This is a post-refereed version of an article which has been accepted for publication by Edinburgh University Press:

Larner, S. (forthcoming), Using a core word to identify different forms of semantically related formulaic sequences and their potential as a marker of authorship, *Corpora*, 11.2. http://www.euppublishing.com/journal/cor

TITLE: Using a core word to identify different forms of semantically related formulaic sequences and their potential as a marker of authorship

AUTHOR: Samuel Larner

AFFILIATION: Manchester Metropolitan University

CONTACT DETAILS: Department of Languages, Information and Communications

Manchester Metropolitan University

Oxford Road

Manchester

M15 6BH

UNITED KINGDOM

E-MAIL ADDRESS: s.larner@mmu.ac.uk

TELEPHONE: 0161 247 3922

ABSTRACT

Formulaic sequences should make an excellent marker of style because if authors treat them as one lexical choice, they are unlikely to be aware of the individual words contained within. However, there is no clear-cut way to robustly identify all, and only, formulaic sequences in text. If one particular word can be isolated which occurs frequently in formulaic sequences—a core word—then a reasonable sub-set of word sequences will be identified, the majority of which can be expected to be formulaic. Using the core word *way* which occurs in many formulaic sequences (e.g., *in a way, by the way, by way of*), the aim of this research is to establish whether individual authors use different *way*-phrases from each other and, for comparative purposes, whether authors use alternative non-formulaic realisations of the same semantic content. If inter-authorial differences can be found, *way*-phrases may hold potential as a marker of authorship. The results indicate that for one author, the phrase *in a way* appeared to be used distinctively. Therefore, there is potential for formulaic sequences to be used as a marker of authorship, albeit for only one author out of twenty, which limits the usefulness of such a marker in a forensic context.

KEYWORDS

Authorship; core word; idiolect; forensic linguistics; formulaic sequences; recurrent phrases; style; way.

1. Introduction

In the field of forensic authorship attribution, lexis has been well explored as a marker of style (e.g. Chaski 2001; Coulthard 2004; Hoover, 2002, 2003; Kredens 2001). However, authors can make efforts to disguise their linguistic ability (e.g. Shuy 2001) so stronger markers of style are likely to be those which move beyond relatively surface level features such as non-standard spellings, and instead focus on features of idiolect which authors may less easily disguise. Evidence from psycholinguistics (e.g. Hoey, 2005; Wray, 2002), sociolinguistics (e.g. Coulmas, 1979), corpus linguistics (e.g. Moon, 1997, 1998; Sinclair 1991) and both L1 and L2 language acquisition (Pawley & Syder, 1983; Peters, 1983, 2009; Vihman, 1982) shows that when communicating, language users often rely on patterns in language and have "preferred formulations" for expressing ideas (Wray, 2006: 591); a point which is also supported by theoretical viewpoints such as Construction Grammar (Goldberg, 2003).

Wray (2002) coined the term *formulaic sequences* to account for such language, which she defines as 'a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar' (p. 9). Wray (2002: 9) found 57 different terms each describing characteristics of language that can be thought of as formulaic sequences, including collocations, idioms, fixed expressions including idioms, multi-word items, phrasal lexemes, recurrent phrases, and situation bound utterances. It is the fact that multi-word sequences are stored as single lexical items that is an important feature of formulaic sequences (Bannard & Lieven, 2009; Ellis, 1996; Erman & Warren, 2000; Pawley & Syder, 1983; Wray, 2002, 2008) and the underlying principle is that these sequences are not created through analysis of the individual words within a sequence. In fact, Wray

(2002) argues that language users only break down and analyse sequences of words if some need arises—*needs-only analysis*—and that according to this principle, "nothing would be broken down unless there were a specific reason" (p. 130). In this way, needs-only analysis accounts for irregularity in formulaic sequences. Phrases and sequences of words which, if analysed, would be found to contain obsolete vocabulary and ungrammatical structures do not cause problems in daily interaction precisely because "they do not invite analysis" (p. 131) even though they could be analysed if analytical processing were activated. Wray (2002) provides the example of the formulaic phrase *by and large* to illustrate her point: "The word *large* in *by and large* is not associated with the regular word meaning 'big' because there is no demand on native speakers ever to analyze the phrase and assign a meaning to its component parts" (p. 132). The ability for language users to handle both novel material and formulaic sequences suggests a part-analytic and part-holistic processing of language (Wray, 2002) and by exploring the ways in which formulaic sequences sit with other theories of language processing including generative grammar, functional grammar, pattern grammar, frame semantics, and construction grammar, Wray (2008) locates "formulaic language within a comprehensive model of how grammar, use, and psychological and social motivation interact" (p. 73; cf. Chapter 7 for comprehensive discussion).

Given the potential for sequences of words to be processed holistically, Larner (2014) proposed that formulaic sequences may be suitable as markers of authorship. Since the individual lexical items contained in formulaic sequences are less likely to be overtly monitored by the language user, they are likely to escape conscious manipulation, making them a more robust marker of authorship than surface level features of language. The aim of the current research is therefore to develop a corpus-based method for identifying formulaic sequences which may unlock evidence about

habitual and characteristic authorial style. In order to do this, it is firstly necessary to discuss the existing literature which explores the potential link between formulaic sequences and authorship before discussing methods for identifying formulaic sequences in texts.

1.1 Formulaic Sequences and Authorial Style

Literature which empirically investigates the relationship between formulaic sequences and authorial style is sparse, with perhaps Larner (2014) being the only research which specifically investigates their potential as a marker of authorship in a forensic context. Nonetheless, both Kuiper (2009) and Schmitt, Grandage and Adolphs (2004) describe research which more generally supports the individualised use of formulaic sequences. Kuiper (2009), focussing on supermarket checkout operator interactions with customers found that, based on 200 recordings, interactions could be broken down into a series of stages, with each stage being characterised by specific formulaic sequences (routine formulae in his terms). For example, the interactions typically began with a greetings formulae phase consisting of routine formulae such as *hi*, *hello*, or *gidday*, and were followed by a "start phase" with routine formulae such as *how are you today?* Focussing specifically on greetings formulae, Kuiper found that some of these routine formulae were shared between all checkout operators, including *How are you?* and *How are you today?* whilst others were used more regularly by only one checkout operator, leading Kuiper to argue that the use of particular formulae is 'equivalent to a signature' (p. 114). In this way it should be possible to identify a checkout operator on the basis of the routine formulae they use much like a forensic linguist attempts to identify an author

based on similar patterns of language in texts. It should, however, be borne in mind that this was an extremely restricted and task-oriented context.

Schmitt, Grandage and Adolphs (2004), although investigating whether recurrent clusters identified using corpus linguistics methods held psycholinguistic validity, like Kuiper (2009), found that some formulaic sequences were linked to idiolect. They presented a selection of twenty-five frequent and infrequent recurrent clusters from existing reference lists and corpora frequency counts, interspersed in dialogue, to thirty-four native speakers (an additional forty-five non-native speakers took part in the study but the results are not discussed here). The participants were required to repeat back what they had heard in a dictation task. Schmitt et al. reasoned that if stretches of dictation were long enough, participants' working memories would be overloaded and content would need to be reconstructed using their own linguistic resources rather than rote memory. Therefore, any of the recurrent clusters recited back were likely to be holistically stored and therefore psycholinguistically valid as formulaic sequences. As predicted, some recurrent clusters were produced less frequently by the participants (e.g. *in the same way as*, *to give you an example*) suggesting that they were not stored holistically, whereas others were reproduced correctly by most participants (e.g. *to make a long story short*, *I don't know what to do*), implying that they may be stored as formulaic sequences.

