Analyzing co-occurrence data

Stefan Th. Gries
University of California, Santa Barbara &
Justus Liebig University Giessen

Philip Durrant
University of Exeter

Abstract
In this chapter, we provide an overview of quantitative approaches to co-occurrence data. We begin with a brief terminological overview of different types of co-occurrence that are prominent in corpus-linguistic studies and then discuss the computation of some widely-used measures of association used to quantify co-occurrence. We present two representative case studies, one exploring lexical collocation and learner proficiency, the other creative uses of verbs with argument structure constructions. In addition, we highlight how most widely-used measures actually all fall out from viewing corpus-linguistic association as an instance of regression modeling and discuss newer developments and potential improvements of association measure research such as utilizing directional measures of association, not uncritically conflating frequency and association-strength information in association measures, type frequencies, and entropies.

1 Introduction

1.1 General introduction

One of the, if not *the*, most central assumptions underlying corpus-linguistic work is captured in the so-called distributional hypothesis, which holds that linguistic elements that are similar in terms of their distributional patterning in corpora also exhibit some semantic or functional similarity. Typically, corpus linguists like to cite Firth's (1957:11) famous dictum "[y]ou shall know a word by the company it keeps" but Harris's (1970:785f.) following statement actually makes the same case much more explicitly, or much more operationalizably:

> [i]f we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference of meaning correlates with difference of distribution.

That is, a linguistic expression *E* – a morpheme, word, construction/pattern, … – can be studied by exploring what is co-occurring with *E* and how often. Depending on what the elements of interest are whose co-occurrence is studied, different terms have been used for such co-occurrence phenomena:

- lexical co-occurrence, i.e. the co-occurrence of words with other words such as the strong preference of *hermetically* to co-occur with, or more specifically, be followed by, *sealed*, is referred to as *collocation*; for collocations, it is important to point out the locus of the co-occurrence and Evert (2009:1215) distinguishes between (i) surface co-occurrence (words that are not more than a span/window size of *s* words apart from each other; often *s* is 4 or 5), (ii) textual co-occurrence (words in the same clause, sentence, paragraph, …), and (iii) syntactic co-occurrence (words in a syntactic relation);

- lexico-grammatical co-occurrence, i.e. the co-occurrence of words with grammatical patterns or constructions such as the strong preference of the verb *regard* to be used in the passive of the *as*-predicative (e.g., *The Borg were regarded as the greatest threat to the Federation*) is referred to as *colligation* or *collostruction* (see McEnery, Xiao, & Tono 2006:11 or Stefanowitsch & Gries 2003).[1]

Different studies have adopted different views on how collocation in particular, but also co-occurrence more broadly, should be approached – how many elements are considered (two or more)? Do we need a minimum observed frequency of occurrence in some corpus? is a certain degree of unpredictability/idiosyncrasy that the co-occurrence exhibits a necessary condition for collocation status? etc. Also, co-occurrence applications differ in their retrieval procedures: Studies that target a word or a construction may retrieve all instances of the word/construction in question and explore its co-occurring elements; other studies might approach a corpus with an eye to identify all (strong) collocations for lexicographic, didactic, contrastive, or other purposes. For the sake of generality, we will discuss here a somewhat atheoretical notion of co-occurrence that eschews commitments regarding all of the above questions and is based only on some mathematical relation between the observed co-occurrence and non-co-occurrence frequencies of l elements in a corpus; it goes without saying that different research questions or practical applications may require one or more commitments regarding the above questions (see Bartsch 2004, Gries (2008b), and Evert (2009) for more discussion of the parameters underlying co-occurrence and their historical development).

The simplest possible way to explore a linguistic element (such as *hermetically* or *regard*) would be by raw co-occurrence frequency – how often do I find the collocation *hermetically sealed* in my corpus? – or, more likely, conditional probabilities such as p(contextual element(s)|E) – how likely is a verbal construction to be an *as*-predicative when the verb in the construction is *regard*?

While obtaining sorted frequency lists that reveal which collocates or constructions occur most often or are most likely around an element E is straightforward, much corpus-linguistic research has gone a different route and used more complex measures to separate the wheat (linguistically revealing co-occurrence data) from the chaff (the fact that certain function words such as *the, of*, or *in* occur with everything a lot, *qua* their overall high frequency. Such measures are often referred to as *association measures* (AMs) simply because they, typically, quantify the strength of mutual association between two elements such as two words or a word and a construction. In the following section, we discuss fundamental aspects of the computation of some of the most widely-used AMs.

2 Fundamentals

For decades now, AMs have typically been explained on the basis of co-occurrence tables of the kind exemplified in Table 1, which contain observed frequencies of (co-)occurrence of a linguistic expression E (for instance a particular word) and one of the l types of contextual elements X (e.g. other words or constructions $X_{1\text{-}l}$ can occur with/in); for instance, if the ditransitive construction is attested with l=80 verb types in a corpus, one would generate 80 such co-occurrence tables. In each such table, cell a is the frequency with which E is observed with/in element X, cell b is the frequency with which E is observed without X, this means the overall frequency of E is a+b, etc. Often, such a table would also contain or at least refer to the corresponding expected frequencies in the same cells a to d, i.e. the frequencies with which X

and *E* would be observed together and in isolation if their occurrences were completely randomized; these frequencies are computed from the row and column totals as indicated in Table 1 as they would be for, say, a chi-squared test.

