DOCTOR OF PHILOSOPHY

To cut a long story short

an analysis of formulaic sequences in short written narratives and their potential as markers of authorship

Samuel Tomblin

2013

Aston University

Some pages of this thesis may have been removed for copyright restrictions.

If you have discovered material in AURA which is unlawful e.g. breaches copyright, (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please read our Takedown Policy and contact the service immediately

‘TO CUT A LONG STORY SHORT’: AN ANALYSIS OF FORMULAIC SEQUENCES IN SHORT WRITTEN NARRATIVES AND THEIR POTENTIAL AS MARKERS OF AUTHORSHIP

SAMUEL DAVID TOMBLIN

Doctor of Philosophy

ASTON UNIVERSITY

MAY, 2012

© Samuel David Tomblin, 2012
Samuel David Tomblin asserts his moral right to be identified as the author of this thesis.

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without proper acknowledgement.

Aston University

‘TO CUT A LONG STORY SHORT’: AN ANALYSIS OF FORMULAIC SEQUENCES IN SHORT WRITTEN NARRATIVES AND THEIR POTENTIAL AS MARKERS OF AUTHORSHIP

SAMUEL DAVID TOMBLIN

Doctor of Philosophy

MAY, 2012

Thesis Summary:

Previous research into formulaic language has focussed on specialised groups of people (e.g. L1 acquisition by infants and adult L2 acquisition) with ordinary adult native speakers of English receiving less attention. Additionally, whilst some features of formulaic language have been used as evidence of authorship (e.g. the Unabomber’s use of *you can’t eat your cake and have it too*) there has been no systematic investigation into this as a potential marker of authorship. This thesis reports the first full-scale study into the use of formulaic sequences by individual authors.

The theory of formulaic language hypothesises that formulaic sequences contained in the mental lexicon are shaped by experience combined with what each individual has found to be communicatively effective. Each author’s repertoire of formulaic sequences should therefore differ. To test this assertion, three automated approaches to the identification of formulaic sequences are tested on a specially constructed corpus containing 100 short narratives.

The first approach explores a limited subset of formulaic sequences using recurrence across a series of texts as the criterion for identification. The second approach focuses on a word which frequently occurs as part of formulaic sequences and also investigates alternative non-formulaic realisations of the same semantic content. Finally, a reference list approach is used. Whilst claiming authority for any reference list can be difficult, the proposed method utilises internet examples derived from lists prepared by others, a procedure which, it is argued, is akin to asking large groups of judges to reach consensus about what is formulaic.

The empirical evidence supports the notion that formulaic sequences have potential as a marker of authorship since in some cases a Questioned Document was correctly attributed. Although this marker of authorship is not universally applicable, it does promise to become a viable new tool in the forensic linguist’s tool-kit.

KEYWORDS: authorship attribution, formulaic language, forensic linguistics, idiolect, style markers

Acknowledgements

Firstly, I extend my deepest thanks to the 20 authors whose narratives form the data used in this research. In committing their time to helping me and sharing their often private stories, these people have made this research possible.

I am thankful for the time that I spent studying with Janet Cotterill and Alison Wray at Cardiff University. Both shared their extensive expertise in forensic linguistics and formulaic language respectively, and whilst they may not agree with the approaches I have taken in these pages, their perspectives have undoubtedly influenced my own.

I would like to thank all of my friends in the Centre for Forensic Linguistics—I have had a great time at Aston University. I would particularly like to mention Nicci MacLeod and Rui Sousa-Silva for their friendship and support.

Tim Grant spent many hours supervising me in the statistical aspects of my research. I left every meeting with a far better understanding of statistics and a throbbing headache. I do now have an outstanding appreciation for the ‘jelly bean’ problem and am extremely grateful for his patience with my self-diagnosed dyscalculia.

Next to thank is the person who gave me the opportunity to carry out my research and who allowed me to explore my ideas in my own way and in my own time, Malcolm Coulthard. Malcolm, I have thoroughly enjoyed working with you and feel both honoured and privileged to count myself as one of your students.

There are two people to thank, without whom my research would never have been completed, and nor would I be the same person. I consider myself to be incredibly fortunate to have such wonderful friends, advocates and parents, Andrew and Sharon Tomblin. Your unwavering support and your ability to never flinch, regardless of what I've thrown at you over the years, is what makes me love and respect you so much.

And finally, my partner, Simon Larner, has given up a great deal to enable me to finish my research. Simon, you've never complained and you've been nothing less than supportive through my many, many mood swings. I'm sorry that our life has been on hold for the past few years. It’s time for us to work on making your dreams come true now and I promise to support you with all the things that will make you happy.

Contents

List of tables

List of figures

Chapter 1

'A good beginning makes a good end': prelude

This research arises from one very simple question: if an author goes to great lengths to remain anonymous, why would they then use a phrase so distinctive that it could lead to their identification? Consider the details of the following case:

Between May 1978 and April 1995, 16 bombs targeted at individuals working in universities and airlines were detonated in America. The specific industries targeted led to the FBI codename UNABOM (Fitzgerald, 2004: 193—4) and as a result of these bombings, three people were killed and many more were injured. In 1995, *The New York Times*, along with three other recipients including *The Washington Post*, received a manuscript entitled "An industrial society and its future" which was claimed to have been written by the Unabomber to outline his ideological position—a terrorist's manifesto. Along with the manifesto was a deal—publish the manuscript in full and the bombing would stop. *The Washington Post* eventually published the manuscript in September 1995 (p. 206).

The manifesto constituted a wealth of evidence "in the form of the extensive and detailed writings of a bomber who had avoided identification and eluded investigators for so many years" (Fitzgerald, 2004: 198). Supervisory Special Agent James Fitzgerald of the FBI Behavioral Analysis Unit was tasked with a significant undertaking. He analysed numerous documents sent by the Unabomber over the years in order to ascertain clues about his identity. These texts included various ruse letters and other ideological writings as well as the manifesto. Fitzgerald was intrigued by one expression in particular: "you can't eat your cake and have it too" which he found to be historically correct but not in popular usage (p. 213):

> While not a mistake in spelling, grammar or punctuation, it was very interesting that in spite of how careful this writer was in crafting his manifesto, he made what seemed to be a simple mistake. Perhaps it was his haste or just carelessness, but for whatever reason, he transposed the two verbs in this well-known proverb (p. 205).

Upon reading an internet version of the manifesto, Linda Patrik was unnerved. Although she had never met her brother-in-law, there was something about the text that seemed familiar. She asked her husband, David Kaczynski, to read it and then urged him to compare the manifesto with her brother-in-law's known writings. David sceptically complied, but started to suspect that his older brother Theodore may indeed be the Unabomber. The occurrence of one phrase in particular convinced him: *cool-headed logicians*. David recalled his brother "using that distinctive term on numerous occasions" and as a result, he contacted the FBI (Fitzgerald, 2004: 208). David Kaczynski and his mother made available many documents known to have been written by Theodore. In one of these documents, Fitzgerald also found the historically correct expression “we can’t eat our cake and have it too”.

In April 1996, after reviewing all of the evidence, including the report of an extensive comparison of documents known to be written by Theodore Kaczynski and documents written by the Unabomber, a federal judge signed a warrant to search Kaczynski’s cabin in Montana. Inside the cabin was “a virtual treasure trove of evidentiary materials” including a fully assembled bomb and numerous bomb parts (Fitzgerald, 2004: 215). In 1999, Kaczynski pleaded guilty to being the Unabomber (p. 219).

So herein lies the mystery. Since Kazcynski went to such great lengths to evade detection—he even stopped his bombing campaign for seven years because he thought he may have been identified (Fitzgerald, 2004: 196—7)—why then did he use phrases which were so distinctive and so characteristic of him? It is of course possible that he wanted to get caught, although the FBI agent overseeing the text comparison project does not accept this proposition (Fitzgerald, personal communication). Surely then there is only one other reason: Kaczynski was unaware that these phrases were so idiolectally distinctive. So how can an author use phrases and not be aware of their distinctiveness? What if these phrases were used so formulaically that they simply seemed normal?

In this research, it will be proposed that phrases such as these are examples of formulaic language, “words and word strings which appear to be processed without recourse to their lowest level of composition” (Wray, 2002: 4)—in other words, sequences of words that are holistically processed as single items. If authors treat a sequence of words as one lexical choice, they are unlikely to be aware of their own idiolectal preferences or of the words contained in such holistic sequences. It therefore follows that a low-level lexical feature such as formulaic language may hold the potential to differentiate between different authors.

1.1 Thesis overview

To investigate this claim, the key issues relating to forensic linguistics, specifically forensic authorship attribution, and those relating to the field of formulaic language will be presented. In Chapter 2, the underlying assumptions of authorship attribution are questioned and the evidence in support of idiolect is assessed. Qualitative and quantitative approaches to authorship attribution are then described with the pros and cons of each being evaluated. The evidential status of authorship attribution evidence is then discussed in order to contextualise the ensuing empirical research, paying particular attention to the *Daubert* evidential standard. Chapter 3 moves the focus to the field of formulaic language and begins with an outline of what it actually is, and its underlying theory is evaluated. This theoretical groundwork, in conjunction with the findings of Chapter 2, enables a full consideration of why formulaic language *should* hold potential to differentiate authors and enable a Questioned Document to be successfully attributed. Also in Chapter 3, the key issues surrounding identification of formulaic language in written text are presented. Chapter 4 outlines the research design for the empirical work: this includes describing three approaches that should be fruitful avenues to explore as well as a full account of the data used.

Chapters 5—8 then describe the results of these three approaches. Specifically, Chapter 5 outlines a corpus-based, frequency driven approach that identifies identical forms across a series of texts. In Chapter 6, one core word is focussed on in detail—a word which is central to a subset of formulaic sequences. Crucially, this chapter also investigates alternative non-formulaic realisations of the same semantic content. In Chapter 7, a reference list approach is investigated. What differentiates this approach from previous attempts at using reference lists is the way in which the list was created—by collecting examples of formulaic language from pre-existing lists available on the internet rather than relying on any one individual's intuition. At the end of each of these three empirical chapters, full consideration is given to the results, paying particular attention to whether the methods are valid, reliable, and feasible for forensic purposes.

Chapter 8 attempts to draw the three empirical chapters together in order to answer the central research question of whether formulaic sequences can be used as a marker of authorship. In this chapter, the empirical work is reviewed and the limitations of the data and the methods are considered. The research is finally concluded in Chapter 9, where future directions for the investigation of formulaic language as a marker of authorship are proposed.

Chapter 2

‘The more we learn, the less we know’: forensic authorship attribution and idiolect

This thesis sets out to explore the use of formulaic language as a marker of authorship. The investigation therefore requires an understanding of relevant research in both forensic authorship attribution and formulaic language.

Questions relating specifically to forensic authorship attribution will be dealt with in this chapter, namely:

- How robust is the theory underlying forensic authorship attribution?
- How robust are existing approaches to forensic authorship attribution?
- How is forensic authorship attribution evidence received by the courts?

Chapter 3 will then address the issues relating to formulaic language before attempting to draw the two fields together in order to consider its potential as a marker of authorship.

2.1 How robust is the theory underlying forensic authorship attribution?

There are two assumptions underpinning authorship attribution:

i) Every author has a unique variant of the language(s) they use, their *idiolect*; and

ii) Idiolectal features can sometimes be identified in texts through comparison with other texts.

This research is concerned with authorship attribution in the forensic linguistics context, as opposed to the attribution of literary texts (e.g. Clemit & Woolls, 2001; Covella, 1976; Hoover, 2003b; Mannion & Dixon, 2004; Putter, 2004; Smith, 1986). By necessity then, these assumptions have been simplified and do not take into account the practical issues inherent in forensic investigations, namely, that the linguist typically has no choice over what texts are available for analysis, that some authors may deliberately disguise their style, that the length of texts will undoubtedly vary, and that the texts may not be comparable in terms of genre and date of composition. Each of these factors alone will affect how many idiolectal features may be identified and whether they are, in a given case, distinctive for a particular author. In this way, just as some physical characteristics are closer between some people such as identical twins (e.g. Künzel, 2010; Mollet, Wray, Fitzpatrick, Wray, & Wright, 2010), so too will some idiolects be closer, meaning that texts may be explainably non-distinctive. The result is that some texts may be better or worse exemplars of idiolect and furthermore that in some cases, attributions are more successful whilst in others, they are impossible. These are practical issues which will be returned to at various points throughout this research.

2.1.1 On defining *idiolect*

Although the term *idiolect* was first coined by Bloch (1948), Sapir (1927) laid the groundwork in his discussion of the relationship between speech and personality. Sapir outlined five levels of speech that were indexical of individual personality including voice, dynamics, pronunciation, vocabulary and style. Of these, *vocabulary* and *style* are the most relevant precursors to the concept of idiolect. Sapir argued of vocabulary that:

> We do not all speak alike. There are certain words which some of us never use. There are other, favorite, words which we are always using … Individual variation exists, but it can properly be appraised only with reference to the social norm. Sometimes we choose words because we like them; sometimes we slight words because they bore or annoy or terrify us. We are not going to be caught by them. All in all, there is room for much subtle analysis in the determination of the social and individual significance of words (p. 903).

Here, Sapir clearly draws out the complex relationship between the individual and society, further exemplified through his consideration of individual style:

> We all have our individual styles in both conversation and considered address, and they are never the arbitrary and casual things we think them to be. There is always an individual method, however poorly developed, of arranging words into groups and of working these up into larger units. It would be a very complicated problem to disentangle the social and individual determinants of style, but it is a theoretically possible one" (p. 903-4).

Some linguists have given a more prominent place to writing alongside speech than Sapir (e.g. Coulthard, 2004) whilst for others, the term 'style' is instead a recognised term for 'idiolect in writing' (e.g. Kredens, 2002). In the context of this research, idiolect will be considered to include written manifestations.

A good early definition for the discussion of idiolect is Hockett's (1958): "the totality of speech habits of a single person at a given time constitutes an idiolect" (p. 321). Hockett's definition raises two issues: potentially, one might need to observe and catalogue every single speech habit before one could fully characterise an individual's idiolect and that idiolect will change over time.

In so far as *totality* means *complete* and *entire*, Hockett appears to suggest that idiolect is the entire repertoire of speech habits available to a single person. However, it is impossible to collect a totality, although for Hockett's purposes, this would not have been an issue. In fact, in a later paragraph, Hockett notes that the entire idiolect cannot be observed, only examples of the linguistic output that it generates (1958: 322). In other words, rather than being able to observe the totality of habits, all that the linguist can observe is what a speaker or writer actually does at the particular point of observation.

The second implication of Hockett's definition, that idiolectal features can only be described *at a given time*, implies that idiolect is organic and evolutionary in nature and will differ when observed at different times. This raises the question of by how much and whether the difference is significant. Related to this is the issue of the rate at which such change occurs; a question which so far has received no definitive answer with the exception of Bel *et al*. (2012) who show that the use of bigrams and trigrams do not vary substantially across a span of between 6—10 years for individual authors (ages unknown), indicating that this feature remains sufficiently stable for this limited period of time at least. Corroborative evidence is provided by Barlow (2010) who found that bigrams were used consistently over the shorter period of one year in the spoken language of White House Press Secretaries (the issue of fossilisation of idiolects is discussed in Section 2.1.5). The problem with such a definition for forensic purposes is that not only would the idiolect of one individual differ from that of another (as assumed in authorship attribution), but would also be subject to variation between the same individual when observed at different points. If this definition is to be accepted, it would make the comparison of documents in the forensic context very difficult because Known Documents (those documents whose authorship is attested) are rarely authored at the same time as each other or as the Questioned Documents (those documents whose authorship is unknown or under suspicion).

Sixty years later, Louwerse (2004) claimed that writers "implicitly leave their signature in the document they write" and that idiolects "are person-dependent similarities in language use" (p. 207). He explains that if idiolect exists, texts composed by one author will show more similarities in language than texts composed by different authors (p. 207). However, a potential problem arises in relation to Hockett's definition. Louwerse states that similarities between texts produced by one author will be greater than texts produced by different authors. Hockett proposes that idiolect will change over time. Unless the individual signatures upon which Louwerse's definition relies remain static, the similarities between two pieces of writing by the same individual at different times could be no greater than the similarities between two individuals with similar linguistic backgrounds (a common assumption e.g. Loakes, 2006). It seems then that the temporal dimension could indeed be a confounding variable in forensic authorship attribution. Through examining a third definition of idiolect, a clearer picture may be gained.

Coulthard (2004) also says that "every native speaker has their own distinct and individual version of the language they speak and write, their own *idiolect*" and that "this *idiolect* will manifest itself through distinctive and idiosyncratic choices in texts" (p. 431—2, original emphasis). The main difference here is between what Coulthard refers to as *choice* and what Hockett refers to as *habit*. Insofar as c*hoice* implies conscious decision, *habit* implies an involuntary behaviour pattern. If idiolect is based on habit, it is reasonable to argue that a person's linguistic patterns will remain constant, until such a time when that habit is changed. In this scenario, texts produced during a period when the habit remains the same should be comparable. Choice, however, is more volatile and dependent on many extra-linguistic factors (e.g. mood of the individual, genre of the text, audience of the text, time available to compose the text and indeed recency) as well as conscious attempts to disguise identity. As such, any features of language that are subject to choice could result in differences between texts produced by the same author, regardless of when they were authored. Choice is a confounding variable, so forensic linguists may benefit from researching markers of authorship which are beyond an author's conscious control which might include, for example, a concentration on grammatical items (e.g. Mosteller & Wallace, 1963) and formal features such as word length (e.g. Grant, 2004).

