John Benjamins Publishing Company

Collocations in context

A new perspective on collocation networks*

Vaclav Brezina, Tony McEnery and Stephen Wattam
Lancaster University

The idea that text in a particular field of discourse is organized into lexical patterns, which can be visualized as networks of words that collocate with each other, was originally proposed by Phillips (1983). This idea has important theoretical implications for our understanding of the relationship between the lexis and the text and (ultimately) between the text and the discourse community/ the mind of the speaker. Although the approaches to date have offered different possibilities for constructing collocation networks, we argue that they have not yet successfully operationalized some of the desired features of such networks. In this study, we revisit the concept of collocation networks and introduce *GraphColl*, a new tool developed by the authors that builds collocation networks from user-defined corpora. In a case study using data from McEnery's (2006a) study of the Society for the Reformation of Manners Corpus (SRMC), we demonstrate that collocation networks provide important insights into meaning relationships in language.

Keywords: collocation networks, collocations, statistics, *GraphColl*, swearing

1. Introduction

The linguistic research on word associations is vast. Firth's (1957:6) suggestion to look at the "company that words keep" has been operationalised in a number of different ways (see Evert 2004, 2008, 2010) and has been explored in a number of different contexts (e.g. Baker et al. 2008, Xiao & McEnery 2006, Syanova

& Schmitt 2008). However, more than fifty years into the research on colloca-
tions, many of the lessons learned from this research have yet to be systemati-
cally evaluated and fully implemented in the tools that corpus linguists use (see
Gries 2013 for an important discussion about how research into collocations can
be improved). Traditionally, three criteria for identifying collocations have been
proposed. These are: (i) distance, (ii) frequency, and (iii) exclusivity. The distance
specifies the span around a node word (the word we are interested in) where we
look for collocates. This span is called the 'collocation window'. The distance of
the collocate from the node can be as little as one word if we are interested, for
instance, in the adjectives immediately preceding a noun in English, or as much
as a span of four or five words on each side of the node, if we are interested in
more general associations (for a debate on collocational distance, see Sinclair et al.
2004: 42–48). The second criterion, frequency of use, is an important indicator of
the typicality of a word association. For instance, the noun *love* occurs frequently
with the preposition *in* and therefore *in love* is an important 'chunk' in the English
language. However, *in* can also appear in front of many other nouns, such as *case*,
fact, or *school*. Consequently, the relationship between *love* and *in* is not exclusive.
On the other hand, *love* is much more strongly and exclusively connected with the
noun *affair*; when the word *affair* appears in text, there is a large probability that
the preceding word is *love*. In addition to the three criteria discussed above, Gries
(2013) points out three other criteria that should be considered: (iv) directionality,
(v) dispersion and (vi) type-token distribution among collocates.

Directionality refers to the fact that the strength of the attraction between two
words is rarely symmetrical. For example, the word *affair* has a stronger relation-
ship with the word *love* than *love* with the word *affair* because *love* co-occurs with
other words than *affair* more often than vice versa. Yet the traditional association
measures do not capture this difference because the majority of those commonly
used in corpus linguistics are symmetrical measures.[1] Gries (2013) therefore sug-
gests using Delta P as a measure that takes directionality into account by producing
two different values of collocational strength for any pair of words. Dispersion is
the distribution of the node and the collocates in the corpus (cf. Gries 2008). For
example, in a general corpus of British English such as the BNC the word *affair* col-
locates with *love* in 189 cases distributed across 151 texts. This is a relatively even
distribution compared to another potential collocate *agape* (a Greek term for non-
romantic love), which occurs 9 times but only in 2 texts. Finally, Gries (2013) raises

---

1. This is the case with symmetrical collocation windows (e.g. the span of three words on the
left and three words on the right). On the other hand, asymmetrical collocation windows (e.g.
zero words on the left and three words on the right) produce asymmetrical results with any col-
location measure.

type-token distribution as a desirable criterion which has been partly operationalized through the lexical gravity G measure in Daudaravičius & Marcinkevičienė (2004). This criterion takes into account not only the strength of a given collocational relationship (say between *love* and *affair*), but also the level of competition for the slot(s) around the node word from other collocate types. In the BNC, there are about 13 thousand different collocate types which compete with *affair* for a slot near the word *love*.

To these criteria we should add a seventh feature: the connectivity between individual collocates. Collocates of words do not occur in isolation, but are part of a complex network of semantic relationships which ultimately reveals their meaning and the semantic structure of a text or corpus. For example, in the BNC the word *affair* does not collocate with words such as *unrequited*, *undying* or *madly* but is connected with these through the word *love* which collocates with both *affair* and the three terms mentioned above (among others). As we argue in this paper, collocates should not be considered in isolation but rather as part of larger collocation networks (this notion is discussed in detail in Section 2.1).

When we consider the desirable criteria for identifying collocates outlined above on the one hand, and the (lack of) implementation of these criteria in current corpus tools, on the other hand, we see a large discrepancy between theory and practice. Most corpus tools offer users only a handful of pre-defined collocation measures, which implement only some of the desirable criteria; this considerably limits the study of different properties of collocations. Moreover, there are very few tools available for investigation of collocation networks i.e. collocations in context.[2] This article introduces *GraphColl* ("graphical collocations" tool), a new flexible software for investigating collocation networks. *GraphColl* implements a range of different collocation measures (including the directional Delta P), and also allows the user to define their own statistics via a simple interface. *GraphColl* can thus be used to uncover meaning connections in text and discourse that may otherwise pass unnoticed.

---

2. One of such tools is *CONE* (Gullick et al. 2010), which, however, implements only a limited number of options for collocation identification and setting of the threshold values, which makes the exploration of different aspects of the collocational relationship difficult. Moreover, *CONE* does not work directly with corpora but requires a data file with pre-computed associations values between words as an input. This pre-computing is done with a library available for download which generates the required input for *CONE*.

2.   Collocation in context: Basic principles

This section discusses previous research dealing with the concept of collocation networks including different options for operationalization of these networks. In addition, the underlying principles of identification of collocations are discussed with the focus on systematization of the notation that will allow full replicability of the results.

2.1   Collocation networks

The idea that text in a particular field of discourse is organised into lexical patterns, which can be visualised as networks of words that collocate with each other, was proposed by Phillips (1983, 1985, 1989) and later explored in a number of studies using both general and specialised corpora (e.g. Alonso et al. 2011; McEnery 2006a, 2006b; Baker 2005; Williams 1998). These studies indicate that collocation networks have the potential to provide us with an insight into important lexical connections in discourse. These connections, however, can be analysed on a systematic basis only with appropriate computational technology that allows us to run multiple comparisons of mutual attraction between different pairs of words in large datasets.

The theoretical basis for the notion of collocation networks is provided by Phillips (1983, 1985, 1989), who discusses the relationship between collocates and textual macrostructure. Phillips (1989) argues that collocation networks, or 'lexical networks' (the latter is Phillips's preferred term), can be used to operationalize the psychological notion of the 'aboutness' of a text. Phillips (e.g. 1989) proposes that these networks constitute a distinct level of linguistic analysis, which cannot be explained by traditional linguistic theory, as it requires a deeper understanding of lexical processes and their interconnections through collocation networks. Phillips (1989) also considers different options for analysing and visualizing the relevant lexical relationships. He proposes using cluster analysis (specifically Ward's method) for the primary analysis of text, combined with a display of the results as simple digraphs (i.e. directed graphs). Phillips (1989) demonstrates the use of these methods in a study of university textbooks; however, replicating Phillips's (1989) methodology and applying it in other analyses is to a large extent problematic. This is due to three interconnected issues. (i) Although Phillips (1989) clearly intends to identify syntagmatic lexical sets (i.e. sets of words that co-occur in sentences/discourse) (e.g. Phillips 1989: 52), a cluster analysis employed in the way that Phillips (1989) describes, in fact, reveals paradigmatic rather than syntagmatic relationships between words. This means that the words which Phillips (1989) identifies are items that occur with a similar set of collocates and can be thus

considered 'pseudo-synonyms' rather than members of a collocation network. (ii) It is also unclear how the output of the cluster analysis (a dendrogram) can be transformed into a digraph, and where the information about the directionality of the collocations comes from. (iii) The software used by Phillips (1989) is no longer available and his description of the method does not allow full replicability.

Another important contribution to the study of collocation networks is Williams' (1998) paper dealing with the lexical structure of research articles on plant biology. Instead of performing cluster analysis on the whole dataset, Williams (1998) suggests a stepwise procedure, which starts with a single initial node and its collocates, and from there gradually builds a complex collocation network by considering each of the collocates as a new node and adding a network of collocates around each such node. The initial nodes are taken from the first 50 lexical words in the frequency list based on the text or corpus. To identify collocates, Williams (1998) uses the Mutual Information (MI) score with various cut-off points (4, 5 and 6). Although replicable in principle, the precise parameters of Williams' (1998) procedure are somewhat unclear. For example, Williams (1998: 157) suggests calculating the MI statistic only for "collocates with a frequency of co-occurrence of 8". This would, however, randomly limit the pool of collocates specifically to those that happen to occur with a particular frequency in the corpus (perhaps "8 or more" was meant). Also, it is not clear what the span (size of the collocation window) was in which the collocates were considered. In addition, the choice of the MI statistic is not justified in any great length other than saying that it is sufficient for identifying relationships "between lexical items whether they form a 'term' or not" (Williams 1998: 155). Moreover, Williams' (1998) approach to collocation networks also takes no account of directionality, which, as we saw above, is one of the desirable characteristics of collocational relationships. Similar approaches have been used in a number of other studies exploring collocation networks in scientific and professional English texts as well as spoken data (e.g. Williams 2002, Alonso et al. 2011, Jhang & Sung-Min Lee 2013).