Table 1 Schematic co-occurrence frequency table

	Co-occurring element *X*	Other elements (not *X*)	Row totals
Element *E*	obs.: *a* exp.: $^{(a+b)\times(a+c)}/_{n}$	obs.: *b* exp.: $^{(a+b)\times(b+d)}/_{n}$	*a*+*b*
Other elements (not *E*)	obs.: *c* exp.: $^{(c+d)\times(a+c)}/_{n}$	obs.: *d* exp.: $^{(c+d)\times(b+d)}/_{n}$	*c*+*d*
Column totals	*a*+*c*	*b*+*d*	*a*+*b*+*c*+*d*=*n*

As mentioned above, such a co-occurrence table is generated for every element type $X_{1\text{-}l}$ ever occurring with *E* at least once or, if the element analyzed is *X*, then such a co-occurrence table is generated for every element type $E_{1\text{-}l}$ ever occurring with *X* at least once. For instance, if one studied the *as*-predicative construction, then *X* might be that construction and elements $E_{1\text{-}l}$ could be all verbs occurring in that construction at least once and one could use the values in each of the *l* tables to compute an AM for every one of the *l* verb types of *E* co-occurring in *X* to. These results could then be used to, for instance, rank-order and then study them by strength of attraction which is often interesting because of how expressions that co-occur with *X* reveal structural and/or functional characteristics of *E* (recall the Firth and Harris quotes from above).

A large number of AMs has been proposed over the last few decades, including (i) measures that are based on asymptotic or exact significance tests, (ii) measures from, or related to, information theory, (iii) statistical effect sizes, various other measures or heuristics; Evert (2009) and Pecina (2010) discuss more than altogether 80 measures and since then even more measures have been proposed. However, the by far most widely-used measures are (i) the loglikelihood measure G^2 (which is somewhat similar to the chi-squared test and, thus, the *z*-score, and which is highly correlated with the *p*-value of the Fisher-Yates exact test as well as the *t*-score, (ii) the pointwise Mutual Information (*MI*), (iii) the odds ratio (and/or its logged version), which are all exemplified here in Table 2 on the basis of the frequencies of the co-occurrence of *regard* and the *as*-predicative in the British Component of the International Corpus of English reported in Gries, Hampe, & Schönefeld (2005).

Table 2 Co-occurrence frequencies of *regard* and the *as*-predicative in Gries, Hampe, & Schönefeld (2005)

	As-predicative	Other constructions	Row totals
regard	80 exp.: $^{99\times687}/_{138,664}$	19 exp.: $^{99\times137,977}/_{138,664}$	99
Other verbs	607 exp.: $^{138,565\times687}/_{138,664}$	137,958 exp.: $^{138,565\times137,977}/_{138,664}$	138,565
Column totals	687	137,977	138,664

(1) $$G^2 = 2\sum_{i=1}^{4} obs \times \log\frac{obs}{exp} \approx 762.196$$

(2) $$t = \frac{a - a_{\exp}}{\sqrt{a}} = \frac{80 - 0.49}{\sqrt{80}} \approx 8.889$$

(3) $$\textit{pointwise Mutual Information} = \log_2 \frac{a}{a_{\exp}} = \log_2 \frac{80}{0.49} \approx 7.349$$

(4) $$\textit{odds ratio} = \frac{a}{b} / \frac{c}{d} = \frac{80}{19} / \frac{607}{137958} = \frac{a}{c} / \frac{b}{d} \approx 956.962 \quad (\log \textit{odds ratio} \approx 6.864)$$

All four measures indicate that there is a strong mutual association between *X* (the *as*-predicative) and *E* (*regard*); if one computed the actual *p*-value following from this G^2, one would obtain a result of $p<10^{-167}$.[2] However, this sentence also points to what has been argued to be a shortcoming of these measures: The fact that they quantify *mutual* attraction means that they do not distinguish between different kinds of attracted elements:

- instances of collocations/collostructions where *X* attracts *E* but *E* does not attract *X* (or at least much less so);
- instances where *E* attracts *X* but *X* does not attract *E* (or at least much less so);
- instances where both elements attract each other (strongly).

Based on initial discussion by Ellis (2007), Gries (2013a) has shown that each of these three kinds of collocations is common among the elements annotated as multi-word units in the British National Corpus:

- *according to* or *upside down* are examples of the first kind: If one picks any bigram that has *to* or *down* as its second word, it is nearly impossible to predict which words will precede it, but if one picks any bigram with *according* or *upside* as the first word, one is quite likely to guess the second one correctly;
- *of course* or *for instance* are examples of the second kind: If one picks any bigram with *of* or *for* as the first word, it is nearly impossible to predict which word will follow, but if one picks any bigram with *course* or *instance* as the first word, one is quite likely to guess that *of* or *for* are the first word correctly;
- *Sinn Fein* and *bona fide* are examples of the third kind: each word is very highly predictive of the other.

Crucially, all of the above examples are highly significant – in the spoken part of the BNC, all have G^2-values of >178 and *p*-values of $<10^{-40}$ – but they are clearly different in the structure of the association between the two words, which none of the measures in (3) to (4) (can) reveal. This may in turn be the reason why it is not uncommon to find that bi-directional AMs are not as highly correlated with uni-directional psycholinguistic gold standard data such as reaction times or elicitation tasks (see, e.g., Mollin 2009). Therefore, one proposed 'fix' to research on co-occurrence phenomena has been to rely less on bi-directional, or symmetric, AMs, but rather use uni-directional, or asymmetric, ones such as simple conditional probabilities (see (5)) or the ΔP measures (see (6)).