These three definitions, somewhat representative of the many that could have been reviewed (e.g. Labov, 1972b; Trudgill, 1974, 2003; Wardhaugh, 2006) capture between them the key issue for authorship attribution, namely, the extent to which an individual's idiolect really is a reliable signature irrespective of stylistic choice and change over time. Additional issues that must be borne in mind include the question of whether idiolect really exists at all—whether each individual's language really has its own signature—and how much linguistic output would be needed in order to capture a sufficiently idiolectal profile.

Kuhl (2003) offers a theoretical account of idiolect by approaching the notion from a cross-disciplinary perspective—that is, he applies the concepts and theories from Natural Systems Theory and Complexity and Chaos Theory (rooted in physics, chemistry and computing) to theoretical linguistic notions including language, language contact, and idiolect (p. 57). In this way, Kuhl talks of human beings as being 'open systems' which evolve through interaction with the physical and cognitive environment:

> We not only take in food and excrete waste, but we apply energy to the world in the form of work and existence in general. We act on and are acted upon by the world and its many others, and are designed by nature to change and evolve our behaviors, and prime among these behaviors is language (p. 260).

This natural sciences approach to language enables Kuhl to argues that language contact is an exchange of energy leading to change, adaptation and innovation:

> The idiolect is bounded by the sum total of one's lived language experience, and its particular form is "shaped" as such. The language choices one makes and is able to make [sic] is also determined by linguistic situation, that is, language contact with other idiolects. We bring to bear the linguistic resources, which include pragmatic knowledge and worldview, that seem appropriate to negotiation [sic.] a particular conversation or discourse. We are engaged in a process of feedback with our interlocutor and these idiolects exert influence over one another throughout the duration of the contact. The influence of other idiolects, depending on frequency of contact, may lead to long-term changes and alterations of a particular idiolect while other changes may be short-term or become stored as a part of the larger idiolectal resources of individuals (p. 261).

Clearly, for Kuhl, idiolect is not static but rather dynamic, changing and adapting both short-term and long-term in response to contact with other interlocutors—other idiolects. Such a view immediately creates problems for the forensic linguist seeking static features of idiolect for the purposes of authorship attribution since texts created at other times are unlikely to be fully comparable, unless the author has had virtually no contact with other idiolects. Nonetheless, it is interesting that Kuhl opts for the term "choices" in the above quotation implying, as with the previous definitions of idiolect, that conscious decisions are made.

It is readily acknowledged that the theory of idiolect, to date, lacks empirical investigation (e.g. Kniffka, 2007; Kredens, 2001, 2002; Louwerse, 2004), and, as discussed above, the totality of linguistic habits for each person can never fully be observed. This point is echoed by Coulthard (2004): "any linguistic sample, even a very large one, provides only very partial information about its creator's idiolect" (p. 432). However, there is a growing sense that without empirical testing, it will prove increasingly difficult to get expert testimony on authorship admitted to the courts (e.g. Howald, 2008; Kniffka, 2007; Kredens, 2001, 2002) and so, despite Kuhl's (2003) novel approach to idiolect, it is important to concentrate specifically on empirical, rather than theoretical, investigations.

Despite a general lack of research, three studies in particular have attempted to find empirical support for the existence of idiolect. Louwerse (2004), focussing on semantics, used a corpus of literary texts to find evidence of idiolect. Kredens (2001) investigated idiolect at the lexical and grammatical levels in spoken language and Mollin (2009) compared collocations in language produced by former British Prime Minister Tony Blair against a reference corpus to demonstrate how idiolectal collocations can be identified. These three pieces of research will be considered in detail in an attempt to weigh up the evidence.

2.1.2 Evidence in support of *idiolect*: a literary approach

Based on previous research which found that literary authors writing in the same period shared pragmatic, syntactic and semantic similarities, Louwerse (2004) reasoned that perhaps language features might also be shared across different groups based on idiolect and so formulated the hypothesis that linguistic features in the texts of one author should not be significantly different whilst texts by different authors should (p. 209). Additionally, a *sociolect-time* hypothesis was proposed (texts written in the same time period should not significantly differ whilst those written in different time frames should) which may have been useful in determining whether authorial style changes over time as implied by Hockett (1958). However, Louwerse grouped authors from one time period and compared them with groups of authors from another time period rather than looking at individual author-specific changes over time. As such, only the idiolect hypothesis is of interest.

Sixteen literary texts were used in the analysis, four from each of four Modernist authors (George Eliot, Charles Dickens, Virginia Woolf and James Joyce) (Louwerse, 2004: 210). The total corpus consisted of 3,327,487 words with the shortest text consisting of 20,863 words and the longest text containing 363,323 words. Louwerse reports that Fokkema and Ibsch (1987) identified thirteen semantic fields characteristic of the Modernist period: *consciousness, observation, detachment, agriculture, criminality, economy, industry, nature, psychology, religion, science, sexuality* and *technology,* so he too used these categories which he populated with 592 lemmata created from two sources: *Roget's Thesaurus* (each of the semantic fields was used as a keyword in the thesaurus so that semantically related words could be identified) and the *WordNet* database (using the label of the semantic field as a hyponym allowed all related hyponyms to be selected). For each of the 592 lemmata, corresponding derivations and inflections were generated to produce a total of 1,461 word forms (p. 210—1). Louwerse accounted for the different text sizes by converting the raw frequency data into *a per 1,000 words of text* score (p. 211).

Between-groups (texts between authors) and within-groups (texts by one author) comparisons were made to test the idiolect hypothesis. The word frequency of semantic fields did indeed differ significantly between some of the authors. Whilst there were significant differences between Dickens with Eliot, Dickens with Woolf and Dickens with Joyce, there were none found between Eliot and Joyce, Eliot and Woolf or Woolf and Joyce. Furthermore, the within-group testing showed that each of the four texts by Eliot, Woolf and Joyce differed in the word frequency of semantic fields. Louwerse concluded that the idiolect hypothesis was only confirmed by the texts of Dickens, "resulting in very limited support for the idiolect-hypothesis" (2004: 212).

A possible explanation for this result is that it relies on precise semantics—a word form is either present in a text or not. It is feasible that the general semantic field may be present but not in the exact word-forms expected. Therefore, Louwerse carried out a second test using Latent Semantic Analysis (LSA), "a statistical, corpus based, technique for representing world knowledge" (2004: 213) where quantitative information about co-occurrences of words in paragraphs and sentences are entered into a matrix so that the semantic relationship between words can be estimated (Louwerse provides more detail about the statistical method behind LSA). According to Louwerse, LSA is special because it not only determines semantic relatedness between words, but also the semantic relatedness of words that accompany words. He provides the example that *consciousness* and *mind* are semantically highly related not because they occur in the same sentences and paragraphs together but because "words that co-occur with one equally often co-occur with the other" (2004: 214).

Applying the LSA method to the same 16 texts and using the same 13 semantic fields, Louwerse found that between-author groups differed from each other as predicted in the idiolect hypothesis. However, the four texts written by Eliot also differed significantly from each other, as did each of the four texts by Dickens and Joyce. Only the texts by Woolf seemed to be homogenous (2004: 215). Louwerse proposed that the 13 semantic fields selected may be masking idiolect; in other words, semantic analysis may still reveal idiolectal behaviour, but the analysis is restricted by the predefined list of semantic fields. His third study accounted for this possibility.

All paragraphs and sentences in each text were compared with all paragraphs and sentences in all other texts, rather than against a word list as in the previous studies. For Dickens and Joyce, each of their four texts differed more between themselves than between the texts by other authors. The texts produced by Eliot and the texts produced by Woolf showed homogeneity amongst themselves. Louwerse concluded that for Eliot and Woolf, "similarity in content can be found, supporting an idiolect hypothesis" whilst for Dickens and Joyce, the texts differed within one author (2004: 217). Louwerse acknowledged that his analyses assume that the semantic fields within each text remain the same.

Therefore, a final investigation was required to check this assumption. Using the same LSA methods and the same texts as outlined for the previous studies, an ANOVA test showed significant differences between the four authors but also differences were found between each of the four texts for each of the authors. Louwerse concluded again that no support for the idiolect hypothesis could be found (2004: 218). A summary of the four investigations described and the results are shown in Table 2.1.

Table 2-1 Summary of investigations carried out by Louwerse (2004)

Study	Investigation	Result
1	Predefined list of semantic fields compared to words in each of the texts	Very limited support (for Dickens only)
2	Predefined semantic fields compared using LSA	Very limited support (for Woolf only)
3	All sentences and paragraphs compared with each sentence and paragraph in the other texts using LSA	50% success (texts by Eliot and Woolf)
4	Comparisons to determine semantic homogeneity within texts using LSA	No support for idiolect hypothesis

To summarise, Louwerse tested a hypothesis which presupposed that authors use the same semantic fields and that the frequency of the contents of those fields would predict patterns in idiolect (2004: 219). After finding only very limited support for the idiolect hypothesis, Louwerse argued that drawing conclusions based on semantic similarities is problematic because authors change their style and semantic space between texts (2004: 219—20). Furthermore,

> [The] lack of internal homogeneity in one text, between texts and between authors can be explained by the (semantic) deviation from the norm the author tries to establish. These variations are exactly what makes the idiolect … of literary texts unique, and is in fact what makes those texts literary (2004: 220).

Louwerse inadvertently implies that literary texts are not suitable texts on which to find support for idiolect because of the variations inherent in them—otherwise known as 'elegant variation' whereby alternative expressions are used as a replacement for previously used expressions to avoid repetition (Leech and Short, 2007: 197). If authors monitor their idiolects so easily when constructing a text, it seems unjustified to label this as idiolectal when really it is a case of an author making a stylistic choice for literary effect and the point should perhaps be made that literary novelists are by no means normal, everyday writers[1].

This point is made by Grant (2008):

> [T]here do seem to be some features of forensic texts which distinguish them from those texts typically analysed in literary, historical or scriptural authorship work. The texts of these more literary analyses are of course also diverse. It might be tentatively argued, however, that this non-forensic caseload concerns texts which are in some way crafted; the author may have spent some time and thought in their composition. Further to this, this crafter feel may be because many of these texts were intended for a wider readership. It also may be the case that these texts are generally written by professional or at least educated writers. Finally it might be thought that these texts are written to impress the reader, in some way. If

[1] Although Carter (1999) argues that "everyday conversational discourse" contains literary properties and that "ordinary" language *is* creative (p. 195).

> any of these assertions can be accepted it might also be accepted that they tend to be points of difference with forensic texts (p. 216—7).

Louwerse's investigation is more in line with Coulthard's (2004) definition of idiolect being about choice rather than habit and in light of this, the same results could perhaps not be expected from a set of non-literary texts produced by the same authors. It is therefore questionable whether testing idiolect on literary texts is an appropriate way to identify idiolect, especially for forensic purposes.

Moreover, it is also questionable whether Louwerse's approach was a suitable way to explore idiolectal variation. The investigation relied solely on semantic fields and it may be the case that using a small set of semantic fields is not the most efficient way to investigate idiolect. This point is indirectly made by Louwerse himself, when he argues that "similarity in content can be found, supporting an idiolect hypothesis" for the texts produced by Eliot and Woolf. Supposedly, a method based on semantic fields could only ever expect to find similarity in content. After all, Louwerse's initial inclination to explore semantic fields was because of earlier work by Fokkema and Ibsch (1987) who found that authors writing during the Modernist period displayed similarity in semantics. It is therefore questionable whether Louwerse actually investigated idiolect in this research, or whether he simply identified that authors writing at the same point in history shared an enthusiasm for the topics about which they wrote. To investigate this possibility, a researcher would need to apply the semantic field analysis to a variety of texts, preferably those written in a non-literary genre. Only then could the relationship between semantic fields and idiolect be established.

In conclusion then, Louwerse's study is not necessarily damaging to the idiolect hypothesis; rather, it is more likely that his approach was too narrow. Empirical research based on non-literary texts, using a broader range of variables may provide more fruitful avenues for investigation.

2.1.3 Evidence in support of *idiolect*: a forensic approach

In a 2001 paper based on his PhD thesis, Kredens aimed to improve understanding about idiolectal variation specifically in the forensic context by identifying and describing differences between two sociolinguistically similar spoken language corpora (p. 406). The fact that he focused on lexis *and* grammar, and that his data were not literary, make this approach very different from that of Louwerse (2004) and indicates that his findings should be more generalizable.

Kredens explored the notion of idiolect by comparing spoken data from two people with similar biological (e.g. gender, age) and social backgrounds, both performing the same function, namely, speaking during radio and television interviews (2001: 407). He reasoned that if idiolectal differences could be found even between biologically and socially similar people, the variation would be greater between speakers with dissimilar characteristics (2001: 406). Seven criteria were proposed to ensure that the speakers were linguistically homogeneous:

i) The speakers should have the same socioeconomic and educational background;
ii) be similar in age; and
iii) be of the same gender.
iv) The language should originate in a similar type of speech event; and
v) similar subject matter should be discussed.
vi) The tape-recorder should not influence delivery.
vii) Points i, ii, iii and vi should also apply to any other speakers whose idiolects are not being studied (i.e. the interviewers) (2001: 407).

Kredens further stipulated that only small corpora should be used since the amount of spoken material available in the forensic context is typically very limited (2001: 406). He found suitable material in media interviews with two contemporary English musicians Robert Smith and Steven Morrissey. Both had been born in 1959 in the North of England into working class families. Their formal education had ended at 18 whereupon they both co-founded successful bands and moved to London where they lived for several years and both enjoyed successful careers in the music industry as songwriters and performers, which Kredens took as evidence that they were " affluent individuals" (2001: 408). On this basis Kredens reasoned that their idiolectal development would have been conditioned by similar factors.

A total of six radio and television interviews were selected (three for each author) which contained only two participants in each and where similar subject matter was discussed. The Smith corpus contained 3,243 words and the Morrissey corpus contained 3,167 words. Kredens supposed that any effect of being taped would be the same for both and furthermore that since they were seasoned performers for over twenty years, they would not have paid particular attention to their language in the interviews. Kredens argued that "it can be quite safely assumed that the selection contains a representative sample of their idiolects" (2001: 409) although he noted that the data probably lacked certain linguistic features such as informal register and profanities (2001: 408—9).

A null hypothesis was formulated: no significant differences between the speech styles of Smith and Morrissey would be found. To test the hypothesis, a set of linguistic features with potential to differentiate between the speakers were selected based on impressionistic judgements of the corpora and successful discrimination in other authorship studies. Kredens selected the following potential discriminators for analysis:

a) Lexical level

   i. Most frequent words
   ii. Average word length
   iii. Type/Token ratio
   iv. Hapax legomena

b) Grammatical level

   i. Contracted forms
   ii. Frequency of adverbs
   iii. Use of adverbs for intensification
   iv. Frequency of adjectives
   v. Emotive adjectives
   vi. Discourse markers
   vii. Relative clauses

The frequency and usage of each of these features was analysed and compared across both corpora. A table displaying Kredens' results is reproduced below as Table 2.2.

Table 2-2 Summary of Kredens' results (2001: 440)

Aston University

Illustration removed for copyright restrictions

Table 2-2 shows that the most frequent words, the frequency of adverbs and discourse markers (such as *you know*) discriminated between the speech of Smith and Morrissey at the $P<0.001$ significance level. The use of intensifying adverbs, the frequency of adjectives and the use of emotive adjectives discriminated between the two speakers at the $P<0.05$ level of significance. Kredens therefore rejected the null hypothesis and concluded that the differences enabled the contrast in speech styles to be shown:

> Robert Smith's is largely incoherent, disfluent and rather informal; Steven Morrissey's—characterised by explicitness, emotiveness and a certain degree of formality. Some degree of idiolectal variation has thus in this case become an empirically proven fact; it is the case that similar biological, social and interactional characteristics do not preclude the possibility of there being a considerable difference between two speech styles (2001: 440—1).

Kredens states that on the basis of this study, "certain scientific grounds seem to exist for supposing that some elements of an individual's idiolect are always present in his speech" (2001: 442—3). The impact and importance of this conclusion for authorship attribution research should be apparent, albeit with a few potential limitations.

In assessing this approach and the assumptions made, Kredens pointed out that the topics of conversation in each interview were similar (e.g. musical influences, performing live, the music industry and work on new albums) which adds further credibility to idiolectal differences emerging even when the same topics were being spoken about. It is important to consider though that these topics of conversation are very common amongst all musicians giving interviews and the influence of both coaching and repetition is a variable which may have been overlooked. It is quite possible that their utterances contained pre-fabricated responses and therefore might not be characteristic of Smith and Morrissey's normal conversational habits, or choices, since their language may have been rehearsed and manipulated by others. The effect of the tape-recorder may also have been slightly more evident than assumed since Kredens acknowledges that certain features such as informal register and profanities were missing from the data, which might also explain Kredens' conclusion that Smith's speech style was incoherent and disfluent.

