Studii de lingvistică 7, 2017, 125 - 144

Collocate networks in the language of crime journalism

David Brett[1]

Abstract: Standard procedures for the treatment of collocates, which involve the elaboration of lists of collocates on a two-by-two basis, are far from optimum for the study of connectivity, *i.e.* observing whether these collocates in turn display a tendency to co-occur or not. This paper explores an alternative strategy that has garnered considerable interest in recent years: that of using Social Network Analysis procedures.
Lists of collocates (concgrams) were extracted from a one million word corpus of crime journalism using standard techniques. Gephi software was then used to transform the list of collocates into a network.
A small number of collocate pairs were seen to be isolates, *i.e.* collocating only with each other, while the majority belonged to the giant component, composed of pairs in which at least one member collocates with at least one other word. Modules (clusters of highly interconnected collocates) were identified; these were seen to pertain to specific subject areas. The corpus was then re-examined to see where these clusters of collocates occurred, and co-occurred, and to gauge how much this technique may tell us about the 'aboutness' of particular texts.

Key words: social network analysis, collocation, collocate networks, newspaper language, crime journalism.

1. Introduction

1.1. Collocation

Collocation has been the object of scientific investigation for over half a century, yet, as Gries (2013:137) observes, this "does not mean that we as a field have arrived at a fairly unanimous understanding of what collocations are (in general), how they are best retrieved/extracted, how their strength or other characteristics are best measured/quantified, etc." Confusion concerning the first point, what collocations are, is discussed

[1] Università degli Studi di Sassari, Dipartimento di Scienze Umanistiche e Sociali; dbrett@uniss.it.

by Evert (2008: 1212-4), who makes a distinction between: "the *empirical* concept of recurrent and predictable word combinations" and "the *theoretical* concept of lexicalised, idiosyncratic multiword expressions". The first notion concerns those properties of word combinations that can be observed when subjecting large bodies of text to appropriate statistical procedures. Such types of analysis provide candidates for consideration as full-blown collocations in the second sense (*i.e.* non-compositionality, non-substitutability and non-modifiability, Manning and Schutze 1999: 172-3). In the following pages, when the term *collocation* is used, it is meant in the first sense. Similarly, the term *collocate* is used to indicate a member of such a combination.

In a recent paper, Brezina *et al.* (2015: 140) outline the criteria adopted for the automatic extraction of collocations as being based on:

1) *distance*: the maximum number of words around the node word in which collocates are to be searched for. This is often referred to as the *window*, or *span*, and is frequently set at four or five words to the left and right;

2) *frequency*: how often a given type appears within the collocation window;

3) *exclusivity*: the proportion between how often a given type occurs inside and outside the collocation window.

Dispersion is also another point of interest. Church and Gale (1995) use the expressions "bunchiness" and "burstiness" to describe the phenomena in which high frequency items may be concentrated in small sections, and hence be less typical than other items, which, while having lower frequency values, are more evenly distributed throughout the corpus. While this has generally been discussed in relation to single lexical items, the very same notion applies to combinations of two or more words, *i.e.* a collocation that is found consistently throughout the corpus is of greater importance than one whose frequency is concentrated in small sections.

Directionality is another aspect to be taken into consideration. The force of attraction between the two lexical items in a collocate pair is rarely equal, for example, consider the word pair *unleavened bread*. Having a mutual information score of 14.36 in the BNC[2], there is clearly a strong attraction between the two words, even though there are only 19 instances of the two words together. However, this attraction is not evenly distributed as there are 24 instances of *unleavened* in the BNC, *bread*, on the other hand, with its 3621 instances, is far more frequent. Therefore *unleavened* is far more attracted to *bread* than vice versa, as roughly four out of five times it occurs with *bread* in the co-text. On the other hand, co-occurrence with *unleavened* accounts for only one out of every 200 instances of *bread*. Sinclair (1991: 115-16) describes and provides terminology for this phenomenon*:* "When *a* is node and

[2] Data from British National Corpus: http://corpus.byu.edu/bnc/.

b is collocate, I shall call this *downward collocation* – collocation of *a* with a less frequent word *b*. When *b* is node and *a* is collocate, I shall call this *upward collocation".* Gries (2013) evaluates a series of tests that have been developed in recent years to account for directionality, concluding that deltaP[3] is the most suitable measurement to account for inequality in forces of attraction.

A final feature discussed by Brezina *et al.* (2015: 141) is that of *connectivity*: "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". Therefore, the fact that much of the work conducted on collocation up to now has focused on the co-occurrence of word pairs may mean that important properties have been and continue to be overlooked. Since corpus linguistics as a field does not avail itself of theoretical constructs and practical tools to deal with complex networks, we need to search for a field that does. One strong candidate is the field of Social Network Analysis, one whose history and development have a great many parallels to that of corpus linguistics. The objective of this paper is to illustrate the application of such techniques to a corpus of crime journalism articles, in order to evaluate the insight that the adoption of these tools provides. The results will be compared with those from a previous study conducted by the author on a corpus of travel journalism.