However, they also observed that whilst some recurrent clusters were always produced by participants, or at least attempted, suggesting holistic storage, and some were never produced or attempted, suggesting no holistic storage, some recurrent clusters were in the middle of this cline: some speakers appeared to store some recurrent

clusters as formulaic sequences whilst others did not, a finding which they link directly to idiolect:

> Every person has their own unique idiolect made up of their personal repertoire of language, and as part of that idiolect, it seems reasonable to assume that they will also have their own unique store of formulaic sequences based on their own experience and language exposure (Schmitt, Grandage, & Adolphs, 2004: 138).

In this way, they argue that the mental lexicon contains a majority of formulaic sequences that are shared across the speech community, but also a 'unique inventory of formulaic sequences' (p. 138) based on individual abilities in fluency and powers of expression, which may also be linked to topic and discourse situation. Schmitt et al.'s conclusion is based on the results of a relatively small sample of participants and indeed only a relatively small selection of recurrent clusters. However, it is interesting that in a more general context, idiolectal differences were found, lending further support to Kuiper's (2009) context-specific research.

On this basis, Larner (2014) reasoned that if formulaic sequences are linked to idiolect, they should be useful in distinguishing patterns in texts produced by different authors. In order to investigate this claim, Larner developed a reference list of 13,412 formulaic sequences compiled from a variety of internet sources. Using a corpus of 100 short personal narratives produced by twenty authors (five texts per author), Larner applied an automated approach which compared each text to the reference list and highlighted

any matches. Through statistical testing, Larner found that in terms of formulaic sequence types (rather than tokens), inter-author variation was greater than intra-author variation; that is, the five texts produced by the same author where more similar than those produced by other authors. Turning next to the normalised count of formulaic sequences (i.e. the number of words making up a formulaic sequence per 100 words), Larner again found inter-author variation to be greater than intra-author variation. However, whilst statistically significant variation was found between each sub-corpus of texts produced by the authors, Larner found that qualitative analysis was not successful and that the patterns of formulaic sequence types found across each author's texts were not strong enough for application in a forensic context. In this regard, Larner argued that the results supported Kuiper's (2009) research in that individual variation could be identified, but not with the same 'signature' potential, leading to the conclusion that 'there seems to be potential for formulaic sequence usage to differ between individuals, but the method outlined … has not been able to capture those differences sufficiently' (2014: 20).

The limitation of Larner's (2014) research is that the method is predicated on the basis that authors will either use, or not use, particular forms of formulaic sequences—that is, that with the exception of some small degree of pronoun variation which his automated approach could tolerate, the same content words were expected to be used in fixed sequences. What his study does not accommodate is the fact that authors' mental lexicons may contain formulaic sequences which are individual to them—in other words, authors may have individual preferred formulations for expressing semantically-related ideas. For instance, Mollin (2009) found that former UK Prime Minister Tony Blair idiosyncratically used the collocation *entirely accept,* whilst *totally agree* is a typical collocation in general speech (according to the BNC), and *entirely endorse* is a

more typical collocation of a parliamentary style, showing that whilst semantically related for conveying maximal agreement, different forms can be used to express a similar meaning. The question, then, is how can different realisations of semantically related formulaic sequences be reliably identified?

1.2 Identification of Formulaic Sequences

A variety of approaches to the identification of formulaic sequences can be found in the literature with the most appropriate method being selected based on the particular characteristics of formulaicity under investigation. For instance, since formulaic sequences are not always fixed in form and do not always have firm borders, surveying members of the same speech community for their intuitions about whether a given string is formulaic or not, or whether they can finish a string that is started for them, can offer useful insights into potential formulaicity. Applying structural analysis—where formal criteria including non-compositionality (that a literal interpretation is not possible) and fixedness (the degree to which word order can be changed, and lexical insertions, inflections and replacements are possible) are examined—can be useful in determining whether a sequence is formulaic, particularly with idioms. Since some types of formulaic sequence are linked to specific functions (for instance, Kuiper (2009) as described above), pragmatic and functional analysis may be most appropriate for determining which sequences lack transparency when tied to specific social settings. However, despite the variety of approaches, they are notoriously difficult to identify, leading Wray (2008) to comment that '[i]dentifying formulaic sequences in normal language can be rather like trying to find black cats in a dark room: you know they're there but you just can't pick them out from everything else' (p. 101). Erman and Warren (2000) caution that whilst some formulaic sequences ('prefabs' in their terms) are less

inconspicuous and are more easily identifiable, 'the identification of "all and only" the prefabs in a text is in practice impossible' (p. 33).

Read and Nation (2004) refer to the computer analysis of texts as a 'powerful new tool' for the identification of formulaic language (p. 30). Under this category, two techniques are available. Firstly, if an investigator has a sense that a particular string of words is formulaic, corpus software can be used to extract all examples of the word string for further analysis (e.g. Danielsson, 2003). Alternatively, a purely statistical approach can be used to identify sequences of words which 'regularly co-occur throughout the corpus beyond a threshold level of probability' (Read & Nation, 2004: 30) and the speed with which a computer can generate frequency counts make it an attractive technique to use (Wray, 2008). This latter technique can be incredibly useful for gaining insight into formulaic sequences that would normally be missed by intuition alone (e.g. Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004; Schmitt, Grandage, & Adolphs, 2004); however it conversely generates a large amount of data which are not formulaic (Read & Nation, 2004: 31). Therefore, for both approaches to corpus analysis, Read and Nation argue that data need to be evaluated by the investigator through human judgement, or through checking that the formulaic sequences can be classified into a classification system, if such a system is being used (p. 31). In contrast, Wray (2002) argues that applying "ad hoc intuitive decisions" (p. 27) potentially undermines the objectivity brought about by automated analysis.

To tackle the question of how best to identify semantically related formulaic sequences which take different forms, if one particular word can be isolated which occurs predominantly and frequently in formulaic sequences—a core word—then a reasonable subset of sequences will also be identified, the majority of which could be expected to be formulaic. The rationale behind using a core word is that a frequent

content word will have fragmented meaning (Wray 2002: 29) and therefore will rely on other words for the construction of a unified meaning. Wray (2002) discusses this in relation to Willis (1990):

> Willis (1990) nicely illustrates this fact with reference to the word *way*, which he argues could usefully be a key vocabulary item in ESL teaching. This is not because *way* in the sense of 'minor road', or even 'direction', is particularly frequent, but because *way* figures in numerous expressions (e.g. *in a way, by the way, by way of, ways and means*) which, between them, propel the word virtually to the top of the frequency counts in a large corpus (Wray 2002: 29).