McEnery (2006a) uses collocation networks as one of many tools for exploring discourses related to swearing in English. In contrast to Williams (1998), McEnery (2006a) constructs directional collocation networks (the directional orientation is marked by an arrow) starting with specific nodes of interest (identified via the keyword procedure). The association measure used by McEnery (2006a) is the squared version of MI (MI2), with a cut-off point of 3 and a span of +/- 5 words around the node. McEnery (2006a: 234, footnote 44), admits that this choice of statistic was partly motivated by practical considerations, namely the availability of this measure in *WordSmith Tools* (version 3, Scott 1999), the corpus tool he used for the research (for more details see Section 5.1). Although fully replicable, due to its social and linguistic rather than methodological focus, McEnery's

(2006a) study does not discuss in detail the full implications of the methodological decisions taken in the course of building the collocation networks. This paper builds directly on McEnery (2006a), aiming to replicate and elaborate on that study's findings; it discusses the concept of collocation networks in general, as well as the operationalization of collocation networks via the *GraphColl* software.

2.2 Association measures and collocation parameters notation (CPN)

Since the main concern of this paper is systematizing the study of collocations and offering a new perspective on collocation networks, we also need to briefly discuss association measures that are used for the automatic identification of collocations.[3] The collocational relationship is a complex one and no single association measure can capture all of its aspects. In essence, individual association measures differ in how much emphasis they put on the different criteria discussed in Section 1. For the majority of association measures, the statistical procedure for identification of collocates involves two steps: (i) establishing a random co-occurrence baseline (expected frequencies), (ii) comparing observed frequencies with the random co-occurrence baseline. All the widely used collocation measures are therefore based on comparison of (some of) the values in two tables: the contingency table with observed frequencies (Table 1) and the contingency table with expected frequencies (Table 2). The latter table is derived entirely from the former, using the equations stated below, and indicates the frequencies which we would expect to observe if the words in a text or corpus were randomly arranged, with no associations between words. The shaded cells in Table 1 represent values which we need to collect directly from the corpus (using an appropriate piece of software). These are:

i.   Number of tokens in the whole corpus: N
ii. Frequency of the node in the whole corpus: R_1
iii. Frequency of the collocate in the whole corpus: C_1
iv. Frequency of the collocation (i.e. node + collocate) in the collocation window: O_{11}

---

3. The approach to identification of collocations presented in this article (and implemented in *GraphColl*) follows the practice in corpus linguistics where collocations are identified in texts using the collocation window method where no grammatical relations between words are assumed. An alternative approach, often used in NLP, identifies collocates as items in dependency relations with the node and involves dependency parsing (see Wermter & Hahn 2006). However, as shown in Bartch & Evert (2014) this approach to a large extent depends on the quality of the parser and might be therefore problematic for languages for which reliable automatic parsing is not available.

Table 1. Observed frequencies[4]

	Collocate present (*affair*)	Collocate absent	Totals
Node present (*love*)	O_{11}	O_{12}	R_1 (first row)
Node absent	O_{21}	O_{22}	R_2
Totals	C_1 (first column)	C_2	N

Table 2. Expected frequencies: random occurrence baseline

	Collocate present (*affair*)	Collocate absent	Totals
Node present (*love*)	$E_{11} = \dfrac{R_1 \times C_1}{N}$	$E_{12} = \dfrac{R_1 \times C_2}{N}$	R_1
Node absent	$E_{21} = \dfrac{R_2 \times C_1}{N}$	$E_{22} = \dfrac{R_2 \times C_2}{N}$	R_2
Totals	C_1	C_2	N

Association measures can be understood as different ways of comparing the observed and expected values, putting different weight on different aspects of the collocational relationship. The default association measures implemented by *GraphColl* are listed, together with their formulae and other software implementations, in Appendix 1. A more detailed discussion of individual association measures can be found in Evert (2004, 2010) and Pecina (2010).

Three things need to be noted at this stage. First, to compensate for the problem of small expected frequencies when calculating collocates, Evert (2008) proposes a correction for calculating R_1:

$$R_{1\ correct} = R_1 \times \text{window size.}$$

Association measures based on contingency tables can therefore have two forms: the uncorrected and the corrected one (see Appendix 1 which distinguishes between these two versions in most statistics). Second, association measures that take into consideration dispersion are based on comparison of multiple contingency table pairs, each for an individual corpus section (or subcorpus). Third, in any scientific approach, replicability of results is crucial. For this reason, the notation below (see Table 3) is introduced for specification of the full set of parameters

---

4. The notation in the contingency tables is based on Evert (2004, 2010). O_{11} stands for the observed frequency in the first row and the first column of the first contingency table; O_{12} stands for the observed frequency in the first row and the second column of the first contingency table. E_{11} is a symbol for the expected frequency in the first row and the first column of the second contingency table, and so on. R_1 is the total of the first row of both tables and C_1 is the total of the first column of both tables.

for collocate identification/extraction. This is intended to make comparing the results of different analyses easier for the reader. As can be seen from Table 3, seven different parameters are used to determine the specific settings for identification of collocates. Statistic ID refers to the number in column 1 of Appendix 1, which is a unique identifier referring to a specific version of an association measure. This is followed by the name of the statistic and the cut-off value used, the span of the left and the right context, the minimum required frequency for the collocate in the whole corpus, and the minimum required frequency for the collocation (i.e. the co-occurrence of the node and the collocate). The last parameter, the filter, optionally specifies any further procedures within the collocation extraction process, for example any removal of certain words from the results (e.g. based on word class membership), or a minimum dispersion value.

Table 3. Settings for identification of collocations

Notation categories	Statistic ID	Statistic name	Statistic cut-off value	L and R span	Minimum collocate freq. (C)	Minimum collocation freq. (NC)	Filter
Example	4a	MI2	3	L5-R5	5	1	function words removed
In-text notation (example)	4a-MI2(3), L5-R5, C5-NC1; function words removed						

3. *GraphColl*: Software description

GraphColl 1.0 is a free tool that is available from the project's website (http://www.extremetomato.com/projects/graphcoll). It was developed with both novice and advanced users in mind, providing full control over the statistics and methods used to build collocation networks, whilst also offering sane defaults for casual users.

The system runs locally on a desktop computer, with a graphical user interface. The interface is structured around a series of tabs (see Figure 1), which may be followed in a wizard-like manner to construct, explore and export a collocation graph (network). Graphs are presented as detachable tabs, allowing multiple graphs to be generated and examined at once. The "Stats" tab offers advanced users the facility to tweak the statistical procedures used during graph calculations.

The first step in a collocation network analysis using *GraphColl* is to import one or more corpus files (as UTF-8 encoded plaintext) and enter a name for the imported corpus. This is done under the "Import" tab (see Figure 1). Here the user can also adjust settings for word delimiter and punctuation characters, which
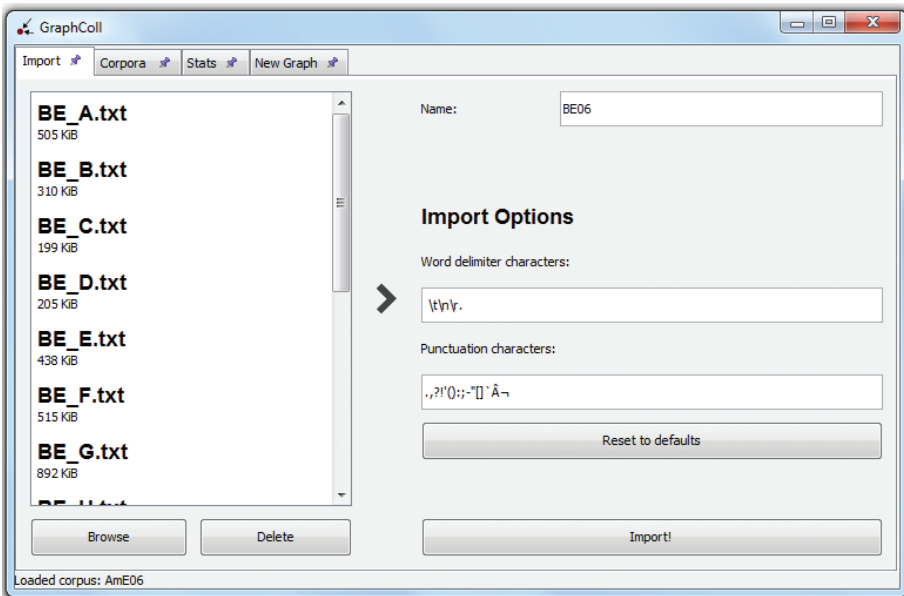
Figure 1. *GraphColl*: Loading corpus files