1.2. Social Network Analysis

A great deal of attention has been paid in recent years to the ways in which people behave and interact in offline, and ever more frequently in online environments. This has given rise to a whole new methodology called Social Network Analysis (SNA). One of the earliest studies of this type can be found in Moreno (1960: 35), in which 26 schoolgirls were asked to express who their first and second choices for dining partner were. As can be imagined some girls were more popular than others and were chosen more frequently than the average of two, others, conversely, were seen to be not very sought after. Some groups of two or more girls expressed a mutual preference, suggesting that in the real world these would constitute highly connected components through which information would travel more quickly than in other parts of the network. Modern information technology greatly facilitates the creation of networks with thousands of components, and the subjects that have been studied range from Facebook friend networks,

[3] DeltaP is "a simple directional association measure derived from the domain of associative learning" (Gries 2013: 140) that provides results that can be plotted on a scale from -1 to 1 to show directionality, against a measurement of the strength of collocation, such as log-likelihood, or mutual information.

to connections between bloggers, to links between Wikipedia pages (Easley & Kleinberg 2010).

The potential of such methods for providing insight into what we will term "communities of collocates" is most certainly an avenue worth exploring. Indeed the oft-cited phrase by Firth (1957: 11), "You shall know a word by the company it keeps", would suggest that any methodology developed for the quantitative analysis of social interaction should at least be explored by linguistics to verify its applicability to the observation of such a phenomenon as collocation. Conversely, social network analysis tools and methods display a great deal of versatility and robustness, allowing them to be used to explore a plethora of phenomena, some even far removed from the original scenarios of social interaction, such as international trade (Krempel & Plümper 2003) and airport connections (Cheung & Gunes 2012). Therefore its application to corpus linguistics is, at the very least, promising.

2. Materials and methods

2.1. Materials

The corpus of texts examined in this work is a collection of articles from the 'UK Crime' section of the British daily *The Guardian* called the Guardian Crime Corpus (GCC)[4]. The corpus is comprised of 1820 articles, amounting to one million tokens. The articles appeared in the online version of the newspaper[5] over a period of four months in late 2011.

2.2. Methodology

2.2.1. Annotation for part-of-speech

The corpus was annotated for part-of-speech (PoS) using Tree Tagger[6], a tool that not only attributes a PoS tag to each token in the text, but also provides its lemma. Thereafter, the lemmas constituted the focus of the work, *i.e.* rather than dealing with a series of different frequencies for the word forms *find* [...] *body*, *find* [...] *bodies, found* [...] *body*, *found* [...] *bodies*, etc., working with the list of lemmas allowed us to search for a single entry FIND [...] BODY. Since all inflected types were normalised to their base forms via lemmatisation, the PoS tags were also simplified to indicate merely whether the token was a lexical verb (VV), adjective (JJ) or common noun (NN).

[4] The texts were gathered at the University of Sassari, Italy, in 2011, by the author and his colleague Prof. Antonio Pinna.
[5] theguardian.com.
[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

2.2.2. Collocate extraction

The first part of the analysis, concerning the extraction of the collocates, was conducted using scripts written by the author in perl[7]. The procedure is essentially that described for the extraction of 'concgrams', *i.e.* pairs of collocates that are fixed in terms of neither position nor constituency (Cheng *et al.* 2009, Greaves 2009). The statistical measurement of collocation strength decided upon was that of log-likelihood (Cantos Gómez 2013, Dunning 1993, Oakes 1998)[8].

Brezina *et al.* (2015: 146) stress the importance of providing clear outlines as to the statistical procedures applied, so that replicability be assured. Following their recommendations, the details are listed below:

1) the statistical procedure adopted was that of log-likelihood (LL),

2) the cut-off value for significance was 10.87, corresponding to $p<0.001$,

3) the window around each node was L4-R4,

4) minimum collocate freq. = 5,

5) minimum co-occurrence = 5,

6) only types belonging to the three main lexical parts-of-speech (Lexical Verb, Adjective, Common Noun) were included in the calculations[9]. Identical lemmas belonging to different parts of speech were kept separate by way of the following notation: LEMMA_PoS (*e.g.* HAND_NN, HAND_VV). No filters were utilised in the application of the L4-R4 window, *i.e.* punctuation was included.

Lemma A	**Freq. A**	**Lemma B**	**Freq. B**	**Freq. AB**	**LL**
COURT	2658	CROWN	541	494	2456.85
POLICE	5584	OFFICER	2018	598	1674.43
COURT	2658	HEAR	1016	419	1598.71
YEAR	2448	LAST	1321	401	1428.47
GUILTY	544	PLEAD	225	199	1269.05
MINISTER	384	PRIME	183	171	1172.1
FIND	2019	BODY	823	284	1093.92
LAST	1321	WEEK	902	261	1060.63
PEOPLE	2419	YOUNG	1044	301	1053.84

[7] The perl programming language has been indicated as an optimum choice for corpus linguistics researchers desiring to create *ad hoc* tools (Danielsson 2004).

[8] Evert (2008: 1218) notes the tendency of the MI score to highlight rare occurrences, indicating log-likelihood as being a more balanced measurement, especially when combined with PoS filters to block function words. Log-likelihood also provides clear cut-off points, enabling the researcher to adopt widely accepted criteria for distinguishing between significant and non-significant co-occurrences.

[9] In an initial phase prepositions were also included, however the results obtained featured clusters dominated by prepositions collocating with lexical words which mutually had little in common. Note that BE and HAVE were excluded as they have their own tags, starting with VB and VH respectively.

