The shapes of collocation

The tool GraphColl (Brezina et al 2015) allows collocational networks to be identified within corpora, enabling corpus analysis to go beyond two-way collocation. With the creation of this tool, more complex forms of collocation emerge, encompassing three or more words. This paper aims to illustrate the types of relationships that can appear when more than two words are considered, using graph theory to account for the different types of collocational 'shapes' that can be formed within GraphColl networks. Using the reference corpus, the BE06, examples of different types of graphs were elicited and then analysed in order to form an understanding of the sorts of relationships between words that occur in particular shapes. For example, it was found that for the graph C_4, two of the non-collocating words were likely to be related grammatically or semantically, either being forms of the same lemma, coming from the same grammatical or semantic class or being synonyms or antonyms of one another. The analysis indicates the need for concepts from graph theory to be introduced into corpus analysis of collocation as well as showing the potential for a more sophisticated understanding of the company that words keep.

Introduction

This paper aims to introduce corpus linguists to graph theory, as a way of helping analysts to interpret collocational networks more easily, and thus enabling a more sophisticated analysis of collocates. The recently available freeware tool GraphColl (Brezina et al 2015) introduces a new dimension to corpus-based analysis of collocation, plotting networks between multiple words, rather than simply showing relationships between two words at a time, as is done with most popular corpus software. In this paper, after describing earlier work on collocational

networks, as well as the tool GraphColl, I demonstrate how collocational networks can enable a more sophisticated and detailed form of analysis, by carrying out an analysis of the word *troops* in a corpus of newspaper articles, first using AntConc (which employs traditional collocational procedures) and then by using GraphColl. I show how the network produced by the GraphColl analysis produced a variety of different shapes or graphs which appeared to show specific relationships between words, and how this raises a question about whether such relationships can be generalised as being typical of certain graphs.

In the method section, in order to test whether such relationships are the province of particular types of graphs I describe how I interrogated a reference corpus, the BE06, in order to derive examples of six types of graphs. The analysis section delineates how the graphs were analysed via concordancing in order to understand the positions of words in each graph, and the extent to which generalizable patterns can be found. Finally, the conclusion section argues that GraphColl requires corpus linguists to incorporate graph theory into their studies of collocation, and argues for further study of collocational graphs, as well as outlining future work that could be carried out.

A central concept within corpus linguistics is collocation, described famously by Firth as 'the company that words keep' (1957: 6). If two words collocate with each other, then they co-occur (appearing next to or reasonably near one another) in some way, usually more often than would be expected if all of the words in a corpus were presented in random order. Collocates help to imbue words with meaning as words can begin to take on aspects of the meaning of the words that they collocate with. This is a phenomenon which is aptly illustrated by the concepts of semantic preference and discourse prosody (Stubbs 2001), where a word collocates with a set of words which belong to either a specific semantic group or appears in the vicinity of words (or phrases) which indicate positive or negative affect. For example, Stubbs (ibid) has showed how the lemma CAUSE tends to collocate with negative

words like *accident*, *anger*, *chaos*, *crisis*, despite not having a negative meaning in itself. Hoey (2005) has argued that such prosodies may prime people who encounter words. For example, if a person hears or reads the word *cause*, they may be primed to expect a description of something negative, or even to evaluate what comes next as intended to be negative by the speaker/author.

Popular corpus tools like WordSmith Tools and AntConc allow collocates to be derived for a node (any word which the user wishes to interrogate). Such tools offer a range of settings to be altered, depending on the analyst's requirements. For example, they may offer a range of measures of calculation, as well as allowing the user to alter settings such as the minimum frequency (the number of times two words must appear together in a corpus for them to be considered as a collocate), the span (the number of words either side of the node that are considered as candidate collocates) or the value of the statistic (e.g. collocational 'strength'). However, until fairly recently, such tools have tended to place a limitation on analysis of collocates, by forcing analysts to consider collocates in terms of two words at a time. This is despite proposals from Philips (1983, 1985, 1989) that words occur as networks of collocates (he referred to them as 'lexical networks'), and a small number of studies which aimed to explore such networks (Williams 1998, AUTHOR 2005, 2014, McEnery 2006, Alonso et al 2011). Research by Philips (1989) used cluster analysis to reveal what Brezina et al (2015) call 'items that occur with a similar set of collocates and can be thus considered "pseudo-synonyms" rather than members of a collocation network'.

Williams (1998), on the other hand, examined lexical structure in a corpus of research articles on plant biology by undertaking a stepwise procedure which began with a single node, acquiring its collocates and then treating each new collocate as a node in itself to obtain new collocates. AUTHOR (2005) and McEnery (2006) took a somewhat different approach, calculating networks based on nodes that were also keywords in their corpora, and focussing

mainly on relationships between different keywords. These early examples of collocational networks were achieved painstakingly, with tools like WordSmith Tools 3 (Scott 1999) used to identify collocates of each node separately, and networks needing to be created and represented visually by hand, rather than automatically formed.

Described in Brezina et al (2015), the creation of the tool GraphColl resolves many of the issues of working with collocational networks. It has the advantage of being free as well as easy to use. After loading in a corpus and specifying collocational settings, a word (the node) is typed into a search box and this produces a visual representation of its collocates. Words are shown as attached to small coloured circles and the collocational relationships between them are indicated with lines. The length of the line between two words indicates the 'strength' of collocation if an effect size measure is used. In order to make it easier for users to visualise relationships, any word can be selected and dragged to a different part of the screen, enabling analysts to simplify 'messy' networks which contain numerous crossing lines. As with Williams' method, any word in an existing network can be clicked on to produce its collocates. For some of the collocational measures used, the tool takes into account directionality of collocation, with arrow heads showing which direction collocation occurs in. Revisiting the work by McEnery (2006) on discourses of swearing, and using a number of different measures of collocation that the tool allows, Brezina et al (2015) show how GraphColl identified collocates that were not found in the earlier study, which illuminate the religious context of the debate on swearing in the corpus used. Additionally, the tool indicated collocates which alluded to personalization of the discourse, along with explicit labelling of offenders against morality.

The analysis which precipitated this paper was concerned with the representation of social actors in a 630,000 word corpus of newspaper articles about Muslims collected in the Sun (a conservative British tabloid newspaper) in the year 2010. In order to find the most frequently

mentioned social actors, I created a frequency list and read down the list, noting potential words that indicated social actors. One of the earliest words which appeared was *troops* (occurring 226 times), and this is the word that I decided to base my analysis round. As a way of getting an idea of how *troops* are constructed in this corpus, I began by obtaining collocates of the word. Table 1 shows the strongest collocates of the word. I used the tool AntConc (version 3.4.3) which allows collocates to be calculated using the MI (mutual information) or T score statistic. For this paper I used MI (Church and Hanks 1990, Stubbs 1995). This statistic tends to favour relationships between lexical words while eschewing high frequency grammatical words like *of* and *the*. It is an effect size measure, showing collocational strength rather than one which gives a p value to indicate the amount of confidence we can state that a relationship exists. Each pair of words is assigned an MI score which indicates strength of collocation – the higher the score, the stronger the relationship. The score takes into account the number of times that two words occur together and away from each other. The threshold for labelling two words as collocates was based on them having an MI score of at least 6. A threshold of 3 has been previously viewed as indicating a 'strong' collocate (see for example Hunston 2002: 71-2) although more recently, work by Durrant and Doherty (2010: 145) has indicated that for a collocation to be 'psychologically real' e.g. one word to trigger the thought of another, an MI of 6 would be required. In order to focus on reasonably high frequency patterns, I specified that a collocational pair must occur at least 20 times before I would consider it for analysis. This resulted in 3 collocates of *troops*, as shown in Table 1.