It follows that identifying all instances of *way* in a corpus should provide a direct path to a range of formulaic sequences, albeit a very limited subset.

This research begins to investigate whether *way*-phrases are used by individual authors to the extent that texts produced by a relatively disparate closed sample of authors can be differentiated. Two stages are outlined in this paper. The first stage assesses whether any of the authors appear to have preferences for particular *way*-phrases. In the second stage, an attempt is made to establish whether, on the occasions that the authors have reason to express the same meaning, they use *way*-phrases or alternatives that do not include this core word. Given the investigative nature and potential forensic application of this paper, it is important to stress that the aim is to engage with testing and evaluating a method rather than outlining an exhaustive investigation into every single formulaic sequence that occurs in text.

2. Data

The data comprise 100 texts written by twenty authors, with each author producing five texts. Authors were provided with a daily structured writing task over a five day period.

Each morning, the authors were sent two essay-style questions and were required to answer whichever one they felt most comfortable answering. Open-ended questions which elicited personal narratives were asked since by asking emotionally-charged questions, it is hoped that the likelihood of participants focussing on their language use was reduced (Labov & Waletsky 1997). If participants were unable to answer either of the two questions, they were provided with a list of five substitute questions, from which they could select one to answer (see appendix for full list of question prompts). The decision to solicit five texts was motivated by the need to balance gaining sufficient data for authorial patterns to emerge against not going beyond the realms of feasibility for the forensic context, or indeed asking too much of the participants. Chaski (2001) deemed three texts to be sufficient for testing markers of authorship and Grant (2007) used 175 texts composed by 50 authors—an average of 3.5 texts per author. Hänlein (1999) used between 13 and 17 texts per author. Using five texts falls within this range and ensures that at a rate of producing one text per day, participants could complete the task in less than a week.

Deciding on the required length of the texts to make the research legitimate for forensic purposes may be somewhat arbitrary, since the lengths of authentic forensic texts vary, as do the number of texts available for analysis. Other empirical research into markers of authorship for forensic purposes has been conducted on short texts (e.g. Chaski, 2001; Grant, 2004; Winter, 1996) although a lower word-limit threshold has not yet been established for the minimum amount of text required for analysis. Therefore, the issue of feasibility needs to be the main criterion. In order that participants did not find the task too cumbersome, they were asked to write approximately 500 words. Since researchers have found formulaic patterns in texts shorter than this (Chenoweth, 1995), 500 words is a reasonable length of text on which to establish whether patterns of

formulaic sequences can reliably differentiate authors. All participants signed a consent form and were fully debriefed, in accordance with institutional ethical guidelines. The author corpus of 100 texts contained 65,113 words with each author producing an average of 3,325 words across their five texts. The average text length was 651 words with the shortest being 485 words and the longest being 822 words.

3. Stage 1: Authorial Preferences for *Way*-Phrases

Since *way* is expected to form part of numerous formulaic sequences, the first stage seeks to establish if this is in fact the case and, if so, whether authors demonstrate a preference for certain *way*-phrases over others. Clearly, if patterns of preference can be determined for any or all of the authors, then formulaic sequences which rely on the core word *way* may be markers of style.

3.1 Method

Using *WordSmith Tools* (Scott 2008), *way** was entered as the node and 105 occurrences were extracted from the 100 text author corpus (ninety-four instances of *way* and eleven instances of *ways*). From here on, *way* will be used for brevity but should be understood to include *ways*. Of the 105 concordance lines, two were excluded from the analysis on the grounds that neither were instances of the author's original words and therefore cannot be taken as characteristic of their authorial style, as shown in lines 1 and 2.

1	good food and my father singing 'My	**way**	' on the karaoke. It was a typical
2	that we were leaving. She replied 'no	**way**	' and continued dancing. I rang mum

For the remaining 103 concordances, it was necessary to isolate all of the words that could be considered to form a *way*-phrase. For this purpose, the decision was made to include all of the words surrounding *way* that would need to be removed if an alternative formulation was to be used. Five examples (underlined) are provided below (lines 3-7):

3	chronic diarrhoea and I drove all the	**way**	down to Oxford (where he lived at the
4	my masters, it is linked in several	**ways**	, and the experience and life
5	Santa doesn't exist. I suppose in a	**way**	I must have done, as when I was
6	120 miles north of Liverpool a long	**way**	from Deeside and when John got a job
7	mind he's still alive and that's the	**way**	I want it to stay. I miss him so much

In line 3, *all the way* is considered to be a *way*-phrase since this entire group of three words could conceivably either a) be removed entirely (e.g. *I drove down to Oxford*), or b) would need to be removed and replaced to convey the same meaning whilst keeping the sentence grammatical (e.g. *I drove the long distance down to Oxford*). The same is true for line 4, where *in several ways* constitutes the *way*-phrase. In line 5, the sentence could have been written as *I suppose I must have done*, indicating that *in a way* is the *way*-phrase. Similarly, line 6 contains the phrase *a long way* and line 7 contains *the way*.

Of course, the *way*-phrase was not easily extracted from every concordance line. In line 8, there is no clear-cut solution to the question of whether *right* is part of the phrase *in the way*, or whether it is an adjective which pre-modifies, but is not holistically stored alongside *in the way*.

8	knew that I was standing right in the	**way**	. What I didn't know was that the

In this case, the decision was made to exclude *right* on the basis that the single word *right* could be removed from the sentence without altering meaning, whereas the

sequence *in the way* could not (**I was standing right.* compared to *I was standing in the way.*). This suggests that the three words *in*, *the* and *way* in this sequence are more closely bound than the word *right*, which is more likely an optional addition, although admittedly an important one included for rhetorical effect. All 103 *way*-phrases were sorted according to author, in order to establish patterns for specific *way*-phrases.

Comparative data can be drawn from the BNC, a 100 million word corpus of British English, where *way* occurs 107,692 times (equivalent to 1.08 occurrences per 1,000 words). The frequency of *way* across each author sub-corpus per 1,000 words is shown in Table 1.

Table 1: Occurrences of *way* per 1,000 words across the author corpus

Author	**Occurrences of *way***	**Size of sub-corpus**	**Occurrences per 1,000 words**
Judy	1	3427	0.29
David	1	3058	0.33
Melanie	2	2879	0.69
Thomas	3	3824	0.78
Michael	2	2516	0.79
Sue	3	3716	0.81
John	3	3119	0.96
Mark	3	2844	1.05
Nicola	4	3021	1.32
Elaine	4	2941	1.36
Rick	6	3583	1.67
Greg	5	2980	1.68
Carla	6	3217	1.86
Keith	6	3067	1.96
Hannah	7	3559	1.97
Sarah	6	2957	2.03
June	7	3151	2.22
Jenny	9	3518	2.56
Alan	12	3916	3.06
Rose	15	3820	3.93

In comparison to the BNC, it can be seen that some authors (e.g. Judy and David) use *way* less frequently, some at roughly the same level (e.g. Sue and John) and some who

use *way* more than twice as frequently (e.g. Jenny, Alan, and Rose). The overall frequency of *way* in the author corpus is 1.55 per 1,000 words, showing that *way* occurs 47% more frequently in the author corpus than in the BNC.