affect how the corpus is tokenized. The *GraphColl* tokenizer works in two stages. First, the input text is split according to the "word delimiter characters". The tokens thus generated then have any "punctuation characters" trimmed from each end. In
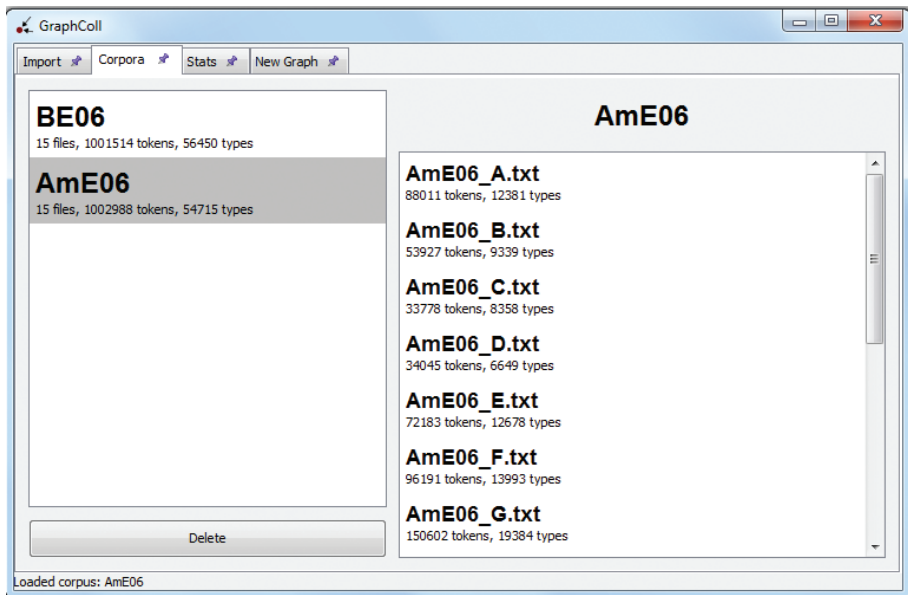
Figure 2. *GraphColl*: Overview of loaded corpora

most cases, it is recommended to use the default settings for word delimiter and punctuation characters.

The "Corpora" tab allows users to inspect the corpora that have been imported into *GraphColl* (see Figure 2), and view their overall size and constituent files. It is possible to remove corpora, but not to edit them: once imported, corpora are considered atomic.

When the user is satisfied with the imported texts, they then move to the "New Graph" tab (see Figure 3). This panel allows the user to define properties of the collocation graph to be produced, such as:

i.   the size of left and right collocation windows (span);
ii.   the association measure;
iii. the minimum collocate and minimum collocation frequency;
iv.   advanced users can employ "advanced thresholds", which are boolean expressions written in the Groovy scripting language that can perform complex operations;
v.   whether or not to use the corrected R_1 value

Once set, these properties cannot be changed for a particular graph; this makes it impossible for users to create meaningless graphs where the collocation parameters vary between nodes.

Figure 3. *GraphColl*: New graph

As mentioned above, *GraphColl* allows multiple graph views on different tabs for parallel analyses. The graph view (Figure 4) works in a manner similar to a search engine. It creates graphs based on a word entered in a search box. It is possible to explore and edit the graph by repeatedly entering search terms, or by manipulating it using the mouse. We can thus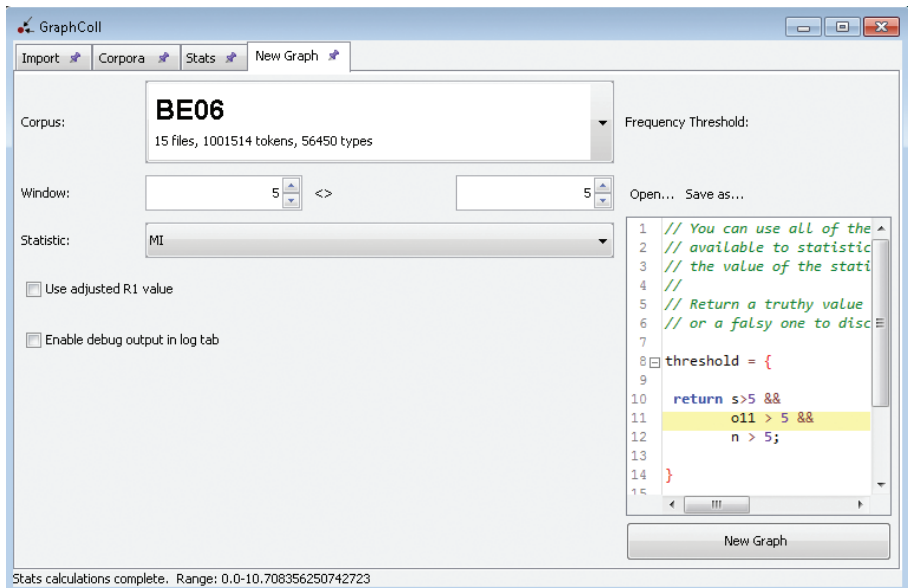 create a collocation network at any level of complexity, including for instance first-, second-, third-, etc. level collocations (counting from the original node).

The process by which collocates are identified is as follows. A collocation search is performed in the corpus for the given node. This computes all word frequencies within the collocation window for the specified node. Then, a statistical comparison is run between the frequencies of words within the collocation window and those outside of the window. Each point ('vertex') in the graph (displayed as a circle) represents a word type in the corpus. Lines ('edges') run from the node to its collocates, their length representing the strength of the collocation. Shorter lines indicate higher values of the association measure, and thus stronger collocational bonds. The spatial arrangement of the individual collocates and their relative position in the graph is motivated solely by the clarity of display and does not have implications for the analysis of the collocational relationship. Collocates and connecting lines are only added to the graph if the statistical comparison reads above a user-defined threshold. This is used to reduce the impact from hapax legomena and unusual word combinations, which tend to overpopulate graphs built using automatic comparison methods.
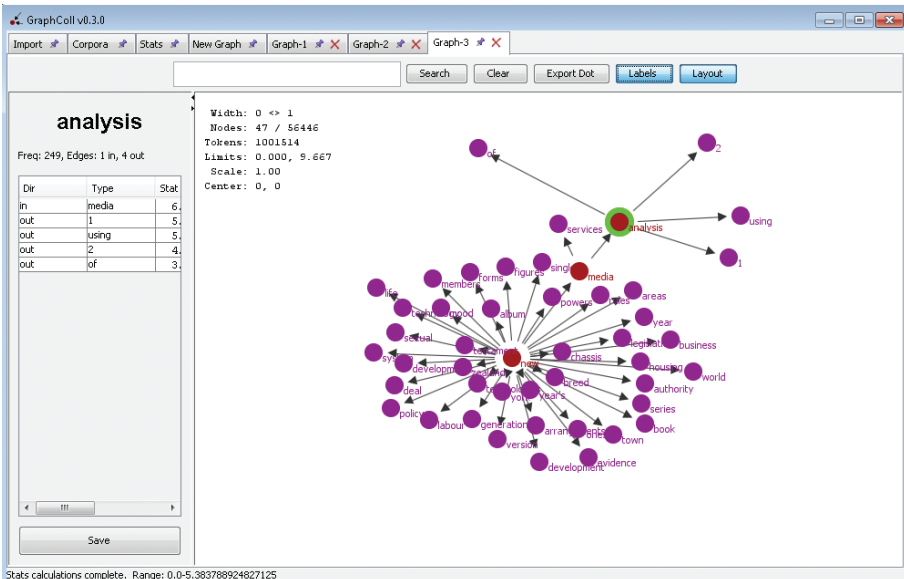
Figure 4. *GraphColl*: Graph view

The collocates of each node may be viewed in the table on the left of the graph view. Selecting a node presents its immediate collocates, their frequencies and association measure scores in a searchable table; selecting an item in this table selects the corresponding node in the graph, and vice versa. It is possible to export this list for any node as a CSV file, or to export the whole graph in *GraphViz* dot format or as an image file.

The graph is colour-coded to indicate the completeness of the nodes. The types which have had a full collocation search computed for them are coloured red. Types which have been identified as collocates, but have not had their own collocates computed, are purple. A graph that is entirely red represents a complete view of all collocational relationships in the corpus.

The "Stats" tab (see Figure 5) exposes the algorithms used to compute each association measure, allowing advanced users to modify the formulae used. By default, *GraphColl* implements 14 collocation statistics, most of which have two versions (26 different equations altogether) — for details and references see Appendix 1. The user can also define new statistical measures — for details see Appendix 2.
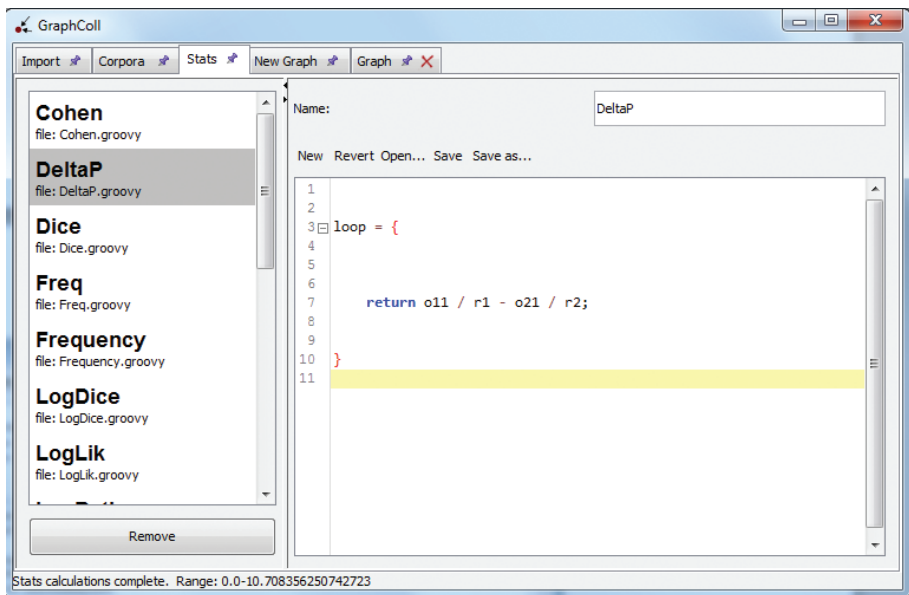
Figure 5. *GraphColl*: Stats definition

4. Collocation networks in *GraphColl*: Concept demonstration

This section briefly demonstrates the concept of collocation networks as operationalized in the *GraphColl* software. The aim is to show how connectivity as the seventh dimension of the collocational relationship (see Section 2.1) operates in practice. To do this, we explore collocation networks of different levels of complexity starting with the node *time*, the most frequent noun in the English language. BE06, a one-million word corpus of current written English that follows the Brown family sampling frame (Baker 2009) is used to demonstrate the type of collocation networks that can be identified in a standard general corpus. For collocate identification, we have selected MI score because it is an association measure commonly used in corpus studies and implemented in a large number of corpus tools (see Appendix 1).