Collocate	Frequency as collocate	MI score
Afghanistan	46	8.65
British	49	7.67

our	48	7.2

Table 1. Collocates of *troops*.

Exploration of concordance lines that contain these collocational relationships helps to identify why the words occur together. *Afghanistan* and *troops* tend to co-occur as the troops being mentioned are described as being in Afghanistan (30 out of 46 cases have this pattern). Additionally eight out of 46 lines refer to troops being killed in Afghanistan and three refer to *troops* serving or fighting in Afghanistan. However, both *British* and *our* act as more straightforward modifiers to *troops* with the sequence *British troops* occurring 42 out of the 49 times that the words occur together and *our troops* occurring 34 out of 46 times (although *our brave troops* occurs 7 times). Further examination of concordance lines indicates a discourse which is generally supportive of British intervention in Afghanistan and particularly supportive of the people involved in the fighting e.g.:

> I AM disgusted by the vile rants and display of hatred shown towards our brave troops now serving in Afghanistan. The people involved in burning the giant poppy on Armistice Day should be deported. (The Sun, November 16, 2010)

> CHAMP TO CHUMP; Muslim who abused our troops is ex-British boxing title holder (The Sun, June 19, 2010)

The term *our troops* suggests a narrative voice which aims to create a shared perspective between writer and reader – it assumes that the reader agrees with the views put forward, of British troops being brave and doing a good job. About half the cases of *our troops* are from published letters to the newspaper, although the Sun also uses this construction in its own news reporting too.

Differently, *British troops* does not occur in letters to the Sun, although similar constructions to *our troops* are made, with references to *hero British troops* and *brave British troops* in news stories. *British troops* are often described as coming under attack from various sources who are negatively represented:

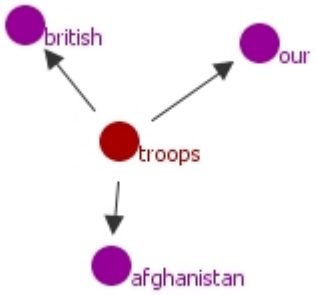
> Both belonged to a gang which gloated over terror bombings and urged the murder of British troops in Iraq and Afghanistan. (The Sun, January 11, 2010)

> British troops were also criticised by US chiefs for what they called a failure to impose security in Afghanistan. (The Sun, November 29, 2010)

This last example is from an article called US Secrets Exposed which details how the Wikileaks website has shown how US officials have made 'vicious slurs' about British politicians and members of the Royal family. The part about American criticism of British troops is thus situated as part of a wider set of offensive claims about the UK, and the Sun distances itself from this claim with the wording 'for what they called…' US chiefs are thus problematized for being critical of British troops in this article.

At this stage we might want to stop the analysis, having examined the three strongest collocates of *troops*, concluding that the Sun's stance of British troops is supportive, especially its use of the construction *our troops* by letter writers and its own journalists. However, the analysis that was carried out only considers three words that directly collocate with *troops*. Let us move on to an analysis of *troops* using the tool GraphColl with same corpus and collocation settings as those used for Antconc above.

Figure 1 shows what is produced when the word *troops* is first entered into GraphColl's search box.



This figure is a visual representation of the table that was produced by AntConc, albeit with the MI scores and frequencies missing (although they can be obtained via a table in the GraphColl interface). The length of each line is representative of the MI score of the two words under consideration, with shorter lines showing words that have stronger relationships to one another. Despite the appearance of arrow heads in Figure 1, the MI score calculation does not actually take into account directionality of collocation. Gries (2013) has noted that collocation can have two directions. In other words, A may collocate with B, but B may not collocate with A. McEnery (2005) gives an example of this: *red-herring*. While it is likely that we will see the word *red* if the word *herring* occurs in a text, we probably are not so likely to see *herring* if we see the word *red*. In GraphColl directionality is nominally shown for some methods of collocation via the use of arrow heads. However, for the MI calculation, the arrow heads do not denote directionality of collocation, they simply indicate which node words have been expanded on. A word which has no arrow heads pointing away from it has simply not been clicked on to determine its collocates.

When the three collocates of *troops* are clicked on in turn, their collocates are obtained and we begin to see a network. This is shown in Figure 2.

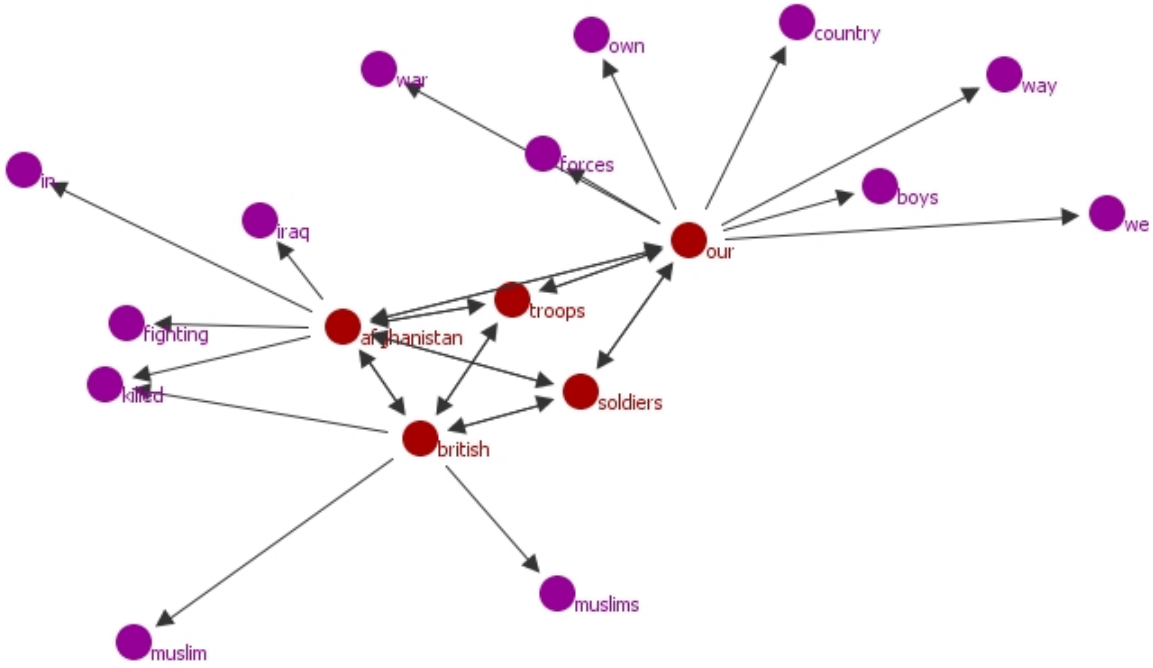


Figure 2. Collocational network of *troops*, expanded to show second-order collocates

While Figure 1 contained 4 words, Figure 2 has 18, showing a more complex set of relationships. Words like *Muslim*, *boys* and *Iraq* could said to be 'second order collocates' of *troops*. The figure could be expanded further by clicking on the new collocates that have appeared (to obtain third-order collocates), but I wish to focus only on the 'second order collocates' and also discuss some of the links between collocates that have now emerged in this figure. How does this figure add to our knowledge of how the word *troops* is used in the corpus?