3.2 Results

The 103 instances were made up of fifty-five different phrases. The range of phrases used is presented in Table 2 (organized from most frequent to least frequent), alongside their total frequency across the corpus and the number of authors who used a particular phrase. All twenty authors used at least one phrase.

Table 2: Fifty-five *way*-phrases identified in the 100 text author corpus

***Way*-Phrase**	**Frequency across entire corpus**	***N* authors using *way*-phrase**
in a way	19	8
the way	6	4
way	6	4
all the way	4	4
on my way	3	3
on the way	3	1
the only way	3	2
a way	2	2
both ways	2	2
in a strange way	2	2
in so many ways	2	2
made my way	2	1
made our way	2	1
way of dealing	2	1
my way	2	2
only one way	2	2
out of the way	2	2
the same way	2	1
there is no way	2	2
The remaining 36 *way*-phrases occurred only once in the corpus and were therefore used by only one author: a certain way; a long way; along the way; any other way; any way; by the way; either way; for ways to; gave way; get out of the way; go out of my way to; half way; in a different way; in a roundabout way; in any serious way; in any sordid way; in many other ways; in many ways; in several		

ways; in some way; in such a kind way; in the way; let's put it that way; make their way; making his way; one way or the other; some ways; the exact way; the only way to; the other way around; the rest of the way; the ways; the whole way; ways; way of releasing; worked my way

The first important observation is that no phrase was used by every author. Table 2 shows that the majority of phrases were used just once by only one author (e.g. *in a roundabout way, some ways, the other way around, the whole way*). Other phrases such as *made my* way and *made our way* are used by only one author, which could be characteristic of authorial style, but with such low occurrence in the corpus (twice for each) this cannot be demonstrated convincingly. By contrast the phrase *in a way* is used by less than half of the authors (eight) and occurs nineteen times, warranting further investigation. Table 3 shows which authors use this phrase, how frequently, and in how many of their five texts.

Table 3: Authors using *in a way*

Authors using *in a way*	**Frequency of use of *in a way***	**Number of texts containing *in a way***
Rose	10	5
Alan	2	2
Jenny	2	1
Carla	1	1
Hannah	1	1
John	1	1
Keith	1	1
Melanie	1	1

Only Rose used *in a way* consistently across all five texts. For the remaining seven authors, *in a way* occurs typically only once, except for Alan and Jenny who use it twice. Therefore, the frequency and consistency with which it is used may indicate that this phrase may be a marker of style for Rose. In the BNC there are only 2,751 occurrences of *in a way*. As such, this *way*-phrase appears to be relatively rare adding more significance to the fact that Rose uses it consistently and frequently in comparison

to the other authors and the BNC. This phrase occurs 0.29 times per 1,000 words in the author corpus and 0.03 times per 1,000 words in the BNC, meaning that *in a way* is 26% more frequent in the author corpus. There is no other evidence of any authorial patterns. It therefore seems that the remaining phrases hold little potential to be characteristic of any other author's style.

As an additional measure, phrases were grouped and reduced to their underlying structures (e.g. *in a/any ADJ way* as a single variable phrase, rather than the four individual phrases *in a different way, in a roundabout way, in any serious way* and *in any sordid way* and instances of *way of dealing* and *way of releasing* were treated as *way of X-ing*). Again, no patterns emerged across the entire corpus or for any individual author sub-corpus.

In some respects, given the supposed prominence and importance of *way* in texts, it is surprising that stronger patterns have not emerged, either for individual authors, or for the group of twenty authors as a whole. However, *way* does seem to be prominent in many formulaic sequences as evidenced by the fact that—with the exception of *way* and *ways* as single words—the meaning behind all other phrases was contained within a two or more word sequence. Moreover, *way* seems to be the core word of these sequences since it is largely surrounded by function words (e.g. *in a way, all the way, on the way*) and therefore can be considered an essential component for whatever meaning the authors wished to express. Therefore, it does appear to have been possible to identify a range of formulaic sequences by using the core word *way*, as suggested by Wray (2002) and Willis (1990). However, in determining whether authors have consistent patterns in the *way*-phrases they use, there is some, but only very limited, evidence since one author (Rose) out of twenty used *in a way* ten times across all five of her texts demonstrating that texts produced by Rose do appear to be marked as different from all

other texts in the corpus due the frequency and consistency with which *in a way* occurs. Of course, the point needs to be made that in some respects, the bar is set very high—necessarily so, in fact, for potential application in the forensic authorship context. Given that, with the exception of *in a way*, no other phrase is used in such a way that might suggest an idiolectal preference, the second stage seeks to dig beneath the forms that these particular *way*-phrases take, and instead focuses on the meanings that are conveyed in order to determine whether authors express the meaning behind *in a way*, but in different forms.

4. Stage 2: Alternatives to *Way*-Phrases

It was established through Stage 1 that focussing on the form of *way* sequences may be limited, at least in short texts. Authors may instead express similar meanings but in different forms which do not contain the word *way* and so will not be identified through the use of this core word. Therefore, in order to continue this investigation, phrases used to express similar meanings are the next focus. The rationale for this stage is described by Wray (2002):

> To capture the extent to which a word string is the preferred way of expressing a given idea (for this is at the heart of how prefabrication is claimed to affect the selection of a message form), we need to know not only how often that form can be found in the sample, but also how often it *could* have occurred (p. 30, *original emphasis*).

Outside of the formulaic language literature, the same point has been made, for example, by Kredens (2001), dealing specifically with the forensic context: '[A]

forensic analysis needs to allow for the fact that different speakers can favour different lexical means for expressing the same attitude' (p. 426). However, to date, there has been a lack of empirical research which sets out to investigate this issue. Stage 2 therefore aims to contribute to this research gap by categorising phrases by their meaning. In this way it may be possible to establish where authors show a preference for expressing meaning in a particular way compared to other authors. If preferences can be established, there may be some potential to highlight aspects of authorial style based on this method.

4.1 Method

Following on from Stage 1, a gloss was produced for each of the fifty-five *way*-phrases in order to determine the meaning being conveyed. A selection of nine *way*-phrases (underlined) which occurred a total of 26 times in the author corpus is presented below organized under four clearly discernible glosses.

'=do more than necessary/expected' (1 occurrence)

9	person – I really do go out of my	**way**	to prevent hurt. That was all behind

'=not a possibility, option' (2 occurrences)

10	Well, there is no	**way**	I'm telling you my most embarrassing
11	if I lived in England there is no	**way**	he'd even have been with Ian, but as

'=on several levels, for different reasons' (5 occurrences)

12	but this made it worse in many	**ways**	as he was searching for an excuse and
13	masters, it is linked in several	**ways**	and the experience and life
14	this gap and while this is some	**ways**	positive it may mean significant

'=to some extent, in some respects' (18 occurrences)

15	doesn't exist. I suppose	in a	**way**	I must have done, as when I was
16	any attention to myself and	in a	**way**	didn't see why they should know. This
17	friends in the evenings so	in a	**way**	I was leading a double life. I

In total, twenty-nine different glosses were derived from the fifty-five way phrases which accounted for all 103 uses of way. For each of the glosses, a series of synonyms were extracted from the dictionary and thesaurus components available through *Oxford Reference Online* (Stevenson, 2010). Drawing on the glosses provided in lines 9-17, Table 4 shows the synonyms that were identified (quite whether these are in fact synonyms, or even near-synonyms is discussed later).