Figure 6 shows the node *time* with its first-order collocates as defined through the MI statistic and the application of various threshold (cut-off) values specified in the caption using CPN. While interpreting the graph, we need to focus on the length of the arrow from the node to the collocate which represents the strength of the collocational relationship as expressed by the association measure; the layout of the individual collocates is motivated by the clarity of display and has no meaning in itself (see Section 3). Similarly, as we are using the MI score, we need to be careful not to infer (uni)directionality of the collocational relationship from the arrow pointing from the node to the collocate. Directionality of the relationship between the node and a collocate can be observed in the graph only under two conditions: (i) a directional measure such as Delta P has been selected for collocate identification; (ii) the collocation search has been computed for the collocate in question (i.e. both the node and the collocate are coloured red). Neither of these conditions is met in the graph in Figure 6.

Figure 6. First-order collocates: 3a-MI(6), R5-L5, C5-NC5; no filter applied

Nonetheless, the graphical display in Figure 6 allows convenient visual inspection of the results and comparison of the strength of the relationship between the node and the individual collocates as established by the association measure.

With *GraphColl*, we can move easily beyond first-order collocates to explore connectivity between collocates at various levels of the collocational relationship. Figure 7 displays second-order collocates around the node *spend* which is one of the prominent collocates of the original node *time* not only because of the strength of the association measure, but also because different forms of the same verb (*spent*, *spending*) occur in the set of first-order collocates. We can see that, apart from *time*, *spend* is connected with another temporal expression (*hours*), the adverb *together* (which figures in the pattern *spend* [temporal expression] *together*), as well as the noun *money*. The connection between *time* and *money* is well-established in the literature on conceptual metaphors (see Lakoff & Johnson
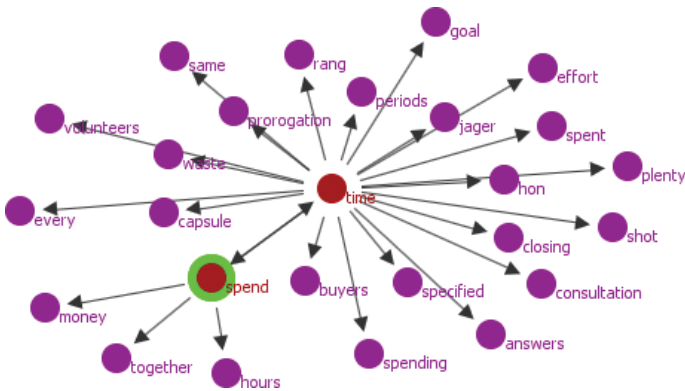
Figure 7. Second-order collocates: 3a-MI(6), R5-L5, C5-NC5; no filter applied
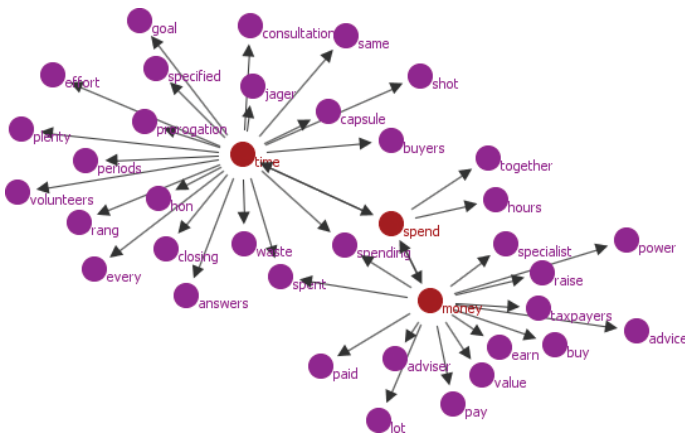
Figure 8. Third-order collocates: 3a-MI(6), R5-L5, C5-NC5; no filter applied

1980, Li 2014), the basic claim being that we understand the abstract notion of *time* in terms of money-related concepts.

GraphColl provides empirical evidence about the connectedness of these two concepts based on a corpus of general English which shows that this connection is most strongly established through various forms of the verb *spend* as can be seen from Figure 8. At the same time, we can observe that both *time* and *money* also have a number of unique associations that create a complex network of meanings that surpasses the one-to-one mapping "time is money" originally suggested by Lakoff & Johnson (see also Li 2014 for a critical evaluation of the original theory).

The complex network of associations can especially be seen in Figure 9 which displays fourth-order collocates (counting from the original node *time*). Here we can see financial terms around the node *money* and a connection between *money* and *power* and a set of collocates around the latter node. It can be hypothesised that our understanding of a word or concept in the collocation network to a certain degree affects the understanding of other concepts in the same network, though the type of the impact and the exact principles of activation of different associations needs to be further investigated.

Finally, it needs to be noted that an absence of a link between two types in the graph does not mean an absence of the collocational relationship or association in the discourse or language; the graph displays collocates defined through a combination of parameters (as specified in CPN) and excludes collocates that do not meet these specifications. For example, in Figure 9 neither *buy* nor *waste*
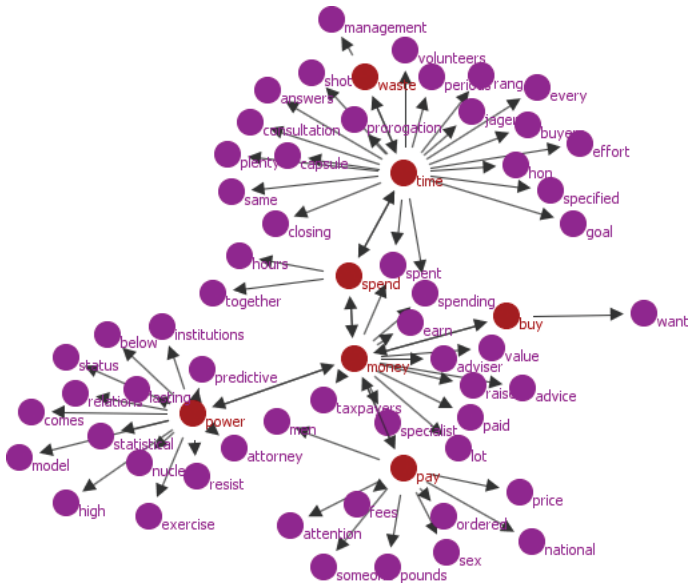
Figure 9. Fourth-order collocates: 3a-MI(6), R5-L5, C5-NC5; no filter applied

are connected with both *time* and *money*. However, when we relax the criteria (or change the statistic) for collocate identification both *buy* (co-occurring 4 times with *time* and 7 times with *money*) and *waste* (co-occurring 16 times with *time* and 4 times with *money*) will appear with a link to both *time* and *money*. In this sense, the graphs produced by the tool are exploratory in nature rather than providing a single answer to the question of connectedness between words, as is collocation itself.

5. Collocation networks and swearing in English: A case study

This section reports a case study that further demonstrates the use of *GraphColl* with a specialised corpus. Here, the focus is on the analysis of the discourse about the "reformation of manners" in Britain in the late 17th and early 18th centuries to show how a topic in corpus-based discourse analysis can be explored using collocation networks. The goal is to replicate McEnery's (2006a) original analysis of the collocation networks associated with words such as *swearing* and *drunkenness*, but also to show what new insights *GraphColl* can bring to this topic.

McEnery (2006a, 2006b) analyses a moralistic discourse on swearing which was initially promoted by religious societies such as the Society for the Reformation of Manners (SRM) in the 17th and 18th centuries, and which has had a lasting effect on our modern sensitivities. In so doing, he applies the sociological concept of 'moral panic' to discuss the social processes underlying general attitudes towards bad language. A moral panic is typically associated with positioning a social phenomenon (or group of people associated with it) as "as a threat to societal values and interests" (Cohen 1980: 9, cf. also Altheide 2009). Typically, for a moral panic to spread there needs to be a medium of public communication. In present-day societies, this role is played by different media of mass communication, including the internet. In the 17th and 18th centuries, many essays, reports and pamphlets were published and disseminated to forge and guard what was perceived as morality. The excerpt below is taken from one such pamphlet and shows how a moral panic about swearing was created and propagated.

(1) Secondly, Common *Swearing is a Vice* dangerous to our selves, when we consider what may be expected from Man. Our Reputation is blasted by it, it sinks our Credit in the World, and proves prejudicial to our Estates (Walker 1711 — taken from SRMC).

Linguistically, a moral panic rests on word associations as the mechanism of creating strong links between a target word (which becomes the subject of moral panic) and its evaluations. In the example above, swearing is labelled as a *vice*, a religious

term for inclination to wrongdoing. By repeating this and similar labels in a number of different contexts, the authors of the pamphlets against swearing create a lexical (and social) framework of associations in which swearing is perceived in strongly negative terms, as an act that undermines the social and religious order.