In terms of biological, social and economic factors, the subjects were extremely well matched. What can less well be accounted for is how closely the speakers were matched in the interactions which would have shaped their linguistic development. One may have interacted with adults more than peers as a child whilst one may have enjoyed crosswords, reading and language puzzles more than the other. These factors alone may have shaped the level of sophistication, and variety and complexity of lexis available to each subject. Of course, it would be impossible and unreasonable to create an inventory for each person based on the history of their interactions, but the point should perhaps be made that in forensic data, there are always factors which one cannot control.

When contrasting these two empirical studies, differing conclusions may be expected. Louwerse, on one hand, used much longer texts. If idiolect does exist, it should be more prevalent in longer texts because there will be more opportunity to capture it. On the other hand, although Kredens used considerably shorter texts, he used a wider variety of markers than Louwerse who explored only semantic fields. Given Kredens' results, it would certainly be fruitful to build upon this approach, noting in particular that convincing evidence of idiolect was found for two non-literary authors who were far more closely matched than might usually be encountered in the forensic context. Therefore, adopting a similar approach with larger groups of authors may be the next logical step in establishing idiolect.

2.1.4 Evidence in support of *idiolect*: a reference corpus approach

Rather than attempting to demonstrate that certain linguistic features differ between a closed set of authors and then attributing those differences to manifestations of idiolect, Mollin (2009) explored a large data set for one person and compared the occurrence of one linguistic feature, maximizer collocations (collocates of intensifiers "which express the very highest point on the scale" e.g. *entirely understand, completely understand, fully understand* (p. 373)), against a reference corpus. Using this approach, Mollin was able to show: (i) differences between one speaker against the general language community, ii) that those differences are not attributable to factors such as genre, and iii) that speakers make choices in the language they use and some choices, although not necessarily striking, may be preferred by some speakers and dispreferred by others and are therefore part of that speaker's idiolect.

Mollin selected public speeches, statements and interviews given by Tony Blair (2009: 370). She collected 3,119,931 words. Less than 1% of the data were taken from newspaper articles written by Tony Blair, so this research is predominantly concerned with idiolect as it relates to spoken data. Mollin wanted to include as much spontaneously spoken language as possible: 52% of the corpus contained naturally occurring speech but 48% was potentially preformulated, meaning it could have been written by someone other than Blair (p. 371). Mollin argued that although maximizer collocations were more frequent in the spontaneously spoken portion of the corpus, the proportions between different maximizer collocations in both sections of the corpus was substantially the same (p. 371), so the high proportion of preformulated language was not a problem for her. The data spanned the years 1988 to 2007 (p. 372). Unfortunately, Mollin did not carry out a longitudinal analysis to determine if, and whether, Blair's idiolect developed and/or fossilised—a point which will be returned to in Section 2.1.5.

Eleven maximizer adverbs were used to detect collocational patterns that could be associated with Tony Blair. Maximizers were three times more common in Tony Blair's speech than in the BNC, representative of British English, and she reasoned that this was likely to be a characteristic of political speech rather than being a feature of Blair's idiolect (p. 374).

Mollin identified three steps that researchers must take to isolate real differences in frequencies between corpora:

i) Statistically identify the words that co-occur more frequently than would be expected by chance across both corpora (in this case, Blair and the BNC);

ii) Identify collocations that statistically appear to be typical of idiolect (in the Blair corpus in this case); and

iii) Test the identified collocations qualitatively using a synonym test, taking into account choice of equivalent synonymous collocations and register-specific collocations.

Without these three steps, Mollin argued, "all that the researcher would be able to say is that there are differing frequencies of collocations in two different corpora, but not which differences in frequency really make a difference" (2009: 376). It is this ability to describe which frequencies really make a difference that separates Mollin's research from that of Louwerse (2004) and Kredens (2001) and which, crucially, moves closer towards the evidential standards required of forensic authorship attribution methods by testing markers of authorship against a reference corpus (Grant, 2007).

Mollin identified 578 maximizer collocations that occurred in both the BNC and the Blair data. Using Sheer Frequency scores, Mutual Information scores and Log Likelihood measures, Mollin narrowed down to 42 maximizer collocations which had statistically different frequencies in the Blair corpus compared with the BNC and which therefore held potential to be idiolectal. Having completed the quantitative steps, a qualitative analysis was carried out on the basis that some collocations may be the same as those found in the BNC or characteristic of particular registers (2009: 382). In other words, the 42 maximizer collocations may well be statistically different in the Blair corpus compared to the BNC, but they could be features of politician-speak rather than indexical of Tony Blair's idiolect. The 42 collocation candidates were subjected to a synonym test where each collocation is viewed "as a variant of a variable" (p. 382). It is not sufficient to demonstrate statistically that *absolutely central* is used 35 times more frequently in the Blair corpus than it is in the BNC. Instead, the researcher needs to ask: "[I]f speakers want to express that something is maximally central, which maximizer will they choose?" (p. 382). For example:

Blair Corpus: maximizer + *central* <*absolutely* 100%> (15 tokens)

BNC: maximizer + *central* <*absolutely* 92%, *fully* 8%> (12 tokens)

This example shows that in the Blair corpus *absolutely* collocates with *central* 100% of the time that he wanted to express that something was maximally central. Yet in the BNC, *absolutely central* is also the predominant choice (used 92% of the time) with *fully central* being only a minority choice. Therefore, even though *absolutely central* occurs 35 times more in Blair’s speech than it does in the BNC, it cannot be argued to be idiosyncratic to Blair since it is also the majority choice for expressing the same concept in the BNC. Blair just happened to talk about things being maximally central more often (p. 383). For authorship attribution purposes, it would be wise to consider that forensic linguists such as Grant (2010) might alternatively argue that 100% is significantly different from 92% and that by combining other variables on which there is also variation, a strong opinion about authorship may still be reached. However, since Mollin was not concerned with authorship, *absolutely central* could be discounted as an idiosyncratic choice for Blair.

From this analysis, Mollin identified 25 collocations that were potentially idiolectal. She finally checked that they were not specific to certain registers or groups of speakers and she tested whether any of the 25 collocations were typical of three registers in the BNC: speech, newspaper writing or academic writing. Collocations were then assigned to one of four categories as shown in Table 2.3:

Table 2-3 Mollin's (2009) results of maximizer collocations (p. 387)

Aston University

Illustration removed for copyright restrictions

Mollin concluded that through these analyses, “a mere sixteen collocations remained—and these ought to be ‘entirely accepted’ as idiosyncratic preferences” for Tony Blair (p. 389).

Rigorous methods were used in this research to demonstrate which collocations can fairly be attributed to the idiolect of Tony Blair and combining statistical measures of frequency with qualitative analyses in comparison to a reference corpus moves closer to capturing aspects of idiolect than perhaps Louwerse and Kredens were able to achieve. However, there are some limitations which may affect the level of support towards empirical evidence of idiolect. Mollin claimed to focus on a more ordinary user of language (2009: 369) and in comparison to Louwerse (2004) who focussed on literary authors, this may be the case. However, as with Kredens' (2001) choice of subjects, it is debateable whether public figures using language in a media context are ordinary language users. Considering that Blair needed to project a particular public image of himself (presumably confident, intelligent, trust-worthy and motivated by the desire to be re-elected) and given that almost half of the corpus was scripted to a greater or lesser extent, and given that Blair is a practised and accomplished public speaker, it is less convincing to argue him to be an ordinary user of language.

Related to this point is the question of whether Blair can truly be considered the author of his texts in this specific political context. This is highlighted by Grant (2008) who describes an example from D. Foster's (2001) British edition of *Author Unknown*:

> … some newspaper articles 'signed' by UK Prime Minister, Tony Blair, were in fact written by his then press secretary, Alistair Campbell. The suggestion is that although Tony Blair is the declarative author, Alistair Campbell was the executive author[2]. A parallel may be drawn with political speech making where it is typical that these two authorship functions are be [sic.] separated (between the speechwriter and the politician who delivers the words) but with written texts such a division of labour is perhaps more controversial and less frequently acknowledged." (Grant, 2008: 218)

Clearly, the argument that the 16 maximizer collocations are idiolectal for Tony Blair is somewhat devalued if the corpus was contaminated by the choices and habits of one or more other authors—an important consideration given that almost of half of Mollin's data were potentially preformulated. On the other hand, Mollin argued "it is assumed that the phenomenon under research, maximizer collocations, is stable enough in the speaker for it to surface" even when the contents "may have been pre-established" (2009: 371).

Next to consider is Mollin's opinion that the 16 maximizer collocations classified as idiolectal to Blair were not register specific. Comparing Blair's speech to the BNC and eliminating collocations that occurred in other parliamentary texts does increase the likelihood that the collocations associated with Blair were idiolectal rather than shared amongst the parliamentary language community. What Mollin was not able to demonstrate is whether Blair would use the same collocations when performing the role of husband or father, or rather how generalizable Blair's idiolect is outside of his work persona, although this of course would be a tremendous undertaking. This point really relates to Hockett's definition of idiolect in that Mollin was not able to capture the

[2] These are Love's (2002) terms. The 'declarative author' is the person who takes responsibility for the content, "the validator" (p. 44), whilst the 'executive author' is the person who actually writes the text, "the compiler of the verbal text up to the point where it is judged suitable for publication" (p. 43).

totality of Blair's idiolect, only Blair's idiolect as a politician. The relevance to forensic linguistics is whether two texts produced by Blair (e.g. a political speech and a private family letter) would be comparable.

A further complication with Mollin's research is the view that certain maximizer collocations are idiolectal for Blair because they are not the predominant choices for authors in the BNC. It may be true that in comparison to a general reference corpus such as the BNC, certain collocations do appear to be idiolectal to Blair. However, since no author in the BNC contributed 3 million words (as Blair did), quite whether there was sufficient opportunity for other authors' idiolectal patterns to be detected is unlikely. For this reason, Kredens (2001) is able to more convincingly show how two idiolects differ from each other but without the benefit of showing how the idiolects of Morrissey and Smith differed in relation to other idiolects.

Finally, it should not go unnoticed that there is an element of intra-author variation within Blair's choice of maximizer collocations. From the 16 maximizer collocations attributed as idiolectal to Blair, *absolutely committed* and *completely committed* occurred, as did *entirely understand* and *totally understand*. This means that when Blair wished to convey that someone or something is maximally committed or maximally understands, he used two variants for each of the variables. Since Mollin is not primarily concerned with authorship attribution, she did not need to acknowledge this point. However, for the forensic context this is a salient piece of information. Without knowing what motivates one choice over another and without knowing whether one variant occurs more frequently than another, it could be potentially problematic to compare texts if one variant occurred alone in some texts whilst another variant occurred in other texts.

In conclusion, Mollin presented a different approach to establishing idiolect by comparing a corpus of Known Documents against a reference corpus. Although this has some limitations, as outlined above, the approach moves closer towards providing robust evidence of idiolect. The complication is that in forensic investigations, finding appropriate reference corpora may not always be possible and, as described above, even though the BNC is comparable in terms of the English language, the opportunity for other authors' idiolects to surface is less likely. This is a not unimportant point given that in forensic investigations, typically smaller corpora (such as those outlined by Kredens) are available so markers of authorship need to be prevalent.

Although it is widely acknowledged that more empirical research is needed to prove or disprove a theory of idiolect (Howald, 2008) there is limited investigation into whether idiolects actually exist and the research that does exist fails to demonstrate the existence of idiolect for ordinary language users across a range of different text types. None of the three studies described above have attempted to identify features of idiolect in texts that can be considered to be entirely naturally occurring. With the exception of Kredens (2001), the texts used have also been far longer than is practical in the forensic context so the results of Louwerse (2004) and Mollin (2009) are less generalizable. It is also interesting that Kredens and Mollin selected markers of authorship which were not content specific; that is, they could theoretically be applied to any texts written in any genre, unlike Louwerse's investigation which focussed only on semantic fields relevant to Modernist literary authors. It is surely not coincidental then that Kredens and Mollin gained more support for the existence of idiolect than Louwerse and this would indicate that future investigations into idiolect, and indeed authorship attribution, would do well to identify features which are more likely to be content-free rather than genre-specific.

Therefore, whilst it would be too early to disregard the concept of idiolect, it is necessary to think more carefully about whether a sound theory of idiolect is really necessary for authorship attribution. It should perhaps be assumed instead that forensic linguists can only compare features which occur at the particular moment of text creation as evidenced in the available documents. Whether this is idiolect is arguable.

2.1.5 Is a theory of idiolect really necessary for authorship attribution?

Grant (2007) suggests that with all other potential sources of variation (such as ideological variation, cultural variation, register variation, dialectal variation, and variation due to the relationship between the participants), forensic linguists may not need "the concept of an idiolect" (p. 4—5). Instead, forensic linguists need to demonstrate linguistic consistency and distinctiveness between texts (Grant, 2010: 509). The focus on style rather than idiolect is more formally known as forensic stylistics which McMenamin (2010) defines as "the analysis of linguistic variation" when "applied to items of written language in dispute" such as documents of unknown authorship (p. 492).

McMenamin (2010) describes *style* as, "in part ... the sum of the recurrent choices that become subconscious habits of choice" (p. 488). McMenamin explains that *choices* are the selection of particular forms over other possible variants and *style markers* are therefore "the observable result of the habitual and usually unconscious choices an author makes in the process of writing" (p. 488). This definition shares some features with the definitions of *idiolect* provided in Section 2.1.1 (such as choice and habit). However, it is an atheoretical definition. There is no underlying theory that presupposes each individual's language is different because they have received different input (cf. Mollin, 2009: 368); simply that authors make choices and whilst we cannot claim to understand the processes behind those choices, we can tangibly see which choices are selected when a text is created. A good marker of style will therefore be one which can be demonstrated to be idiosyncratic although not necessarily idiolectal.

Before moving on to discuss the work of Johnstone (1996) in relation to idiosyncracy, it is important at this point to acknowledge Turell's (2010) concept of 'idiolectal style'. Turell (2010) argued that in forensic linguistic approaches to authorship problems, the concept of idiolect has been used "maybe without acknowledging sufficiently that there is a theoretical controversy over the existence of idiolects in present-day linguistic discussion" (p. 217) and therefore proposes the term *idiolectal style*. This, she argues, has more relevance to forensic contexts since it can better be demonstrated in texts than the theoretical construct of idiolect. Defined as "the set of options that writers take from the linguistic repertoire available to them as users of a specific language" (p. 217), *idiolectal style* has less to do with the system of language that an individual has and is instead concerned with how the system of language which is shared amongst lots of people (e.g., a dialect, sociolect, genderlect) is used distinctively by an individual, how the individual's language production appears to be unique, and how various choices are selected from the total set of options. In this way, the definition of *idiolectal style* sits comfortably on the cline between the theoretical but controversial concept of idiolect and the less theoretical but more demonstrable practice formally known as forensic stylistics.

Johnstone (1996) argued that idiosyncracy is a cultural, psychological, and in some cases social, requirement for many speakers and set out to demonstrate that idiosyncracy is common in texts. Johnstone explained that linguistic differences between people are especially evident in narratives: "No one would suppose that two different people would ever produce identical stories in identical words" (p. 56). Indeed, this is the theory underlying plagiarism detection since we have the potential to create infinitely many novel sentences and the chances of two texts with a high proportion of shared lexical strings between them is low unless one text has been derived from the other (Coulthard, 2004, 2010; Johnson, 1997). For Johnstone, the fact that no two people produce identical narrative texts is because people are unique and that "it is precisely in narrative that people's individuality is expressed most obviously, because the purpose of narrating is precisely the creation of an autonomous, unique self in discourse" (1996: 56). The crucial point is that in creating narratives, people draw on the range of resources available to them—such as language, dialect and gender—but a narrative does not turn out in a particular way as a result of those resources. To use Johnstone's example, a narrative does not take a particular shape because the narrator is African American or female. Rather, the narrative takes the shape it does *because* the narrator has drawn on her resources as an African American female, in culmination with the other resources available to her, to create her individual voice:

> The influence of society, situation, and psychological differences is that they provide differential resources for talk. Social, psychological, and rhetorical facts are mediated by the individual, who selects and combines linguistic resources available in his or her environment to create a voice, not just a voice with which to refer to the world or relate to others but a voice with which to be a human (p. 58).

Johnstone argued that a person is recognisable to others partly as a result of being consistent in how they use language and her interest was in linguistic choices which “are in some way consistent no matter what the purpose of the talk is, who the audience is, or how planned and edited the speech or writing is” (1996: 129). To investigate this claim, Johnstone examined the language of two people who shared many of the same linguistic and cultural resources, but one had a remarkably consistent style whilst the other did not (p. 129).