Table 4: Examples of synonyms and search nodes for glosses

Gloss	**Synonyms**	**Search nodes**
=do more than necessary/expected	put myself out; go out on a limb; do more than I need to; should; required to be done; needed; essential; obligatory; requisite; required; compulsory; mandatory; imperative; vital	myself; limb; more than; should; required; needed; essential; obligatory; required; compulsory; mandatory; imperative; vital
=not a possibility, option	chance; likelihood; probability; hope; risk;, hazard; danger; fear; possibility	chance; likelihood; probability; hope; risk; hazard; danger; fear; possibility
=on several levels, for different reasons	on several levels; for different reasons; ground(s); basis; purpose; point	levels; reasons; ground*; basis; purpose*; point*
=to some extent, in some respects	respect; regard; aspect; facet; sense; detail; a little; somewhat; rather; sort of; kind of	respect*; regard*; aspect*; facet*; sense*; detail*; a little; somewhat; rather; sort of; kind of

As can be seen from the final column in Table 4, based on these synonyms, a series of nodes with which to search the author corpus were created. Many of the items recurred throughout the process. For example, in the first row of Table 4, *required* occurs twice.

Duplicates were therefore removed leaving 242 search nodes that could potentially convey the same meaning as any one of the identified *way*-phrases which generated a total of 2,458 concordance lines. These concordances were then manually checked. If the phrase surrounding the node did not convey the same meaning as the *way*-phrase, it was discarded. If it did convey the same, or at least similar meaning the phrase was retained. The process for determining which words constituted the phrase was the same as that outlined for Stage 1, i.e. all the words necessary for meaning and/or the words that could be removed leaving behind a grammatical sentence. For clarity, a worked example for the gloss '=in a certain manner, fashion' follows.

4.2 Worked Example

A range of *way*-phrases that could be glossed as '=in a certain manner, fashion' were identified, including: *in a way, in such a way,* and *way* as a single word (which for the present purposes is being treated as a formulaic sequence; see section 5 for discussion), as indicated in examples 18-20:

18	well at least never in a	**way**	that would ordinarily be thought of
19	easily be behaving in such a	**way**	. After that incident it wasn't quite
20	an incredibly unfair and brutal	**way**	to do anything but I seemed left with

Melanie is the only author to use the phrase *in a way* in the sense of '=in a certain manner, fashion' (line 18, compare later with Rose's use of *in a way* glossed as '=to some extent, in some respects'). In line 19, the phrase *in such a way* occurs only once in the author corpus, used by Jenny in her second text. The word *way* to convey this meaning, as in line 20 occurs twice in the corpus by Sue, in her second and fourth texts. On the surface then, it would be tempting to argue these three phrases as being indicative of individual style—no other author uses these phrases to convey this

meaning. However, before such a claim can be confidently made, the following needs to be established: 1) whether any other authors actually express this meaning (after all, it is not sufficient to argue that an author does not use a particular phrase if they have no need to express the meaning behind that particular phrase) and 2) if they do express this meaning, which phrase(s) do they actually use?

As outlined above, a series of synonyms were identified for the gloss '=in a certain manner, fashion' as the basis for identifying other phrases that express the same meaning in the author corpus. A selection of twenty-five potentially synonymous phrases, in concordance lines organized alphabetically by node, is presented as lines 21-45a. The near synonymous expressions which convey this meaning are underlined.

21	was the last cast member to arrive	**as I did**	not need any make up. I pulled
22	Society and still went out as much	**as I did**	in the first two years (it's a
23	interested in. The more creative	**aspects**	of my life I decided to keep
24a	most afford to drop, as was the	**convention**	in my school - I had decided
25	feelings known to him or anything	**like that**	! Luckily, i think some people
26	complete concrete! I really didn't	**like that**	and that's what impressed me
27a	Josh wouldn't have wanted to exist	**like that**	; to have been such a burden
28a	him but obviously I didn't see it	**like that**	. Unfortunately I wasn't
29a	never knew I could betray someone	**like that**	. The next day I went over to
30	Being lanky	**means**	there have been many
31a	I had achieved AABC - by no	**means**	bad results, but over the last
32	lies. I will tell a lie if that	**means**	I won't hurt somebody's
33	I kind of went into proactive	**mode**	and went straight home to work
34a	Suddenly thankful for my hands-on	**nature**	I took over and after two
35	in and saw her, because of the	**nature**	of the operation she was lying
36	and used to call me names as they	**regarded**	me as one of the 'clever'
37	spine, I did cry, but carried on	**regardless**	. The kindness of the girls
38	that he didn't have the decency,	**respect**	, courtesy or balls to tell me
39	He was joking about it. I lost	**respect**	for him then. I texted him a
40	of her mother's and my teacher's	**respect**	. I also argued with my friend
41a	we fell straight back into the old	**routine**	. He said the right things to
42	leaving my room he had the exact	**same**	profile from the rear as my
43a	I had no longer felt quite the	**same**	about the relationship for
44	was blurred at first. I was in a	**state**	of shock, I sat down and was
45a	in. At this point I was in such a	**state**	that my sister ran out to save

At this stage of the analysis, it would be beneficial for a second-rater to assess the data so that a level of inter-rater reliability could be established. However, in the forensic context, linguists are often required to work in isolation due to the sensitive and confidential nature of the data, and the time pressures involved in producing forensic evidence (Shuy, 2006) may further preclude this from being a possibility. Therefore, given the applied focus of this paper, whilst it is possible to argue that some of these

examples may be at least related to the meaning '=in a certain manner, fashion', the decision was made to include only clear-cut examples in the analysis rather than consulting a second-rater to discuss the grey areas, which would not always be feasible in practice. From this selection of twenty-five concordances, sixteen can be discarded since they do not appear to explicitly convey the meaning '=in a certain manner, fashion'. The remaining nine concordances do more clearly express this meaning and can replaced with the following *way*-phrases, whilst still retaining a similar meaning as shown in lines 24b-45b.

24b	most afford to drop, as was the	*way*	in my school - I had decided
27b	Josh wouldn't have wanted to exist	*in that way*	; to have been such a burden
28b	him but obviously I didn't see it	*in that way*	. Unfortunately I wasn't
29b	never knew I could betray someone	*in that way*	. The next day I went over to
31b	I had achieved AABC -	*In no way*	bad results, but over the last
34b	Suddenly thankful for my hands-on	*way*	I took over and after two
41b	we fell straight back into the old	*ways*	. He said the right things to
43b	me. I had no longer felt quite the	*way*	about the relationship for
45b	in. At this point I was in such a	*way*	that my sister ran out to save

Through this process, it is possible to ascertain which of the authors express this particular meaning, and more importantly, how they actually express it. Comparisons can then be carried out across authors to determine whether there are any patterns in how this meaning is expressed and if there are, whether they are shared by all authors (i.e. a certain phrase is the common form to express a meaning) or whether they are more distinctive (i.e. a certain phrase is less often used by other authors to convey a particular meaning). The results are presented below.