(5) a. $$p_{E|X} = \frac{a}{a+c} = \frac{80}{687} \approx 0.116$$

b. $$p_{X|E} = \frac{a}{a+b} = \frac{80}{99} \approx 0.808$$

(6) a. $$\Delta P_{E|X} = \frac{a}{a+c} - \frac{b}{b+d} = \frac{80}{687} - \frac{19}{137977} \approx 0.116$$

b. $$\Delta P_{X|E} = \frac{a}{a+b} - \frac{c}{c+d} = \frac{80}{99} - \frac{607}{138565} \approx 0.804$$

As is obvious from the equations, ΔP is essentially an adjusted conditional probability. In this case, and this is not atypical, the difference between the conditional probabilities and the corresponding ΔP-values is quite small and may even seem to be negligible. However, ΔP appears more useful for theoretical reasons (its exact form has proven useful in research on associative learning (Ellis 2007) and it seems reasonable that a co-occurrence percentage of an element (*X*) with another (*E*) gets normalized or adjusted by 'what *E* does in general' as most other AMs do anyway) as well as empirical reasons (it performed better than conditional probability in Schneider to appear).

Another relevant issue is concerned with the fact that some AMs – in particular those ultimately related to significance tests such as G^2, chi-squared, *z*, … – conflate frequency (how often are the elements *X* and *E* observed together and in isolation?) and effect size (how strong is the attraction between *X* and *E*?). From a seemingly moderate to high G^2-value in isolation, it is not obvious whether that value reflects medium frequencies of (co-)occurrence and high association or very high frequencies of (co-)occurrence and a medium degree of association. This also means that AM-values involving different overall frequencies (*n* in Table 1) such as from differently frequent elements and/or differently large corpora) cannot be compared if the computation of the AM is sensitive to *n*: if multiplying all values of a table such as Table 1 or Table 2 changes the AM, differences between AMs cannot be interpreted meaningfully – one would need to use measures instead that keep frequency and association strength separate and only reflect association strength the latter (such as the odds ratio or ΔP, see Chap. 5 for a similar recommendation regarding frequency and dispersion).

As mentioned above, a fully-fledged application of AMs to co-occurrence can involve computing one or more AMs for each element *E* co-occurring with a fixed element *X* at least once and rank-ordering them. Often, studies focus on either the top *t* elements (with *t* taking on different number such as 20 or 100 depending on the application) or focus on all elements that meet a particular threshold value for the AM and/or also just the observed frequency value (e.g. a researcher might study only collocations whose *MI*-score is ≥3 and whose observed co-occurrence frequency *a* is ≥5; see e.g. Ackermann & Chen 2013); it is important to realize that such threshold values are hardly ever motivated by a robust theoretical or psycholinguistic perspective but are usually practical stop gaps; the goal of such a stop gap might be to trim down a list of co-occurrence elements by discarding collocates for the evidence of the strength of attraction is more shaky (because the AM was computed on the basis of very few actual co-occurrences). In addition, some researchers focus on the actual numeric values of the AMS of X_{1-l} whereas others might only focus on their ranks, as has been done in most collostructional research. Finally, some AMs have somewhat well-known characteristics: For instance, *MI* often returns very low-frequency but nearly deterministic co-occurrences (such as proper names) whereas *t* and G^2 usually return higher-frequency co-occurrences, which has not only led some researchers to consider both the AM and the observed frequency *a* as mentioned above (to, for instance, avoid having to deal with many infrequent proper names returned by *MI*) but has also led them to use more than one AM with complementary characteristics (such as *MI* and *t*) at the

same time. It is worth reiterating, however, that such decisions are typically pragmatically rather than theoretically motivated.

We now discuss two representative studies in which many of the above methodological decisions are reflected.

Box 1 Durrant (2014)

A prominent theme in recent second language research has been the learning of collocations. Historically, there has been a perception that second language learners find collocations difficult to acquire. This led Wray (2002) to propose an influential model which suggests that the mechanisms of L2 learning systematically focus learners on individual words and prevent them from acquiring collocations. Recent work appears to undermine this picture however. There is evidence that second language learners do acquire collocations from exposure and that they make extensive use of such collocations in their language production (see Siyanova-Chanturia 2015 for a recent review). Within this work, there is evidence that different AMs offer different and complementary perspectives on collocation learning (Ellis, Simpson-Vlach, & Maynard 2008; Durrant & Schmitt 2009; Bestgen & Granger 2014), which will need to be integrated to provide a rounded picture.

Durrant (2014) is a recent example of such work, attempting to determine the relationships between second language learners' knowledge of English collocations and various measures of collocation frequency and association. He re-analyzed the results of 19 different tests of collocation knowledge conducted in eight different countries, as identified through a systematic review of the literature. Frequency and AMs for items on each test were retrieved from the British National Corpus (BNC) and the Corpus of Contemporary American (COCA) and correlated with the number of learners who answered the corresponding test items correctly. Correlations were summarized for the 19 tests through a meta-analysis (see Chap. 27).