TAKE	2179	PLACE	545	253	1038
ARREST	1128	SUSPICION	354	194	950.83
CUSTODY	389	REMAND	159	138	913.98
PHONE	432	MOBILE	203	147	912.55
COURT	2658	TELL	2109	324	910.07
COURT	2658	MAGISTRATE	386	218	905.37
SAY	9305	SPOKESMAN	321	245	850.31
FIND	2019	GUILTY	544	205	805.24
HOME	1361	SECRETARY	291	167	797.12
CHIEF	356	CONSTABLE	166	120	764.57
DO	3835	KNOW	1412	279	755.88

Table 1: The twenty most statistically significant collocate pairs in the GCC

After the above procedure was applied to the data, 8689 statistically significant collocate pairs were found. Of great interest is the fact that these collocate pairs were composed of only 2663 lemmas. Therefore, on average each lemma collocates with 3.26 other lemmas[10]. In other words, there is considerable connectivity between the collocates. This becomes evident simply by observing the twenty collocate pairs in Table 1. While most lemmas occur only once on the list, some are repeated: the noun COURT collocates with the nouns CROWN and MAGISTRATE, and the verbs TELL and HEAR; the verb FIND collocates with the noun BODY and the adjective GUILTY; the latter also collocates with the verb PLEAD; finally, LAST collocates with the nouns WEEK and YEAR.

In the terminology of Social Network Analysis this would be described as follows: the noun COURT has a degree of 4; the verb FIND, and the adjectives GUILTY and LAST have a degree of 2; all the other nodes have a degree of 1. The lemma COURT therefore emerges as a *hub*, a node which is highly interconnected. In Corpus Linguistics terms it would be described as a lemma that has a marked tendency to form collocations within the genre under examination.

2.2.3. The importation of data to network analysis software

As described up to this point, the methodology applied is more or less standard Corpus Linguistics practice. When considering how to analyse the connectivity of the results we are venturing into somewhat uncharted territory with few previous studies providing an example to be followed. The software of choice in this work is Gephi[11], an open-source and free package, which runs on all the main operating systems. It is a particularly powerful and flexible tool, both in terms of the vast array of filters and statistical procedures that can be applied to the data and the range of

[10] Alternatively, to use Social Network Analysis terminology, average degree is 3.26.
[11] http://gephi.github.io/.

options that can be availed of when elaborating graphical representations of the networks created. The downside for the corpus linguist is that, unlike Brezina *et al.*'s (2015: 141) GraphColl, it is not a "one-stop" tool. To the contrary, there are four separate steps to the procedure in this study: tagging for PoS and lemma; collocate extraction; formatting of data for importation to Gephi; elaboration of networks in Gephi. However, despite the fact that the process is far from straightforward, the pay-off once the data are imported to Gephi is considerable. A considerable number of further options are available to the researcher, concerning filters to focus on particular nodes, or groups of nodes, and various ways to arrange the nodes (*e.g.* expand, reduce, circular format, etc.). Finally, numerous algorithms are available to allow further analysis of the networks. This final point will be returned to below.

The process of preparing the data obtained from the collocation extraction process for importation to Gephi is relatively straightforward. Two plain text files must be prepared: an 'Edges' file (see example in Table 2), which adds a column containing the word 'undirected' to the data. A second plain text file must contain a simple list of the nodes. The lemma appears both attached to ('id') and detached from ('label' and 'PoS') the PoS tag to provide for a distinction between, for instance, ABUSE_NN and ABUSE_VV, and to allow PoS data to be treated separately (see Table 3).

source	**target**	**type**	**freq.**	**LL**
COURT_NN	CROWN_NN	undirected	494	2456.846
POLICE_NN	OFFICER_NN	undirected	598	1674.43
COURT_NN	HEAR_VV	undirected	419	1598.71

Table 2: An example of collocate data ready for importation to Gephi as an edges table

id	**label**	**PoS**	**freq.**
COURT_NN	COURT	NN	2658
POLICE_NN	POLICE	NN	5584
OFFICER_NN	OFFICER	NN	2018
Etc.			

Table 3: Collocate data concerning nodes ready for importation to Gephi

3. Results and discussion

3.1. Node degree, the giant component and isolates

When applying SNA procedures to the analysis of collocation, some important decisions must be made at the outset. The first concerns the treatment of strength of collocation. We can simply consider collocation

as a Boolean property, *i.e.* two terms collocate = TRUE/FALSE, setting the value at TRUE when a word or lemma pair exceeds a threshold value in a statistical test such as log-likelihood. A second option, that which is adopted in this paper, involves taking into consideration how much the word pair exceeds the threshold, *e.g.* providing each edge with a weight based on the result of the statistical test: this allows stronger collocations to be readily visible in the resulting graphs.

A second decision is related to the property that determines the size of the nodes. In this case there are three options. The first, rather reductive strategy, is to merely leave all the nodes the same size. The other two possibilities involve scaling the node size of a given lemma on the basis of either its raw frequency or its degree, the latter being the number of other lemmas with which it collocates. In some cases, such as the analysis of the collocates of a given word, an indication of raw frequency may be particularly useful. In this paper, where interconnectivity is the main focus, node size is set to give an indication of degree.

Provided with the raw data concerning links on a two-by-two basis, software for the analysis of social networks, such as Gephi, can instantly construct a network composed of edges and nodes. With a few simple queries, considerable information can be obtained. For example, the number (and percentage) of nodes belonging to the *giant component*[12] can be calculated. The actual members of the giant component are of less interest than the isolates that lie without. The latter are by definition collocations composed of words that collocate only with each other. These, which to use a term inspired by Easley & Kleinberg (2010)[13] we may term "desert island collocations", are composed of two elements that are generally low frequency items, but with a very strong force of attraction. Some examples from the GCC can be seen in Table 4. The proportion of these isolates with respect to the giant component (1.5% in the GCC) may well differ across genres and constitute a fruitful field for further study. In fact, a similar study (Brett, in press), conducted on a corpus of travel journalism, found a proportion of collocate pairs not belonging to the giant component that was more than three times greater (4.7%) the above value.