4.3 Results

From the 2,458 concordance lines generated from 242 nodes, a total of 141 concordance lines contained words or expressions which were considered to be alternatives or near-synonyms for one of the *way*-phrases identified in Stage 1. When these 141 alternatives

are added to the 103 *way*-phrases, twenty-nine different meanings were expressed a total of 244 times across the author corpus. All of the *way*-phrases and alternative expressions were plotted on a grid to enable clear cross-referencing. Table 5 below, organized according to frequency of occurrence, summarizes how many times each meaning occurred in the corpus, along with how many authors expressed that meaning.

Table 5: Glosses for *way*-phrases ranked by frequency of occurrence

Meaning	**Total Occurrences**	**Used by *N* Authors**
=to some extent, in some respects	35	11
=method, how to achieve an objective	31	14
=emphasis	29	15
=in a certain manner, fashion	24	12
=in a certain manner, how	21	9
=embarked on a route, journey	18	12
=the entire distance, journey, time	15	8
=particular direction, towards an outcome (metaphorical)	11	6
=method, no options/possibilities	8	6
=mid-point	7	6
=in each direction, left and right	5	3
=on several levels, for different reasons	5	3
=do more than necessary/expected	4	3
=devising plans, solutions	3	3
=embarked on a route, journey (metaphorical)	3	3
=great distance, far	3	3
=like, in a similar fashion	3	2
=move to safety, away from path of danger	3	2
=a different situation, alternative scenario	2	2
=broke, collapsed	2	2
=from available options	2	2
=in any condition, state	2	2
=vice versa	2	2
=helped through alternative means	1	1
=in the direct path of danger	1	1
=manner, in different ways	1	1
=move to safety, away from path of danger (metaphorical)	1	1

=remainder of the journey	1	1
=tactfully express	1	1

It can be seen from Table 5 that the meaning '=to some extent, in some respects' occurs the most frequently, a total of thirty-five times, and is used by eleven of the twenty authors. The second most frequently occurring meaning, '=method, how to achieve an objective', occurs thirty-one times and is used by fourteen authors. The third most frequent category, '=emphasis', occurs slightly fewer times, twenty-nine, but is used by slightly more authors, fifteen. At the bottom end of the table is a selection of meanings which are expressed only once in the corpus, and by only one author, including '=in the direct path of danger', '=remainder of the journey' and '=tactfully express'. It should be apparent that those meanings towards the top end of the table will be more useful as evidence of authorial style since there will be more comparative data, compared to those at the bottom end of the table which are used so infrequently that meaningful patterns cannot be established. Examples of the range of expressions for the top five most frequently expressed meanings found in the author corpus are presented in Table 6.

Table 6: Range of expressions used to convey the top five meanings

Gloss	***N* potential expressions**	**Expressions used to convey meaning**
=in a certain manner, fashion	16	*by no means; by the way; convention; in a way; in a/any ADJ way; in some way; in such a kind way; in such a way; like that; nature; quite the same; routine; sense of style; style; such a state; way*
=emphasis	10	*far; far too; get myself back; much; much more; on the journey; rather; significantly; so much; way*
=method, how to achieve an objective	9	*a chance; a way; how; my best course of action; my*

		way; only one way; option; the only way; way of X-ing
=to some extent, in some respects	6	*in a way; in that respect; in the other sense; kind of; somewhat; sort of*
=in a certain manner, how	3	*How; manner in which; the way*

It is now possible to determine whether 1) authors have a preference, and 2) how distinctive that preference is in comparison to other authors. Dealing with the first part, no strong preferences for all authors were found—indeed only two authors expressed the same meaning consistently at least once in all five of their texts: Alan ('=emphasis') and Rose ('=to some extent, in some respects'). Of these two authors, Alan expressed '=emphasis' in a different way each time (*much, far, way, significantly, far too)* as shown in lines 46-50, so there is no evidence of a patterned preference when expressing this meaning.

46	I can do. Forgetting is always	much	harder and if someone has done
47	And see dozens of people who are	far	worse off than yourself.
48	fake, the real Santa would be	way	too busy to fly down to a
49	I tend to go for older ones, I'm	significantly	poorer so it would be quite
50	a-few-days-at-a-time-because-I'm-	far-too-	-fat-stage) and I had eaten

Rose, however, expressed '=to some extent, in some respects' consistently across her five texts, using the expression *in a way*. This therefore seems to be a preference for her. However, in her fifth text, Rose also used the expressions *kind of, in that respect,* and *in the other sense* (lines 51-53), along with *in a way* three times—in other words, although she does have some variation in the forms she used to express this meaning, there is a predominant form, *in a way*.

51	year. As a result I like to	kind of	blend into the crowd so I
52	Me in such a kind way, so	in that respect	I didn't mind. Especially
53	on it with hindsight! But	in the other sense	I really wished he hadn't of

Of the meanings that are expressed by only one author, they are not expressed with enough frequency to suggest that they may be linked to authorship (see Table 5): '=helped through alternative means' is expressed by only one author (Hannah, *in a different way*), '=in the direct path of danger' (Greg, *in the way*), '=manner, in different ways' (Jenny, *in many other ways*), '=move to safety, away from path of danger (metaphorical)' (Judy, *out of the way*), '= remainder of the journey' (Rick, *the rest of the way*) and '=tactfully express' (Alan, *let's put it that way*). It would be tempting to argue that these expressions are markers of style due to their uniqueness, but of course, this is impossible due to the limited data. To make such claims, other authors would need to express these same meanings in order to determine the potential alternative expressions.

For none of the meanings studied is there a set expression. That is to say that the authors have a variety of choices available to them when they wish to express any of these meanings. Two expressions come close to having limited choices: '=mid-point' (either *half way* or some variation of *in the middle of*) and '=in each direction' (where authors use either *both ways* or *in the other direction*). However, these meanings were only expressed seven and five times respectively, so it may just be that there was insufficient data to explore alternatives.

The archetypal situation would be if each meaning was expressed in a particular form consistently across each author's five texts and in ways different from all other authors. Such a situation did not occur meaning that there were no clear patterns for how authors chose to express particular meanings. As was demonstrated in Table 5, there are a range of forms used to express the same, or at least similar, meanings, some of which use the core word *way*, and others which do not. This supports the claim that specific meanings—those identified in this research at least—can be expressed in different

forms and on the limited available data it appears that there is no one form for expressing any one of the selected meanings. The expression *in a way* again seems to be characteristic of Rose's style by being both a preferential choice and a consistent choice.

5. Discussion

There are three key issues that need to be addressed in evaluating the method presented in this paper. Firstly, a set of alternative phrases were identified during Stage 2—are these alternative phrases really synonymous? Secondly, regarding the corpus itself, what would be the effect of working with a larger, or indeed smaller, set of data? Finally, are the *way*-phrases identified in Stage 1 valid as examples of formulaic sequences? Each of these issues will be dealt with in turn.

