	0	2199	2240
Totals	**125**	3	3	2	1	9793	**9927**

Figure 5. Schematic representation of the required zooming in and zooming out

Given all the variability we are missing with this strategy, it is a legitimate question to ask why we still often get reasonably good results. We get reasonably good results – but often also not so good results or results that clash with experimental or other data – because the cells in our regular 2×2 co-occurrence tables are a proxy for dispersion and entropy, just not a very good and reliable one. Why is that so? Three reasons: first, the larger the observed frequencies are, the more widely dispersed our data will be, and the quality of our results will be proportional to the sizes of the frequencies in our tables.

Second, in corpus linguistics, nearly everything we count is Zipfian distributed. That means, the numbers of tokens represented by *b*, *c*, and *d* will come from, nearly irrespective of how large *b*, *c*, and *d* actually are, a Zipfian distribution of types and their frequencies that, when summed over, make up *b*, *c*, and *d*: few types will have high frequencies, and many types will have low frequencies. Thus, when *b*, *c*, and/or *d* are large, then, in general, we are more likely to have a larger number of types, and when they are small then we are more likely to have a smaller number of types. Thus, the quality of our results will be proportional to the degree that the Zipfian distributions (and their entropies) that underlie the frequencies in cells *b*, *c*, and *d* are similar.

Third, there is recent work that strongly supports the notion that simple low-dimensionality frequency counts are much less important than the kind of higher-dimensional data made of multidimensional conditional probabilities, entropies, etc.: in a truly pioneering study, Baayen (2010) provides comprehensive evidence for the assumption that the kind of simple frequency effects corpus linguists and psycholinguists often invoke merely arise out of learning a wide range of more specific distributional patterns in the context of expressions, and given my above pleas, the following is worth quoting in detail (my emphases):

> A principal components analysis of 17 lexical predictors revealed that most of the variance in lexical space is carried by a principal component on which *contextual measures* (syntactic family size, *syntactic entropy, BNC dispersion*, morphological family size, and *adjectival relative entropy*) have the highest loadings. Frequency of occurrence, in the sense of pure repetition frequency, explains only a modest proportion of lexical variability. Furthermore, the principal component representing local syntactic and morphological diversity accounted for the majority of the variability in the response latencies. (Baayen 2010: 456)

Findings like these have the potential to bring about no less than a paradigm shift in corpus linguistics and support the above argumentation and my plea to advance corpus linguistics to the next level, one where we begin to (i) leave behind simplistic frequencies of (co-)occurrence and take the high dimensionality of our data as seriously as it needs to be taken and (ii) embrace the connections of corpus linguistics to psycholinguistics that Baayen's work so beautifully highlights.

In sum, often results we obtain for phenomenon *P* are reasonably good because of a favourable interplay of (i) large corpus frequencies of *P* (increasing the chance that *P* is nicely dispersed) and, something much less under our control, namely (ii) that the type-token distributions, which we simplistically conflate in cells *b*, *c*, and *d* and then conveniently ignore, happen to be Zipfian and similar enough to not cause problems.

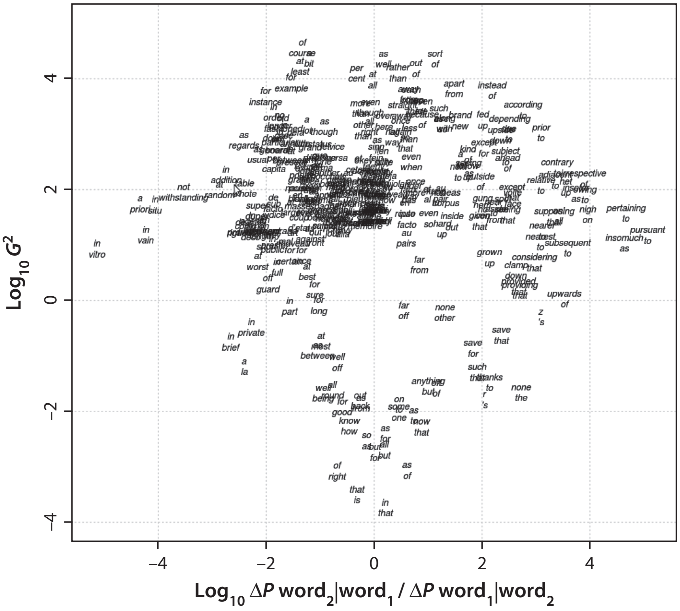
Thus, I believe we can drastically increase our chances (quite literally) of improving both our results and the match of our methods to current cognitive and psycholinguistic theories – exemplar-based models being the most obvious candidates; cf. Gries (2012) – by following the above plan of research. If we were to pursue even only one of these avenues, that would already breathe much new and needed life into this old topic – now just imagine we began to tackle them all!

Notes

* This paper is dedicated to Michael Stubbs in grateful recognition of how his contributions in the field of phraseology and corpus-based lexical semantics (in particular his 2001 book *Words and Phrases*) shaped my thinking. I thank two reviewers and the special issue editors for comments and feedback. The usual disclaimers apply.