The study focuses in particular on how different measures of frequency and association differ in their ability to predict learner knowledge. The predictors assessed differed in five key aspects:

- the choice of corpus: frequency data were retrieved separately from the BNC and COCA, and from each of the main register-based sub-corpora of each, i.e. the five written sub-corpora titled academic, fiction, magazine, newspaper and non-academic (this last appears in BNC only) and the spoken sub-corpus;
- the choice of measure: collocations were quantified in terms of raw frequency, *t*, *MI*, and conditional probability;
- the span within which words had to appear to be counted towards an item's frequency of collocation. Two spans were used: four words either side of the node and nine words either side of the node;
- whether counts were based on lemmatized or non-lemmatized counts. In the former, for example, *arguing strongly* and *argued strongly* would both count as cases of the collocation *argue strongly*. In the latter, these counts would be kept separate;
- to account for possible effects of the evenness of dispersion of a collocation within the corpus, each item was also quantified with a *DP* value (see Chap. 5).

A number of key findings emerge from these data. First and overall, frequency and association data were found to be reliable predictors of learners' knowledge of collocation.

Frequency and *t*-values from COCA achieved correlations with knowledge of between ρ=0.24 and ρ=0.27. Second, there was a large difference in predictiveness between frequency and *t*, on the one hand, and *MI* and conditional probability on the other. For these foreign language learners, it seems that the number of times a collocation occurs is a far more important factor than the strength of association between components. This tallies with the psycholinguistic work of Ellis, Simpson-Vlach & Maynard (2008), who found that the accuracy and speed of processing of lexical bundles by ESL learners in the US was predicted by frequency but not by *MI*. In contrast, the accuracy and speed of processing of native speakers was predicted by *MI* but not by frequency. These findings show with great clarity both the importance of the differences between different types of measures and the fact that no measure can, in any absolute sense, be regarded as 'the best'. Different measures work suit different purposes and in some cases, using multiple, contrasting measures, can bring out important patterns that would be missed by the use of a single measure.

Third, frequency data derived from COCA were substantially better predictors of knowledge than those from the BNC. Participants in the tests analyzed came from Denmark, France, Japan, Jordan, Saudi Arabia, Spain and Sweden. While it is possible that students in these settings are more influenced by US than by British English, the impact of this is likely to have been marginal: Given the widespread view that learners' overall knowledge of collocation is weak (see, e.g., Wray 2002), the idea that they have picked up a particular British or US 'collocational accent' (if such a thing exists) seems unlikely. A more plausible explanation is the more contemporary nature of COCA, which continues to be updated on a yearly basis. In contrast, the BNC includes mostly texts produced in the 1980s and 1990s. Since collocation is a highly context-sensitive phenomenon, it is likely that the 20-30 years which separate today's students from the BNC texts will make it a less good guide to the sorts of language to which they are exposed.

Fourth, within the two national corpora, there were also substantial differences in the predictiveness of data from different registers. In both the BNC and COCA, fiction showed the strongest correlation and academic writing the weakest.

Finally, the dispersion of a collocation across a reference corpus had only a weak relationship with knowledge (more widely spread collocations were better recognized), and this relationship was significant only in the BNC.

Box 2 Hampe and Schönefeld (2006)

Hampe and Schönefeld (2006) use AMs to understand *syntactic creativity*. This refers to examples such as those in (7), in which a verb is used in an argument structure with which it is not usually associated.

(7) a. Social media bore her stupid
b. The boiler shuddered to a halt

Hampe and Schönefeld's study asks how such instances should be accounted for in linguistic theory, comparing in particular two possible accounts. One, attributed to Goldberg (1995), holds that verbs maintain their usual meanings while inheriting syntactic slots and a generic meaning from the abstract argument-structure construction (ASC). This account is contrasted with Hampe and Schönefeld's own model, in which creative uses are described in terms of the syntactic blending of two verbal expressions. On this model, the unusual structure

(e.g. Noun *bored* Adjective, as in (7a)) triggers the retrieval of another verbal concept which is more usually associated with the ASC (e.g. Noun *makes* Noun Adjective, as in *Social media made her stupid*). They argue that the intended meaning of the creative form is reached through conceptual integration or blending of the two concepts.

Hampe and Schönefeld evaluate the plausibility of these models through a detailed description, first, of the typical verbal associates of a complex-transitive ASC and, second, of the syntactically creative uses of four verbs. AMs – in particular, $p_{\text{Fisher-Yates exact test}}$ (a measure very highly correlated with G^2, see Section 2) – are central to both analyses. These analyses are particularly useful for illustrating the uses to which AMs can be put in that, though they rely on the same test, each makes use of a rather different type of association and for different purposes. The first analysis looks at associations between an abstract construction and the verbs which instantiate it in order to understand the range of meanings which the construction can carry. The second looks at associations between verbs and their accompanying collocations in order to understand restrictions on the use of particular syntactically creative forms.

The first analysis focuses on the ASC illustrated above in (7a), i.e. constructions in which the verb is followed by a direct object and an adjectival phrase acting as object predicate (usually referred to in construction grammar as the *resultative* construction). Retrieving all cases of this construction from the syntactically parsed ICE-GB corpus, Hampe and Schönefeld use the Fisher-Yates exact test to identify verbs which are associated with it. These verbs are taken to indicate the meanings in which the ASC is most characteristically used. The associated verbs are classified into three semantic groups. The first is most centrally characterized by *make* (the strongest associate of this ASC), which is used to indicate causation (as in *He* made *John angry*). Other associated verbs of this type are *render, get,* and *set*. Closely related to these are verbs most centrally represented by *keep* (as in *She* kept *it safe*), which indicate maintenance of a given state. Other examples include *leave, hold,* and *have*. The third group of verbs has *find* (after *make,* the second mostly strongly associated verb of the ASC) as its central example (as in *He* found *her arrogant*). The group is also represented by *consider* (as described above). This group is rather different from the first two in that it cannot be classified under a broad 'resultative' meaning by which the ASC has been characterized. These are cognition verbs which, Hampe and Schönefeld observe, can be described as 'attributive', rather than 'resultative'. They argue that these should be treated as distinct constructions, pointing out that the generic ASC described by Goldberg (1995) fails to provide the relevant semantics for the attributive uses.