The first issue relates to synonymy. During Stage 2, a range of alternatives to the *way*-phrases were identified in the data. The alternatives were identified through a range of synonyms and near-synonyms using the dictionary and thesaurus tools in *Oxford Reference Online* (Stevenson, 2010). The majority of these 'alternative' concordances were not in fact synonymous with the *way*-phrases. This raises the question of what is meant by 'synonymous'. It is true that a very loose interpretation has been applied in this research—relying upon a subjective synonym test—in other words, whether it was possible to replace the *way*-phrase with an alternative formulation whilst still conveying a similar meaning. This raises the question of whether, for instance, *in a way* meaning '=to some extent, in some respects' is really interchangeable with *kind of* or *sort of*? At a grammatical level, these are of course interchangeable. But is there a change in semantics, no matter how subtle? Hoey (2005) argues that the expressions *around the world* and *round the world* are primed in similar ways since they share the same sorts of collocates (e.g. *halfway* and *markets*) but one is more strongly primed than the other

with his overall conclusion being 'we may hypothesize that synonyms differ in respect of the way they are primed for collocations, colligations, semantic associations and pragmatic associations and the differences in these primings represent differences in the uses to which we put our synonyms' (p. 79). Similarly, Carter (2004) argues:

> [I]dioms are not simply neutral alternatives to less semantically opaque expressions. There is a difference between 'I smell a rat' and 'I am suspicious', or 'She's on cloud nine' and 'She's extremely happy' … In all cases the idiomatic expression is used evaluatively and represents a more intense version of the literal statement' (p. 132).

Although Carter talks exclusively about idioms, the same point can likely be made about all aspects of formulaic sequences. Therefore, it is important to understand the authors' motivations for choosing *kind of*, *sort of* or *in a way.* Is it a matter of formulaicity, with a preferential choice being made, or is there another factor, such as rhetorical style being the stronger force? As Hoey commented, the way that the authors used these synonyms, if they are accepted as synonyms, would need to be taken into greater consideration before attempting to apply this method to the forensic context.

The second issue—that of the corpus itself—also warrants attention. The *way*-phrases identified were extracted from 100 texts. The resulting 'alternative' expressions were based only on the same *way*-phrases. What would have happened if 200, 300, or even just 101 texts had been available for analysis? Would a larger set of formulaic sequences with the core word *way* have been identified, opening up potential for a greater number of alternative expressions? And likewise, five texts were used for each of the twenty authors. Would using only four texts or perhaps ten texts have made a difference? Based on the current data, it is possible to determine the frequency of *way* as

used in fewer texts in each author sub-corpus. This will determine whether having fewer texts for each author will significantly alter the results. Table 7 shows how many occurrences of *way* there are in each author sub-corpus (i.e. all five texts). The occurrences of *way* are then shown for texts 1-4 and in the final column, the occurrences of *way* in just the first three texts.

Table 7: Occurrences of *way* with fewer texts

Author	**Occurrences of *way* (5 texts)**	**Occurrences of *way* (4 texts)**	**Occurrences of *way* (3 texts)**
Alan	12	9	8
Carla	6	5	3
David	1	1	1
Elaine	4	4	2
Greg	5	2	0
Hannah	7	7	7
Jenny	9	8	7
John	3	2	0
Judy	1	1	1
June	7	4	4
Keith	6	4	4
Mark	3	3	3
Melanie	2	2	1
Michael	2	2	2
Nicola	4	3	2
Rick	6	3	1
Rose	15	11	8
Sarah	6	6	3
Sue	3	3	2
Thomas	3	3	3

Table 7 shows, as would be expected, that with fewer texts, so too are there fewer occurrences of *way*. More importantly though is the fact that the frequencies do not decrease for all authors at the same rate. Hannah and Thomas, who use *way* seven and three times respectively, still have the same frequency of use in just three texts as they did in five (in other words, all of their uses occur in the first three texts). Rose, on the other hand, who was the greatest user of *way* in five texts, uses it only eight times in three texts—where once there was a marked stylistic difference, her use is now

comparable to Hannah's. Similarly, Mark and John both use *way* in five texts, three times, but in just three texts, John does not use *way* at all whilst Mark's three uses remain. At one point they used *way* equally, but with fewer texts, one author *appears* to be using it more frequently than the other.

The point really is that *way* is not distributed evenly in these texts and using fewer texts would therefore significantly impact the results. What cannot be determined from the current data, though, is whether using more texts would create the same effect. There is the possibility that an author's use of *way* stabilizes over five texts, but there is no real reason to believe that this should be the case. Of course, the argument was made in Section 2 that the data analysed in this research are typical of the sorts of texts encountered in forensic investigations. In this regard, since a level of ecological validity has been attained, it makes little difference whether additional texts would affect the analysis because they would unlikely be available in an authentic applied forensic context. In light of this, the approach is unsuitable for forensic investigations.

Although not yet developed sufficiently for application in the forensic context, the results presented here may provide a good foundation on which future research can build. The core word *way* has been presented here as a case study and it is important to remember that it would not be prudent to view *way* as a magic bullet—that is, it could not be expected that simply using *way* and the formulaic sequences associated with it would reveal something about all authors in all text types. However, it may be fruitful to carry out an investigation of a variety of different core words in order to determine whether combinations of formulaic sequences provide more convincing results since it is likely to be the combination of a variety of features that is more indicative of authorship than the patterns of usage for any one word. Consistent combinations of formulaic sequences would certainly provide stronger evidence of authorship. In this

regard, Willis (1990) argues that high frequency nouns may be a suitable candidates, and *thing* in particular looks especially promising for future analysis given its high frequency and incorporation in a variety of formulaic sequences (e.g. *one thing after another, the shape of things to come*) (p. 39).

The final issue, validity, refers to whether the *way*-phrases identified actually are formulaic. Read and Nation (2004) argue that this is a particularly problematic criterion, since 'storage as a whole unit' is difficult to operationalize (p. 35). Whilst it is not possible to claim that this set of authors did process these *way*-phrases as holistic sequences based only on the external evidence of written output, it is reasonable to argue that they are likely to be formulaic on the basis that in almost all cases, a combination of two, three or more words were required in order to convey meaning. That is, the phrase *in a way* is a likely formulaic sequence since neither word on its own conveys the meaning '=to some extent, in some respects' and therefore holistic processing is required to understand the meaning. On the other hand, there are several instances of *way* and *ways* as single words that are less likely to be formulaic since they rely less on the words around them for their meaning to be understood. As discussed above, quite where the dividing line between the literal and the non-literal occurs is not clear.

6. Conclusion

This paper has described a method comprising two stages for identifying a small subset of formulaic sequences in an authorship corpus to tease out potential stylistic differences between individual authors and in doing so has extended the work of Larner (2014). Each stage has its limitations, but it is intriguing that in each case, the same result was found for Rose: that the phrase *in a way* appears to be used distinctively by

her. It is certainly encouraging that both stages achieved the same result indicating a level of support through triangulation. In this way, further support can be provided to Schmitt, Grandage and Adolphs (2004) and Kuiper (2009), that even in a less experimental and less routinized situation, the use of formulaic sequences do appear to be used idiosyncratically, albeit for only one author out of twenty. Likewise, in line with Larner (2014), the approach to identifying formulaic sequences in written language outlined in this paper does illuminate some very limited authorial differences. Whilst Larner (2014) was unable to capture those differences sufficiently beyond statistical testing, the current approach captures those differences in a qualitative way, but reveals noticeable consistency between texts and variation compared to other authors, for only one author in the corpus.