First, it is useful to notice some potential equivalencies between words in the figure. For example, consider the word *our*. As we saw with the AntConc analysis, it collocates with *troops*, but now we see that it also collocates with *forces*. Both *troops* and *forces* are plural nouns and we may postulate that the two collocational relationships *our-troops* and *our-forces* are equivalent to each other, and indeed, concordance analyses indicates that this is the

case – compare the supportive uses of *our forces* below, with the examples of *our troops* given earlier in the paper:

> FIRING at civilians 21 times in four years equates to about five times a year and shows remarkable restraint from our forces. (The Sun, November 2, 2010)

> In Britain, on Remembrance Day when we give thanks to our war heroes, jeering fanatics hurl insults at our forces while police let them. (The Sun, November 16, 2010)

We may also notice another plural noun collocate of *our* in the network – *boys*. The literal meaning of *boys* is male children, although readers who are familiar with the Sun's rhetoric may be aware that the construction *our boys* also refers to soldiers:

> The Sun saw for itself just what Our Boys have been up against when we joined one of the last foot patrols by 40 Commando - in the very centre of the town which boasts a population of 20,000. (The Sun, September 21, 2010)

> OUR Boys are in high spirits after successfully pulling off the largest helicopter assault in British military history. (The Sun, February 15, 2010)

One interpretation of *boys* here is that it foregrounds the youth of the men who are engaged in battle, perhaps conjuring up images of them as sons. However, boys does not literally need to mean young, and I would argue that it is more likely used as an affectionate term in the newspaper. Also, notably, it is a male construction, so the term does some ideological work in backgrounding female soldiers. The capitalised use of *Our Boys* in the example above indicates how the Sun marks this term as a kind of official designation (of the 53 cases of *our boys*, 50 of them occur with first initial capitals). Table 2 indicates the frequencies in the

corpus of different combinations of terms in the collocational network which appear to have similar or equivalent meanings.

Term	Frequency
British troops	41
British soldiers	42
British boys	0
British forces	10
Our troops	33
Our soldiers	12
Our boys	53
Our forces	14

Table 2. Frequencies of terms relating to British troops.

Our boys is thus the most frequent way that the Sun refers to British soldiers (in the corpus at least), a point which would not have been immediately clear had we simply focussed the analysis around the word *troops*.[1] We may have thought to look for related words (perhaps *forces*), but *boys* may not have come to mind, particularly if we were not familiar with the Sun's discourse prior to the analysis. The collocational network therefore helps us to find a related and important linguistic construction which tells us more about the Sun's preferred construction of British soldiers.

Further value to the collocational network approach is shown by the fact that there is another term which suggests an equivalency with *troops* - the word *soldiers*. It collocates with both

[1] As a further indication of how *our boys* is an ideological choice of The Sun, the phrase only occurs five times in an equivalent Guardian (a British liberal broadsheet) newspaper corpus of the same period, and three of these use scare quotes to be critical of the term.

our, *British* and *Afghanistan*, thus collocating with all the words that *troops* collocates with too. However, it does not collocate with *troops*. We might argue then that *troops* and *soldiers* are collocationally similar, although they repel each other. Intuitively, it would appear that *troops* and *soldiers* are synonymous, at least in this corpus – we are unlikely to see these words in the same sentence because their meanings are so similar – one term is usually sufficient on its own. Considering that *troops* and *soldiers* have such a similar set of collocates, it would perhaps make sense to expand our original analysis of *troops* to look at *soldiers* as well. A concordance analysis of *British soldiers* shows that it is used in a similar way to *British troops*, in stories which focus on constructing such soldiers as brave, and contrasted against other social actors who are viewed as villainous:

> Evil Abdul Ghani Baradar, 42 - who has the blood of 261 British soldiers on his hands - was tracked down by the CIA and Pakistani intelligence after FLEEING Afghanistan. (The Sun, February 17th, 2010)

> Anjem Choudary may despise this country and all it stands for but that doesn't stop him trousering an obscene amount of taxpayers' money. He actually receives £8,000 a year MORE in handouts than many British soldiers earn risking their lives in Afghanistan. (The Sun, Janaury 9, 2010)

The presence of different terms for the same concept (*British troops*, *British soldiers* etc), does not have to be ideological, it could simply be the case of writers wishing to avoid repetition. Similarly, *our troops* or even *our boys* could have non-ideological uses (e.g. where a general refers directly to troops who he is leading). However, I would argue that in this network, *our troops* is ideological, and by extension, the related terms *British troops*, *British soldiers* etc, also carry with it some of that ideology. A pertinent point to make about the collocational network is that we can see that *British* and *our* seem to have a similar sort of

relationship to *troops* and *soldiers*. *British* and *our* have similar collocates to each other, but do not collocate with one another. A simplified diagram of the relationship between the four words is shown in Figure 3.

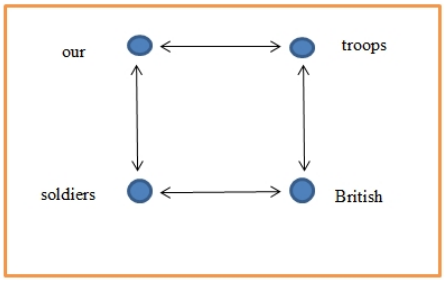


Figure 3. The relationship between *our*, *troops*, *soldiers* and *British*

So a further question arises now in relation to the meaning of the word *our*. Ideologically, the plural person possessive pronoun *our* is interesting because its referent can be ambiguous. When the Sun writes about *our troops* and *our soldiers*, we could interpret *our* as only referring only to the journalists who work in the Sun, or we could view *our* as indicating a combination of the Sun and its readers, or we could even see *our* as encompassing everyone in Britain, whether they read the Sun or not. As noted above, *our* could occur in a quote from an army general, so may not be related to the Sun's narrative voice at all. Concordance analysis shows, however, that the *our troops/boys/soldiers/forces* construction tends to mainly occur in the Sun's narrative voice, rather than being attributed to the quotations of others.