The language of Barbara Jordan (b.1936—1996), a politician, public orator, media personality and professor of public policy was explored. Johnstone examined six transcripts and in order to identify the elements of Jordan’s speech that remained consistent across speech tasks, the transcripts ranged from her most formal political speeches (given in 1976) through to a relatively relaxed, spontaneous interview (conducted in 1992). In the middle of this cline occurred what Johnstone termed “edited interviews” which were interviews with Jordan that were published in magazines and therefore were subjected to editorial changes. In total, Johnstone examined 11,867 of Jordan’s words (p. 131). Johnstone also examined the language used by Sunny Nash, a degree educated African American woman, who, in addition to working as a musician, photographer, journalist, editor and TV producer, predominantly worked as a free-lance writer. As with Jordan, a range of texts were used for analysis, which for Nash included informal articles about her personal history and personal memoirs published in a newspaper in 1986 when Nash was in her thirties (Johnstone, personal communication) and 1993 when Nash was in her forties (Johnstone, personal communication). Johnstone also examined three formal historical pieces all published in 1992 and an unedited interview which was conducted in the same fashion as the Jordan unedited interview. The Nash corpus comprised 3,972 words (p. 135—8).

Consistency in linguistic style was measured as the amount of variability in the frequency of a set of linguistic features across different genres. She identified 17 features that were typical of informational and non-involved discourse (which would create an authoritative effect) including the frequency of nouns, words of four or more syllables, prepositions, attributive adjectives and a type-to-token ratio of long words. She also identified eight features that reflect personal stance such as the frequency of first-person singular pronouns, certainty verbs (e.g. *ascertain, know, prove*), emphatics (e.g. *a lot, for sure*) and predictive modals (e.g. *will, would, shall*). Using the frequency of occurrence of these features, Johnstone compared the Jordan texts with each other, the Nash texts with each other, and the group of Jordan texts with the group of Nash texts.

Johnstone found notable differences between the styles. Jordan used longer words, twice as many attributive adjectives and slightly fewer contractions than Nash (although in the edited texts, Nash used more attributive adjectives). According to Johnstone, the differences in degree of consistency were striking with some features being more consistent than others. She found that Jordan's style remained consistent across her texts and that "once she chooses the *mot juste* for a concept, she uses the same word again and again" (p. 150). She concluded that:

> As an individual decides, sometimes consciously and sometimes not, how to be, act, and sound, he or she selects from among the available linguistic resources. There are many ways to choose among and utilize the resources that are at hand. Ways of acting and talking provided by regional, ethnic, vocational, and gender models (among others) can be adopted or resisted, used predictably or creatively, as can ways of acting and talking provided by certain audiences, situations, or topics. An individual's style at any moment is the result of a complex set of calculations and choices (1996: 155).

Although Johnstone was not interested in authorial consistency from a forensic linguistics point of view, her work raises some interesting issues which are of direct relevance. Before discussing this relevance, some of the limitations must firstly be acknowledged.

Johnstone found that the features she investigated showed Jordan to have a more consistent style, in most cases, than that of Nash. This finding, though perfectly valid for Johnstone's purposes, is open to criticism on the grounds that her two corpora are not directly comparable. The Jordan corpus contained 11,867 words whilst the Nash corpus contained only 3,972 words. Given the difference in size between the two corpora, it could be argued that Nash's style was not given sufficient opportunity for consistency to be established. Johnstone also acknowledges that her research design would have been more robust if she had compared the same genres between Jordan and Nash (in other words, if published interviews and prepared speeches existed for Nash). However, she argued that this was not a problem for her because she was interested in how the two authors chose to express themselves and through which media (1996: 135—6). Whilst this indeed may be the case, a far stronger argument about style consistencies and inconsistencies could have been constructed if the genres and texts were more comparable. After all, Johnstone's selection of Jordan and Nash was motivated by the fact that they shared many of the same cultural and linguistic resources. If she had compared their language choices across similar texts in the way that Kredens

(2001) did, Johnstone would have been better positioned to argue that Jordan had a very striking style whilst Nash, having many of the same choices to make as Jordan, did not.

Johnstone's choice of linguistic variables is also interesting with her selecting a variety of objective measures such as the frequency of nouns, words of four or more syllables, type/token ratio of long words etc. It is questionable whether these measures are appropriate for capturing elements of style. Certainly the frequency of nouns and 1st person singular pronouns are frequently utilised as part of the Linguistic Inquiry and Word Count text analysis software (Pennebaker & Lay, 2002) which has enjoyed variable success at, for example, differentiating deceptive from truthful styles (Bond & Lee, 2005; Hancock, Curry, Goorha, & Woodworth, 2004). Other measures selected by Johnstone carry less validity as a marker of style, such as words of four or more syllables, which is reminiscent of the much maligned Cusum technique (Farringdon, 1996). Presumably, Johnstone reasoned that a measure such as this would be linked to style on the basis that one author may use longer, more morphologically complex words than another, but this assumption requires far more investigation (although see Grant (2004) who found that whilst words containing six letters or less are generally insignificant, words over seven letters in length do seem to vary between authors).

What is particularly interesting about Johnstone's investigation is the time period over which the texts were composed. It is debateable in the field of forensic linguistics whether texts composed at different times are comparable (as indicated by the discussion in Section 2.1.1). By using texts produced from 1976 to 1992 and still finding consistency over this 16 year period, Johnstone demonstrated that perhaps this may be less of a concern than has previously been assumed. Similarly, in her investigation of Tony Blair's idiolect, Mollin (2009) used data spanning a period of 19 years (1988–2007) which may suggest that Blair's use of maximizer collocations remained constant over this period, although Mollin does acknowledge that 91% of the corpus consisted of texts produced during the last ten years (p. 372); this is nonetheless still a substantial period of time. Alternatively, it is also possible that Jordan and Blair are an exception rather than the rule, since the texts produced by Nash ranged from 1986 to 1993 and were not found to be consistent over this shorter period.

A further question that arises, then, is whether style fossilizes and if so, at what age? It would be unreasonable to read too much into Johnstone's results since this was not her concern, but certainly there may be something interesting in the fact that Jordan's style remained more similar during her forties and mid-fifties than did Nash's, whose texts were produced during her mid-thirties and forties. This may indicate that perhaps style does fossilise, or at least the rate of change decreases, with the onset of middle age. Whilst the comparison between authors may not be conclusive, the findings for Jordan, or rather, the specific finding that intra-author style remains consistent in spite of genre, length of text and date of composition supports the notion that authors' styles are consistent on some level (in this case lexical and syntactic) and that, therefore, texts spanning different periods may still be comparable for authorship purposes.

For Johnstone, then, *choice* is crucial, rather than *habit* and her argument is that language output is shaped by an individual's resources. It is interesting that even if the focus shifts away from idiolect towards idiosyncrasies, many of the same problems still exist. The key conclusion is that an author's style is the "result of a complex set of calculations and choices" (1996: 155) and so even if the notion of style consistency is adopted in lieu of idiolect, there is still little evidence that the same choices will be made in different texts, or similar texts produced under different circumstances. As mentioned previously, a marker of authorship which occurs at the subconscious level may therefore be incredibly useful. It should also be noted that Johnstone deliberately identified an author (Jordan) who had a noticeable style, and given the previous consideration of ordinary language users, it is unclear whether consistency could be found for other users; it certainly was not for Nash. As a result of these limitations, the generalizability of these results to the forensic context is diminished.

To conclude this section, the work of Grant (2010) must be considered. For Grant, the central question to be investigated was whether authorship attribution actually requires a strong theory of idiolect or whether the "detection of degrees of consistency and the determination of degrees of distinctiveness" without a strong theory is still valid (p. 509). This is in light of the fact that some types of text, particularly SMS text messages which are extremely short, may be "too short to allow the possibility of idiolectal analysis" (p. 509). In other words, a theory of idiolect is useful conceptually, but when it comes to the task of comparing texts, the forensic linguist can only compare consistent patterns and determine distinctiveness and therefore "[p]ractical authorship analysis may depend less on a strong theory of idiolect than on the simple detection of consistency and the determination of distinctiveness" (p. 509).

Through a re-analysis of previous forensic casework which involved text messages, Grant developed a quantitative method based on features in the texts (such as abbreviations, spacing, lexical choices etc.). He argued that it was possible to conduct authorship attribution based on stylistic consistency and distinctiveness but cautioned that his results do not demonstrate that "individuals are absolutely consistent" or "that every author will be consistent in the same way" (p. 521). Therefore, Grant argued that any authorship attribution work which identifies consistency and distinctiveness without explaining what accounts for these features, is not sufficient for providing a theory of idiolect—a theory needs to explain the findings:

> Theories have to have explanatory power. Any investigation limiting itself to observation and description of consistency and distinctiveness in authorship style might fairly be considered idiolect free authorship analysis. (p. 521).

The research described in Sections 2.1.2—2.1.5 can therefore be assessed in this way. Johnstone (1996) and Kredens (2001) identified patterns of consistency and distinctiveness between data produced by different speakers and, crucially, attempted to explain what accounted for such differences (the combination of various social and biological resources). In contrast, Mollin (2009) and Louwerse (2004) also identified patterns of consistency and distinctiveness, yet were unable to offer any explanation for what might cause such patterns, preferring instead to simply label the patterns as "idiolectal". Therefore, whilst their findings were clearly argued to be evidence of idiolect, for Grant, the latter two might be termed "idiolect free authorship analysis". If all that has been identified are consistent and distinctive patterns, with no explanation, there is no basis for claiming that aspects of idiolect have been demonstrated. Grant concludes:

> To the extent that it can be shown that one individual's language is measurably unique in the population of all language users, this is, or would be, an astounding fact. Even less extreme individual linguistic distinctiveness demands a combination of cognitive and social investigation and demands a combination of cognitive and social explanations. Observable individual linguistic uniqueness demands a theory of idiolect. (Grant, 2010: 522).

In this way, Grant argued for a re-think about the conceptualisation and practice of authorship attribution and that markers of authorship which are grounded in sociolinguistics and cognitive or psycholinguistics will have the upper hand in contributing to a theory of idiolect.

At this juncture, whilst a review of idiolect and related concepts has been presented, it is now necessary to be explicit about how idiolect will be understood and applied in this research. From reviewing the definitions of idiolect provided by Hockett (1958), Louwerse (2004) and Coulthard (2004), the consensus seems to be that either choice, habit, or both are intrinsically linked to idiolect. The definition to be used in this research is as follows:

> Determined and conditioned by a wide and immeasurable range of biological, sociological, cognitive and environmental factors (including *inter alia* age, IQ, occupation, friendship networks, language contact), idiolect is the combination of language choices (planned features) and habits (subconscious features) made by an individual, the sum of which creates a distinctive, albeit oftentimes overlapping, range of choices and habits from another individual.

In this conceptualisation of idiolect, the goal of the forensic linguist is to identify those features of idiolect which overlap less with others in order to demonstrate the similarity or difference between a series of authors. It will not be possible to determine in this research what constitutes a choice and what constitutes a habit, but one might surmise that a feature such as using a specific lexical item to mark identity (such as youth vernacular words commonly found in the school playground) may be a choice whereas final /g/ clipping in the case of spoken language may be a habit. Furthermore, choices and habits should be viewed as being on the same cline. A feature may start out as being a conscious choice but over time moves into being a habit. An example may be when a person moves to a new geographical region and starts to use a dialectal term of endearment in order to fit in but over time more naturally and automatically uses that term. Any reference to idiolect in the remainder of this research should be understood against this definition.

At the outset of this chapter, it was stated that there are two assumptions underlying authorship attribution. The first was that each author has an idiolect, and the second was that this idiolect remains constant across texts. Based on the limited empirical research into idiolect there is some evidence that specific linguistic features may be idiolectal for a very small set of authors, sometimes only one, on a narrow range of texts. Naturally, without clearer evidence for the first assumption, it is impossible to engage with the second. However, as an alternative to the theory of idiolect, idiosyncracies and markers of style may be sufficient for authorship attribution work and there is some evidence that they remain constant across genre, length of text and date of composition. Next to consider are the methods available to the forensic linguist for establishing the authorship of texts.

2.2 How robust are existing approaches to forensic authorship attribution?

Current methods for establishing the author of a text can generally be subsumed under quantitative and qualitative. McMenamin (2002) explains the distinction:

> The work is qualitative when features of writing are identified and then described as being characteristic of an author. The work is quantitative when certain indicators are identified and then measured in some way e.g. their relative frequency of occurrence in a given set of writings (p. 76).

2.2.1 Qualitative approaches to forensic authorship attribution

In qualitative approaches, the forensic linguist typically engages in a close-reading of all Known Documents and Questioned Documents to determine which features, based on expertise, seem most characteristic of authorial style so that a comparison of styles can be made (e.g. Kredens, 2001). This type of analysis occurs at, but is not limited to, the lexical level, for example, shared lexical items between texts (Fitzgerald, 2004), shared lexical fields (D. Foster, 2001) and misspellings (and conversely, correct spellings) (Eagleson, 1994; Shuy, 2001). Analysis can also occur at the level of syntax and discourse analysis (McMenamin, 1993) and formatting preferences (McMenamin, 2002;

2010). According to McMenamin (2002), qualitative analysis will not result in absolute conclusions but it does enable forensic linguists to say plausible things about authorship based on the discovery, description and categorisation of linguistic elements. This can be a tricky distinction to draw between qualitative and quantitative analysis, since quantitative analysis does not necessarily result in absolute conclusions either.

Qualitative analysis allows the linguist to become acquainted with the texts through the close-reading process. Idiosyncrasies often present themselves at this stage and the analysis often relies on impressionistic and intuitive judgements about the data which are applied on a case-by-case basis. For example, in Fitzgerald's (2004) analysis of the Unabomb Manifesto (as described in Chapter 1), the use of the phrase "you can't eat your cake and have it" was identified as occurring in the known writings of Theodore Kaczynski. Quantitative analysis, where the forensic linguist typically explores the data generated by their original search parameters, would not necessarily be sensitive to such idiosyncratic uses of language. On the other hand, it can be very easy for a forensic linguist to notice a feature which seems marked, and afford more significance to its occurrence than it deserves (Mollin, 2009; Solan & Tiersma, 2005: 155—6).

McMenamin (2002) explains that qualitative evidence is more demonstrable in courtrooms than quantitative evidence because it is the actual language data that is presented, rather than statistics. Qualitative results "appeal to the nonmathematical but structured sense of probability held by judges and juries" (p. 129) and is often more accessible to the jury since jurors can easily be confused by quantitative evidence expressed as statistical values (Coulthard, 2010). Working exclusively in either a quantitative or qualitative approach can also create problems in the courtroom with respect to comparing like with like. If, for example, the defence's expert uses quantitative methods whilst the prosecution's expert uses qualitative methods, it can be very difficult for the judge and jury to compare the evidence. Whilst it is most likely that judges and juries will better comprehend qualitative results, it is becoming harder for evidence based solely on qualitative analysis to be admitted into courtrooms as evidence (cf. Section 2.3, p. 39).

2.2.2 Quantitative approaches to forensic authorship attribution

Quantitative approaches to literary authorship attribution have a long history (e.g. Mosteller & Wallace, 1963) and quantitative methods in the field of forensic authorship attribution are gaining in importance. One motivation behind this is that quantification, when conducted properly, is more scientific since analyses are typically objective and replicable. In quantitative methods of authorship attribution, the forensic linguist uses a pre-defined selection of authorship markers, identifies their frequencies across all known and questioned documents and then, using statistical analyses, establishes the significance of similarities and differences and consequently, the most likely (or unlikely) author of the Questioned Document. Such markers might include average word and sentence length (Farringdon, 1996; Mannion & Dixon, 2004), frequency of function words (Bagavandas & Manimannon, 2008) syntactic complexity (Chaski, 2001; Grieve, 2007), frequency of n-grams and bi-grams (Bel, Queralt Estevez, Spassova, & Turell, 2012; Clement & Sharp, 2003; Feiguina & Hirst, 2007) and vocabulary richness (Baker, 1988; Chaski, 2001; Grieve, 2007; Holmes & Forsyth, 1995; Mosteller & Wallace, 1963).

The natural corollary of quantitative analysis is an automated method relying on minimal human introspection. Research which explores automated methods is largely conducted on literary texts rather than forensic materials (e.g. Clement & Sharp, 2003; Hoover, 2001, 2004a, 2004b; Juola, Sofko, & Brennan, 2006) although work with non-literary texts does exist (e.g. Burrows, 2002; de Vel, Anderson, Corney, & Mohay, 2001). Whilst quantitative analysis reduces potential subjectivity from the forensic linguist, the analytic methods as opposed to the statistical presentation of the results often come in for criticism. Chaski's (2001) work was severely criticised by Grant & Baker (2001) based on the reliability and validity of the authorship markers investigated whilst the Cusum Technique has been largely discredited, because, amongst other reasons, the analysis was not replicable (Grant, 1992; Hardcastle, 1997; Sanford, Aked, Moxey, & Mullin, 1994; Smith, 1994). It seems to be a characteristic of quantitative analysis that it should be subjected to more criticism and peer-review than qualitative analyses seem to attract. It is this kind of peer criticism that strengthens the field of forensic authorship attribution when seeking to admit evidence to the courts under the *Daubert* criteria (cf. Section 2.3, p. 39).