References

Bannard, C., & Lieven, E. (2009). Repetition and reuse in child language learning. In R. Corrigan, E. Moravcsick, H. Ouali & K. Wheatley (eds), *Formulaic Language: Acquisition, loss, psychological reality, and functional explanations* Vol. 2. Amsterdam: John Benjamins Publishing Co. 299—321.

Biber, D., & Conrad, S. 1999. 'Lexical bundles in conversation and academic prose' in H. Hilde & S. Oksefjell (eds) *Out of Corpora: Studies in honour of Stig Johansson*, pp. 181-190. Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. 2004. '*If you look at ...*: lexical bundles in university teaching and textbooks', *Applied Linguistics, 25* (3), pp. 371-405.

Carter, R. 2004. *Language and Creativity: The art of common talk*. London: Routledge.

Chaski, C. 2001. 'Empirical evaluations of language-based author identification', *Forensic Linguistics: The International Journal of speech, Language and the Law, 8* (1), pp. 1-65.

Chenoweth, N. A. (1995). Formulaicity in essay exam answers. *Language Sciences, 17*(3), 283—297.

Coulmas, F. (1979). On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics, 3*, 239—266.

Coulthard, M. 2004. 'Author identification, idiolect, and linguistic uniqueness', *Applied Linguistics, 25* (4), pp. 431-447.

Danielsson, P. 2003. 'Automatic extraction of meaningful units from corpora: a corpus-driven approach using the word *stroke*', *International Journal of Corpus Linguistics, 8* (1), pp. 109-127.

Ellis, N. (1996). Sequencing in SLA: phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18*, 91—126.

Erman, B., & Warren, B. 2000. 'The idiom principle and the open choice principle', *Text, 20* (1), pp. 29-62.

Goldberg, A. (2003) Constructions: a new theoretical approach to language. *TRENDS in Cognitive Sciences* 7(5): 219–224.

Grant, T. (2004). *Authorship Attribution in a Forensic Context.* Unpublished Ph.D., University of Birmingham, Birmingham.

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *The International Journal of Speech, Language and the Law, 14*(1), 1—25.

Hänlein, H. 1999. *Studies in Authorship Recognition-A corpus-based approach*. Frankfurt: Peter Lang.

Hoey, M. 2005. *Lexical Priming: A new theory of words and language*. Abingdon, Oxon: Routledge.

Hoover, D. L. 2002. 'Frequent word sequences and statistical stylistics', *Literary and Linguistic Computing, 17* (2), pp. 157-180.

Hoover, D. L. 2003. 'Frequent collocations and authorial style', *Literary and Linguistic Computing, 18* (3), pp. 261-286.

Kuiper, K. 2009. *Formulaic Genres*. Basingstoke: Palgrave MacMillan.

Kredens, K. 2001. 'Towards a corpus-based methodology of forensic authorship attribution: a comparative study of two idiolects', in B. Lewandowska-Tomaszxzyk (Ed.), *PALC 2001: Practical Applications in Language Corpora*, pp. 405-446. Frankfurt: Peter Lang.

Labov, W., & Waletsky, J. 1997. 'Narrative analysis: oral versions of personal experience', *Journal of Narrative and Life History, 7* (1-4), pp. 3-38.

Larner, S. 2014. 'A preliminary investigation into the use of Fixed formulaic sequences as a marker of authorship', *The International Journal of Speech, Language and the Law,* 21 (1), pp. 1–22.

Mollin, S. 2009. '"I entirely understand" is a Blairism: the methodology of identifying idiolectal collocations', *International Journal of Corpus Linguistics, 14* (3), pp. 367-392.

Moon, R. (1997). Vocabulary connections: multi-word items in English. In N. Schmitt & M. McCarthy (eds), *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press. 40—63.

Moon, R. (1998). *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (eds), *Language and Communication*. New York: Longman. 191—226.

Peters, A. (1983). *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Peters, A. (2009). Connecting the dots to unpack the language. In R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (eds), *Formulaic Language: Acquisition, loss, psychological reality, and functional explanations* Vol. 2. Amsterdam: John Benjamins Publishing Co. 387—404.

Shuy, R. (2006). *Linguistics in the Courtroom: A practical guide*. Oxford: Oxford University Press.

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

Stevenson, A. (Ed.) (2010) *Oxford Reference Online*, Oxford: Oxford University Press. http://www.oxfordreference.com/pub/views/home.html (accessed January 2012).

Read, J., & Nation, P. 2004. 'Measurement of formulaic sequences', in N. Schmitt (ed.), *Formulaic Sequences*, pp. 23-25. Amsterdam: John Benjamins Publishing Co.

Schmitt, N., Grandage, S., & Adolphs, S. 2004. 'Are corpus-derived recurrent clusters psycholinguistically valid?', in N. Schmitt (ed.), *Formulaic Sequences: Acquisition, processing and use*, pp. 127-151. Amsterdam: John Benjamins Publishing Company.

Scott, M. 2008. WordSmith Tools Version 5. Liverpool: Lexical Analysis Software.

Shuy, R. 2001. 'DARE's role in linguistic profiling', *DARE Newsletter, 4* (3 (Summer)), 1-5.

Vihman, M. (1982). Formulas in first and second language acquisition. In L. Obler & L. Menn (eds), *Exceptional Language and Linguistics*. London: Academic Press Ltd. 261—284.

Willis, D. 1990. *The Lexical Syllabus: A new approach to language teaching*. London: HarperCollins Publishers.

Winter, E. 1996. 'The statistics of analysing very short texts in a criminal context', in H. Kniffka (Ed.), *Recent Developments in Forensic Linguistics*, pp. 141-179. Frankfurt am Main: Peter Lang.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2006). Formulaic language. In E. K. Brown (ed.), *The Encyclopedia of Language and Linguistics*. Oxford: Elsevier. 590—597.

Wray, A. 2008. *Formulaic Language: Pushing the boundaries*. Oxford: Oxford University Press.

Appendix

	Structured Writing Task (Participants answered one question per day)
Day One	What has been the best moment of your life?
	When did you last cry and what made you cry?
Day Two	Have you ever told a lie and what were the consequences?
	What has been the worst moment of your life?
Day Three	How did you find out that Santa Claus doesn't exist?
	What is the biggest decision you have ever made and did you make the right one?
Day Four	What is the most life-threatening situation you have ever been in?
	What is the angriest you have ever been?
Day Five	What has been the most embarrassing moment of your life?
	How close have you ever got to having your heart broken?

If participants were unable to answer either question from each day's set, they were provided with the following list of five substitute questions, from which any one could be selected:

i) If you could change anything in the world, what would it be and why?

ii) Who you do admire and why

iii) If you could be invisible for a day, what would you do?

iv) What would you do if you won £1,000,000?

v) Would you like to be a housemate on *Big Brother* and what are your reasons?