The fact that *our* and *British* appear to share the same collocates in the network, would perhaps suggest that ideologically, the Sun intends *our troops* to address everyone in Britain – that when it uses *British troops*, it is using this term synonymously with *our troops*, thus imbibing British troops with a more positive stance. I would argue that when readers (uncritically) engage with the Sun's discourse on a regular basis, they will eventually come to

internalise these collocational relationships, not just binary relationships like *our* and *boys*, but the fact that the collocates operate in networks. I would not claim that readers will automatically think of the words *our boys* if they read the term *British troops*, but that the internalisation of collocational relationships will stretch beyond the linking of two words to encompass a network where *soldiers*, *troops*, *forces*, *boys*, *our* and *British* are linked together. Readers will thus have a more positive understanding of a seemingly neutral term like *British soldiers*, because they have encountered other terms like *our boys*, which have a more explicitly positive meaning, share collocates with *British soldiers* and are frequent.

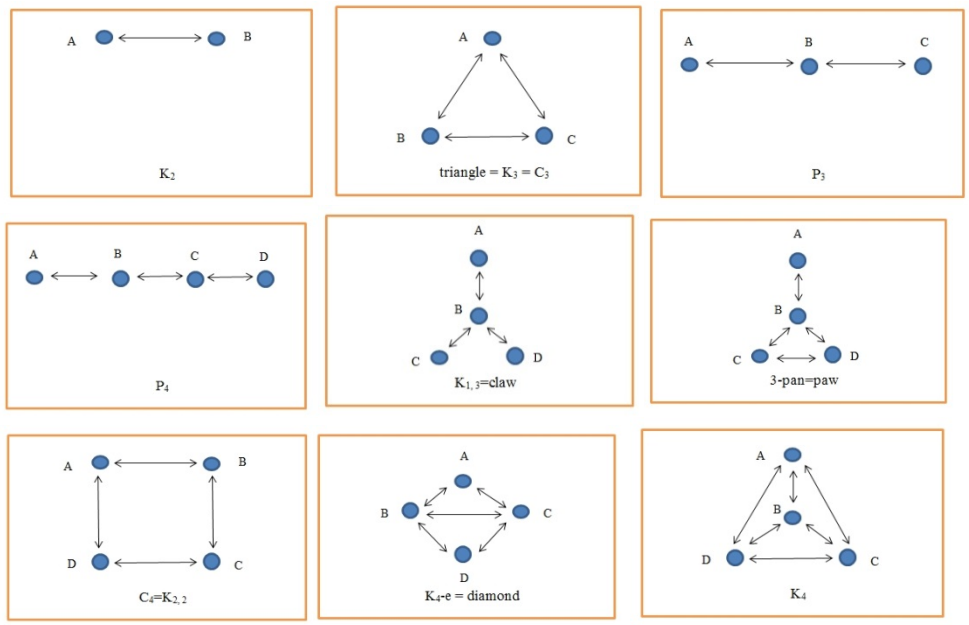
Therefore, I want to argue that collocational networks give 'added value' to corpus analysis by indicating relationships between multiple words which can help to suggest equivalencies, synonyms, rewordings or related terms and concepts, which (in the case of a discourse-based analysis) may have ideological significance. They can also help to suggest relevant terms which may not have been considered for analysis in the first instance (in this case the terms *boys*).

A closer look at the collocational network in Figure 2 indicates that it is made up of numerous 'shapes', the lines connecting *British*, *Afghanistan* and *soldiers* form a triangle while those connecting *our*, *troops*, *British* and *soldiers* make a four-sided (quadrilateral) shape (as shown in Figure 3). In mathematics these shapes are referred to as graphs (hence the name GraphColl for the tool) and are considered as being made up of vertices (nodes or points) – in this case words, along with arcs or lines which connect them (see Harris 2000 and Douglas 2001 for detailed accounts of graph theory). The graph containing just *British*, *Afghanistan* and *soldiers* has several names, it can be named as C_3 – the "C" denotes it as cycle graph (which are graphs consisting of a single cycle, or some number of vertices in a closed chain), while the $_3$ denotes the number of vertices. It can also be called K_3, where K denotes a complete graph – one where every pair of distinct vertices is connected by a unique

edge (in other words, everything is connected to everything else). The C_3 or K_3 graph also has a third (less formal) name: triangle. However, if we consider three other words: *Iraq*, *Afghanistan* and *British*, we can see that these nodes do not consist of a triangle. *Afghanistan* collocates (connects) to both *Iraq* and *British*, but the latter two words do not connect to each other. This is referred to as a path (or linear) graph and would be denoted as P_3. Similar labels can be applied to graphs that consist of four points, so that *our*, *troops*, *British*, *soldiers* would be C_4, and *Iraq*, *Afghanistan*, *British*, *Muslims* would be P_4.

Figure 4 indicates the range of different types of graphs consisting of 2, 3 and 4 vertices. As I am interested in cases of words which do have a relationship with each other (as opposed to those that do not), I have only included cases of graphs where every node is connected to at least one other node. Therefore I do not include graphs such as that formed by *British*, *Muslim* and *we* in Figure 2, as *we* does not connect to either *British* or *Muslims*.

Figure 4. Graphs containing 2, 3 and 4 vertices



From examining some of the graphs which occurred in the collocational networks around the Islam news corpus (Figure 2), I noticed that different graphs seemed to be connected to

different types of semantic or grammatical relationships. The graph C_4 which looks a square and contains *British*, *troops*, *our* and *soldiers* was discussed above where I argued that *British* and *our* did not collocate because they acted as equivalents to one another, as did *soldiers* and *troops*.

Consider another type of graph in Figure 2 which contains the words *British*, *troops*, *Afghanistan* and *soldiers*. This is a K_4-e or diamond graph with four nodes and five edges: A-B, B-C, C-D, D-A and A-C. While *troops* and *soldiers* do not collocate, all of the other words do. So in the previous graph, *British* and *our* did not collocate, but *British* and *Afghanistan* do collocate. This time *British* and *Afghanistan* do not mean the same thing. However, they are related words – both belong to a semantic set we could call 'nationalities', and there is a further link: along with the United States, Britain invaded Afghanistan in 2001. So this helps to explain why *British* and *Afghanistan* co-occur, when *British* and *our* did not.

Next consider a K_4 graph taken from an analysis of the same corpus, where every word collocates with every other word (not shown in Figure 2): *hate*, *cleric*, *anjem*, *choudary*. Two of the words operate as the name of a person, *anjem* and *choudary*, while the other two describe him, *hate* and *cleric*. There are 24 cases of *hate cleric* in the corpus, showing how these two words have become a distinct term. In fact, the fixed sequence of words *hate cleric Anjem Choudary* appears 11 times in the corpus, indicating it is a relatively frequent four word lexical bundle, defined by Biber et al 2004: 376 as 'the most frequent recurring lexical sequences in a register'.