However, cases such as Example (1) above show only the immediate associations (collocates) of the word *swearing*. To appreciate the complexity of the moralistic discourse we also need to look at how the immediate associations are connected with one another and, more importantly, how these are connected to other (more distant) associations. Thus, for instance, in 17th and 18th century discourse, *swearing* is connected with a whole range of religious evaluations, through further associations of *vice* with notions such as *prophaneness* and *irreligion*, as in Example (2).

(2) we do most humbly beseech Your Majesty, That all *Vice, Prophaneness and Irreligion*, may in a particular manner be Discouraged in all those who have the Honour to be Employed near Your Royal Person; (Yates, 1699 — taken from SRMC).

Ultimately, any discourse rests upon a large network of associations, where each one activates a number of others (cf. Hoey 2005), that produces social meaning through multiple cross-associations. These cross-associations, however, cannot be observed even with careful reading of source documents, but need to be analysed using a tool that allows simultaneous multiple comparisons of word frequencies and co-occurrences. *GraphColl*, which employs different statistical measures (each highlighting different aspects of the collocational relationship) to analyse and visualize collocation networks, is a tool designed specifically for this type of research.

The following two research questions were formulated to guide this study:

RQ1: Is McEnery's (2006a) analysis replicable with *GraphColl*?
RQ2: What additional insights into the discourse of moral panic can we gain using *GraphColl*?

These research questions are primarily methodological, intended to demonstrate the variety of uses of *GraphColl*. However, RQ2 will further enhance our understanding of the complex processes behind the creation of a moral panic around swearing, which has had a lasting impact on the general perception of swearing until the present day.

5.1 Method

The data for this case study was The Society for the Reformation of Manners Corpus (SRMC). This corpus was compiled by McEnery (2006a) for the study

summarised above. The same corpus is used here to make possible the replication of McEnery's (2006a) results. It should be noted that there are several versions of the corpus, which differ with regard to the texts they contain and the presence or absence of normalization of Early Modern English spelling. In this study, we used a version which retains original (non-standardized) spelling and comprises four texts. Two of these are early core texts for the Reformation of Manners movement; the other two come from the end of the period of the society's activities. The total size of the corpus is 120,532[5] tokens (see Table 4).

Table 4. SRMC structure

Text	Tokens	Date	Reason for inclusion
Yates	43,016	1699	core texts: widely cited during the period
Walker	63,515	1711	
Anon	4,201	1740	diachronic representativeness: end of the period
Penn	9,800	1745	
Total	120,532		

McEnery (2006a) uses *WordSmith Tools* (Scott 1999) to identify prominent collocates, which he later employs to create collocation networks. The process of building these networks is largely manual, and involves a rather painstaking comparison of the different associations between the node and its collocates. By contrast, all the analyses reported in this article were carried out automatically using *GraphColl*. McEnery's (2006a) original study is also constrained by the limited number of association measures available in *WordSmith Tools* (ver. 3). As McEnery (2006a: 234, footnote 44) notes, his preference would have been to use the cubed variant of the mutual information statistic (MI3), but this option was not available in that version of *WordSmith Tools*.[6] McEnery (2006a) therefore opts for MI2, the squared variant of the MI metric. The parameters used are: a span of 5 words on each side of the node; statistic threshold 3; minimum collocate frequency 5; and no minimum collocation cut-off point (see Table 5).

---

5. This is the word count provided by *GraphColl* based on the default tokenization options as described in Section 2. McEnery (2006a) quotes 120,709 words.

6. MI3 is available in newer versions of *WordSmith Tools*; but overall the range of statistical measures in this tool is fairly limited. For example, newer versions of *WordSmith Tools* discontinue support for the MI2 statistic, making it impossible to replicate work such as McEnery's using *WordSmith Tools* v. 4 and above.

Table 5. McEnery's (2006) settings for identification of collocation

Statistic ID	Statistic name	Statistic cut-off value	L and R span	Minimum collocate freq. (C)	Minimum collocation freq. (NC)	Filter
4a	MI2	3	5L-5R	5	N/A	function words removed; strongest collocates considered

This can be written in the collocation parameters notation (CPN) as 3a-MI2(3), L5-R5, 5–1; function words removed; strongest collocates considered (see Section 2.2). Because *GraphColl* does not limit the user's choice of the association measure, in this study we explore the properties of five different statistics and their possible contributions to discourse analysis — including MI3 (which was McEnery's (2006a) original preferred choice).

5.2 Results and discussion

This section reports on the results of a series of analyses using five different collocation metrics. First, the replication of McEnery's (2006a) study is presented using the MI2 statistic followed by new results using MI3, log-likelihood, Delta P and Cohen's *d*.

5.2.1 *MI2: Replication of McEnery (2006a)*

Using the same settings as in McEnery's (2006a) original study (see Table 5) *GraphColl* produces the results displayed in Figure 10. For practical reasons, Figure 10 displays only 100 strongest collocates (the total number of collocates identified was 245). We can see that all collocates from McEnery's (2006a) collocation network are also present in Figure 10, and have been highlighted in the graph with rectangles (*cursing*, *common*, *prophanation*, *lewdness*, *parliament*, *drunkenness*, *excessive*, *blasphemy* and *damning*). Since these nine collocates are discussed in detail in McEnery (2006a: 177ff), the focus here will be on the remaining collocates found by *GraphColl*, which were not (for practical reasons) explored in McEnery's (2006a) original study. These are mainly collocates that illuminate the strong religious context of the debate, such as *prophane/profane*, *vain*, *sinful*, *conscience*, *sin* (*against god*), *damn*, *condemn*, *jews*, (*god's*) *name, lord's-day* and *temptation*. In addition to these, two other groups of collocates can be observed: (i) collocates with general negative associations such as *dismal*, *drinking* (as another "sinful" activity), *false*, *contemptuous*, *abominable*, *wantonness*, *lying* and *negligent*; and (ii) descriptive collocates such as *conversation*, *effects*, *land*, *examin'd*, *causes*, *essay*, *civility*, *engagement*, *caution* and *act*. The former create an additional layer

of general pejorative evaluations and associations, as concretely exemplified by the chapter title "The dismal Effects of prophane SWEARING" (Walker 1711). The latter set consists of other key terms contributing to the general shaping of the discourse around the nature of swearing, its *causes* and *effects*, and its legal consequences (which are described in the pamphlets with reference to the *Act of Parliament against Swearing and Cursing*).

Figure 10. Collocates of *swearing*: Replication of McEnery (2006a) — 4a-MI2(3), R5-L5, C5-NC1; function words removed

Let us, however, now focus on some of the major religious contexts in which the term *swearing* occurs.

(3)   […] and put any Stop to this public *Dishonour of God's Name*, by exposing the *Sin and Folly of common Swearing*, I shall think my self happy in my Endeavours and Studies (Walker).

(4)   And as common *Swearing is heinously Sinful*, as it is a *Dishonouring of God*, which Dishonour is attended with Circumstances and Consequences of Guilt: (Walker)

(5)   Sit down a while and consider these things, vain Man, lay them to Heart and ponder them in thy Mind, and then think thy vain and rash *Swearing by the God* that made thee, and the Christ that Redeem'd thee, Innocent, if thou canst (Walker).

(6)   there can be no agreement between he and the swearing damn christian of this age who be so far from obey he whose name they take that they be

> not come to the righteousness of the law that *condemn all vain swearing* but lie under the heavy judgement of the lord for the breach of his third commandment thou shalt not *take the name of the lord thy god in vain* (Penn).

All four Examples (3) to (6) above refer to a situation in which god's name is directly evoked (dishonouring of god, swearing by god, taking the name of god in vain). Example (6), in addition, refers to the perceived divine source of the prohibition of swearing, the third commandment from the Old Testament. This would indeed be a strong source of authority for the 17th and 18th-century pamphlet writers and moralists. At the same time, these examples show that *swearing* was originally a fairly narrow concept, involving a specific type of (what was perceived as) sinful behaviour, which over time became generalised to any uses of bad/foul language, not necessarily those directly connected with evoking God's name. The collocation network in Figure 10 shows the origin of this semantic generalisation — a strong association of *swearing* with other types of immoral behaviour, as perceived by the Society for the Reformation of Manners authors, such as *drunkenness*, *cursing* and *lewdness*.

Finally, to return to RQ1, we see that McEnery's (2006a) results can indeed be fully replicated using *GraphColl*. In addition, *GraphColl* also provided new perspectives on the development of attitudes to swearing in English, highlighting some of the dominant themes of the 17th/18th-century debate, such as the strongly religious dimension, and other themes discussed below.