	Lemma A	**PoS A**	**f A**	**Lemma B**	**PoS B**	**f B**	**f AB**	**LL**
1	PROTEST	VV	44	INNOCENCE	NN	53	18	131.4
2	BASEBALL	NN	16	BAT	NN	14	13	128.72
3	AGGRAVATING	JJ	21	FACTOR	NN	83	15	113.03
4	MASSAGE	NN	12	PARLOUR	NN	13	11	110.99
5	BROAD	JJ	26	DAYLIGHT	NN	22	13	109.97
6	AID	VV	19	ABET	VV	11	11	91.72

[12] The giant component has been defined as "a connected component that contains a significant fraction of all the nodes" (Easley & Kleinberg 2010: 31).
[13] Easley & Kleinberg (2010) make reference to people living on tropical islands as being potential non-members of a hypothetical giant component.

7	UNDERGO	VV	43	SURGERY	NN	47	13	91.59
8	HEROIN	NN	58	ADDICT	NN	22	12	89.88
9	SHOTGUN	NN	50	SAWN-OFF	JJ	10	10	71.33
10	PURE	JJ	24	CRIMINALITY	NN	90	10	67.43

Table 4: Some examples of "desert island" collocates[14]

Conversely, we may also be interested in seeing which words are the most "promiscuous" in their formation of collocations. These are to be found in the giant component and are instantly identifiable on a graph of the network in which node size is set to mirror degree[15]. Table 5 displays the top twenty lemmas in the GCC in terms of degree. One feature that is readily apparent is the fact that the list is dominated by nouns (16), with relatively few verbs (4) and no adjectives. This is to be contrasted with the findings in a similar study of travel journalism reported in Brett (in press), which yielded a far more balanced list in terms of proportions of nouns, verbs and adjectives, these being 8, 8 and 4, respectively.

	Lemma	**PoS**	**Degree**		**Lemma**	**PoS**	**Degree**
1	POLICE	NN	71	11	CASE	NN	32
2	OFFICER	NN	46	12	DO	VV	32
3	CRIME	NN	45	13	OFFENCE	NN	31
4	MAKE	VV	42	14	TAKE	VV	30
5	MURDER	NN	40	15	HOME	NN	29
6	CHILD	NN	39	16	FIND	VV	27
7	COURT	NN	38	17	LIFE	NN	27
8	SENTENCE	NN	35	18	YEAR	NN	27
9	MAN	NN	33	19	EVIDENCE	NN	26
10	PEOPLE	NN	33	20	FAMILY	NN	26

Table 5: The words in the GCC with the highest degree[16]

3.2. Identification of communities of collocates

While Social Network Analysis as a discipline is best known for its visuals that "have provided investigators with new insights about network structures and have helped them to communicate those insights to others" (Freeman 2000), it also offers an array of mathematical procedures to analyse the structure of static networks and track changes to dynamic ones. Some of these are of little interest

[14] Word pairs that collocate exclusively, and hence do not belong to the giant component.
[15] Alternatively once degree is calculated, we may visualise the list of lemmas with the highest degree in Gephi's 'data laboratory'.
[16] Those forming the greatest number of collocate pairs.

to linguists[17], one which may provide considerable insight into the connectivity of collocation pairs concerns the identification of modules or groups of nodes which are more highly connected with each other than with other nodes. Using a software tool like Gephi, the data at our disposal can be elaborated in two ways.

The first is visual: a graph of all the nodes and edges, initially laid out in a random fashion, can be processed using an algorithm such as Force Atlas 2, which brings (strongly) connected nodes closer together and distances unconnected (or weakly connected) nodes.

The second uses an algorithm that attributes nodes to modules, or clusters, of highly interconnected nodes. Each node is then provided with an authority rating, an indication of how much the node contributes to the interconnectedness of the module.

In the current study, running the modularity algorithm[18] on the data provided a large number of modules, some containing a great many nodes, some very few. We chose to focus on those that have a number of nodes that is both manageable and sufficiently informative, *i.e.* from 34 to 44. For obvious constraints relating to space only a small number of these modules will be discussed. The modules chosen roughly relate to three stages of the history of a criminal act: the act itself, its discovery and the legal consequences.

3.3. Module 14: DEATH

Module 14 contains 34 lemmas, which make up 2.64% of the total number of nodes in the network. The most authoritative members are DEATH_NN, DIE_VV, HOSPITAL_NN, SUFFER_VV and INJURY_NN. The collocates in this module relate to a very clear scenario: that of acts of violence and their effects and consequences, be they medical treatment for the victim or the search for the culprit. These semantic fields are summarised in Table 6.

Bodily harm	DEATH, DIE, SUFFER, INJURY, WOUND, FATAL, SEVERE, SUSTAIN
Violent acts	STAB, THREAT, SHOOTING, KICK, GUNSHOT, PUNCH
Medical treatment	HOSPITAL, TREAT, PATIENT, TREATMENT, NURSE
Investigation	CAUSE, CIRCUMSTANCE, ESTABLISH, SUSPICIOUS

Table 6: Selected members of Module 14 organised into semantic fields

[17] For example, the shortest path from one node to another, *i.e.* starting from node A how many edges must one travel down, and how many nodes must one pass through to get to node B? While this would clearly be of interest to those designing transport networks, for linguistic purposes, its use is limited. The same applies to techniques for the study of dynamic networks, as corpora are generally considered to be a closed, static phenomenon, with the sole exception, perhaps, of monitor corpora.