Quantitative analysis reduces the likelihood of variation in results, provided that the same features are examined in the same texts each time the analysis is repeated. However, this is not always the case. Linguists are typically trained in the humanities rather than the sciences, and so it is possible that scientific principles may be misunderstood and/or misapplied, particularly issues regarding sampling, reliability and validity. Grant (2004; 2007) is a pioneer of highlighting the need for linguists to exercise caution in this regard. A striking example of the problem comes from Olsson (2004). In his introductory textbook to the field of forensic linguistics, he explains to students:

> In the course of the book I have shown the importance of acquiring a basic understanding of statistics. This is really important. However, it is not necessary to become an expert in that field. In fact that could be a dangerous policy. By setting yourself up against professional statisticians, you simply prepare yourself for a fall, unless you are an exceptionally gifted mathematician. Either become a linguist or a statistician—the person who can do both is very rare. In any case, it is simply not necessary to become a statistician. Although some courts might penalize you because you used a different formula from some other expert, as

> long as you can give reasonable grounds for your methods, and as long as your methods are as scientific as required, that is all that really matters. In any case, if you are working on a very important case in which statistics abound—just consult with a professional. Do not try to do it yourself. It is not worth it (p. 198).

It is somewhat disquieting on one hand to see students encouraged to have a basic understanding of statistics, but on the other hand be told not to bother learning *too* much because it is both impossible and "dangerous". By suggesting that a linguist will be fine in the courtroom as long as methods are as scientific as required oversimplifies the situation and implies that it is the act of using statistics which makes the analysis scientific—exactly the kind of problem that Grant (2007) cautions against. Furthermore, Olsson suggests that when working on very important cases (all cases in the forensic context where personal liberty is at stake are important), a statistician should be consulted. It is precisely these cases where the linguist should know exactly what statistics to use so that when facing cross-examination, they are able to stand by their analyses rather than deferring to the person whom they consulted. It is not clear what Olsson means by saying "statistics abound". If it is that the statistics are in some way in control, it is questionable whether that is a desirable state of affairs. The linguist should choose to use statistical tests where appropriate and applicable, and certainly within the limits of their own competencies. The linguist can choose to adopt qualitative methods, in which case statistics will not abound. Grant (2007) shares concerns about Olsson's approach to statistics. Commenting on Olsson's (2004) book, Grant explains:

> [M]ethodological design and statistical analysis appear to be something added at the end of an analysis (just checking for statistical differences between individuals). Such an approach leads to weak statistical analysis, doubtful conclusions, and a lack of apparent seriousness in attempts at quantification (2007: 3).

This discussion so far implies two distinct approaches and it is true that some linguists remain faithful to a particular analytical paradigm (Chaski, 2001). However, others combine the use of both (McMenamin, 2002; 2010). In fact, Coulthard and Johnson (2007) comment that "[i]t is not unusual for the expert to use more than one approach" (p. 173) and Solan and Tiersma (2004, 2005) advocate the use of "eclectic approaches" which combine the strengths of both qualitative and quantitative methods and certainly a mixed methodology enables the triangulation of results.

2.3 How is forensic authorship attribution evidence received by the courts?

As discussed previously, the lack of a sound theory of idiolect underlying forensic authorship attribution methods does not prevent the work being carried out. However, some linguists are in fact beginning to question whether the field of forensic authorship attribution is sufficiently developed to be used as evidence in court (Kniffka, 2007; Shuy, 2006) and so it is necessary to briefly consider rules for the admissibility of evidence.

The admissibility of forensic authorship attribution evidence has been described in detail for only two countries (several legal jurisdictions can exist in a country, so 'country' is here being used as a vague umbrella term to contrast the general rules for admissibility): The United States of America (Howald, 2008; McMenamin, 2004; Solan & Tiersma, 2004, 2005; Tiersma & Solan, 2002) and Germany (Kniffka, 2007), although some discussion of general linguistic evidence in the UK context is provided by Coulthard (2005a) and in the Polish context by Kredens (2006). Challenges to admissibility seem to be similar across the UK, USA and Germany although it should be acknowledged that linguistics experts are increasingly gaining acceptance in courts (Solan, 2010) and in fact it is generally accepted that the UK will most likely adopt a version of the American *Daubert* criteria (Coulthard, 2004; Grant, 2007). The ruling in *Daubert* states that an expert must be able to provide answers to the following four criteria when describing the methods used to arrive at their expert opinions:

1) Whether the theory offered has been tested;
2) Whether it has been subjected to peer review and publication;
3) The known rate of error; and
4) Whether the theory is generally accepted in the scientific community.

Tiersma and Solan (2002) argue that in theory at least, linguistic evidence should fare well because linguistics "is a robust field that relies heavily on peer-reviewed journals for dissemination of new work" (p. 225). They provide examples of areas of linguistics which seem to be routinely admitted to courts, including dialectology, comprehensibility and readability, linguistic proficiency, and linguistic issues in trademark cases. They point out that some areas are more problematic, including phonetics and speaker identification, discourse analysis, the meaning of words and phrases in such legal texts as contracts and statutes, the comprehensibility of parts of jury instructions, and, of utmost importance to the current discussion, disputed authorship.

These are 'problematic areas', they suggest, because they cannot withstand the rigour of the *Daubert* criteria (although see McMenamin (2004) for an objection to this claim). However, Coulthard (2004) argued that the methods of forensic authorship analysis do in fact meet the *Daubert* criteria (p. 444). For Coulthard, the occurrence of shared lexical items between texts is conclusive evidence that they have not been independently created. He therefore reasoned that the theory has been tested (criterion 1). He acknowledged that further work is required to determine how many shared items are necessary to support a decision (p. 444). However, this assertion is limited only to the area of plagiarism detection and does not extend to cases where the author of a Questioned Document needs to be identified.

Tiersma and Solan (2002) argued, and Coulthard (2004) independently asserted, that the second criterion can easily be met because many publications in linguistics are subjected to peer review. The third of the *Daubert* criteria (the known rate of error) is more problematic. Rather than arguing that there is a known rate of error, Coulthard challenged the academic community to test the rate of error (p. 445). On assessing whether the theory is generally accepted in the scientific community (criterion 4), Coulthard claimed that the general theory of idiolectal variation is "generally accepted across the whole linguistic community" although the discussion in Section 2.1 shows this to be not quite the case. Whilst there may be general acceptance of the theory of idiolect there is considerable variation over definitions and how it can and should be identified. For the present at least, it should be assumed that the courts will continue to have grounds to refuse linguistic testimony until an established set of markers of authorship have been validated by peer-review in the linguistics community and until a known rate of error for all of the analytical approaches can be calculated.

2.4 Summary

This chapter has raised several key questions about the practice of forensic authorship attribution. Firstly, because of the lack of empirical evidence, differences between texts cannot simply be claimed to be due to idiolect—although it may be the case that differences *are* due to idiolectal variation, there is a need for considerably more investigation into the ways in which idiolect may manifest and the factors which invoke an idiolectal choice on one occasion compared to another. Instead, it may be safer to talk about authorial habits and choices which remain constant, although there is still a need to explain findings in relation to idiolect to avoid the danger of "idiolect free" authorship attribution (Grant, 2010). Therefore, methods of authorship attribution need to demonstrate authorial habits and choices across texts and in light of the evidence reviewed in this chapter, data should be non-literary.

One potential area for authorial differences may be the differential retrieval of words from the mental lexicon: "the empirical psychological studies suggest that we create associations and so fall into linguistic habits" (Grant, 2007: 5). Lexis has been well explored as a marker of authorship (e.g. Chaski, 2001; Coulthard, 1994; 2004; Fitzgerald, 2004; D. Foster, 2001; Hoover, 2002, 2003a; Johnson, 1997; Kredens, 2001) and Solan and Tiersma (2005) advocated lexis as a fruitful area of research in the authorship context (p. 173). Therefore, a method for identifying authors based on the subconscious, habitual use of lexis should provide a good opportunity to capture authorial consistency and distinctiveness, if not idiolect. Empirical investigations might therefore consider focussing on those areas of lexis which occur frequently in texts so that regardless of text size, there may potentially always be something to analyse. Formulaic language, “[w]ords and word strings which appear to be processed without recourse to their lowest level of composition” (Wray, 2002: 4), is believed to be abundant in discourse (Schmitt & Carter, 2004) and seems to fit all of these criteria as a potential marker of authorship. Before it can be demonstrated how the theory of formulaic language might hold the key to discriminating between texts produced by different authors, an account of formulaic language is required. Such an account is provided in the next chapter.

Chapter 3

'Breaking new ground': formulaic language as a marker of authorship

In this chapter, the research surrounding formulaic language is discussed before the potential relationship between formulaic language and authorship attribution is considered. In Chapter 4, the data that will be used for the empirical work is described. However, at various points throughout this chapter, illustrative examples of naturally occurring data will be drawn from the corpus, referred to here as the author corpus.

3.1 What is formulaic language?

Language enables us to express our ideas in many different ways and the opportunity for novelty is vast:

> There is no doubt that essentially all speakers of a language are free to produce sentences they have never heard or produced before. Very few people, on seeing two blue rabbits in a fish-bowl, are going to be poorly equipped, linguistically, to express their experience, even though the sentence they would need to create for the task would undoubtedly be completely novel to them (Fillmore, 1979: 95).

Speakers are free to choose which lexical items to use and how to arrange them (Corrigan, Moravcsik, Ouali, & Wheatley, 2009: xi). However, whilst the potential for novel utterances is limitless, speakers appear "to renounce the great freedom that the language offers" (Corrigan, Moravcsik, Ouali, & Wheatley, 2009: xiii). Nattinger and DeCarrico (1992) suggest that "just as we are creatures of habit in other aspects of our behaviour, so apparently are we in the ways we come to use language" (p. 1).

Evidence from psycholinguistics (e.g. Hoey, 2005; Wray, 2002), sociolinguistics (e.g. Coulmas, 1979), corpus linguistics (e.g. Moon, 1997, 1998a, 1998b) and both L1 and L2 language acquisition (Pawley & Syder, 1983; Peters, 1977, 1983, 2009; Vihman, 1982) shows that when communicating, we often rely on patterns in language and have "preferred formulations" for expressing ideas (Wray, 2006: 591). This results from the fact that much of our everyday activity is routine: "As similar speech situations recur, speakers make use of similar and sometimes identical expressions, which have proved to be functionally appropriate" (Coulmas, 1981: 2). In fact, mastering the balance between novel language and routine language is a key characteristic for sounding like a competent, fluent and native speaker (Coulmas, 1981; Ellis, 1996; Fillmore, 1979; Howarth, 1998; Pawley & Syder, 1983).

Such routine language can in a global sense be termed *formulaic* which was defined in the prelude to this thesis as being "[w]ords and word strings which appear to be processed without recourse to their lowest level of composition" (Wray, 2002: 4). Wray provides the example of the breakfast cereal *Rice Krispies*. During an advertising campaign for television, people were asked what they thought the product was made of and were surprised to learn that it was rice. According to Wray, people had "internalized this household brand name without ever analyzing it into its component parts" (2002: 3). For these people, *Rice Krispies* appeared to be stored and produced as a single lexical item, rather than two separate items. The fact that multi-word sequences may be stored as single lexical items is an important feature of formulaic language (Bannard & Lieven, 2009; Ellis, 1996; Erman, 2007; Erman & Warren, 2000; Pawley & Syder, 1983; Wray, 2000, 2002, 2008).

Given that *formulaic language* is an umbrella term, a survey of the literature soon reveals that many other terms exist to describe different aspects of formulaic language. These include Collocations (Gledhill, 2000; Herbst, 1996; Stubbs, 1995), Idioms (Grant & Bauer, 2004; Simpson & Mendis, 2003), Fixed Expressions including Idioms (e.g. Moon, 1998a) Formulaic Sequences (e.g. Schmitt & Carter, 2004; Wray, 2002), Multi-word Items (e.g. Moon, 1997), Phrasal Lexemes (e.g. Moon, 1998b), Recurrent phrases (Stubbs & Barth, 2003) and Situation Bound Utterances (e.g. Kecskés, 2000), to name just a few. In fact, Wray (2002: 9) found 57 different terms each describing what can be thought of as formulaic. These terms, though related, denote slightly different characteristics associated with formulaic language. Some definitions emphasise the importance of context and register (Cortes, 2004; Kecskés, 2000) whilst others focus on the physical distance between words (Hoover, 2003a) or whether sequences of words are contiguous (Hoover, 2002; Stubbs, 2002; Stubbs & Barth, 2003). Whilst definitions vary, the underlying principle is that some word combinations are not created through independent selections.

Estimates vary regarding how much of the language produced is formulaic. Erman and Warren (2000) claim that 55% of spoken and written language may consist of prefabs ("a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization" (p. 31)) which includes such examples as *out of date, at the time, in the end, here and there, a waste of time, for some reason* and *all over the place*. Chenoweth (1995) found 77% of each answer to essay style exam questions, regardless of length, consisted of formulaic expressions (p. 292) where formulaicity was identified according to frequency and intuition. Pawley and Syder (1983) argue that "the largest part of the English speaker's lexicon consists of complex lexical items including several hundred thousand lexicalized sentence stems" (p. 215) which they define as "a unit of clause length or longer whose grammatical form and lexical content is wholly or largely fixed" (p. 191). Examples provided by Pawley and Syder (1983) include *it's on the tip of my tongue, some people are hard to please, call me after work, would you like some more?* and *speak for yourself*, again to provide only a representative few and which, notably, are longer than the examples provided by Erman and Warren. A lack of consensus over the exact proportion of formulaic language compared to novel language in everyday usage results from differences in definitions, methods of identification and contexts of use. However the overriding claim is that formulaic language is ubiquitous and prevalent in language (Wray, 2002).

3.2 How robust is the theory of formulaic language?

The following discussion explores the main theories for how formulaic language might be stored and processed as holistic, single lexical items compared to other sequences of words which appear to be novel (i.e. constructed from individual constituent parts). An understanding of these processes will contribute to the discussion in Chapter 2—that language use may be more about habit than choice—and will help to contextualise the argument presented in this chapter that formulaic language holds the potential to be a marker of authorship since authors should be less aware of their use of holistically produced sequences.

3.2.1 How is formulaic language processed and stored?

Sinclair (1991) argued that no single model of language processing has been able to account for how meaning arises in language in a satisfactory way (p. 109). He therefore proposed the ‘open choice principle’ and the ‘idiom principle’ to describe our dual ability to create and understand novel expressions and preferred formulations. The open choice principle is a way of seeing texts as being produced by the result of a very large number of complex choices made by the language user at different positions in a clause. Sinclair suggests that the open-choice principle is the normal way of seeing and describing language: texts are seen as a series of slots which have to be filled from the lexicon (p. 109). However, according to Sinclair, words do not occur randomly in texts and the open-choice principle alone is not sufficient to explain restrictions on the choices that the language user may like to make. Therefore, the second processing principle, the idiom principle is required: “The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments” (p. 110). Sinclair suggests that most “normal” text (as opposed to, for example, legal documents which are highly constrained and poems which may be “crafted” rather than written, cf. Grant (2008) discussed in Section 2.1.2, p. 19) is formed by using the idiom principle with only occasional “switching” to the open-choice principle (p. 113). This “switching” is necessary because the two principles are entirely incompatible (p. 114). He concludes by arguing that a “slot-and-filler” model of language which relies on the grammar to provide options for lexical choice such as the open choice principle, is secondary to the idiom principle: “The open-choice analysis could be imagined as an analytical process which goes on in principle all the time, but whose results are only intermittently called for" (p. 114). To demonstrate the open choice and idiom principles in practice, the following extract is taken from the author corpus with underlined text highlighting "single choices" (idiom principle) and the remaining text arising from the open-choice principle. It should be pointed out that this analysis is speculative, since it is impossible to actually know how these items were retrieved from this author's lexicon.

1 I think / the worst moment of my life / was when / a family member / died. / It was a shock /
2 when it happened / as / she / hadn't been / ill. / I remember / the whole thing / very vividly /
3 even though / it happened / quite a few / years ago / now. / It was / early on a / **Sunday** /
4 morning, / about / **7** / o' clock / I think, / and / the / phone was ringing. / I knew / that / was
5 a bad sign / from the start / because / no one / rings / that early on a / Sunday. / I heard / my
6 / Dad / get up / and / answer it / and / my heart was beating in my chest / because / I was so
7 / worried about / what had happened. / I / couldn't / really hear / what was said / on the
8 phone / but / I knew / it was something bad. / I then / heard / my / Mum / get up / and / go
9 downstairs / to find out / what was / going on (Rose-2).

In this short paragraph, it can be seen, as Sinclair argues, that the majority of the text is made up of semi-preconstructed phrases (i.e. idiom principle) with those items classed as 'open choice' indicating only specific details that are relevant to the story including people (e.g. *Dad, Mum*). In line 5, the author could have selected the word *as* rather than *because* (as she did in line 2) and *calls* instead of *rings,* so I have suggested that these are open choices (although these supposed open choices may have been *primed* for this author, see below). However, other examples such as *a family member* (line 1) do not permit variability (e.g. **a family person*) and so have been labelled as being generated through the idiom principle. In lines 3 and 4 are two examples of semi-fixed frames: *early on a Sunday morning* and *about 7 o' clock*, where *Sunday* and *7* (indicated in bold) can be slotted into these frames depending on the specific details of the story.