GraphColl requires analysts to perform analysis of collocational networks – an undertaking which many corpus linguists are unlikely to have done before. I would suggest that a good starting point for such an analysis would be to identify different graphs within the network, as such graphs are likely to suggest potentially interesting connections between groups of

words. It would be sensible for analysts to form hypotheses about why words appear in a particular graph. Therefore, knowledge that the typical types of relationships between words in a K_4 graph may be different to those in a C_4 graph (if that is the case) would be useful for analysts. Clearly though, so far we only have one example of each of these types of graph under discussion, so while it may be tempting to draw conclusions that particular graphs are due to certain relationships between words, we would need more evidence in order to claim that a particular graph is the result of certain words having synonymous or related meanings or being the result of lexical bundles. Therefore, it was decided to try to derive and examine more examples of different types of graphs. The research question which drives the remainder of this paper is thus: *do certain graphs within collocational networks lend themselves to particular linguistic relationships between nodes?*

Method

While the subset of the Muslim newspaper corpus I worked with was useful in initially identifying different types of graphs within collocational networks, it was decided to try out the technique on a second corpus, a reference corpus which would contain a wider range of topics and registers. Any findings from this corpus would therefore be more generalizable than one which is only based on a specific register (news) and topic (Muslims). I chose the BE06 corpus (AUTHOR 2009) which contains a million words of written published British English from around 2006, containing 15 genres of writing (including fiction, news, academic writing and official documents). In order to decide which node words I would attempt to derive shapes from, I created a frequency list of the corpus and then, starting at the 25th most frequent word, I took every 25th word, until I had reached the 1000th most frequent word. The 40 nodes examined were *you, so, other, get, day, each, et, great, help, child, full, you're, music, whole, behind, play, light, effect, yes, pay, makes, areas, account, lives, material, involved, compared, specific, costs, worked, seven, james, talking, reached, aged,*

shall, *forces*, *ensure*, *concerned* and *suggest*. In terms of frequency, the most frequent word in the list, *you*, occurred 4,386 times in the BE06 while the least frequent word, *suggest*, occurred 109 times. Going lower than 100 words in frequency tended to provide few collocates (with the settings used – see below), and it was felt unwise to lower the minimum frequency beyond 5 as this would give less reliable collocates.

Again, I used the Mutual Information (MI) statistic (with a cut-off of 6) in order to calculate collocates. Views about the 'best' measure of collocation have tended to vary over the years, and GraphColl currently offers 14 ways of calculating collocation. I have chosen MI for several reasons, first it is a well-known measure used by many other researchers including Williams (1998) described above; second, I have used it in other studies where it has produced believable results (e.g. AUTHOR 2005), third, having tried some of the other measures available within GraphColl, it was the one which appeared to give me collocates which looked credible. The minimum frequency that two words must appear together was set at 5, although in a few cases this resulted in many collocates that were difficult to view onscreen, so in such cases the minimum frequency was raised until only 10 collocates of the node word remained.[2] The collocational span was left at GraphColl's default setting of 5 words either side of the node.

Once the 'first order' collocates for each node were obtained, they were expanded in order to obtain 'second-order' collocates of the node. In some cases, this resulted in additional links between words in the network as some first order or second order collocates collocated with one another. In carrying out this process, different types of graphs were identified and catalogued. Once all of the first and second order collocates of a node had been exhausted I

[2] An issue arises when working with GraphColl which is to do with the amount of information that is displayed on the screen and the extent to which different collocates and relationships can be visually identified. I found that working with more than 10 first order collocates and then expanding them to second order collocates resulted in networks that were almost impossible to interpret due to the presence of so many vertices and edges. Ultimately, GraphColl is perhaps most effective when working with smaller numbers of (high frequency and/or high saliency collocates).

moved on to the next node. In order to avoid having several graphs containing the same node (thus limiting generalizability), I took only one type of graph from each network. In cases where a network contained more than one type of graph, I decided which one to include by throwing a dice. Once I had obtained examples of different types of graphs, I aimed to analyse them in order to address the research question outlined above. AntConc was again used in order to interrogate the specific relationships between the words in a graph.

Analysis

Table 3 shows the numbers of graphs that were collected via the procedures described above. A point which is worth raising here is that the graphs that are identified within collocational networks are not separate and unconnected to other words, they are almost always nested within other sequences of words. For example, looking back to Figure 3, we can see that triangles occur as smaller parts of paw, diamond and K_4 graphs. Triangles do not have to occur within the larger diamond and K_4 graphs, but in almost all cases they will appear at least within one paw graph. The only exception would be a (rare) case where a node word has only two collocates, which also happen to collocate with one another but with nothing else. In order to consider triangles which are 'truly' triangles and not part of larger graphs then, I have not considered any triangles that are parts of diamond or P_4 graphs, but acknowledge that triangles are connected to other graphs (particularly paw graphs) within collocational networks. The collection criteria in Table 3 shows how I have tried to place limits on the graphs so that as far as possible they do not appear as part of other graphs.

Graph	Number elicited	Collection criteria
Triangle	9	Graph must contain the node and two first order collocates (that do not collocate with any other first order collocates)

P_3	14	Graph must either consist of a node and two of its collocates (that do not connect to anything else), or a node and any one collocate that only has one other collocate, or a node with only one first order collocate and then one of its second order collocates
C_4	15	Graph contains node, plus two collocates which do not connect to each other, but do connect to a fourth word which is not a collocate of the node
Diamond	15	Graph contains node and and at least two of its first order collocates; everything in the graph connects, except for any two words.
Claw	9	Graph must contain a first order collocate which collocates with the node and only two other collocates (which do not connect to each other or the node).
K_4	3	Graph contains node and three other collocates that all collocate with one another

Table 3. Number of graphs obtained.

Using the criteria in Table 3 some graphs were easier to elicit than others; in particular a small number of K_4 graphs were found. It would have been possible to supplement the approach, for example, by identifying the most frequent four word lexical bundles in the corpus and then testing them to see if they resulted in K_4 graphs. However, while such a method might identify lexical bundles as K_4 graphs it is somewhat limiting in that it does not take into account K_4 graphs that may not be lexical bundles, so I have remained with the

more exploratory approach described above, allowing 4-word lexical bundles to emerge more naturally (if they are present).[3]

Table 4. Triangles (relationships between A-B, B-C and A-C)

A	B	C
child	parental	leave
you're	I'm	going
music	laptop	live
compared	men	women
costs	total	per
james	hellebore	butcher
Britain's	armed	forces
studies	suggest	results
you	don't	know

Table 4. Triangles.