[18] The settings for the Modularity Community Detection Module were the following: Resolution=0.5 (the default value, 1.0, provided few modules that were too large and heterogeneous for our purposes); "Randomize" and "Use weights" were checked.

Figure 1: A graphic representation of the collocates in Module 14[19]

Example 1 shows selected concordance lines containing the lemma SUFFER as a verb and other collocates from Module 14. The verb co-occurs with two collocate pairs, STAB+WOUND and HEAD+INJURY, which in turn tend to be mutually exclusive[20]. The injuries sustained are often qualified by an adjective such as *multiple, severe* and *fatal*; the first showing a preference to co-occur with STAB+WOUND (lines 1-4), the second with HEAD+INJURY (lines 8-11), while the third can be found with both. Another interesting pattern to emerge is that of SUFFER+(a)+[type/severity]+WOUND+*to*+*the*+[body part], as can be seen in lines 15-21.

Example 1: Selected concordance lines for Module 14 containing the lemma SUFFER as a verb[21]

1 off the resort of Kusadasi. They **suffered multiple stab wounds,** including having their
2 victims, aged 18 and 19, **suffered multiple stab wounds.** One is in
3 the PCSO and the officer, who **suffered multiple stab wounds** which were initially thought
4 Five young men **suffered multiple stab wounds** today in an apparent
5 from Wythenshawe, south Manchester, **suffered fatal stab wounds** in the struggle.

[19] In this and in the following figures: node size is indicative of degree; node colour is indicative of Part-of-Speech (red, green and blue for adjectives, nouns and verbs, respectively); and edge size is indicative of strength of collocation (log-likelihood).
[20] In the GCC corpus there are 21 instances of STAB+WOUND and none of STAB+INJURY. Similarly there are 60 instances of HEAD+INJURY and only 7 of HEAD+WOUND.
[21] In this and in the following examples, the tokens belonging to the module in question are in bold.

6 the householder before one of the assailants **suffered fatal** knife **injuries.** Floral tributes to
7 in Italy. Both victims had **suffered** horrific **stab wounds** to their chest and
8 The jury heard that Thomas had **suffered severe head injuries** when he fell out
9 and painful death". Tom Inglis **suffered severe head injuries** when he fell out
10 A post-mortem examination showed she had **suffered severe head injuries.** On Monday,
11 subjected to repeated abuse. Tom Inglis **suffered severe head injuries** when he fell out
12 killing a Wales football fan who **suffered fatal head injuries** outside Wembley stadium.
13 on Sunday 13 March. Ashton had **suffered fatal head injuries.** A £20,000 reward
14 Tong on Tuesday night. Both had **suffered** serious **head injuries.** The women,
15 said. Trevor Ellis, 26, **suffered** a **gunshot wound** to the head after
16 was loaded. The IPCC said Duggan **suffered gunshot wounds** to his chest and right
17 night as Trevor Ellis, 26. He **suffered** a **gunshot wound** to the **head**.
18 who lived in Edmonton, had **suffered stab wounds** to the chest and thigh
19 for the prosecution, said their father **suffered stab wounds** to the face, neck
20 told the court that Antoni Robinson **suffered stab wounds** to his face, neck
21 locally as Marvin Henry, was found **suffering** a **fatal wound** to his torso after

Another group of highly interconnected collocates that can be observed in Figure 1 concerns the lemmas DEATH, TREAT and SUSPICIOUS. Example 2, which shows selected concordance lines for TREAT as a verb, provides a very clear example of the extra insights that can be obtained by adopting SNA-inspired techniques for the study of collocations. Traditional analysis would not automatically favour the recognition of the clear pattern that the concordance lines display. While DEATH+TREAT, TREAT+SUSPICIOUS, DEATH+SUSPICIOUS would all appear on a list of collocate pairs, the repeated cooccurrence of all three would not be underlined. Furthermore, *n*-gram analysis would be of little more help, as it would pick up on the 4 instances of the 4-gram *being treated as suspicious* in lines 1-4, as well as the 3 instances of the 5-gram *treating the death as suspicious*, the slight variability in the slots to the left of both would suffice to obfuscate the connection with *death(s)* in the former case and *treating* in the latter.

Example 2: Selected concordance lines for Cluster 14 containing the lemma TREAT as a verb

1 time. "The **death** is being **treated** as **suspicious** and police are currently following
2 Yard said. The **death** is being **treated** as **suspicious** and a postmortem examination is
3 said: "The **deaths** are being **treated** as **suspicious** and an investigation is now
4 two. Their **deaths** are being **treated** as **suspicious** by Leicestershire police, although
5 Though officially the **death** was still being **treated** as "**suspicious**", privately detectives
6 at Inverness. Officers say they are **treating** the **death** as **suspicious**. The body
7 frozen, but police said they were **treating** the **death** as **suspicious**. The postmortem
8 and serious crime command are investigating and **treating** the **death** as **suspicious**. Reports
9 Joanna Yeates. Officers said they were **treating** the **death** as "**suspicious**" and
10 M4. Police said they were not **treating** his **death** as **suspicious** or looking for
11 "He added that officers were **treating** her **death** as **suspicious**. A source
12 "Officers are, however, **treating** Joanna's **death** as **suspicious** at this

3.4. Module 64: FIND

Module 64 contains 38 lemmas, which make up 2.95% of the total number of nodes in the network. The most authoritative members are FIND_VV, BODY_NN and MOTHER_NN. The collocates in this module relate to the discovery of the victim of an act of violence. Many of the collocates belong to a limited number of semantic fields, which are summarised in Table 7.