In Sinclair (2004) the concept of the open-choice and idiom principles is further developed. Sinclair explains that the open-choice principle leans towards a "terminological tendency":

> [T]he tendency for a word to have a fixed meaning in reference to the world, so that anyone wanting to name its referent would have little option but to use it, especially if the relationship works in both directions (p. 29)

The idiom principle, on the other hand, has a tendency towards the phraseological, meaning that "words tend to go together and make meanings by their combinations" (p. 29). In this way, Sinclair shifts the focus away from the choices that the language user is required to make and focuses more specifically on how meaning arises from single words or phrases. This is to the extent that Sinclair (2004) hypothesizes: "[T]he notion of a linguistic item can be extended, at least for English, so that units of meaning are expected to be largely phrasal" (p. 29—30). The effect of this is that the "idea of a word carrying meaning on its own would be relegated to the margins of linguistic interest" (p. 30).

Whilst this line of reasoning is in keeping with Sinclair's (1991) notion of the idiom principle, there is a slight discrepancy with the open-choice principle. Sinclair (2004) explains that the association between a word and its meaning is so tightly bound that anyone wanting to make reference to a particular object would have very little choice over its use. This seems to be at odds with a principle centred on a language user having a "very large number of complex choices" to make (1991: 109).

Of relevance to the discussion of *choice* is Hoey's (2005) theory of *lexical priming* whereby he argues that collocations are "a psychological association between words (rather than lemmas) up to four words apart … [which] is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution" (p. 5). He refers to collocation as a psycholinguistic phenomenon and argues that "every word is mentally *primed* for collocational use" (p. 8, original emphasis). In other words, as words are learned through multiple encounters, they become "cumulatively loaded with the contexts and co-texts" in which they are encountered (p. 8). Hoey does not restrict his theory to that of individual words and claims that word sequences too can be strongly loaded with contextual information, a property which he calls *nesting* (p. 8). This, he argues, leads to the creation of lexical items and lexical bundles; both of which are considered to be aspects of formulaic language. Hoey provides the example of the word *word* which collocates with *say*: "*say a word* in turn collocates with *against*, and *say a word against* collocates with *won't*" (p. 11).

Lexical priming creates something of a problem in relation to authors having a *choice* over the words and word sequences they use. Clearly, if priming associations are strong, then any word sequences used by an author cannot be considered to have been a *choice*. In this way, lexical priming relates to habits surrounding language use. This may create problems for a theory of idiolect based on words and/or word sequences since they cannot be claimed to be idiolectal if everybody's primings are the same. However, Hoey explains that "[w]ords are never primed *per se*; they are only primed for someone" (p. 15) and since all individuals have a different collection of life experiences, the primings themselves may be indicative of idiolect:

> [E]verybody's language is unique, because all our lexical items are inevitably primed differently as a result of different encounters, spoken and written. We have different parents and different friends, live in different places, read different books, get into different arguments and have different colleagues, and therefore there is next to nothing that is shared in the data on the basis of which words get primed for us. (p. 181)

In this way, then, lexical priming offers a satisfactory explanation for word sequences which appear to be idiolectal. If an author selects words and word sequences on the basis of habit, those selections are likely to be outside the realm of conscious thought, and, as identified in Section 2.1.1 (p. 15), such markers are likely to be useful for the forensic authorship attribution context.

Unlike the dual-processing system proposed by Sinclair (1991), where although one processing route may dominate either can be activated, Swinney and Cutler (1979), who focus on reception rather than production, describe two opposing processing models for one particular aspect of formulaic language, idioms. An idiom in this research is to be understood as “a string of two or more words for which meaning is not derived from the meanings of the individual words comprising the string” such as *kick the bucket* and *by and large* which is in line with Sinclair’s definition of idioms—that meaning arises from the combinations of words (Sinclair, 2004). The first is the Idiom List Hypothesis. Under this model idioms are not part of the normal lexicon and are instead stored on, and accessed from, a separate list. Swinney and Cutler assert that a literal analysis of a wordstring occurs first and is only then followed by an idiom mode of processing which permits access to the special list of idioms. The second idiom processing model is the Lexical Representation Hypothesis. This hypothesis sees idioms stored in, and retrieved from, the mental lexicon in the same way as other words; therefore there is no requirement for a specialised processing route to a particular list of idioms. Instead, literal and idiomatic analyses occur concurrently as soon as the first word in an idiom string is encountered: “Thus individual words are accessed from the lexicon and structural analysis is undertaken on these words at the same time that the lexical access of the entire string (which is merely a long word) is taking place” (Swinney & Cutler, 1979: 525).

Swinney and Cutler argued that the amount of time it takes for participants to judge the acceptability of idiomatic and non-idiomatic phrases can be used to provide support for either the Idiom List Hypothesis or the Lexical Representation Hypothesis. They reasoned that if the Idiom List Hypothesis is correct, then it should take participants longer (or at least the same time) to judge an idiom to be a meaningful and natural phrase than a literal wordstring since “idiomatic meanings are computed by reference to a special idiom list, via some special mode of processing which is instigated following an attempt at literal computation” (p. 526). Alternatively, if participants judge idioms to be meaningful and natural phrases faster than literal counterparts then evidence will exist in support of the Lexical Representation Hypothesis since “[t]he access of the lexical interpretation should conclude far more quickly than the access and computation of the relationships among the several lexical items in the literal interpretation of the idiom” (p. 526).

Swinney and Cutler therefore carried out empirical research to test this. Participants were required to read a series of word strings as units and make decisions about whether or not they formed meaningful and natural phrases. Crucially, rather than just looking at how many strings the participants accurately identified, Swinney and Cutler used response latency as an indicator of which processing model was the more likely to be in operation.

Twenty participants were randomly presented with 152 word strings comprising 23 grammatical idioms (e.g. *break the ice, pain in the neck, out of line*), 23 grammatical control strings (e.g. *break the cup, pain in the foot, out of food*) and 30 non-idiomatic but acceptable strings which were all balanced with 76 ungrammatical phrases (e.g. *stranger is during, destroy be however*). Swinney and Cutler found that grammatical idioms were judged to be acceptable faster than the matched controls which provides support for the Lexical Representation Hypothesis:

> [A]s recovery of any acceptable meaning was sufficient for a positive classification response, and as the access of any single lexical item (the lexicalised idiom) can undoubtedly be accomplished more quickly than the access and computation of the relationships among the several words in a (control) phrase, the results support a model in which idioms are stored and accessed as lexical items (1979: 528).

It is important to remember that Swinney and Cutler dealt only with idioms in this research. The defining characteristic of an idiom is that the meaning cannot be literally derived from its constituent parts and it was the time it takes to reach either a literal or idiomatic meaning that Swinney and Cutler used as their variable. Whilst the empirical evidence presented does lend support for the Lexical Representation Hypothesis, it cannot be inferred that all formulaic language is processed in this way, since, although some formulaic language is understood idiomatically, not all of it is (such as common collocations, which Hoey (2005) would argue are primed). For this reason, it seems that a dual-systems approach to processing which allows switching between analytic and holistic processing mechanisms offers the most comprehensive account for how formulaic language can be produced alongside novel language.

Erman and Warren (2000), without explicitly saying so, seem to favour the Idiom List Hypothesis by arguing that the mental lexicon can no longer be looked at “as a store of single words with the odd idiom thrown in” (p. 56). They propose a preliminary, “rough sketch” model (p. 56) to account for the existence of prefabs (as previously defined in Section 3.1, p.434). They argue that formulaic language is stored as a separate entity to single words. Therefore, in addition to the lexicon (specific knowledge of single words) Erman and Warren postulate that there is a ‘phrasicon’ which contains prefabs. The lexicon and the phrasicon interact with our knowledge of grammar, our wider knowledge of the world, and the specific context of the language use in order to create meaning, particularly from non-compositional and opaque language.

In this way, Erman and Warren argue most of the messages we encode and decode will be discarded. However, sometimes we will encounter new senses of words which will be added to the lexicon and sometimes we encounter combinations of words with “transparent syntactic-semantic structures” which are added to the phrasicon (2000: 56). By virtue of having clear form-meaning mapping, the links with the lexicon and knowledge of the grammar are maintained which "make prefabs syntactically and sometimes semantically flexible in ways which are not possible in the case of items in the lexicon" (p. 56). Of course, this model is presented as a rough sketch and does not arise from empirical evidence. They also found in their data that when prefabs had slots, the lexical options were often (though not always) semantically related: "[I]t is certainly a clear tendency and is an indication that what we store in some cases is a meaning rather than a specific word." (p. 41) and they provide examples such as *go to X* where *X* can be filled with semantically related items such as *lectures, class, seminars, meetings* and *with X in common* where *X* can be filled with items from the same semantic field such as *little, much, a lot* etc. This relates to Sinclair (2004) who argued that "units of meaning are expected to be largely phrasal" (2004: 29—30).

Wray (2002, 2008) aims to provide an explanation for the status of formulaic language in the lexicon. She too provides an account of formulaic language that envisages "strings of words to have [their] own identity as an entry in the mental lexicon" (Wray, 2008: 10) but unlike Erman and Warren, Wray (2002) adds a sociolinguistic component to her psycholinguistic account of the lexicon. She proposes that formulaic language is "more than a static corpus of words and phrases which we have to learn in order to be fully linguistically competent" (p. 5) and instead is "a dynamic response to the demands of language use and as such, will manifest differently as those demands vary from moment to moment and speaker to speaker" (p. 5).

Wray (2002) reviews a wide range of data from different types of language users and language users at different stages of acquisition (i.e. normal adult language, child first language acquisition, child second language acquisition, teen and adult second language acquisition and people with aphasia). She then models how formulaic language operates for each of these groups before drawing together each of her models to form an integrated model: the Heteromorphic Distributed Lexicon. Rather than viewing the lexicon as one entity containing all lexical entries, the Heteromorphic Distributed Lexicon consists of five separate lexicons: Grammatical; Referential; Interactional (routine); Memorized and Reflexive. Each of these lexicons is divided to represent three sizes of formulaic unit: the morpheme (which can be bound or free and includes monomorphemic words); the formulaic word (holistically stored polymorphemic words) and the formulaic word string (strings of words which are stored and processed holistically). Examples from each of these lexicons are provided in Table 3.1 below:

Table 3-1 Examples of the Heteromorphic Distributed Lexicon (Wray, 2002: 263)

Aston University

Illustration removed for copyright restrictions

Wray (2002) explains that lexical items can be represented in multiple lexicons depending on whether a string is segmented into its component parts. She uses the string *Take it slowly!* by way of example. *Take it slowly!* may be stored holistically in Lexicon III as an interactional holistic word string. Additionally, *take* and *slowly* (and the lemma *slow)* may be stored in Lexicon II (referential) and the grammatical *it* and *–ly* could be stored in Lexicon I. The result is that *take it slowly* could be created by rule as well as holistically. However, it would mean something different. Whilst holistically the string means 'perform your action with care', *take it slowly* created by rule would mean 'grasp the object at a slow speed' (p. 263—4). In this way, the Heteromorphic Distributed Lexicon is not streamlined, since there is large amount of duplication of items (p. 268). However, Wray argues that this does not cause a problem since efficiency is still retained: "the lexicon lists only those units—large or small—which direct experience has identified as communicatively useful" (p. 268).

Each lexicon represented in the model is a "repository of all linguistic units which are not subject to further segmentation, and which are therefore handled as holistic units" (2002: 264). According to Wray (2002), what separates this model from other models is what qualifies as not requiring further segmentation: "In this model, a polymorphemic word or word string can qualify simply by virtue of its not *needing* to be segmented in normal use, rather than it being *unable* to undergo segmentation" (p. 264, original emphasis). This is what Wray calls *needs-only analysis*; that we only break down and analyse sequences of words if some need arises. Wray explains 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 language. 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" even though they could be analysed if analytical processing were activated (p. 131). Wray 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. Its meaning and functions are stand-alone, so no analysis is necessary (p. 132).

Wray (2002) argues that a key function of formulaic language is the promotion of self and therefore, the lexicons are organised for each individual in a way which best promotes their interests. She proposes that our individual repertoires of formulaic language are not the same and that they contain what each individual has found to be useful for them in order to meet their needs:

> Highly literate people with a love of words may have a large and active store of morphemes alongside their stores of words, phrases and texts, affording them the luxury of constructing and understanding novel words and sentences that may be beyond the easy competence of someone whose lexicon has a smaller store of such units (p. 286).

However, whilst the repertoires of formulaic language may vary from person to person, Wray (2002) asserts that individual inventories of holistically stored sequences are heavily influenced by the speech community:

> We have stored them because they 'sound right' to us, that in turn being because they have often been heard in the speech of others. And by using them we, in turn, contribute to what others hear most often and therefore store in their own inventories. (p. 74)

The impact that this may have on our idiolect, and the potential application of formulaic language theory to forensic authorship attribution, will be discussed later (Section 33.4, p. 56). It would appear that Wray's theory aligns with that of Hoey (2005), since she argues that our store of holistic sequences is determined through our data input—the language that we encounter from other people and places. However, Wray (2008) points out a crucial point of difference: for Hoey, "the word is the fundamental currency of processing" whereas in Wray's needs-only analysis model, it is not the word, rather it is whatever lexical unit (either bigger i.e. a sequence, or smaller, i.e. a morpheme) "that constitutes the largest form-meaning mapping so far found adequate to handle the effective manipulation of input and output" (2008: 67).

Whilst there may be disagreement about *how* formulaic language is processed and stored, the key issue which gains consensus, is that language which is processed holistically, that is, as a single unit, is regarded as formulaic. It is on this basis that formulaic language was identified as a suitable candidate for a new marker of authorship, since if authors are less aware of the formulaic choices they make (if indeed they are choices), then these choices may hold clues about their authors, particularly if the same choices occur frequently. The stage is now set to explore in more detail the possibility of using formulaic language as a marker of authorship and three main questions about the inter-relationship between forensic authorship attribution and formulaic language need to be answered:

i) How should formulaic language be defined for use in the forensic context?

ii) Why might formulaic language be a reliable marker of authorship?

iii) Can formulaic language be identified in ways sufficiently robust for forensic purposes?

The remainder of this chapter engages with these questions.

3.3 **How should formulaic language be defined for use in the forensic context?**

3.3.1 What are the current definitions?

In Section 3.1, the point was made that a variety of definitions exist to account for different aspects of formulaic language. Coulmas (1979) for example, coined the term *routine formulae* which he defined as:

> expressions whose occurrence is closely bound to specific social situations and which are, on the basis of an evaluation of such situations, highly predictable in a communicative course of events. Their meaning is pragmatically conditioned, and their usage is motivated by the relevant characteristics of such social situations (p. 240).

His examples of routine formulae include *don't mention it, my pleasure,* and *I'm sorry.* This definition highlights the link between formulaic language and social context as well as the need for pragmatic insight. The term *lexical bundle* is defined as "the most frequently recurring sequences of words in a register" (Biber, 2009: 282). Such a definition adopts a frequency-driven approach and directly relates the recurrence of word sequences to specific registers such as academic writing (Biber, Conrad, & Cortes, 2004; Cortes, 2004). Moon (1998a) coined the term *Fixed Expressions and Idioms* (FEI) to subsume fixed expressions ("holistic units of two or more words") and idioms ("semi-transparent and opaque metaphorical expressions such as *spill the beans* and *burn one's candle at both ends*") as a broader category (p. 2—3). Using the variables of institutionalisation (the extent to which a formulation is accepted as a lexical item), lexicogrammatical fixedness (particularly lexicogrammatical "defectiveness" as in *kith and kin* where the word *kith* is redundant in present day English outside of this expression) and non-compositionality (the extent to which meaning arises from the string as a whole rather than the constituent words), Moon was able to assess the extent to which a string of words could be considered an FEI. Importantly, Moon's criteria are variable so although they enabled her to identify FEIs from novel strings, they are not present to an equal extent in all items (p. 9). These three definitions, including the *prefabs* and *lexicalised sentence stems* as defined in Section 3.1 (p. 44) capture different aspects of formulaicity and different ways of identifying formulaic material.

Clear definitions are necessary to enable identification in texts. However, restricted definitions such as the small sample so far described often lead to exclusivity so that some strings which fall outside the definition are not counted, even though they may appear to have something formulaic about them (Wray, 2002: 44). Restricted definitions make it difficult to capture the essence of formulaicity or describe general characteristics of formulaic language. To achieve this, a more inclusive definition is required. Such a definition is provided by Wray (2002) who defines the *formulaic sequence*:

> [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's definition of the *formulaic sequence* is intended to be as inclusive as possible so that it can be used as a coverall term for any part of language that has been considered formulaic by previous definitions (p. 9) such as prefabs and lexicalised sentence stems as previously discussed. However, whilst the definition of the formulaic sequence is intended to be inclusive, it is not intended to be a definition that enables identification of formulaic material in texts: "Although the formulaic sequence can be used for identification at the general level of items that 'appear to be prefabricated', what appears to be prefabricated needs its own clear definition" (Wray, 2008: 97).