5.2.2   *MI3: Reduction of low frequency bias*

Let us have a look at some of the other options for association measures. First, the results with the cubed version of the MI statistic (MI3) will be explored. MI3 was suggested to further reduce the low frequency bias of a simple MI score (Daille 1995). As is generally accepted (Evert 2008), the simple MI score emphasizes the exclusivity of the collocational relationship and thus has a propensity to highlight unusual combinations (including even typos and non-standard spellings) that occur only once or twice in the corpus. For this reason, MI (and MI2) are often combined with a minimum frequency threshold for the collocate and/or collocation. In the previous analysis using MI2, the threshold of C5 was applied, following McEnery (2006a). However, even larger and more specific (NC) threshold would have been desirable to weed out rare co-occurrences appearing only once in combination with the node *swearing*.[7] With MI3, no such threshold is typically necessary, because the measure gives more weight to observed frequencies and thus

---

7. Using *WordSmith Tools*, McEnery was not able to apply the minimum collocation frequency threshold (NC).

gives high scores to collocations which occur relatively frequently in the corpus. Uncommon collocates such as *open-lewdness* or *sabbath-breaking* that figure at the top of a simple MI collocate list (but occur only once in the corpus) are pushed down by more frequent and therefore more typical collocates such as *cursing, common* and *prophane*.
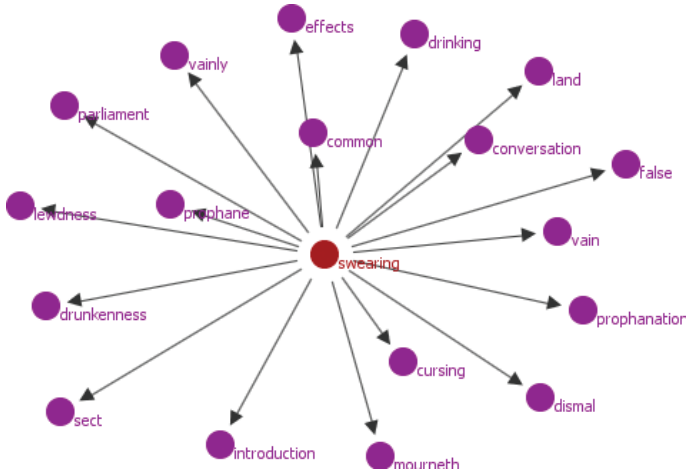
Figure 11. Collocates of *swearing*: 5a-MI3 (11), R5-L5, C1-NC1; function words removed

As can be seen from Figure 11, the graph produced with the MI3 statistic has a similar core to the one produced with MI2 (see Figure 10); however, Figure 11, displays much fewer collocates due to higher statistic cut-off value (11). The parallel between graphs in Figure 10 and Figure 11 is unsurprising because both measures come from the same family of statistics. The relative position of individual collocates (i.e. their closeness to the node in the graph), is, however, different. As noted above, MI3 gives more weight to more frequent collocations, i.e. those with larger observed frequencies. For instance, in Figure 11, *drunkenness* (7 co-occurrences) shows a closer collocational relationship with *swearing* than *lewdness* (3 co-occurrences), whereas Figure 10 displays the converse. Similarly, MI3 gives a higher score to *vain* (8 co-occurrences) over *prophanation* (4 co-occurrences) and downgrades *parliament* (3 co-occurrences). Overall, these rather subtle differences between Figure 10 and Figure 11 point to an important dimension of the moral panic discourse, which is repetition (Cohen 1980). No matter how suggestive, an association which is not repeated enough will be less influential than an association that is more firmly established in the discourse. MI3 can thus be a useful measure for highlighting this feature of collocations.

5.2.3   *Log-likelihood: Evidence against the null hypothesis*

In contrast to the previous two measures (MI2 and MI3), the log-likelihood statistic works within the null hypothesis significance testing paradigm. This means that the question we are asking is not *how large* the difference is between the observed and expected values, but rather whether we have *enough evidence* in the data to reject the null hypothesis (which says that there is no difference between the observed and expected values). Whereas for MI2, MI3 and other similar effect-size statistics the answer to the question we are asking is the value of the statistic (larger value showing a larger effect), for log-likelihood, strictly speaking, the answer is either yes or no. In Figure 12, all statistically significant collocates are shown with the alpha level set to 0.0001 (LL cut-off point 15.13).

Figure 12. Collocates of *swearing*: 6a-LL (15.13), R5-L5, C1-NC1; no filter applied

We can see that when the alpha level is set to 0.0001, we obtain 51 collocates (note that no collocates were filtered out from Figure 12). We can also see that there is a large overlap between the collocates displayed in Figure 12 and those discussed in previous sections. In fact, all but one (*damning*[8]) of McEnery's (2006a) original collocates appear also in Figure 12; these have been highlighted in the graph with rectangles. Although this graph does not identify a new semantic dimension to add to the findings discussed above, it confirms the centrality of those collocates discussed previously, and provides further evidence for the key themes of the moralist debate against swearing — which are (i) connection to other vices (especially

---

8. In the corpus, *damning* co-occurs with *swearing* only once; there is therefore not enough evidence to reject the null hypothesis.

drinking) and (ii) religion (*blasphemy*, *conscience*, *prophane*, *prophanation*, *sin*, *sinful*, *vain* and *vainly*).

5.2.4   *Delta P: Directionality — larger network explored*

With Delta P, directionality of the collocational relationship can be directly explored (cf. Gries 2013). Unlike traditional metrics such as MI, MI2, MI3, log-likelihood, T-score, etc., which in a symmetrical collocation window produce the same association value for node-collocate, collocate-node combinations (i.e. they consider the concurrence of the node and the collocate as one probability), Delta P calculates two different probability values for the co-selection of two words. The first value is the value for the node co-selecting the collocate and the other one is the value for the collocate co-selecting the node. As discussed in the introduction, not all collocational relationships are symmetrical; considering mutual symmetries and asymmetries can therefore help us to better understand the complexities of discourse.

Figure 13 shows the results of a complex analysis of the collocational relationship between the initial node *swearing* and its collocates, which are in this case explored up to the fifth order to show a larger collocation network.[9]
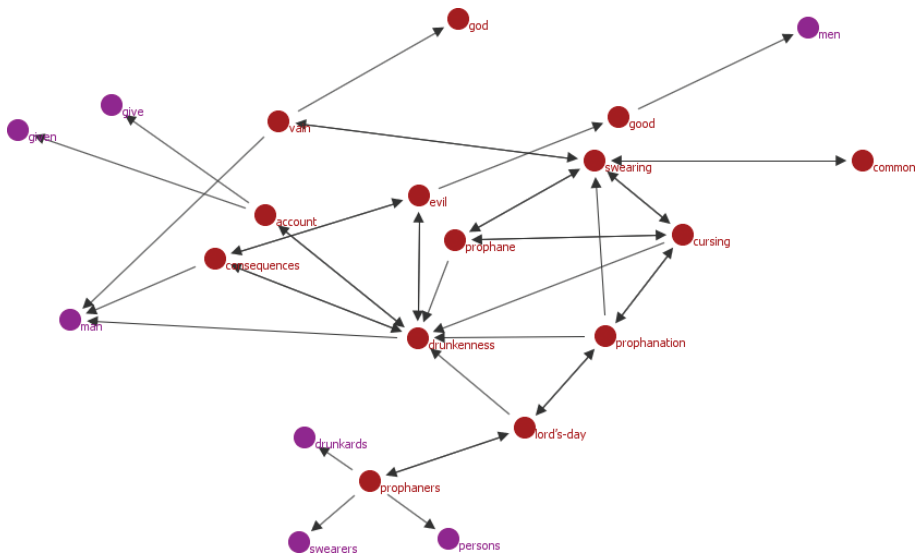

Figure 13. Collocates of *swearing*: 13a-Delta P (0.1), R5-L5, C1-NC4; function words removed

---

9. The directionality in this graph is indicated by the arrows which point either one or both directions. A symmetrical arrow is drawn if both Delta P values are above the threshold value (0.1). Asymmetrical arrow indicates that one of the values of Delta P is below the cut-off point.

As can be seen from Figure 13, *swearing* is closely associated with a number of concepts such as *drunkenness*, *prophanation* and *cursing* that were also highlighted in the analyses above (using different metrics). Here, however, the relationship is explored from the perspective of mutual connectivities of the nodes and collocates and their possible asymmetries.

In the graph, *swearing* is symmetrically connected with collocates such as *vain*, *common*, *cursing* and *prophane*. Interestingly, the noun derived from the adjective *prophane*, *prophanation*, has a stronger relationship with *swearing* than vice versa. This means that *prophanation* would more readily trigger the association with *swearing* than *swearing* would with *prophanation*. In this context, symmetry can be seen as cross-association force that operates between collocates.

Swearing is also connected through *cursing* (its strongest collocate) to *drunkenness* and (yet again) *prophanation* and through these in turn to a host of other associations including the people who would be referred to as "prophaners". These would be *swearers*, *drunkards* and (lewd) *persons*. In this collocation network we can thus readily see how the abstract moralist discourse evolves and becomes personalised, with its metaphorical finger pointing to specific offenders. These connections are realised in individual discourse loci as shown in Example (7):

(7) A Second Society is of about Fifty Persons, Tradesmen and others, who have more especially applyed themselves to the *Suppression of Lewdness, by bringing the Offenders to legal Punishment*: These may have actually suppressed and rooted out about Five Hundred disorderly Houses, and caused to be punished some Thousands of *Lewd Persons, besides Swearers, Drunkards, and Prophaners of the Lord's-Day*, as may appear by their Printed Lists of Offenders.

The same collocation network shows also other relationships, such as *vain* pointing to both *man* (as the offender) and *god* (as the target of the offence), or the asymmetry between *good* and *evil*, both of which add further layers to the rich texture of the moralist discourse.

5.2.5 *Cohen's* d*: Dispersion (*GraphColl *experimental measure)*

Finally, to show the potential of *GraphColl* as an experimental tool, a new association measure, Cohen's *d*, is briefly discussed here. Cohen's *d* (Algina et al. 2005, Cohen 1988) is a commonly used measure of effect size outside of corpus linguistics. It is a measure of the difference between two arithmetic means expressed in standard deviation units. Here we demonstrate how Cohen's *d* can be implemented as an association measure which takes into account the distribution of collocates in different texts (or subcorpora) by comparing the values of collocate frequencies in the collocation window and outside of the window in each text/subcorpus. Due

to space limitations and a different focus of this paper, Cohen's *d* as an association measure cannot be discussed in full detail here, but see Brezina (in preparation).