Family relations	MOTHER, DAUGHTER, SON, BROTHER, FATHER, BOYFRIEND, SISTER
Verbs connected with crime	FIND, LIE, DISCOVER, DUMP, HIDE, LODGE
Location	MILE, FLAT, ROADSIDE, VERGE, WOOD, FLOOR

Table 7: Selected members of Module 64 organised into semantic fields

Figure 2: A graphic representation of the collocates in Module 64

Selected concordances for Module 64 including the lemma BODY (Example 3) provide a clear indication of the contexts in which the collocates co-occur. One context is that of describing the condition of the body, or the circumstances in which it was found (lines 1-5); the location in which the body was found is another (6-10); finally frequent reference is made to the kinship relationship between the victim of the crime and the person who finds the body (11-15), presumably relating to crimes that take place within the domestic environment.

Example 3: Selected concordance lines for Module 64 containing the lemma BODY

1 Walton. Her **naked** and badly decomposed **body** was **found** six months later in Yateley
2 sex worker, even though her **naked body** was **found** in his flat in 2002
3 Parnell and Steve Jones **found** six **dead bodies** and seven or eight seriously wounded
4 often the first to **find** a **dead body,** are those of the family Calliphoridae
5 the law student placed his partly-burnt **body** in **plastic bags** before burying the remains
6 abducted and killed her before **dumping** her **body** 25 **miles** away. Bellfield, 43
7 her in his **flat** before **dumping** the **body.** Six months later, her remains

8 not get far. At 9.30pm his **body** was **discovered lying** next to his BMW
9 on the roadside **verge** where Yeates' **body** was **found** on Christmas morning and the
10 a good six **miles** from where the **body** was **found.** He kept saying :
11 horror" of **finding** her **mother**'s **body** and being told by police that she
12 11, **found** their **mother**'s **body** on their return from school. Terry
13 **daughter**, who **discovered** her **father**'s **body** and has suffered post-traumatic stress ever
14 A year after his **daughter**'s **body** was **found** next to a motorway,
15 teams, would **find** his **daughter**'s **body.** "If we could put a

3.5. Module 41: MURDER_NN

Module 41 contains 44 lemmas, which make up 3.41% of the total number of nodes in the network. The most authoritative members are CHARGE_NN, MURDER_NN and OFFENCE_NN. The collocates in this module relate to the legal consequences of the criminal act and the grouping features a number of terms concerned with how the justice system deals with defendants and the latter's reaction. These include ACCUSE, ADMIT, CHARGE, CLEAR, COMMIT, CONVICT, COUNT (as a noun), DENY, DROP, ENTER, GUILTY, OFFENCE, PLEA and PLEAD[22]. Other members of cluster 41 can be ascribed the subset of "criminal act". These include: ABDUCTION, ASSAULT, BURGLARY, DEFRAUD, KIDNAP, MANSLAUGHTER, MURDER, RAPE and THEFT. Example 4 displays a selection of concordance lines containing CHARGE_NN and other members of Module 41. Three patterns emerge: 1) the position of the defence with regards to the accusation (lines 1-3); 2) the alleged crime, which may be associated with CHARGE by way of post-modification (lines 4-5) or compounding (line 6); 3) a combination of CHARGE_NN, the name of the alleged crime and the position adopted by the defence or the judiciary (lines 7-10). It is precisely in these last lines that we see a particularly clear example of the tendency of collocates to cluster. In each case, in short stretches of eight or nine words, we find four or five words that belong to a community that has been identified using Social Network Analysis tools.

Example 4: Selected concordance lines for Module 41 containing the lemma CHARGE as a noun

1 death. Riggi **pleaded guilty** to reduced **charges** of culpable homicide, the Scottish

[22] We may note that ENTER+PLEA and DROP+CHARGES are two rather effective examples of collocation in the strictest sense, as the meaning of the verbs in both cases is considerably different from their base meaning.

2 a former lecturer, initially **denied** the **charges** but **pleaded guilty** at London's Southwark
3 Anderson **pleaded guilty** to one **charge** of causing unauthorised modifications
4 were stolen. His 15-year-old brother faces **charges** of **violent disorder** and burglary over
5 of attempting to **murder** Rathband, one **charge** of **conspiracy** to **murder**,
6 theft, **violent disorder** and **burglary charges.** Ryan Kelly, 20, was
7 He also **pleaded guilty** to a **charge** of **violent disorder,** admitting he overturned
8 required to **enter** a **plea** to the **charge** of **manslaughter** of Michael Dye before being
9 legal reasons, were **cleared** of the **charge** but **convicted** of **manslaughter**. Enoch Amoah,
10 of Camberwell, was **cleared** of both **charges** but **convicted** of **violent disorder**. All