The second part of Hampe and Schönefeld's analysis explores syntactically creative uses of four verbs: *encourage, support, bore* and *fear*. In particular, they look at cases of these verbs in complex transitive patterns in which the direct object noun is followed by either a prepositional phrase (to create a 'caused motion' construction, e.g. *encourage tourists into the area*) or an adjective phrase (to create a 'resultative' or 'attributive' construction, e.g. *the subject bores them stiff*).

For three of the verbs – *encourage, support*, and *fear* – use in one of the searched constructions is rare (less than 1% of occurrences of the verb) and not listed as a possible form in the corpus-based *Collins Cobuild English Language Dictionary*. The resultative use of *bore* (as in *He bore her stupid*) is more common (accounting for around 7% of uses of the verb) and is listed in the dictionary. Importantly for our current focus on AMs, each verb's syntactically creative use appears to come with collocational restrictions. That is to say, they are strongly associated with specific accompanying words. As with the ASC-verb associations described above, strong collocates are identified using the Fisher-Yates exact test.

Hampe and Schönefeld argue that the apparent restriction of these creative syntactic forms to particular lexical contexts cannot be accounted for in terms of the properties of the

relevant ASCs alone. Taking as an example the case of *fear* and its strong association with *dead* and (to a lesser extent) with terms related to death (*drowned, killed, murdered*), they hypothesize that this use could be motivated by similar forms at different levels of abstraction. Specifically, model verbs strongly associated with the attributive construction (most centrally, the verb *find*) provide a template for understanding events in which a feature or quality is attributed to the direct object. At a less abstract level, collocational restrictions can be explained by the collocates of the model verb. Hampe and Schönefeld point out that the partially lexically-specified sequence *X (be) found dead* appears to have served as a model for the *feared dead* pairing, which was shown to be the central instantiation of this form. The novel form *feared dead*, once instantiated, may in turn serve as a basis for creating similar forms (*feared killed, feared murdered*, etc.).

Regardless of whether we ultimately accept Hampe and Schönefeld's conclusions, their paper demonstrates well how AMs can be used to identify two different types of patterning - attractions between abstract constructions and the verbs which instantiate them, and attractions between verbs and the collocates which accompany them. The former provides a way of understanding the meaning potential of a construction; the latter provides a way of understanding restrictions on use. As Hampe and Schönefeld acknowledge, the purely textual nature of this analysis means that strong inferences about the nature of psycholinguistic representations and processes cannot be drawn. What these analyses do provide, however, is a clearer picture of the language use for which any linguistic model would need to account. This picture provides us with a basis both for forming linguistic hypotheses and for evaluating the *prima facie* plausibility of existing models. Most pertinently to our current topic, their use of AMs provides a granularity of description which cannot be convincingly provided through consideration of abstract syntactic forms or of vocabulary items alone, revealing additional levels of complexity in linguistic patterning and hence in the models that are required to explain it.

3 Critical assessment and future directions

Given its nature as a distributional discipline, the discussion of how to best approach the quantification and exploration of co-occurrence is likely to continue for the foreseeable future, in particular as corpus-linguistic methods are used in a wider range of theoretical frameworks and with a wider range of other kinds of data, be they observational, experimental, or simulation data. In this section, we are discussing a few areas that we feel should on corpus linguists' radar; they involve

- the recognition that much current discussion of AMs is more fragmented than it needs to be (Section 3.1);
- candidates for measures that have so far not been explored but rather than just being yet even more different ways to crunch the same numbers, that offer additional advantages that current measures do not provide (Section 3.2);
- additional pieces of information that virtually no current AM includes (Section 3.3).

3.1 Unifying the most widely-used AMs

The above discussion presented AMs as they are typically discussed, namely based on seemingly unrelated mathematical formulae in turn based on 2×2 co-occurrence frequency tables such as Table 1 and Table 2. While this kind of presentation is nearly omnipresent and perhaps useful in particular for studies discussing very many AMs, it has one big disadvantage: The entirely

differently-looking equations obfuscate the fact that the AMs that are used in probably 90% of all studies involving AMs – G^2, *MI*, the odds ratio, ΔP, and even *t* or *z* – can in fact all be unified once a particular statistical perspective is adopted, namely that of (binary logistic) regression models. As discussed in Chap. 21, binary logistic regression is a statistical tool that allows the user to study the behavior of a dependent variable (e.g., the presence of a verb: *any verb* vs. *regard*) as a function of one or more predictors (e.g., the choice of a construction: *any construction* vs *as*-predicative). The results of binary logistic regression models are similar to those of the maybe more straightforward linear regression models and include the following:

- an *intercept*: log odds of the predicted level of the dependent variable (the second, i.e. *regard*) when the predictor is the first level (i.e. any construction);
- a *coefficient*: the change in log odds of the predicted level of the dependent variable (*regard*) when the predictor becomes the second level (i.e. *as*-predicative);
- the intercept and the coefficient can then be used to compute predicted probabilities of the two levels of the dependent variable;
- a *significance test* of the overall regression model, which, in the case of a model with only one predictor, is also the significance test of that predictor (see also Gries 2013b: Section 5.3).