Figure 14 shows the collocates of *swearing* in the SRMC identified using Cohen's *d*. Even with a very new metric, we obtained a stable set of collocates including *cursing*, *drunkenness*, *common* and *vain*. This is a very important signal that the collocational relationship — and collocation networks in particular — are based on the reality of discourse as reflected in language corpora, rather being a function of any particular statistical procedure.
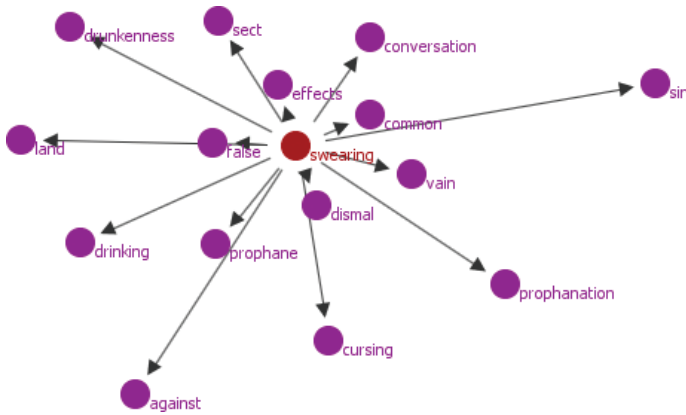

Figure 14. Collocates of *swearing*: 14-Cohen's D (0.5), R5-L5, C1-NC4; no filter

6.   Conclusion: Collocation networks paradigm

The case study demonstrated multiple different ways in which the moral panic discourse around swearing can be explored using collocation networks. Employing different association measures, we established a stable set of collocates that confirm McEnery's (2006a) original findings, yet also extend the scope of the analysis beyond what was possible in the earlier study. In particular, we identified an important religious aspect of the debate and also, via the directional collocation network based on Delta P, the personalization of the discourse and explicit labelling of offenders against morality in the pamphlets. These two findings improve our understanding of the sources and social implications of moral panic respectively. As noted in the literature (Garland 2008, Cohen 1980), a moral panic seeks a strong source of authority, in this case religious authority, and often turns against specific groups of people. The fact that all these complex processes could be summarised in a single image (namely Figure 13) demonstrates the power of this type of analysis. The advantage of using *GraphColl* is thus not only the efficiency with

which it builds collocation networks on the fly, but also its potential to uncover a dimension of linguistic and social research that would otherwise remain unexplored.

On a more general level, the purpose of this article was to demonstrate that connectivity between collocates is an important dimension of the collocational relationship. This was done with both a small (120k) specialised historical corpus as well as a larger (1M) general corpus of current written English. For more evidence about collocation networks in a different context see Baker & McEnery (2015) who use *GraphColl* to explore a 1.5-million word corpus composed of tweets.

In this study, we showed that connectivity can be usefully added "on top" of the remaining six dimensions — distance, frequency, exclusivity, directionality, dispersion and type-token distribution among collocates — to produce informative results. We can thus investigate connectivity in combination with a wide range of association measures, each highlighting different aspects of the collocational relationship by giving different weights to the six dimensions above.

Connectivity as the seventh dimension of the collocational relationship has important implications for our understanding of language and word meaning. Collocation networks show how meanings of words are formed through multiple repeated associations that can be documented only in language corpora. So far, most approaches to word meaning in corpus linguistics have mainly looked at the immediate patterns in a narrow context (that is, first-order collocates). Collocation networks, however, demonstrate that meaningful patterns can be extended beyond this narrow scope and can be identified at the level of the text or discourse. While Philips (1985, 1989) clearly shows that the concept of 'aboutness' is applicable to individual texts, we can extend this notion further to include a broader area of different discourses in general.

In the case study, we saw that a homogeneous specialised dataset (the SRMC) produced a stable set of associations and connections between them, regardless of the association measure used. However, more work needs to be done to investigate the collocation networks evident in different specialised corpora as well as in different general language corpora. Further exploration of this topic will thus shed more light on how individual discourses are connected and how these connections develop over time. In addition, we need to seek discourse-external (psycholinguistic) evidence to establish the extent to which collocation networks are reflected in speakers' mental lexicons.

In very practical terms, collocation networks as an analytical tool have a large potential in a number of areas of linguistic and social research such as discourse studies, psycholinguistics, historical linguistics, second language acquisition, semantics and pragmatics, lexicogrammar, and lexicology. With the free availability of *GraphColl* and its efficient approach to identification of collocation networks,

we hope to see a range of applications that will contribute to our better under-
standing of complex social, cognitive and linguistic processes that shape the ev-
eryday use of language.

References

Algina, J., Keselman, H., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean
    difference effect size: A robust parameter and confidence interval in the two independent
    groups case. *Psychological Methods, 10*(3), 317–328. DOI: 10.1037/1082-989X.10.3.317

Altheide, D. L. (2009). Moral panic: From sociological concept to public discourse. *Crime,
    Media, Culture*, *5*(1), 79–99. DOI: 10.1177/1741659008102063

Alonso, A., Millon, C., & Williams, G. (2011). Collocational networks and their application
    to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci). In I. Kosem, & K.
    Kosem (Eds.) *Electronic Lexicography in the 21st Century: New Applications for New Users:
    Proceedings of eLex 2011, Bled, 10-12 November 2011* (pp. 12–22).

Baker, P. (2005). *Public Discourses of Gay Men*. Abington, UK: Routledge.

Baker, P. (2009). The BE06 Corpus of British English and recent language change. *International
    Journal of Corpus Linguistics*, *14*(3), 312–337. DOI: 10.1075/ijcl.14.3.02bak

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008).
    A useful methodological synergy? Combining critical discourse analysis and corpus lin-
    guistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse
    and Society*, 19(3), 273–306.

Baker, P., & McEnery, T. (2015) Who benefits when discourse gets democratised? Analysing a
    Twitter corpus around the British Benefits Street debate. In P. Baker & T. McEnery (Eds.),
    Corpora and Discourse Studies: Intergrating Discourse and Corpora. Basingstoke, UK:
    Palgrave.

Barnbrook, G. (1996). *Language and Computers: A Practical Introduction to the Computer
    Analysis of Language*. Oxford, UK: Oxford University Press.

Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. *Vernetzungsstrategien,
    Zugriffsstrukturen und automatisch ermittelte Angaben in Internetwörterbüchern 2*, 48–61.
    Retrieved from http://pub.ids-mannheim.de/laufend/opal/pdf/opal2014-2.pdf (last ac-
    cessed March 2015).

Brezina, V. (2014) Effect sizes in corpus linguistics: Keywords, collocations and diachronic com-
    parisons. *35th ICAME conference*, Nottingham 30th April – 4th May.

Brezina, V. (in preparation). Cohen's d: Bringing dispersion into collocations.

Church, Kenneth W., & Hanks, P. (1990). Word association norms, mutual information, and
    lexicography. *Computational Linguistics*, *16*(1), 22–29.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciencies*. Hillsdale, NJ: Lawrence
    Erlbaum Associates.

Cohen, S. (1980) *Folk Devils and Moral Panics: The Creation of the Mods and Rockers*. Oxford,
    UK: M. Robertson.

Daille, B. (1995). *Combined approach for terminology extraction: Lexical statistics and linguistic filtering*, UCREL Technical Papers, No. 15, Department of Linguistics, Lancaster University, Lancaster, UK.

Daudaravičius, V., & R. Marcinkevičienė. (2004). Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2), 321–48. DOI: 10.1075/ijcl.9.2.08dau

Evert, S. (2004). *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. (Unpublished doctoral dissertation). Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, Germany.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (pp. 223–233). Berlin, Germany: de Gruyter.

Evert, S. (2010). Computational Approaches to Collocations. Retrieved from http://www.collocations.de/ (last accessed March 2015).

Firth, J. (1957). *Papers in Linguistics*. Oxford, UK: Oxford University Press.

Garland, D. (2008). On the concept of moral panic. *Crime, Media, Culture*, 4(1), 9–30. DOI: 10.1177/1741659007087270

Gries, S. Th. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437. DOI: 10.1075/ijcl.13.4.02gri

Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next…. *International Journal of Corpus Linguistics*, 18(1), 137–166. DOI: 10.1075/ijcl.18.1.09gri

Gullick, D., Rayson, P., Mariani, J., Piao, S., & Taiani, F. (2010). *CONE: COllocational Network Explorer* [Computer software]. Retreived from http://ucrel.lancaster.ac.uk/cone/ (last accessed March 2015).

Hardie, A. (in preparation). A dual sort-and-filter strategy for statistical analysis of collocation, keywords, and lockwords.

Hardie, A. (2014). Log Ratio: An informal introduction. Retrieved from http://cass.lancs.ac.uk/?p=1133 (last accessed March 2015).

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Abington, UK: Routledge. DOI: 10.4324/9780203327630

Jhang, S., & Lee, S. (2013). Clusters and key clusters in the Maritime English Corpus. 언어과학, 20(4), 199–219.

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Li, S. (2014). A corpus-based analysis of metaphorical uses of the high frequency noun time: Challenges to conceptual metaphor theory. In J Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics 2014* (pp. 79–96). Heidelberg, Germany: Springer International Publishing. DOI: 10.1007/978-3-319-06007-1_5

McEnery, T. (2006a). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. Abington, UK: Routledge.