1. A function for R (R Development Core Team 2012) to compute Δ*P*s for 2×2 tables is available from the author upon request.

2. An alternative representation that is sometimes more useful involves not the actual difference of the Δ*P*-values on the *x*-axis but the logged ratio of the Δ*P*-values. While this does not result in a nicely delimited *x*-axis anymore, which has as its limits the theoretical minimum and maximum of Δ*P*, it stretches the representation more, making the (more interesting) margins more legible.



3. In a corpus-linguistic journal, I cannot resist pointing out how nicely the position of *habeas corpus* supports the present argument: of course there can be many words in front of *corpus* as word$_2$ (*balanced*, *large*, *national*, *representative*, *small*, …), but *habeas* as word$_1$ does not leave open many choices.

4. I determined the maximal frequency of a sentence-internal bigram in the subcorpus *mf*, computed the natural logarithm of *mf*, divided that by eight, multiplied that result with each of the numbers from 1 to 8, and those eight products were antilogged to arrive at a frequency at every order of magnitude. If enough two-word units with that frequency were available, a random sample was drawn from them – otherwise, the missing two-word units were filled up by those with the next closest frequencies of occurrence.

5. Cf. McGee (2009) for a similar study, comparing co-occurrence frequency to the results of a word-association experiment; cf. also Nordquist (2009) for discussion of the mismatch of corpus and elicited data from a usage-based perspective.

6. Cf. also Zhang et al. (2009) for a recent modification of *MI* – *EMI* – and its applicability to multi-word units; a function for R (R Development Core Team 2012) to compute *EMI* for 2×2 tables is available from the author upon request.

7. This kind of application may even relate ΔP and the Minimum Sensitivity statistic *MS* mentioned above, but in a way that seems more appropriate. Essentially, what the approach in Figure 4 amounts to is not choosing the *minimum* of $^{a}/_{a+b}$ and $^{a}/_{a+c}$ (*MS*), but the *maximum* of ΔP $\text{word}_2|\text{word}_1$ and ΔP $\text{word}_1|\text{word}_2$. A student of mine has now used this logic to test whether two kinds of ΔP-values – ΔP $\text{word}_2|\text{word}_1$ or $max(\Delta P\ \text{word}_2|\text{word}_1, \Delta P\ \text{word}_1|\text{word}_2)$ – can distinguish bigrams within an intonation unit from bigrams that span an intonation unit boundary (in the Santa Barbara Corpus of Spoken American English). As was to be expected, the latter approach, $max(\Delta P\ \text{word}_2|\text{word}_1, \Delta P\ \text{word}_1|\text{word}_2)$, fared much better, because it is this approach that can assign a high value to bigrams that will hardly ever span an IU boundary but whose word_2 is more predictive of word_1 than vice versa (such as *of course*); cf. Wahl (2011).

8. The entropy of distribution can be understood as the amount of uncertainty coming with that distribution. If 99 items in an urn are distributed across three colours as {red: 33, green: 33, blue 33}, then this distribution is least informative (relative entropy = 1) because, for instance, if I asked someone what the colour of a ball I draw randomly out of that urn would be, knowing the three colours are equally frequent does not help in guessing the next ball's colour. If, on the other hand, the distribution of colours was this {red: 99, green: 0, blue 0}, then that distribution is very informative (relative entropy = 0) because now everyone would know to "guess" *red*. Thus, entropy quantifies the variability of a distribution.

References

Baayen, R. H. 2010. "Demythologizing the word frequency effect: A discriminative learning perspective". *The Mental Lexicon*, 5 (3), 436–461.

Bartsch, S. 2004. *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Tübingen: Gunter Narr.

Bell, A., Brenier, J. M., Gregory, M., Girand, C. & Jurafsky, D. 2009. "Predictability effects on durations of content and function words in conversational English". *Journal of Memory and Language*, 60 (1), 92–111.

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

Ellis, N. C. 2007. "Language acquisition as rational contingency learning". *Applied Linguistics*, 27 (1), 1–24.

Ellis, N. C. & Ferreira-Junior, F. 2009. "Constructions and their acquisition: Islands and the distinctiveness of their occupancy". *Annual Review of Cognitive Linguistics*, 7, 187–220.

Evert, S. 2005. *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. Ph.D. thesis. Stuttgart: University of Stuttgart.

Evert, S. 2009. "Corpora and collocations". In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, Vol. 2. Berlin/New York: Mouton de Gruyter, 1212–1248.

Ferraresi, A. & Gries, St. Th. 2011. "Type and (?) token frequencies in measures of collocational strength: Lexical gravity vs. a few classics". Paper presented at *Corpus Linguistics 2011, University of Birmingham, UK*.

Firth, J. R. 1957. "A synopsis of linguistic theory 1930–1955". In F. Palmer (Ed.), *Selected Papers of J. R. Firth 1952–1959*. London: Longman, 168–205.