Space does not permit a detailed discussion and exemplification here in prose; for detailed code, computations, and results in R, see the appendix and the companion code file. Suffice it to say here, that

- G^2 is the difference between a regression model that predicts the use of *E* (*any verb* vs. *regard*) given *X* (any construction vs. the *as*-predicative) from a null model that predicts the use of *E* (*any verb* vs. *regard*) given no other information;
- the odds ratio is the exponentiated coefficient in the regression model;
- *MI* is $\log_2$ of the predicted probability of *E* being *regard* happening when *X* is the *as*-predicative divided by the probability of *E* being *regard* in general; etc.

More interestingly, $\Delta P_{Construction \rightarrow Verb}$, the adjusted conditional probability measure from (6), is simply the difference between the predicted probabilities of *regard* being used with and without the *as*-predicative being present.

To reiterate, while corpus-linguistic research into the association between elements has produced dozens of AMs, the frequencies of their use is as Zipfian-distributed as that of words: While there is still a lively discussion of which measure(s) is/are most useful for which specific purpose, a mere handful of (symmetric) measures are used in the vast majority of studies. However, there is now more recognition that at least the symmetry-of-association assumption built into most AMs used is problematic and more uni-directional/asymmetric measures are being explored now. The still intense discussion of AMs notwithstanding, it is instructive to realize that all the most frequent measures – uni- and bi-directional ones – are really only different parts/facets of a simple binary logistic regression trying to predict the realization of one element based on another: Once that is realized, all the seemingly different AMs can be captured under one and the same approach (which is of course part of the reason why many AMs are very highly correlated with each other); not only does that facilitate their teaching, it also naturally bridges the gap between AMs on the one hand and hundreds regression-based studies of alternation phenomena in sociolinguistics or usage-based linguistics or over-/underuse studies in learner corpus research (see Gries to appear).

3.2 Additional (different) ways to quantify basic co-occurrence

As mentioned above, the number of AMs that have been proposed is vast and, ironically speaking, inversely proportional to the number of rigorous and comparative evaluations of many of AMs, which is why it may seem futile to add new measures to the mix. However, Baayen (2011) makes two suggestions regarding how to quantify (directional) co-occurrence that nonetheless appear attractive and merit mention because of how they offer avenues of research or analysis that are as promising as they are underexplored.

The first of these is to use as an AM another general information-theoretic measure, namely the Kullback-Leibler (KL) divergence. The KL divergence is written as D_{KL} (posterior/data || prior/theory), which refers to how much a posterior/data percentage distribution of an element (e.g., *E*) in the presence of another element (e.g., *X*) diverges from the overall/theoretical overall percentage distribution of *E*, ; it is computed as in (8). Equation (9) shows the reverse perspective: how the percentage distribution of *X* in the presence of *E* diverges from *X*'s overall percentage distribution (in both equations, $\log_2 0 := 0$):[3]

(8) $$D_{KL}(p(E|X)\|p(E)) = \frac{a}{a+c} \times \log_2 \frac{a \times n}{(a+b) \times (a+c)} + \frac{c}{a+c} \times \log_2 \frac{c \times n}{(a+c) \times (c+d)} \approx 0.699$$

(9) $$D_{KL}(p(X|E)\|p(X)) = \frac{a}{a+b} \times \log_2 \frac{a \times n}{(a+b) \times (a+c)} + \frac{b}{a+b} \times \log_2 \frac{b \times n}{(a+b) \times (b+d)} \approx 5.483$$

With a bit of simplification, this shows that the presence of *regard* says much more about the presence of the *as*-predicative than the presence of the *as*-predicative says about the presence of *regard* (because 5.483>>0.699), which is more/different evidence that the distribution in Table 2 is better quantified with uni-directional measures.[4] The two versions of this measure are fairly highly correlated with ΔP (r>0.86 in *as*-predicative data, for instance, and >0.8 in Baayen's 2011 comparison of multiple AMs), but an attractive feature of D_{KL} is that (i) it is a measure that has interdisciplinary appeal given the wide variety of uses that information-theoretical concepts have and (ii) it can also be used for other corpus-linguistically relevant phenomena such as dispersion (see Chap. 5), thus allowing the researcher to use one and the same metric for different facets of co-occurrence data.

Baayen's second proposal is to use the varying intercepts of the simplest kind of mixed-effects model (see Chap. 22). Essentially, for the *as*-predicative data from Table 2 used as an example above, this approach would require as input a data frame in the case-by-variable format, i.e. with 138,664 rows (one for each construction) and two columns (one with the constructional choices (*as*-predicative vs. *other*), one with all verb types (*regard*, *see*, *know*, *consider*, …, *other*) in the data. Then, one can compute a generalized linear mixed-effects model in which one determines the basic log odds of the *as*-predicative (-3.4214) but, more crucially, also how each verb affects the log odds of the *as*-predicative differently, which reflects its association to the *as*-predicative. These values are again positively correlated with, say, ΔPs, but the advantage they offer is that, because they too derive from the unified perspective of the more powerful/general approach of regression modeling, they allow researchers to effortlessly include other predictors in the exploration of co-occurrence. For instance, the *as*-predicative is not only strongly attracted to verbs (such as *regard*, *hail*, *categorize*, …) but also to the passive voice. However, traditional AM analysis does usually not consider additional attractors of a word or a construction, but within a regression framework those are more straightforward to add to a regression model than just about any other method.