Table 4 indicates triangles which consist of the node word and two of its first order collocates that also collocate with one another. Looking at this table, four out of nine of the cases contain 2 or 3 words from the same grammatical class e.g. *you're-I'm* (two pronouns connected with enclitics to the BE verb form), *men-women* (two plural nouns), *studies-results* (two plural nouns), *james-hellebore-butcher* (three proper nouns). *James*, *Hellebore* and *Butcher* are actually three characters in the same novel, hence their connectedness in the

[3] Most 4 word bundles tend to contain *the*, *a*, *of*, *to*, *in* or *by* which rarely appear in collocational networks using the MI score. A few K_4 graphs that were also low frequency noun-phrase lexical bundles were found this way though e.g.: *Edinburgh congestion charging proposals*, *Dr Muhmmad Abdul Bari* and *fewer mental health problems.*

corpus. Two of the graphs show words from related semantic groups (*men-women*, *child-parental*). There is not much evidence that triangles are likely to contain words from similar semantic groups then. Could it be the case that triangles are actually the result of three word lexical bundles or idioms? For the purposes of this study, having taken 5 as the minimum frequency of a collocate, I also take the same number to be the minimum times a sequence must appear in order for it in order to be categorised as a lexical bundle. There is a small amount of evidence that some of the triangles occur as three-word bundles. For example, *Britain's armed forces* occurs 4 times, almost making the criteria for a lexical bundle, and if a related case *Britain's conventional armed forces* is added, there would be 5 cases. More convincingly, *you-don't-know* actually does occur as a 3 word bundle, but only 5 times, whereas its 2-word bundles are much more frequent: *you know* (158 occurrences), *you don't* (105), *don't know* (85), *don't you* (33) and *know you* (32). All three words are reasonably frequent and particularly common in conversation where they contribute towards various idiomatic phrases like *I don't know*, *you never know*, and *why don't you*. In fact, what seems more clearly to be the case is that triangles usually contain at least one 2 word bundle. This happens in 6 of the 9 cases in the table, including the two triangles just mentioned above, but also with *I'm going* (30 cases), *parental leave* (24 cases), *results suggest* (8) and *laptop music* (5). *Total-per-costs* has no fixed bundles, although the 5 cases of *total* + *costs* occur in the phrase *total* (optional word) *costs*. Cases where an optional word may appear between two words which occur in a fixed order could perhaps be better thought of as frames Eeg-Olofsson and Altenberg (1994), especially where the optional word may vary e.g. *total ward costs*, *total eviction costs*, *total prosecution costs*. Similarly, *compared* + *men* also occurs in the frame: *compared to/with* (optional percentage + *of*) *men* occur 6 times.

There is thus no single 'rule' which seems to predict the presence of a triangular collocation pattern, although one of the following two factors is likely to be present: having two words

from the same grammatical class and/or having two of the words occurring in a lexical bundle or frame.

A	B	C
reached	hand	until
ok	so	said
parts	other	country
involved	get	those
day	memorial	sunny
each	other's	item
deal	great	good
child	abuse	abduction
fees	pay	tuition
material	study	characterisation
worked	together	hard
music	electronic	pop
travel	costs	air
local	ensure	provide

Table 5. P_3 (relationships between A-B and A-C)

In Table 5, the first row shows the word (A) which acts as the link to the other two. Are there any similarities between words which do not collocate here? This is the case for *deal-great-good*, *child-abuse-abduction* and *music-dance-pop*, where the last two words in each graph function in similar way. So *great deal* occurs 27 times and *good deal* occurs 11, both having the function of quantifying a large amount. *Child abuse* and *child abduction* both refer to similar terrible things that can happen to children and occur 19 and 8 times respectively. *Pop music* and *electronic music* occur 7 and 6 times, both referring to types of music.

Not all cases of the non-collocated pair show such an obvious connection though. Consider *worked-hard-together*. The sequence *worked* (followed by an optional word) followed by *hard* occurs 8 times in the corpus, while *worked* (optional word) *together* appears 5 times. *Together* and *hard* only occur once together so do not count as collocates in this study. However, *together* and *hard* do not appear to have any obvious connection, other than the

fact that they are described as ways of working. Indeed, one explanation for P_3 graphs is that two words can simply be used to modify another one (e.g. *memorial day* and *sunny day*). The words *memorial* and *sunny* are not linked semantically although they both act as modifiers. We may also see cases of 2 sets of 2 word lexical bundles, where there is no relationship between the non-collocating words at all. Consider *get-involved-those*. This is due to *get involved* occurring 8 times and *those involved* occurring 6 times. The words *get* and *those* occur 9 times together (which is not frequent enough for them to count as collocates where MI > 6 as both words are very frequent). Additionally, *get* and *those* do not form any particular sort of lexical bundle when they do co-occur (e.g. *those who get*, *get those eyes*, *get rid of those* etc).

In terms of rules then, as with the triangles, there is no single explanation for P_3 graphs, although there is a tendency towards seeing the two end points having some sort of relationship based on them both modifying the middle word in the graph. However, two sets of 2-word lexical bundles can also result in P_3. Let us now move on to the more complex patterns around graphs containing 4 words.

A	B	C	D
risk	lives	people's	children
specific	ethnic	groups	religious
christmas	day	cold	night
can’t	help	couldn’t	tell
behind	turned	towards	door
I'm	doing	he’s	yes
makes	think	don’t	feel
taken	account	taking	steps
those	compared	group	pain
few	days	seven	months
I	can't	you	know
so	far	too	much
other	variables	between	categories
mrs	james	hellebore	said
children	aged	years	thousands

Table 6. C_4 (relationships between A-B, B-C, C-D and A-D)

Table 6 shows 15 C_4 graphs (having four links, where each word collocates with two other words only). As with the P_3 graphs, it is worth considering whether any of the words which do not collocate show semantic or grammatical relationships. So this involves looking at columns A-C and B-D. Here we see *ethnic-religious, day-night, can't-couldn't, behind-towards, I'm-he's, think-feel, taken-taking, few-seven, days-months, I-you, so-too, variables-categories*. We could view the non-collocating *mrs* and *hellebore* as functionally similar as both act as parts of names. In most of the cases of the C_4 graphs then, two of the non-collocating words show some sort of relationship. Either they are forms of the same lemma or from the same grammatical or semantic class or are synonyms or antonyms. Only three of the fifteen C_4 graphs examined do not contain some sort of obvious relationship between two of the non-collocating words. These are *risk-lives-people's-children*, *those-compared-group-pain*, and *children-aged-years-thousands*.

For the first atypical case, two 2 word lexical bundles were found: *people's lives* (7 occurrences), *children and young people's* (5 occurrences). On the other hand, the relationship between *risk* and *children* (8 occurrences) does not occur in any specific lexical bundle, but in a variety of combinations. Finally, *risk* and *lives* (5 occurrences), has three cases of *risk* (optional word) *their lives*, (although this does not meet the frequency criteria for a lexical bundle or frame). The other two cases contain the words *risk* and *lives* which occur close together but in different sentences.

The second C_4 graph which does not follow the trend is *those-compared-group-pain.* The words *compared* and *those* occur within a frame *compared* [preposition] *those* 8 times. *Group* and *compared* occur 6 times but in no clear pattern. *Group* and *pain* co-occur 23 times, of

which 13 are in the lexical bundle *pain group* (this occurs in a scientific study relating to how people experience pain). Finally, *those* and *pain* occur 13 times but again in no clear pattern.