Gries, St. Th. 2001. "A corpus-linguistic analysis of *-ic* and *-ical* adjectives". *ICAME Journal*, 25, 65–108.

Gries, St. Th. 2003. "Testing the sub-test: A collocational-overlap analysis of English *-ic* and *-ical* adjectives". *International Journal of Corpus Linguistics*, 8 (1), 31–61.

Gries, St. Th. 2008a. "Phraseology and linguistic theory: A brief survey". In S. Granger & F. Meunier (Eds.), *Phraseology: An Interdisciplinary Perspective*. Amsterdam: John Benjamins, 3–25.

Gries, St. Th. 2008b. "Dispersions and adjusted frequencies in corpora". *International Journal of Corpus Linguistics*, 13 (4), 403–437.

Gries, St. Th. 2010a. "Dispersions and adjusted frequencies in corpora: Further explorations". In S. Th. Gries, S. Wulff & M. Davies (Eds.), *Corpus Linguistic Applications: Current Studies, New Directions*. Amsterdam: Rodopi, 197–212.

Gries, St. Th. 2010b: online. "Bigrams in registers, domains, and varieties: A bigram gravity approach to the homogeneity of corpora". In M. Mahlberg, V. González-Diaz & C. Smith (Eds.), *Proceedings of the Corpus Linguistics Conference (CL 2009), University of Liverpool, UK, 20–23 July 2009*. Available at: http://ucrel.lancs.ac.uk/publications/cl2009 (accessed July 2012).

Gries, St. Th. 2012. "Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: Towards more and more fruitful exchanges". In J. Mukherjee & M. Huber (Eds.), *Corpus Linguistics and Variation in English: Theory and Description*. Amsterdam: Rodopi, 41–63.

Gries, St. Th., Hampe, B. & Schönefeld, D. 2005. "Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions". *Cognitive Linguistics*, 16 (4), 635–676.

Handl, S. 2008. "Essential collocations for learners of English: The role of collocational direction and weight". In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, 43–66.

Jelinek, F. 1990. "Self-organized language modeling for speech recognition". In A. Waibel & K.-F. Lee (Eds.), *Readings in Speech Recognition*. San Mateo, CA: Morgan Kaufmann, 450–506.

Kilgarriff, A. 2009. "Simple maths for keywords". Paper presented at *Corpus Linguistics 2009, University of Liverpool*.

Kjellmer, G. 1991. "A mint of phrases". In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honor of Jan Svartvik*. London: Longman, 111–127.

McGee, I. 2009. "Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores". *Corpus Linguistics and Linguistic Theory*, 5 (1), 79–103.

Michelbacher, L., Evert, S. & Schütze, H. 2007. "Asymmetric association measures". Paper presented at the 6th *International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria*.

Michelbacher, L., Evert, S. & Schütze, H. 2011. "Asymmetry in corpus-derived and human word associations". *Corpus Linguistics and Linguistic Theory*, 7 (2), 245–276.

Mollin, S. 2009. "Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations". *Corpus Linguistics and Linguistic Theory*, 5 (2), 175–200.

Nordquist, D. 2009. "Investigating elicited data from a usage-based perspective". *Corpus Linguistics and Linguistic Theory*, 5 (1), 105–130.

Pecina, P. 2009. "Lexical association measures and collocation extraction". *Language Resources and Evaluation*, 44 (1–2), 137–158.

Pedersen, T. 1998. "Dependent bigram identification". In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), July 28–30*, 1197.

R Development Core Team. 2012: online. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org (accessed July 2012).

Raymond, W. D. & Brown, E. L. 2012. "Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish". In St. Th. Gries & D. S. Divjak (Eds.), *Frequency Effects in Language Learning and Processing*. Berlin/New York: Mouton de Gruyter, 35–52.

Shanks, D. R. 1995. *The Psychology of Associative Learning*. New York: Cambridge University Press.

Smadja, F. 1993. "Retrieving collocations from text: Xtract". *Computational Linguistics*, 19 (1), 143–177.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford/Malden, MA: Blackwell.

Tversky, A. 1977. "Features of similarity". *Psychological Review*, 84 (4), 327–352.

Wahl, A. R. 2011. "Intonation unit boundaries and the entrenchment of collocations: Evidence from bidirectional and directional association measures". Unpublished ms, Department of Linguistics, University of California, Santa Barbara.

Wiechmann, D. 2008. "On the computation of collostruction strength: Testing measures of association as expressions of lexical bias". *Corpus Linguistics and Linguistic Theory*, 4 (2), 253–290.

Zhang, W., Yoshida, T., Tang, X. & Ho, T.-B. 2009. "Improving effectiveness of mutual information for substantival multiword expression extraction". *Expert Systems with Applications*, 36 (8), 10919–10930.

Author's address

Stefan Th. Gries
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100
United States of America

stgries@linguistics.ucsb.edu