McEnery, T. (2006b). The moral panic about bad language in England, 1691–1745. *Journal of Historical Pragmatics*, 7(1), 89–113. DOI: 10.1075/jhp.7.1.05mce

Pecina, P. (2010). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2), 137–158. DOI: 10.1007/s10579-009-9101-4

Phillips, M. K. (1983). *Lexical macrostructure in science text* (Unpublished doctoral dissertation). University of Birmingham, Birmingham, UK.

Phillips, M. (1985). *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam, Netherlands: North-Holland.

Phillips, M. (1989). *Lexical Structure of Text* [Discourse Analysis Monograph 12]. University of Birmingham, Birmingham, UK.

Pedersen, T., & Bruce, R. (1996). What to infer from a description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In *7th International Conference on Statistical Analysis of Textual Data (JADT 2004)* (pp. 926–936).

Rychlý, P. (2008). A lexicographer-friendly association score. In P. Sojka & A. Horák (Eds.) *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN* (pp. 6–9), Brno, Czech Republic: Masaryk University.

Scott, M. (1999) *WordSmith Tools Version 3* [Computer software]. Oxford, UK: Oxford University Press.

Sinclair, J., Jones, S., & Daley, R. (2004). *English Collocation Studies: The OSTI Report*. London, UK: Continuum.

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

Stubbs, M. (1995). Collocations and semantic profiles. *Functions of Language*, *2*(1), 23–55. DOI: 10.1075/fol.2.1.03stu

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, *64*(3), 429–458. DOI: 10.3138/cmlr.64.3.429

Wermter, J., & Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge): A qualitative evaluation of association measures for collocation and term extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 785–792). Association for Computational Linguistics, Sydney: The Association for Computational Linguistics.

Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, *3*(1), 151–171. DOI: 10.1075/ijcl.3.1.07wil

Williams, G. (2002). In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, *7*(1), 43–64. DOI: 10.1075/ijcl.7.1.03wil

Xiao, R. & McEnery, T. (2006). Collocation, semantic prosody, and near synomyny: A corpus-linguistic perspective. *Applied linguistics, 27*(1), 103–129

Appendix 1. Default statistical measures

The following statistical measures are implemented in *GraphColl* in the default settings. However, *GraphColl* enables the user to easily add other statistical measures via the "Stats" tab and implement filter options via the "New Graph" tab. More details about the majority of the association measures listed in Table 6 can be found in Evert (2004) and Evert (2010).

Overview of default statistical measures

Association measure	Reference	Equation	A corpus tool that implements this measure
1. Freq (baseline)	Firth 1957, Phillips 1985	O_{11}	BNCweb, CQPweb
2a. Mu	Evert 2004	$\dfrac{O_{11}}{E_{11}}$	
2b. Mu_corr		$\dfrac{O_{11}}{E_{11\,corr}}$	
3a. MI	Church & Hanks 1990, Stubbs 1995	$\log_2 \dfrac{O_{11}}{E_{11}}$	AntConc, Collocate, SketchEngine, WordSmith Tools
3b. MI_corr		$\log_2 \dfrac{O_{11}}{E_{11\,corr}}$	BNCweb, CQPweb
4a. MI2	Daille 1995	$\log_2 \dfrac{O_{11}^{\,2}}{E_{11}}$	WordSmith Tools (v. 3)
4b. MI2_corr		$\log_2 \dfrac{O_{11}^{\,2}}{E_{11\,corr}}$	
5a. MI3	Daille 1995	$\log_2 \dfrac{O_{11}^{\,3}}{E_{11}}$	SketchEngine, WordSmith Tools
5b. MI3_corr		$\log_2 \dfrac{O_{11}^{\,3}}{E_{11\,corr}}$	BNCweb, CQPweb

Measure	Reference	Formula	Tools
6a. Log-likelihood	Rayson et al. 2004	$2 \times \left(O_{11} \times \log \dfrac{O_{11}}{E_{11}} + O_{21} \times \log \dfrac{O_{21}}{E_{21}} + O_{12} \times \log \dfrac{O_{12}}{E_{12}} + O_{22} \times \log \dfrac{O_{22}}{E_{22}} \right)$	Collocate, SketchEngine, WordSmith Tools
6b. Log-likelihood_corr		$2 \times \left(O_{11} \times \log \dfrac{O_{11}}{E_{11corr}} + O_{21} \times \log \dfrac{O_{21}}{E_{21corr}} + O_{12} \times \log \dfrac{O_{12}}{E_{12corr}} + O_{22} \times \log \dfrac{O_{22}}{E_{22corr}} \right)$	BNCweb, CQPweb
7a. Z score	Barnbrook 1996: 95	$\dfrac{O_{11} - E_{11}}{\sqrt{E_{11}}}$	WordSmith Tools
7b. Z score_corr		$\dfrac{O_{11} - E_{11corr}}{\sqrt{E_{11corr}}}$	BNCweb, CQPweb
8a. Dice	Smadja 1993	$\dfrac{2 \times O_{11}}{R_1 + C_1}$	BNCweb, CQPweb, WordSmith Tools
8b. Dice_corr		$\dfrac{2 \times O_{11}}{R_{1cor} + C_1}$	
9a. Log Dice	Rychlý, 2008	$14 + \log_2 \dfrac{2 \times O_{11}}{R_1 + C_1}$	SketchEngine
9b. Log Dice_corr		$14 + \log_2 \dfrac{2 \times O_{11}}{R_{1cor} + C_1}$	
10a. T_score	Stubbs 1995, Barnbrook 1996: 97	$\dfrac{O_{11} - E_{11}}{\sqrt{O_{11}}}$	AntConc, Collocate, SketchEngine, WordSmith Tools
10b. T_score_corr		$\dfrac{O_{11} - E_{11cor}}{\sqrt{O_{11}}}$	BNCweb, CQPweb

11a. Log ratio	Hardie 2014, in preparation	$\log_2 \dfrac{O_{11} \times R_2}{O_{21} \times R_2}$	CQPweb
11b. Log ratio_corr		$\log_2 \dfrac{O_{11} \times R_2}{O_{21} \times R_{2cor}}$	SketchEngine
12a. Minimum sensitivity	Pedersen & Bruce 1996	$\min\left(\dfrac{O_{11}}{C_1}, \dfrac{O_{11}}{R_1}\right)$	
12b. Minimum sensitivity_corr		$\min\left(\dfrac{O_{11}}{C_1}, \dfrac{O_{11}}{R_{1corr}}\right)$	
13a. Delta P	Gries 2013	$\dfrac{O_{11}}{R_1} - \dfrac{O_{21}}{R_2}; \dfrac{O_{11}}{C_1} - \dfrac{O_{12}}{C_2}$	
13b. Delta P_corr		$\dfrac{O_{11}}{R_{1corr}} - \dfrac{O_{21}}{R_2}; \dfrac{O_{11}}{C_1} - \dfrac{O_{12}}{C_2}$	
14. Cohen's d	Brezina in preparation	$\dfrac{M_{\text{in window}} - M_{\text{outside window}}}{\text{pooled SD}}$	

Appendix 2. Working with the statistical equations (advanced users)

In *GraphColl*, association measures are implemented as scripts written in the Groovy scripting language, which perform basic operations on a number of input parameters that represent the contingency tables outlined in Section 2.2 of this paper.

These scripts may be loaded and edited at runtime, and a default set of scripts are loaded at startup from the installation directory to make available the default measures listed in Appendix 1. The scripts are designed in such a manner that it is easy to provide simple statistical association measures, yet it is still possible to perform more complex calculations involving auxiliary data sources and pre-computed resources.

Each association script may define two closures, which are executed at different stages of a collocation calculation:

i. pullup{} is a closure that is executed once after the collocation window frequencies have been computed, but before the statistics are calculated. Its purpose is to provide a space where users may define objects for use in the later stage. It is provided with frequency lists representing the matched and unmatched regions of the corpus for a given node word.
ii. loop{} is executed once for each collocate. It is provided with variables that represent the cells of the contingency tables above, as well as some pre-computed expected values based on these. The value returned from this block is taken as the association value: if it is nil, the node is ignored and not plotted in the graph.

A third closure, threshold{}, is executed to "filter out" any unwanted values from the graph. This is run once for each collocate, and is defined in the "New Graph" tab. It is expected to return a false value to reject a node, or any other value to retain it.

The variables provided for each collocate's calculation phase are provided in accordance with the notation used in this paper, in addition to a number of frequency lists that are the internal representation for such data. The available variables currently are:

Variable names. Raw frequencies for each collocate
r1 = (use_adjusted_r == true ? ((left + right) * all.get(node)) : all.get(node));
o11 = inner.get(collocate);
o12 = r1–o11;
o21 = c1–o11;
o22 = r2–o21;
r2 = n–r1;
c1 = all.get(collocate);
c2 = n–c1;
n = all.n();

Variable names. Expected values
e11 = (r1 * c1) / n;
e12 = (r1 * c2) / n;
e21 = (r2 * c1) / n;
e22 = (r2 * c2) / n;

Authors' addresses

Vaclav Brezina
ESRC Centre for Corpus Approaches to
Social Science
Lancaster University
FASS Building
LA14YD, Lancater
United Kingdom

v.brezina@lancaster.ac.uk

Tony McEnery
ESRC Centre for Corpus Approaches to
Social Science
Lancaster University
FASS Building
LA14YD, Lancater
United Kingdom

a.mcenery@lancaster.ac.uk

Stephen Wattam
School of Computing & Communications
Lancaster University
LA1 4WA, Lancaster
United Kingdom

s.wattam@lancaster.ac.uk