The fourth non-typical case is *children-aged-years-thousands.* Here we see four distinct lexical bundles or frames: *children aged* (9 cases), *aged* (any number) *years* (19 cases), *thousands of years* (6), *thousands of* (optional modifier) *children* (5). There are additionally four cases where *children*, *aged* and *years* co-occur, in phrases like *children aged 5-15 years*.

It appears then that for K4 graphs, there is a strong likelihood that two of the non-collocating words will be similar in some way, and if this is not the case, then it is probable that the graph will consist of between 2 and 4 lexical bundles or frames.

A	B	C	D
harmonius	play	disharmonius	free
pay	men	paid	sex
bills	shall	bill	private
child	health	education	development
2001	et	2002	al
advice	help	further	information
shut	door	closed	behind
eyes	green	light	bright
bias	material	straining	along
compared	men	sex	women
total	costs	pounds	per
years	aged	per	25
muscle	get	ripped	meal
do	you	don’t	want
sorry	i'm	yes	oh

Table 7 K_4-e (diamonds) (relationships between A-B, B-C, C-D, D-A and B-D)

How about diamonds, which contain five relationships (A-B, B-C, C-D, D-A and B-D)? Fifteen cases were found (see Table 7), and here the only words which do not collocate are A-C. Is there anything to link the words in these two none-collocating columns? As with other types of four word graphs some of the rows do indicate relationships (6 out of 15 cases,

although this pattern is not as frequent for the C_4 graphs): *harmonius-disharmonius*, *pay-paid*, *bills-bill*, *2001-2002*, *shut-closed*, *do-don't*, *sorry-yes* (with the latter pair, I class both as 'discourse markers'). The diamond is actually made up of two triangles that are fused together, so perhaps we are seeing two sets of three-word lexical bundles? Let us examine a few cases more closely.

For *child-health-education-development*, *child health* forms a 2 word lexical bundle (7 cases), *health* and *education* often occur together in lists but do not form bundles, *education* and *development* do not show any clear relationship although do occur together 8 times, *child development* does occur as a lexical bundle (also 8 times), while finally *health* [conjunction] *development* is a frame which appears 7 times. In this case then, we are seeing combinations of two words rather than 3, sometimes in lexical bundles or frames, but not always.

How about *muscle-get-ripped-meal*? Everything except *muscle* and *ripped* collocate. The sequence *get muscle* is a lexical bundle (12 occurrences), although this also occurs as part of *get muscle workout* (4 times) and *get muscle meal plan* (6 times). Additionally, *get ripped* occurs 20 times in the corpus, while *get ripped meal plan* appears 11 times. There appears to be an equivalency then between *get muscle meal plan* and *get ripped meal plan*, explaining why *muscle* and *ripped* do not collocate – they are largely interchangeable. These two words do appear together four times (not enough to meet the frequency threshold of 5 for collocation), appearing in phrases like: *follow this plan to get ripped and build muscle* and *Build muscle/get ripped*. In such cases the use of the conjunction *and*, along with the forwards slash also shows the equivalency between the two terms.

Let us consider one of the clearer cases *pay-men-paid-sex*. Here everything collocates apart from *pay-paid*. There are two lexical bundles at play here: *men who pay for sex* (7 cases) and *men who (had) paid for sex* (6 cases). It is clear that this diamond is essentially due to a single

concept, which can be worded slightly different in terms of altering the tense. When we just look at the relationship between *men* and *sex* we find other variations in the corpus such as *men paying for female sex contacts*, *men reported paying for sex* and *men paying for sex*, but again, the concept is the same. Diamonds then have a less obvious relationship between words than C_4 graphs, encompassing a mixture of related non-collocates and related lexical bundles.

A	B	C	D
you'll	get	your	me
people's	other	lives	young
moon	riders	over	full
popular	music	most	very
deficit	account	billion	trade
talking	himself	about	I'm
concerned	far	about	person
areas	urban	where	other
fun	great	bit	it's

Table 8 Claws (relationships between A-B, A-C and A-D)

The nine cases of claws (Table 8) have word A at its centre, collocating with B, C and D, although none of the other words collocate with one another. Potentially, claws are easy to locate, as any word with three collocates that do not connect to one another can be classed as forming a claw. However, in order to focus purely on claws alone, and not claws that were partial sections of larger graphs (e.g. paws), I have limited the identification of claws to cases where a word *only* collocates with three other unconnected words in a network.

Perhaps naively we may predict that claws occur in cases where word A collocates with three words which all share a similar semantic or grammatical meaning, or at least A operates in three similar lexical bundles or frames. I expected then to see claws which potentially demonstrated semantic prosodies of word A. However, this was not the case for any of the

nine claws examined. There are some cases where two of the words which collocate with word A have a similar function e.g. take *popular*. It collocates with *most, very* and *music*. As seen earlier, *popular* modifies *music* in the bundle *popular music* 8 times. However, *most* and *very* commonly occur as modifiers of *popular*, fulfilling a similar function. Similar with *areas*, we see two modifying collocates (*urban* and *other*), and one collocate which functions differently (15 times in the bundle *areas where*). With *fun-great-bit-it's,* we see two bundles *great fun* (6) and *bit of fun* (7) both which function as modifiers of *fun*, although in different ways – *great fun* suggests a qualitative evaluation of the fun, while *bit of fun* is used idiomatically to minimise the amount of fun, and is used often to excuse behaviour labelled by others as problematic (*it was just/only a bit of fun* is a typical construction). Perhaps surprisingly, the collocate *it's* has no bundle or frame associated with *fun* but can be used in a wider range of constructions. Claws then, perhaps show the least clear patterning of relationships between words. When three words collocate with a fourth word, but not each other, they may have very different reasons for doing so.

A	B	C	D
committee	shall	private	bill
aged	over	50	years
get	ripped	workout	meal

Table 9 K_4 graphs (relationships between all words)

Finally, we come to Table 9 which shows K_4 graphs, or cases where every word collocates with every other word. As noted earlier, these were difficult to locate using the speculative method I employed, and only three were found from my searches of the forty words. While this means we must be cautious about making generalisations about K_4 graphs, let us consider

all three in turn. It was noted earlier that the K_4 graph found in the newspaper corpus was mostly due to a four word lexical bundle *hate cleric Anjem Choudary.* Are the other three K_4 graphs lexical bundles then?

First, *committee-shall-private-bill*. There is no four word bundle which occurs here. Instead, we find a set of related shorter bundles and frames. The bundle *the committee of selection shall* occurs seven times, indicating the relationship between *committee* and *shall*. Another bundle *private bill shall* occurs five times, encompassing three of the four words in the graph. While *bill* and *committee* occur together 18 times, they also occur in a lexical bundle *committee on an opposed private bill* (6 times), and this bundle also includes a third collocate, *private*. Therefore, this K_4 graph is not due to a single lexical bundle but three fairly frequent bundles which encompass two or three of the words in the graph.