In later work, Wray (2008) offered a theory-specific definition for what she termed the *morpheme equivalent unit (MEU)*:

> [A] word or wordstring, whether incomplete or including gaps for inserted variable items, that is processed like a morpheme, that is, without recourse to any form-meaning matching of any sub-parts it may have.

The MEU refers to the lexicon of the individual. This is in contrast to Wray's (2002) definition of the formulaic sequence, which can be understood to be an MEU shared by speech communities. Formulaic sequences are those wordstrings which are suspected to be MEUs for an individual. However, it is impossible to know whether this is in fact the case, since we cannot know whether someone has produced a wordstring holistically or analytically:

> The formulaic sequence, crucially, encompasses any material that *appears* to be prefabricated, not just that which *is*. In effect, the formulaic sequences are the set of examples that we think may be MEUs, before we can be sure just what an MEU looks like (Wray, 2008: 95).

Wray points out that the definition of the MEU "reflects specific claims about the nature and, by implication, provenance of formulaic material in a language" (2008: 12) and therefore, whilst it can be used for theoretical orientation, it too is not appropriate for identifying formulaic language in texts.

So how best can formulaic language be defined for use in this research? There are three options: (i) Use a definition that is quite specific in what is counted as formulaic but excludes other areas of formulaicity; (ii) Use a definition that more widely encompasses formulaic behaviour but makes identification more difficult; or (iii) Use a new definition, which perhaps draws on other definitions, but specifically fits the needs of the research. Given that this research is investigative, it seems that the third option is most appropriate. This will ensure that aspects of formulaicity which may characterise authorial style will not be excluded on the grounds that they fall outside of a rigid definition. Additionally, the definition will most clearly fit the needs of the research questions posed.

3.3.2 How should formulaic language be defined for use in this research?

To make clear that the focus of this research is formulaic language which may be used to determine authorship, the term *idiolectal formulaicity* will be used as the theoretical rationalisation for this research and should be understood to mean:

> Orthographic word sequences including gaps for inserted items, that appear to be holistically stored, habitual and consistent across a group of texts—potentially an aspect of that author's idiolect.

In this definition, which borrows somewhat from Wray's (2008) definition of the MEU, idiolectal formulaicity refers to habits or preferences for particular word strings that an author exhibits. Previous discussions of formulaic language have made reference to the formulaic practices of language communities such as auctioneers (e.g. Kuiper, 1996), sportscasters (e.g. Kuiper, 1996; 2004), students (e.g. Chenoweth, 1995) and academics (e.g. Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004). In contrast, the definition provided here states that the focus should be on formulaic language use by the individual. By treating formulaic language as potentially part of idiolect, it follows that it should hold potential to be a marker of authorship based on the distinct patterns of social and individual use.

As established in Chapter 2, accounts remain somewhat unclear about whether idiolects change over time or whether fossilisation occurs. At best then, any evidence of formulaic language found in the data can be said to be formulaic for that author, at that particular point. Quite whether the same author would exhibit the same formulaic behaviour in ten years (or even ten days) is beyond the scope of this research. This is not an unimportant point given that real forensic cases often draw on comparison texts from across a person's lifetime.

Now that some definitions of formulaic language have been reviewed and a specific definition for use in this research has been proposed, it is important to be clear about how the terms will be used throughout the rest of this research. *Idiolectal formulaicity* will be used in line with the definition provided above when referring to individual formulaic repertoires. *Formulaic sequence* will be used as a coverall term for all other aspects of formulaic language, as intended by Wray (2002). In practice, this means that the aim of the empirical research is to identify formulaic sequences in texts which can be used as evidence of idiolectal formulaicity. To ensure accuracy when describing the research of others, their terms will be introduced and defined as necessary.

The definition of idiolectal formulaicity will not enable identification of formulaic material in texts but rather conveys the theoretical perspective underlying it, namely, that some aspects of language are stored holistically and that they recur in everyday language use frequently enough to constitute part of a person's idiolect. Therefore, a clear idea of how idiolectal formulaicity is going to be identified in texts is required which will be described in Section 33.5 (p. 67). The next stage to consider is whether there is any evidence that formulaic sequences might be indicative of idiolectal formulaicity.

3.4 Why might formulaic sequences be a reliable marker of authorship?

Having reviewed the underlying theory and practice of forensic authorship attribution and the theory of formulaic language, it is now possible to make explicit the potential application of formulaic sequences as a marker of authorship. In order to build the case for using formulaic sequences as a marker of authorship, this section will extend the previous discussion by exploring the psycholinguistic and sociolinguistic determinants that suggest formulaic sequences should be an ideal candidate as a marker of authorship in forensic investigations. It will then be possible to explore the limited research that specifically examines formulaic sequences and authorship, assessing the specific conclusions in light of what we might predict based on the theory.

3.4.1 Theoretical basis for formulaic sequences as a marker of authorship

Beginning firstly with the psycholinguistic theory of formulaic sequences, sequences of words are stored in the lexicon as single items (Bannard & Lieven, 2009; Ellis, 1996; Erman, 2007; Erman & Warren, 2000; Hoey, 2005; Pawley & Syder, 1983; Sinclair, 1991; Wray, 2000, 2002, 2008). This gives the speaker a processing advantage by reducing the cognitive burden of producing entirely novel language. If Wray (2002) is correct in her assertion that we only analyse those things which need to be analysed (needs-only analysis, cf. Section 3.2.1, p. 51), then as language users, we will not necessarily focus on the internal constituents of these formulaic word sequences. Since sequences of words are stored in this pre-packaged holistic form then their occurrence in language may not be noticed by authors. Therefore authors will likely produce sequences of words without necessarily thinking about each individual word and it naturally follows that if authors are unaware that they are using particular sequences of words it will be much harder for them to disguise their style. This point is made by Lancashire (1998):

> Word, phrase, and collocation frequencies … can be signatures of authorship because of the way the writer's brain stores and creates speech. Even the author cannot imitate these features, simply because they are normally beyond recognition, unless the author has the same tools and expertise as stylometrists undertaking attribution research. Reliable markers arise from the unique, hidden clusters within the author's long-term associative memory. (p. 299)

Additionally, there is growing support for the argument that even when there is opportunity for variability within a formulaic sequence, speakers holistically store a particular variant which works best for their needs. For example, Erman (2007) investigated the size of linguistic units in the mental lexicon, using pause distribution and pause duration as indicators. She hypothesized that pauses should be rare between prefabricated structures ("prefabs", as defined in Section 3.1, p. 44) since they are stored and therefore produced holistically. Her data provided support for this hypothesis. Pausing occurred only rarely between component parts of prefabricated structures (p. 47). Erman also looked at prefabricated structures which allowed for some lexical variability (e.g. *that's the big question in X*, where *X* can be filled with any discipline such as *linguistics*, *history*, *science* etc.). She found no evidence that speakers paused more or for longer at the point where one of the variable slots needed to be filled. This, she reasoned, suggests that there was no increase in cognitive effort required at the point where a single lexical item needed to be selected. Unfortunately, Erman only provides this one example and it may well be argued that if a linguist is talking about linguistics, then filling this particular slot with the word *linguistics* probably would not require extra cognitive effort, since the speaker will have been primed by the context. However, other variable slots are evident in the author corpus data, and it is presumed that these are equally as appropriate as examples: *X might say* where *X* can be filled with *some* or *you*; and *It sounds* ADV *harsh* where any adverb can be inserted e.g. *very, really, quite, incredibly*. Reflecting on her finding, Erman suggests that

> speakers may well make preferred choices, and the prefab may therefore be fixed and stored as a unit in the individual user's lexicon. In other words, speakers make preferred choices also where the system allows sometimes considerable variation, which suggests that more combinations of words are presumably fixed in the individual speaker's mental lexicon than will be indicated in dictionaries and corpora (p. 46).

Whilst Erman (2007) was concerned with formulaic sequences that contained slots for variability, Peters (1983) highlighted that some sequences of words may be stored holistically for an individual, as opposed to holistically for a particular speech community, which she called "idiosyncratic formulas". The effect is that such instances would be formulaic for an individual, even though they would not necessarily be recognisable as formulaic by the hearer:

> Thus, if I find an especially felicitous way of expressing an idea, I may store up that turn of phrase so that the next time I need it it will come forth as a prefabricated chunk, even though to my hearer it may not be distinguishable from newly generated speech. (p. 3)

An example of this is Tony Blair's use of "entirely accepted" amongst other collocations as described in Section 2.1.4 which based on the evidence is likely to be formulaic for him but less so for others. It is also less likely that he would be so aware of his apparently idiolectal use of such expressions. Likewise, the Unabomber's use of "cool-headed logicians" may be considered to be an idiosyncratic formula.

Turning next to the sociolinguistic aspects of formulaic language, according to Wray (2002, 2008) formulaic language is a linguistic solution to a non-linguistic problem and that problem is how we get our needs met: "Formulaic choices will be made on the basis of this single agenda, by means of the drive to manipulate others' actions, knowledge, or emotions to one's own advantage" (Wray, 2008: 69). According to Wray, we store holistically those sequences of words for which we have a need. Therefore,

> [w]hat ends up in the lexicon is a direct reflection of the way the language is operating for the individual in his or her speech community or communities. The nature of the lexicon is determined not by structural principles which decide whether an item is simple enough to be stored, but by the individual's priorities in handling real language input (Wray, 2002: 267—8).

As such, there is potential for us all to have different inventories of formulaic sequences resulting from, amongst others, differing needs and differing social and linguistic backgrounds.

Therefore, providing that there is an appropriate way to identify it (cf. Section 3.5), formulaic sequences should reliably mark out an individual author. In the following section, the limited research literature into formulaic sequences as a marker of authorship will be assessed to see whether this prediction is correct.

3.4.2 Evidence of formulaic sequences as a marker of authorship

Some researchers have started to empirically investigate the idea that formulaic sequences may be specific to individuals. The research of Mollin (2009) has already been described in Section 2.1.4 (p. 24), which demonstrated that one aspect of formulaic language, collocations, appeared to be idiolectal. Other research draws similar conclusions. Kuiper (2009) demonstrated how even in situations where variation from routinized phrases would not be expected, idiolectal phrases are still evident. Schmitt *et al*. (2004), although not directly interested in formulaic sequences as markers of authorship, argued that they are tied to idiolect. Finally, Waltman (1973) researching in the traditions of literary authorship, argued that based on the use of formulaic sequences, an anonymous poem could be attributed to an author. All of these findings will now be discussed.

Kuiper (2009) offers insight into individual variation in the use of formulaic sequences during routine interactions, specifically, ritual talk at the supermarket checkout. This talk, occurring between the points when the checkout operator scans the first item to when the customer moves away from the checkout, “has no direct practical use other than as a form of social intercourse” (p. 97) with the exception of the section where payment is dealt with, which “is fully functional” (p. 97). Kuiper predicted that checkout operators would, as a result of the routine physical work, have linguistic rituals for interacting with customers (p. 99).

Kuiper’s data were collected in 1991 from two New Zealand supermarkets in different suburbs of Christchurch. One was located in a lower socio-economic suburb whilst the other was located in a higher socio-economic suburb and both were stores from different supermarket chains. Over a period of one month, 200 interactions were recorded from nine checkout operators, two of whom were male and the remaining seven of whom were female. In this study, Kuiper was interested “in the formulaic inventory and discourse structure used by checkout operators” and also how the customers contributed to the interactions (p. 99—100) but Kuiper was not concerned with sociolinguistic variation. Therefore, he did not control for social variables such as age, socio-economic status or gender of the customers, nor were an equal number of interactions recorded for each checkout operator. In this way, Kuiper aimed to comment on ritual talk at the checkout in general terms rather than focussing on differences based on social characteristics. The interactions were recorded, transcribed and entered into a computer.

Kuiper found that Christchurch checkout operators normally initiated talk with a greeting directed at the customer, which he argued is virtually obligatory (p. 101). Upon further analysis, Kuiper found that the entire interaction consisted of stages and each stage had a set of formulae (in his terms). Examples of the stages identified by Kuiper, along with example formulae, are summarised as Table 3.2:

Table 3-2 Examples of formulae associated with stages of checkout interaction (Kuiper, 2009)

Aston University

Illustration removed for copyright restrictions

From this summary, it can be seen that there are clearly marked points in the discourse where formulae may be used and also restrictions on where in a discourse sequence particular formulae may be used (Kuiper, 2009: 109). Given the highly ritualised nature of this interaction, Kuiper comments that:

> The typical interchanges between customers and checkout operators look, on the face of it, as though they have little room for an individual operator to be different from others, in that they are highly formulaic and the discourse structure ... is highly restrictive. In fact this is not the case (2009: 109).

This is an important observation because if Kuiper can demonstrate individuality in the use of formulaic sequences in a situation where, on the face of it, there should be very little opportunity for individuality, then there is a strong basis on which to predict that there will be greater evidence of individual formulaic sequence use in less restrictive interactions. Greater individuality will increase the effectiveness of formulaic sequences as a marker of authorship.

Using the same data, Kuiper produced finite state diagrams for three of the operators, generating many permissible routes through the greeting stage of the interaction. Kuiper cautioned that whilst the diagrams represented all of the formulae used by each of the operators, none of the operators used “every possible route through the greeting system” (2009: 109) and furthermore all operators had “their own preferred routes” which is “strongly suggestive of individuality” (p. 109) even though the data form only a small sample:

> Operators use particular tracks through their diagram with greater or lesser frequency and this pattern of preference creates an individual style ... All speakers are thus idiosyncratic in a limited way. Many have formulae that they alone use. Some operators are clearly also much more flexible in their use of formulae than others (p. 110)

Kuiper tabulated all of the greeting formulae used by each of the nine checkout operators alongside frequency counts for how often each operator used each formula. By presenting his data in this way, Kuiper was able to show that:

i) Some greeting formulae were used frequently by the majority of operators (e.g. *Hi. How are you?, How are you?* and *How are you today?*);

ii) Some greeting formulae were used rarely and only by one operator (e.g. *Gidday*, *Alright?*, and *How are you going?*);

iii) Some greeting formulae were the preferred formulae, from the range of choices, for operators (e.g. The operators Elsie, Di, Shelly and Kris used *Hi, how are you today?* more frequently than any other formulae in the data); and

iv) Some greeting formulae were used more regularly by only one operator (e.g. Only Dusty used *Good morning. How are you?* and *Morning*).

This process revealed that some operators used formulae much more flexibly than others and that all operators used particular formulae, equivalent to a signature. Kuiper argued that this “signature” indexed each speaker’s persona which “evolves differently over time so that some operators are more conservative, maintaining their signature for long periods, whereas other operators are more flexible” (2009: 113). Kuiper concluded that “even within such a tightly constrained environment as that which the routine actions speech of checkout operators imposes, there is room for individuality, idiosyncrasy and even for a small measure of creativity” (p. 114).

Kuiper’s findings and conclusions are important because they indicate that operators do have preferences for particular formulae and that, provided the more idiosyncratic formulae are distinguished from the more general, it should be possible to identify an individual operator based on the formulae they use. In this way, in an imaginary scenario, a customer may have cause to complain about an operator and may not recall their name. If the customer, for some reason, remembered that the operator said "Morning", it may be possible, armed with enough knowledge about each operator's formulaic inventory, to identify Dusty as the speaker, much like Woodhams and Grant (2006) identified perpetrators of rape through case linkage based on utterances produced during the crime. However, the formulae described by Kuiper are limited to only one context, the supermarket checkout and furthermore, his discussion of individual preferences for formulae is confined to only the opening, greeting phase of the interaction. Whilst his data show that individuals do have preferences for formulae which can, on occasion, mark them out from a small sample of their colleagues, further generalisability beyond this context cannot be assumed.

Generalisability is further restricted by the relatively small sample, which, although sufficient for Kuiper's purposes, allows little room for speculation about the individual use of formulae in different contexts or when more data from more speakers are analysed. However, whilst this is a relatively small data sample, the reality is that in a forensic investigation even less data, and certainly more varied data, may be all that is available. Therefore, whilst Kuiper may be right that the operators left their "signature" through their use of formulae, analysing more varied data from less formulaic contexts/genres may diminish the appearance of something as persuasive as a signature, even though formulaic sequences may indeed be a part of idiolect, or at the very least, an idiosyncratic feature.

There is also a temporal issue to consider. Kuiper argued that use of formulae is indexical of the checkout operator's persona, and that this persona may change over time. This relates to Wray (2002) who argued that formulaic repertoires are dynamic, not static, adjusting to meet changing needs. Therefore, even if formulaic sequences can be demonstrated to be useful as a marker of authorship, their application in forensic contexts may always be limited, depending on when Known Documents are composed in relation to Questioned Documents.