In sum, AM research requires more exploration of measures that allow for elegant ways

to include more information in the analysis of co-occurrence phenomena.

3.3 Additional information to include

Another kind of desiderata for future research involves the kind of input to analyses of co-occurrence data. So far, all of the above involved only token frequencies of (co-)occurrence, but co-occurrence is a more multi-faceted phenomenon and it seems as if the following three dimensions of information are worthy of much more attention than they have received so far (see Gries 2012, 2015 for some discussion):

- type frequencies of co-occurrence: current analyses of co-occurrence based on tables such as Table 2 do not consider the number of different types that make up the frequencies in the cells *b* (19) and *c* (607) even though it is well-known that type frequency is correlated with many linguistic questions involving productivity, learnability, and language change. So far, the only AM that has ever been suggested to involve type frequencies is Daudaravičius & Marcinkevičienė's (2004) lexical gravity, but there are hardly any studies that explore this important issue in more detail (one case in point is Gries & Mukherjee 2010);
- entropies of co-occurrence: similarly to the previous point, not only do studies not consider the frequencies of types with which elements co-occur, they therefore also do not consider the entropies of these types, i.e. the informativity of these frequencies/distributions. Arguably, distributions with a low(er) entropy would reflect strong(er) associations whereas distributions with a high(er) entropy would reflect weak(er) associations. Since entropies of type frequencies are relevant to many aspects of linguistic learning and processing (see Goldberg, Casenhiser, & Sethuraman 2004, Linzen & Jaeger 2015, or Lester & Moscoso del Prado 2017), this is a dimension of information that should ultimately added to the corpus linguist's toolbox.
- dispersion of co-occurrence (see Gries 2008a, Chap. 5): given how any kind of AM is based on co-occurrence frequencies of elements in a corpus, it is obvious that the AMs are sensitive to underdispersion. Co-occurrence frequencies as entered into tables such as Table 2 may yield very unrepresentative results if they are based on only very small parts of the corpus under investigation. For instance, Stefanowitsch & Gries (2003) find that the verbs *fold* and *process* are highly attracted to the imperative construction in the ICE-GB, but also note that *fold* and *process* really only occur with the imperative in just a single of the 500 files of the ICE-GB – the high AM scores should therefore be taken with a grain of salt and dispersion should be considered whenever association is.

To conclude, from our above general discussion and desiderata, one main take-home message should be that, while AMs have been playing a vital role for the corpus-linguistic analysis of co-occurrence, much remains to be done lest we continue to underestimate the complexity and multidimensionality of the notion of co-occurrence. Our advice to readers would be

- to familiarize themselves with a small number of 'standard' measures such as G^2, *MI*, and *t*; but
- to also immediately begin to learn the very basics of logistic regression modeling to (i) be able to realize the connections between seemingly disparate measures as well as (ii) become able to easily implement directional measures when the task requires it;
- to develop even the most basic knowledge of a programming language like R to avoid

being boxed in into what currently available tools provide, which we will briefly discuss in the next section.

4 Tools and resources

While co-occurrence is one of the most fundamental notions used in corpus linguistics, it is not nearly as widely implemented in corpus tools as it should be. This is for two main reasons. First, existing tools offer only a very small number of measures, if any, and no ways to implement new ones or tweak existing ones. For instance, WordSmith Tools offers *MI* and its derivative *MI*3, *t*, *z*, G^2, and a few less widely-used ones (from WordSmith's website) and AntConc offers *MI*, G^2, and *t* (from AntConc's website). While this is probably a representative section of the most frequent AMs, all of these are bidirectional, for instance, which limits their applicability for many questions. Second, these tools only provide AMs for what they 'think' are words, which means that colligations/collostructions and many other co-occurrence applications cannot readily be handled by them. As so often and as already mentioned in Chap. 5, the most versatile and powerful approach to exploring co-occurrence is with programming languages such as R or Python, because then the user is not restricted to lexical co-occurrence and dependent on measures/settings enshrined in ready-made software black boxes, but can customize an analysis in exactly the way that is needed; some very rudimentary exemplification can be found in the companion code file to this chapter; also, see http://collocations.de for a comprehensive overview of many measures.

5 Key readings

Pecina (2010) appears to be the most comprehensive overview of corpus- and computational-linguistic AMs focusing on automatic collocation extraction. In this highly technical paper, 82 different AMs are compared with regard to how well they identify true collocations in three scenarios (kinds of corpus data) and evaluated on the basis of precision-recall curves, i.e. curves that determine precision ($^{\text{true positives (correctly identified collocations}}/_{\text{all positives (all identified collocations)}}$) and recall ($^{\text{true positives}}/_{\text{all trues (collocations to be found)}}$) values for every possible threshold value an AM would allow for. For two of the three kinds of corpus data, measures that can be assumed to be unknown to most corpus linguists score the highest mean average precision (cosine context similarity and the unigram subtuple measure); for the largest data set, the better-known pointwise *MI* scores second highest, and some other well-known measures (including *z* and the odds ratio) score well in at least one scenario.