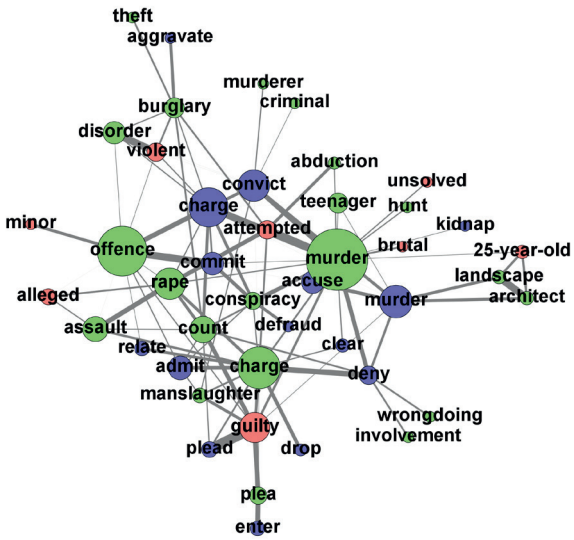

Figure 3: A graphic representation of the network of collocates in Module 41

3.6. The distribution of modules within the corpus

A final stage in the analysis involves returning to the corpus and observing the distribution of the modules. Research questions include: Are the members of certain modules concentrated in certain texts, or are they well dispersed across the corpus? Do some modules show a tendency to co-occur or, conversely, to show complementary distribution?

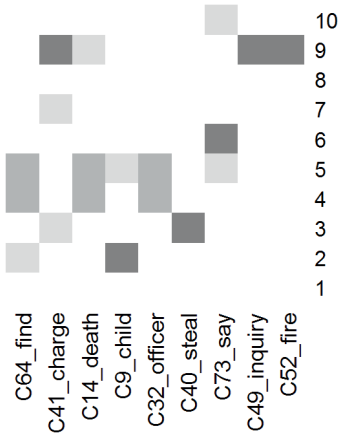

Figure 4: Heatmap showing the co-occurrences of collocates belonging to nine modules in ten texts

Figure 4 consists of a heatmap showing the co-occurrences of collocates[23] of nine modules in the first ten texts of the corpus. It is immediately clear that text five, a short article, about 240 words long, contains a number of co-occurrences of collocate pairs belonging to Modules 9, 14, 32, 64 and 73[24], the most authoritative members of which are CHILD, DEATH, OFFICER, SAY and FIND, respectively. In fact, the article in question is about the investigation into the death of a four-year-old child. Similarly, text nine, which features collocate pairs from Modules 52, 49, 14 and 41 (the most authoritative members of which are FIRE, INQUIRY, DEATH and CHARGE, respectively), is about the trials of individuals found to be involved in the August 2011 riots. Therefore, observing the occurrence and co-occurrence of modules, each of which can be roughly equated to a given scenario and/or semantic field, can provide a considerable amount of information regarding the "aboutness" of individual texts.

4. Conclusions

This article begins with a review of the literature that underlines the complexity of the phenomenon of collocation and outlines the state of the art in current understanding, highlighting several areas that have been identified as being in need of further work. One of these areas is undoubtedly that of connectivity. Up to now, collocates have generally been reported and discussed mainly on a two-by-two basis, an approach

[23] The data concern the co-occurrence of collocate pairs of a given module within the 4L 4R span indicated above, not merely their being present singularly within the same text. The script used for this purpose was developed by the author.
[24] See Appendix A for the modules that are not illustrated above.

that does not facilitate the identification of larger patterns or groupings. Recent research (Brezina *et al.* 2015; Brezina 2016; Brett, in press) has underlined how considerable insight may be gained by presenting collocate pairings in the form of networks. In this way, the terms that are most highly connected emerge immediately, as do the most exclusive pairs. The properties of both may be observed (*e.g.* the PoS categories of the collocates), and compared with their counterparts in other text types. In the case of crime journalism, this study has demonstrated that the list of the most interconnected nodes is dominated by nouns. Work by the author on travel journalism (Brett, in press), suggests that other text types may display more balanced distributions in terms of Part-of-Speech.

Further analysis can be conducted by applying algorithms to identify modules or groups of collocates that are more closely interconnected. Examination of the members of these modules, especially those that are most authoritative, often reveals that the modules correspond to specific scenarios. In the case of the crime articles examined in this paper, these scenarios were the different stages of a criminal investigation, such as an act of violence, the discovery of a body, a suspect being arrested and charged, the trial, etc. Research by the same author on a corpus of travel journalism (Brett, in press) yielded modules that were connected with different types of environment (*e.g.* mountain, river, town), activity (food, wine, festival) and aspects of the travel experience (route, room, tour).

The occurrence and co-occurrence of these modules in individual texts within the corpus may then be examined. Such a procedure has revealed itself to be a reliable indicator of the main subjects of texts. As such the procedure may be a useful tool for automatic sense disambiguation for application in such fields as machine translation and semantic annotation.

The findings of this paper are hence in agreement with those of Brezina (2016), who states that "collocation networks represent an efficient way of analysing complex meaning relationships in discourse" that "enable us to visualize and analyze linguistic practices that give rise to complex meanings of texts and discourses". It is to be underlined, however, that while such procedures may well become a standard part of the corpus linguist's toolkit in the future, the application of such methodologies is at a very early stage, and a great deal of further research is necessary both to confirm their effectiveness and develop guidelines for best practice.

References

Brett, D. (in press), "Social Network Analysis and the Analysis of Collocations in the Language of Travel Journalism", in Baumann, T. (ed.) *Reiseführer - Sprach- und Kulturmittlung im Tourismus / Le guide turistiche - mediazione linguistica e culturale in ambito turistico*, Peter Lang, Bern.