The second case in Table 6 is *aged-over-50-years*, which reads like a bundle. However, this sequence only occurs once in the corpus. *Aged* and *50* do occur seven times together, in sequences like *aged under/over 50*. Another lexical bundle *50 years*, occurs seven times, while *over* and *years* appear 93 times together, notably in the lexical bundle *over the years* (19 cases) and the frame *over the next/past* (any number) *years* (25 cases). Another frame, *aged* (number) *years* appears 18 times. This graph then, is due to a mixture of common bundles as well as frames, rather than occurring in a fixed four word sequence.

Finally there is *get-ripped-workout-meal*. Again, this appears to resemble a four word lexical bundle, but in fact it is due to a different bundle which encompasses seven words and actually contains two sentences: *Get ripped workout Get ripped meal plan* (occurring four times in the corpus). All of the examples of this bundle occur in a bodybuilding magazine where a day by day plan is given for readers to follow. For example:

Thursday

Get ripped workout

Get ripped meal plan

Friday

Get ripped workout

Get ripped meal plan

These short sentences are represented as bullet points with no punctuation between them. However, it should also be noted that earlier we saw how *muscle-get-ripped-real* occurred as a diamond. In fact, six words: *muscle-get-ripped-real-plan-workout* occur as part of a larger graph, with all of these words collocating with at least three of the others. The K_4 graphs therefore do not follow the initially expected pattern of occurring within a 4 word lexical bundle. They may occur due to the presence of a smaller number of lexical bundles and/or frames, or more actually point to a larger graph.

The analysis therefore indicates that some graphs show a preference for words which have some sort of semantic or grammatical similarity (especially the C_4, diamond and triangle graphs). However, other graphs are more likely to be the result of two or more lexical bundles or frames (especially the P_3 and claw graphs).

Conclusion

The creation of the tool GraphColl literally adds a new dimension to collocation, both theoretically and methodologically. The majority of previous research which has used collocation has tended to focus on the relationship between two words. While there is clearly worth in such an endeavour, the ease and speed with which GraphColl can plot collocational networks is a ‘game-changer’ for corpus linguistics research, enabling more sophisticated

analyses to be carried out which focus on links between multiple words, rather than viewing pairs in relative isolation. My initial analysis of news articles about Muslims indicates that the collocational network approach resulted in a richer analysis than that carried out by using a traditional approach to collocation. Collocational networks, and in particular the different types of graphs which are suggestive of certain relationships between multiple words, are therefore a useful way forward for corpus linguistics research. For such work to be effective, analysts would be advised to become familiar with graph theory, helping them to identify different types of graphs within networks more readily. Some graphs show a tendency for their non-collocating words to have related meanings or come from the same grammatical category. Most of the graphs contained at least one lexical bundle or frame. Awareness of how graphs can help to spot words which have equivalent functions, bundles or frames is helpful for analysts who wish to interpret larger collocational networks. Concurrently, more work needs to be done in terms of understanding of how different graphs may outline different types of relationships between words. This paper is the beginning of an attempt to provide a systematic analysis of some of the simple graphs that occur within these networks. However, it should be viewed as a start, rather than an end.

As noted earlier, the analysis did not take into account directionality of collocation, as this was not an available aspect of using the MI score with GraphColl. Using the Delta P (Gries ibid) measure with GraphColl would have shown directionality of collocation although having experimented with this measure using different settings, I found it hard to avoid producing either a) a few high frequency grammatical collocates or b) too many collocates to visualise onscreen. However, using a measure that takes into account one-way directionality of collocation is likely to result in different graphs, and could offer an interesting direction for further research. Additionally, experimenting with different collocational techniques, cut-offs or corpora than those that I used would be worthwhile, allowing for greater (or less)

generalizability of some of the patterns I have described. Another criteria of collocation noted by Gries (2013) is dispersion, which again was not taken into account with the MI measure used. While the BE06 contains 500 texts, concordancing some of the words in graphs found resulted in some cases where collocates mainly or always appeared in a single text (such as the *muscle-get-ripped-meal* example above). Collocates which occur across multiple texts are likely to have greater validity and it would be interesting to see whether the graphs I have identified are more or less common if a cut-off for dispersion is applied.

Additionally, this paper has focussed on small number of three and four word graphs. My analysis found a small number of C_5 graphs containing five links (such as *financial-business-community-costs-help* and *take-off-stop-full-account*). Additionally, while claws did not uncover cases of semantic prosodies, related graphs containing more words (e.g. 5, 6 or 7 words where one word collocates to all the others) may show such prosodies. When working with five or more words, the number of possible graphs is much larger, and further work could try to take into account relationships between words in these cases. Finally, while I tried to place limits on the graphs I collected so that they did not appear to be smaller parts of larger graphs, it is rarely the case that it is completely possible to identify say a 'pure' triangle, where three words only collocate with one another and nothing else. Further work could consider how graphs relate to one another across whole networks.

Traditional collocational analysis takes into account pairwise relationships, and while such research often provides fruitful findings, it is perhaps a simplification of how people actually process language. Only considering pairwise relationships may mean that we miss related words which may play an important role in the construction of meaning. This paper therefore represents an initial stage in terms of thinking about the ramifications of collocation networks. There is clearly more work to be done.

References

AUTHOR (2005)

AUTHOR (2009)

AUTHOR (2014)

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).

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net.

Biber, D. Conrad, S. and Cortes, V. (2004) 'If you look at…: Lexical bundles in University teaching and textbooks.' *Applied Linguistics* 25/3: 371-405.

Brezina, V., McEnery, T. and Wattam, S. (2015) Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*. 20(2), 139-173.

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

Durrant, P. and Doherty, A. (2010), 'Are high frequency collocations psychologically real? Investigating the thesis of collocational priming.' *Corpus Linguistics and Linguistic Theory* 6 (2): 125-155.

Eeg-Olofsson, M. and Altenberg, B. (1994) 'Discontinuous recurrent word combinations in the London-Lund Corpus.' In U. Fries, G. Tottie and P. Schneider (eds) *Creating and Using English Language Corpora. Papers from the 14th ICAME Conference*. Amsterdam: Rodopi, pp. 63-77.

Firth, J. R. (1957) *Papers in Linguistics 1934–1951* (1957) London: Oxford University Press.

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.

Harris, J. M. (2000) *Combinatorics and Graph Theory*. New York: Springer-Verlag.

Hoey, M. (2005) *Lexical Priming. A New Theory of Words and Language.* London: Routledge.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

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

Phillips, M. (1989). *Lexical Structure of Text*. English language research. Birmingham: Birmingham University.

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

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

Scott, M. (1999) *WordSmith Tools version 3*, Oxford: Oxford University Press.

Stubbs, M. (1995) Collocations and Semantic Profiles. *Functions of Language, 2*(1), 23-55.

Stubbs, M. (2001) *Words and Phrases*. London: Blackwell.

West, D. B. (2001) *Introduction to Graph Theory*. Upper Saddle River: Prentice Hall.

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.