Wiechmann (2008) also provides a wide-ranging empirical comparison of association measures, specifically those pertaining to collostruction. He focuses on how well various measures of collostruction strength predict the processing of sentences in which a noun phrase is temporarily ambiguous between being a direct object (*The athlete revealed his problem because his parents worried*) and the subject of a subordinate clause (*The athlete revealed his problem worried his parents*) using cluster and regression analyses.

Notes

1 We are ignoring the lexico-textual co-occurrence sense of *colligation* here.

2 While AMs often agree fairly well in their assessment of the degree of attraction between two elements (or at least their overall ranking), their computation can lead to them having different 'preferences'. For instance, pointwise *MI* is known to return low-frequency but perfectly predictive collocations (e.g. proper names) whereas measures that are ultimately based on significance tests (such as G^2 or t) often rank more frequent items higher; see Evert (2009) for more discussion.

3 The Kullback-Leibler divergence is also already mentioned in Pecina (2010).

4 See Michelbacher, Evert, & Schütze (2007, 2011) and Gries (2013a) for further explorations of uni-directional/asymmetric measures.

References

Ackermann K, Chen YH (2013) Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. Journal of English for Academic Purposes 12(4): 235-247

Baayen RH (2011) Corpus linguistics and naive discriminative learning. Brazilian Journal of Applied Linguistics 11(2): 295-328

Bartsch S (2004 Structural and functional properties of collocations in English. Narr, Tübingen

Bestgen Y, Granger S (2014) Quantifying the development of phraseological competence in L2 English writing: An automated approach. Journal of Second Language Writing 26(4): 28-41

Biber D, Gray B (2016) Grammatical complexity in academic English: Linguistic change in writing. Cambridge University Press, Cambridge

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

Durrant P (2014) Corpus frequency and second language learners' knowledge of collocations. International Journal of Corpus Linguistics 19(4): 443-477

Durrant P, Schmitt, N (2009) To what extent do native and non-native writers make use of collocations? International Review of Applied Linguistics 47(2): 157-177

Ellis NC (2007) Language acquisition as rational contingency learning. Applied Linguistics 27(1): 1-24

Ellis NC, Simpson-Vlach R, Maynard C (2008) Formulaic language in native and second-language speakers: psycholinguistics, corpus linguistics, and TESOL. TESOL Quarterly 1(3): 375-396

Evert S (2009) Corpora and collocations. In: Lüdeling A, Kytö, M (eds.) Corpus linguistics: an international handbook, vol. 2. Mouton De Gruyter, Berlin & New York, 1212-1248

Firth JR (1957) A synopsis of linguistic theory 1930-55. Reprinted in Palmer FR (ed.), (1968) Selected papers of J.R. Firth 1952-1959. Longman, London

Goldberg AE, Casenhiser DM, Sethuraman N (2004) Learning argument structure generalizations. Cognitive Linguistics 15(3): 289-316

Gries ST (2008a) Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics 13(4): 403-437

Gries ST (2008b) Phraseology and linguistic theory: a brief survey. In Sylviane Granger & Fanny Meunier (eds.), Phraseology: an interdisciplinary perspective, John Benjamins, Amsterdam & Philadelphia, 3-25

Gries ST (2013a) 50-something years of work on collocations: What is or should be next …

International Journal of Corpus Linguistics 18(1): 137-165
Gries ST (2012) Frequencies, probabilities, association measures in usage-/exemplar-based linguistics: some necessary clarifications. Studies in Language 36(3): 477-510
Gries ST (2013b) Statistics for linguistics with R, 2nd rev. & ext. ed, De Gruyter Mouton, Boston & New York
Gries ST (2015) More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff (2013). Cognitive Linguistics 26(3): 505-536.
Gries ST (this volume) Analyzing dispersion.
Gries ST (to appear) On over- and underuse in learner corpus research and multifactoriality in corpus linguistics more generally. Journal of Second Language Studies
Gries ST, 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
Gries ST, Mukherjee J (2010) Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes. International Journal of Corpus Linguistics 15(4): 520-548
Harris ZS (1970) Papers in structural and transformational linguistics, Reidel, Dordrecht
Hilpert M, Blasi DE (this volume) Fixed-effects regression modeling
Lester NA, Moscoso del Prado Martín F (2016) Syntactic flexibility in the noun: evidence from picture naming. Paper presented at CogSci 2016
Linzen T, Jaeger TF (2015) Uncertainty and expectation in sentence processing: evidence from subcategorization distributions. Cognitive Science 40(6): 1382-1411
McEnery T, Xiao R, Tono Y (2006) Corpus-based language studies: an advanced resource book. Routledge, Oxon & New York
Michelbacher L, Evert S, Schütze H (2007) Asymmetric association measures. International Conference on Recent Advances in Natural Language Processing
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
Pecina P (2010) Lexical association measures and collocation extraction. Language Resources and Evaluation 44(1): 137-158
Schäfer R. (this volume) Mixed-effects regression modeling
Schneider U (to appear) Delta P as a measure of collocation strength. Corpus Linguistics and Linguistic Theory
Siyanova-Chanturia A (2015) Collocation in beginner learner writing: A longitudinal study. System 53(4): 148-160
Stefanowitsch A, Gries ST (2003) Collostructions: investigating the interaction between words and constructions. International Journal of Corpus Linguistics 8(2): 209-243
Wray A (2002) Formulaic language and the lexicon. Cambridge University Press, Cambridge