Brezina, V. (2016), "Collocation Networks, Exploring Associations in Discourse", in Baker, P., Egbert, J. (eds), *Triangulating Methodological Approaches in Corpus Linguistic Research,* Routledge, London, p. 90-107.

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

Cantos Gómez, P. (2013), *Statistical Methods in Language and Linguistic Research,* Equinox, London.

Cheng, W., Greaves, C., Sinclair, J., Warren, M. (2009), "Uncovering the Extent of the Phraseological Tendency: Towards a Systematic Analysis of Concgrams", *Applied Linguistics,* 30/2, p. 236-252.

Cheung, D. P., Gunes, M. H. (2012), "A Complex Network Analysis of the United States Air Transportation", in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, p. 699-701.

Church, K.W., Gale, W. A. (1995), "Poisson mixtures", *Natural Language Engineering,* 1/2, p. 163-190.

Danielsson, P. (2004), "Programming: Simple Perl programming for corpus work", in Sinclair, John McH. (ed.), *How to Use Corpora in Language Teaching*, John Benjamins, Amsterdam, p. 225-246.

Dunning, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational linguistics*, 19/1, p. 61-74.

Easley, D., Kleinberg, J. (2010), *Networks, Crowds and Markets: Reasoning about a Highly Connected World*, Cambridge University Press, Cambridge.

Evert, S. (2008), "Corpora and collocations" in Lüdeling, A., Kytö, M. (eds), *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, Berlin, p. 1212-1248.

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

Freeman, L. (2000), "Visualizing Social Networks", *Journal of Social Structure*, 1 (on line: https://www.cmu.edu/joss/content/articles/volume1/Freeman.html)

Greaves, C. (2009), *Concgram © 1.0. A Phraseological Search Engine*, John Benjamins, Amsterdam.

Gries, S. (2013), "50-something years of work on collocations: What is or should be next…", *International Journal of Corpus Linguistics*. 20/2, p. 137-166.

Krempel, L., Plümper, T. (2004), "Exploring the Dynamics of International Trade by Combining the Comparative Advantages of Multivariate Statistics and Network Visualizations", *Journal of Social Structure,* 4.

Manning C. D., Schütze, H. (1999), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA.

Moreno, J. L. (1960), *The Sociometry Reader*, The Free Press, Glencoe, Illinois.

Oakes, M. (1998), *Statistics for Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Sinclair, J. (1991), *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

Appendix

The members of five modules identified in the GCC. The most authoritative members of each module are in bold.

Module	Members
9	**CHILD, SEXUAL, OFFENDER, SEX,** ABUSE, INDECENT, ACTIVITY, IMAGE, POSSESS, TRAFFICKING, WORKER, ABUSE, ACT, ADULT, ASSIST, CARE, HUMAN, RIGHT, EUROPEAN, BREACH, DISTRIBUTE, EXPLOITATION, INSTITUTION, NURSERY, PARENT, POSITION, REGISTER, TRUST, APPROPRIATE, CRUELTY, ELEMENT, ENGAGE, MOTIVE, PEACE, PHYSICAL, PORNOGRAPHY, SAFEGUARD, SUBSTANCE, TRADE, WELFARE
32	**OFFICER, KILL, CARRY, SCENE, FORENSIC, WIFE**, SHOOT, EXAMINATION, INJURE, TEST, ARRIVE, FIREARM, KILLING, PRONOUNCE, BIRD, BLIND, ESTRANGE, EXAMINE, EXPERT, FIRE, HUSBAND, INTEND, LIAISON, OTHER, SHOT, UNARMED, SENIOR, ANALYSIS, DEPLOY, DUTY, FRONTLINE, GUNMAN, POST-MORTEM, POSTMORTEM, PREY, RECOVER, SCIENTIST, SPECIALIST, THREATEN, TOXICOLOGY, TRAIN, UNDERCOVER, UNIFORM, UNLAWFUL
73	**SAY, POLICE, CHIEF, STATEMENT,** NAME, SENIOR, CONFIRM, SOURCE, ADDRESS, AGE, SPEAK, INSPECTOR, NHS, BELIEVE, EXECUTIVE, IMPACT, OFFICIAL, READ, REAL, RULING, SPOKESWOMAN, IRA, CARNIVAL, CLOSE, CONFIDENT, COUNCILLOR, DEFEND, DETECTIVE, GOODBYE, HUNT, ISSUE, MITIGATION, ORGANISER, SATISFY, SHOCK, SPOKESPERSON, WELCOME
52	**FIRE, STATION**, BOMB, THROW, SEARCH, LINE, POWER, SET, TRAIN, BUS, PETROL, BOTTLE, EXTINGUISHER, RAILWAY, START, STOP, TUBE, ALIGHT, BRICK, BRIGADE, BUILDING, DESTROY, EAST, ENGINE, EXPLODE, PARCEL, RESCUE, WINE
49	**INQUIRY, BAIL, CALL, FURTHER, RELEASE**, PLEASE, CONDITION, PEND, BAIL, COMMENT, MAIN, AMBULANCE, DETAIL, PUBLICATION, SIDE, WRITE, CONDUCT, MARK, PERSONAL, REFUSE, APPLICATION, ARTICLE, GRANT, PAROLE, PREVENT, ROAD, SWITCHBOARD, AIR, AMEND, BOARD, BREACH, COMMENT, CONDITIONAL, CRITICAL, FREE, FULL, HOUSE-TO-HOUSE, STABLE