Although not the primary focus of their investigation, Schmitt, Grandage and Adolphs (2004) provide a small amount of evidence that formulaic sequence inventories may be linked to idiolect. The aim of their study was to determine whether recurrent clusters which they identified using corpus linguistics methods were psycholinguistically valid; that is, the extent to which recurrent clusters were stored holistically (p. 128). To that end, they drew a distinction between word strings which are identified through corpus analysis but may or may not be stored holistically (e.g. *in a variety of*), which they call *recurrent clusters*, and word strings that are stored holistically, which, following Wray (2002), they refer to as *formulaic sequences* (defined in Section 3.3.1, p. 54).

Schmitt *et al*. used a variety of existing reference lists and corpora frequency counts to identify a range of recurrent clusters which varied from being “relatively frequent to relatively infrequent” (2004: 129). Using several criteria including length, frequency, transparency of meaning and cluster type, they selected 25 recurrent clusters as test stimuli, some of which were more likely to be stored holistically (e.g. *as a matter of fact*) and some that were more questionable (e.g. *in the number of*). Both native and non-native speakers of English were presented with the 25 recurrent clusters interspersed in dialogue and 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 they would need to reconstruct the content using their own resources rather than rote memory. They argued therefore that any of the 25 recurrent clusters that were recited back by participants during the dictation task could be argued to be stored holistically since it would be less cognitively demanding for participants to produce formulaic sequences.

Schmitt *et al*. categorised all of the recurrent clusters in participants' responses according to whether they were a) produced fully intact in terms of lexis and intonation contour; b) attempted but with missing/substituted lexis and/or a not fully intact intonation contour; or c) completely missing. Dealing specifically with the results from 34 native speaker participants, they found that “not all of the clusters were reproduced in a manner which would suggest they were holistically stored in the mind” (p. 135). Some recurrent clusters were produced less frequently (e.g. *in the same way as*, *to give you an example*), suggesting that they were not stored holistically or for some reason were not available to the participants during the dictation task. Others, such as *to make a long story short* and *I don't know what to do*, were reproduced correctly by most of the participants, implying that they may be formulaic sequences.

Between the categories of correct (a) and incorrect (c), were those recurrent clusters that were partially correct (b). This category is particularly interesting since “clusters which were attempted, but not reproduced intact, give the clearest indication that those clusters were somehow not prominent in the mind” (p. 137). Some recurrent clusters were argued to be holistically stored since they were only occasionally produced partially incorrectly or disfluently (e.g. *go away, for example, is one of the most*) whilst others were produced incorrectly and/or disfluently by most of the participants (e.g. *I see what you mean, as shown in figure, aim of this study*), suggesting that they are less likely to be holistically stored. Schmitt *et al*. conclude that recurrent clusters identified through corpus techniques are therefore not always psycholinguistically valid. Instead, “[r]ecurrent clusters vary, with some highly likely to be formulaic sequences on the basis of this evidence, but others quite unlikely to be holistically stored" (p. 138).

However, of particular interest for drawing a relationship between formulaic sequences and idiolect is their observation 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. This suggests, according to Schmitt *et al*., that some speakers stored some recurrent clusters as formulaic sequences whilst others did not: "[I]t is idiosyncratic to the individual speakers whether they have stored these clusters or not" (p. 138). They then make the connection between formulaic sequences and idiolect explicit:

> 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).

Like Wray (2002), they argue that what is stored in our formulaic inventories (the 'formulalect' or 'phrasalect') includes a majority of formulaic sequences that are shared across the speech community. However, there are differences based on individual abilities in fluency as well as individual differences in "powers of expression" which may also be linked to topic and discourse situation. They conclude: "Thus, the bottom line is that just as a person's mental lexicon contains a unique inventory of words, it is likely to also contain a unique inventory of formulaic sequences" (Schmitt, Grandage, & Adolphs, 2004: 138).

That there is growing consensus that formulaic sequence inventories are linked to idiolect is clearly useful for the present purposes. However, Schmitt *et al*.'s conclusion is based on the results of one study which included only 34 native speakers (an additional 45 non-native speaker participants took part in the study but the results have not been discussed here). However, it is interesting that in a more general context, idiolectal differences were still found lending further support to Kuiper's (2009) context-specific research. Taking collocations as a type of formulaic sequence, the work of Mollin (2009), as previously discussed fully in Section 2.1.4 is directly relevant to the argument that formulaic sequences may be linked to idiolect. Recalling that Mollin investigated an extensive corpus of language produced by Tony Blair, she was able to demonstrate that certain collocations occurred repeatedly throughout the corpus when other collocates could have been used and further that these collocations were not register or genre specific. In light of these three studies, more detailed empirical investigation into the relationship between formulaic language and idiolect is warranted.

More specific evidence that formulaic sequences may be markers of authorship comes from Waltman (1973). Waltman's analysis is based on the "Poema de Mío Cid", which situates it in the literary forensics tradition rather than forensic linguistics. Waltman asserted that repetition of "formulaic expressions" throughout the poem indicates single authorship and as such it is necessary to engage with his claims.

Waltman (1973) reported that since 1929, questions surrounding the authorship of the *Poema de Mío Cid,* a Spanish oral epic poem like *Beowulf*, have arisen. Some critics claim single authorship whilst others argue the poem was composed by two or more poets (p. 569). Authorship in this context relates to the number of authors who wrote the poem, rather than identifying a single person with whom common authorship is shared between the poem and their known documents (i.e. attribution). The poem is typically split into two parts so that they can be compared with each other. Literary differences (rather than forensic linguistic markers of authorship) between the halves have been noted which may indicate dual authorship such as differences in style (the first half being sober and historical whilst the second is less serious and more fictitious), differences in the use of verb forms and synonyms between the two parts, higher frequency of assonance in the second part, and genre differences (the first half being a modern novel whilst the second is an epic) (p. 569—570). Others suggest that such variation can be explained by the poem's oral roots. It is asserted that the endings of oral performances vary the most because of the audience's impatience: "The early part of the poem, by frequent performance, becomes relatively stable while the later part requires more powers of improvisation" (p. 570). Waltman investigated the use of formulaic expressions in the *Poema de Mío Cid* claiming that patterns of formulaic expressions would reveal something about the poem's authorship:

> [I]f one can find a large number of formulaic expressions which are constantly used throughout the poem in a developed pattern, this may shed some light on the question of unity [of authorship]. A variance in formulaic expressions in the poem would tend to point toward two composers (Waltman, 1973: 571).

Defining *formulaic expressions* as "a group of words with similarity of vocabulary under somewhat the same metrical conditions", Waltman used concordance software on the entire poem and found 26 formulaic expressions (1973: 571—2). Such phrases included: *El de Biuar* (16 times), *Moros & christianos* (6 times), and *Vala el Criador* (9 times). Waltman found that 24 of the 26 phrases were "fairly evenly distributed" throughout the poem (p. 572).

Waltman then wanted to show that the use of formulaic expressions was linked to authorial style. He took two segments of the poem, each consisting of 20 lines. He selected sections which dealt with the same topic: a parting, farewell scene. He found that both segments contained the same formulaic expressions and of the 40 lines studied, only six contained formulaic expressions that occurred in just one segment. This led Waltman to claim “that there seems to be no great difference between the two parts of the poem in the use of formulaic expressions” (p. 575), concluding that the “appearance of at least 26 different formulaic expressions, which are found in all parts of the poem, is the strongest evidence found in support of only one author” (p. 577).

Waltman demonstrated that formulaic expressions are constant across that particular poem. It should be borne in mind though that Waltman was careful to select segments which were comparable in topic. Therefore, it is possible that formulaic expressions are linked to topic or genre and their reoccurrence throughout the *Poema de Mío Cid* was directly linked to topic rather than authorial style (e.g. Kuiper, 2004). It is also important to consider that Waltman’s definition of *formulaic expressions* is restricted to the field of literature, so although “groups of words with similarity of vocabulary” would likely be accepted by linguists interested in formulaicity, Waltman’s focus on “the same metrical conditions” would appear to be redundant outside of the literary context.

This review of the limited empirical research into formulaic sequences as a marker of authorship has lent support to the theoretical assumptions underlying formulaic language described in Section 3.4.1. Individuals do seemingly have different stores of formulaic sequences which appear to suit their individual needs. Kuiper (2009) demonstrated that in a very specific context, idiolectal variation in the use of formulas was apparent. Similarly, it is interesting that Mollin (2009) found examples of idiolectal collocations, but it should be remembered that she used a very large data sample; far greater than could ordinarily be expected in the forensic context. Schmitt *et al*. (2004) found that some recurrent clusters were more likely to be stored holistically than others and argued, as does Wray (2002) that different social interactions shape formulaic sequence inventories. Finally, Waltman (1973) argued that formulaic sequences may be evidence that different halves of a text were produced by the same author based on the proportion and repetition of various formulaic expressions. However, the conclusion Waltman reaches should be considered cautiously. He is able to show that formulaic sequences occur similarly across the two halves, but he is unable to show how specific to an individual author this feature is and whether we could expect the same pattern across a different set of data.

Therefore, as noted throughout, there are clear limitations which need to be addressed, mainly that 1) the data on which the findings are based are few; and 2) the contexts investigated are very limited. More detailed empirical investigation into the relationship between formulaic sequences and idiolect is therefore required—in other words, there is justification for further exploring the role of formulaic sequences as a marker of authorship, but, crucially, data more relevant to the forensic context (i.e. shorter texts) will need to be used to establish just how much potential this new marker of authorship holds. If formulaic sequences do not occur in sufficient numbers in shorter texts, there is less likelihood that a useful tool for forensic authorship attribution can be developed. All of these issues will be dealt with in the next chapter which outlines the research design for the empirical investigation. Prior to this, it is necessary to consider the approaches for identifying formulaic sequences to ensure that the methods will be appropriate for the forensic context

3.5 Can formulaic sequences be identified in ways sufficiently robust for forensic purposes?

The task of identifying formulaic sequences in texts is not an easy one; indeed Wray (2008) comments "[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 & Warren (2000) talk about two problems with identification: (i) what is a prefab (defined in Section 3.3.1, p. 44) for some members of the community is not necessarily so for others, and (ii) prefabs can be easily overlooked. However, they say, not all prefabs are inconspicuous and are more easily identifiable. They caution that "the identification of 'all and only' the prefabs in a text is in practice impossible" (p. 33). In the following section some of the most commonly used approaches for identifying formulaic sequences in written language are presented before discussing which of these approaches will be the most suitable for the forensic context. In assessing which methods for identifying formulaic sequences are rigorous enough for forensic application, three factors must be considered:

i) Reliability;
ii) Validity; and
iii) Feasibility

Any method of authorship attribution, whether used for investigative or evidential purposes, needs to incorporate analyses which are reliable; that is, analyses which can be repeated to produce the same results. Read and Nation (2004) talk about reliability in the context of identifying formulaic sequences. They state that any methods for identification need to be consistently applied and that any criteria used for identification should be clear. They also emphasize the need for a high level of agreement between judges. A clear description of methods and procedures is also required so that the analyses can be repeated by another linguist (p. 34). Referring back to Erman and Warren (2000: 33), although it may not be possible to collect "all and only" the instances of formulaic sequences in a text, there has to be a high level of confidence that enough have been collected to produce evidence of authorship and additionally, that enough have been identified to ensure that further examples would not significantly alter the results.

Validity refers to whether what has been identified is actually what was intended; in other words, whether the sequences of words identified are actually formulaic. Read and Nation (2004) argue that this is a particularly problematic criterion, since "storage as a whole unit" is difficult to operationalise (p. 35). The aim of this research is to specifically look for examples of formulaic sequences as evidence of idiolectal formulaicity. However, it cannot be known for sure that any identified sequences of words are really evidence of idiolectal formulaicity since there is no way to establish whether they are stored as single items in an author's lexicon. Therefore, examples of formulaic sequences need to be identified which can be convincingly argued to be formulaic for an author. This problem clearly highlights the need for both a definition of formulaic language at a conceptual level and an additional definition for operational purposes. In order to ensure external validity, Read and Nation (2004) argue that data should be representative of the target language and also be large enough to contain sufficient examples which raises the issue of corpora representativeness (p. 35).

Feasibility refers to how well the method can be applied to forensic data, taking into account what is achievable with the often limited resources available to the linguist including time available for analysis, the size and number of forensic texts and the number of people required to carry out the analysis. Methods for identifying formulaic sequences will now be presented and evaluated with respect to these three criteria in an attempt to identify the most robust methods for identifying formulaic sequences in a forensic context. Since this research is concerned with formulaic sequences as a marker of authorship, only methods appropriate for the identification in written language will be considered.

3.5.1 Intuition and shared knowledge

Wray (2002) says that as members of their own speech communities, researchers "often are the self-appointed arbiters of what is idiomatic or formulaic in their data" (p. 20). This is based on the belief that native speakers recognize formulaic language as having special status (Van Lancker-Sidtis & Rallon, 2004: 208). Therefore, intuition can be used as a basis for identifying formulaic sequences. This approach is clearly subjective and for the technique to carry more reliability, at least a second-rater should be used (Read & Nation, 2004: 29). Better still, panels of independent judges rather than individuals or couples can be used to reach consensus about whether a string of words is indeed formulaic (e.g. P. Foster, 2001), for as Wray (2002) comments "there should be a certain resilience in a consensus achieved in this way ... there can be a wide variation in the overall number of sequences spotted by different judges" (p. 22).

Despite a lack of objectivity, using intuition to identify formulaic sequences is ideally suited for a researcher who wishes to adopt an exploratory approach. Formulaic sequences are not always fixed and do not always have firm borders, so it sometimes requires a judgement call to decide whether something is formulaic or not: "[T]he problem with formulaic language is that between the extremes of what is *definitely* formulaic and what is definitely *not* formulaic, there is a sizeable amount of material that may or may not be" (Wray, 2008: 93, original emphasis). This type of discretionary judgement into the 'grey areas' of formulaic sequences can only be performed by researchers (in comparison to automated methods).

Tied into using intuition is the concept of using shared knowledge. If members of the same speech community all share the same knowledge about particular formulaic sequences, then it can be possible to detect which strings of words are formulaic for that community. The method is for one person to start a formulaic sequence and then for other members of the speech community to complete it. Depending on how reliably the sequence is completed by others provides insights into how formulaic the sequence is for that particular speech community. However, the technique is not appropriate for formulaic sequences that allow variation (Wray, 2002: 24—5) since not all members of the same speech community could be expected to complete variable formulaic sequences in the same way.

Reliance on intuition is a commonly used approach for the identification of formulaic sequences (Wray, 2002: 20) although researchers acknowledge that it is at the same time the least scientific: "The status of the intuition of an individual investigator is dubious from a modern "scientific" perspective" (Read & Nation, 2004: 29). This immediately causes problems for any method that might be tested against the *Daubert* criteria (Section 2.3, p. 39). Intuition is not scientific because there is a lack of reliability—what one researcher may judge to be formulaic may not be so for another so there is the danger of significant variation between judges. To complicate the issue further, what may be formulaic for one researcher on one occasion may not even be so for the same researcher on a different occasion for reasons such as tiredness and unintentional changes in how judgements are made (Wray, 2002: 23). Therefore, identifying formulaic sequences using intuition alone cannot offer any reliability.

Claiming validity can be less problematic. If a string of words is intuitively recognised as formulaic, then it has every potential to be stored and processed holistically, particularly if a group of judges can reach consensus. However, intuitions about language are not always correct and in an era of corpus analysis, linguists are often sceptical of intuitive judgements as Sinclair noted over 20 years ago:

> [T]he contrast exposed between the impressions of language detail noted by people, and the evidence compiled objectively from texts is huge and systematic. It leads one to suppose that human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use the language (Sinclair, 1991: 4).

In addition, to make intuitive judgements that are valid, researchers identifying formulaic sequences need to have the same shared knowledge as the people who produced them: "Clearly, any string that is formulaic for, say, the speaker, but not for the hearers, will simply not be understood unless it is transparent" (Wray, 2002: 24).

Intuitive analysis of texts is often restricted to small datasets given that each text has to be read carefully and more than once which can make it a slow and laborious process. It is therefore not feasible to use intuition to identify formulaic sequences in larger texts or indeed in shorter texts if there are many of them as is often the case in forensic investigations. The time pressures involved in producing forensic authorship evidence (Shuy, 2006) may therefore preclude this from being a feasible identification technique. Whilst using a panel of judges may increase reliability, the majority of forensic linguists work in isolation and may not have access to similarly trained linguists who could assist. Furthermore, many forensic materials are confidential, and so the linguist would be unlikely to get permission for a panel of judges to view the texts. It may be possible for linguists to extract word sequences that they believed to be formulaic and to present them to a panel out of context, but the success would rely crucially on the linguist having identified the 'right' sequences of words in the first place. As such, using intuition as a technique for identifying formulaic sequences in forensic texts is not feasible.

3.5.2 Automated approaches: corpus analysis and reference lists

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, there are two techniques 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 certainly 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). For both approaches to corpus analysis, Read and Nation argue that the data need to be evaluated by the investigator in order to determine which instances are actually formulaic sequences.

A limitation of this automated approach to identification is that only fixed formulaic sequences can be identified. The only way to identify variable formulaic sequences is to specify specific lexical components of the wider fixed formulaic sequence. For example, in the formulaic sequence *sort something out*, where *something* is a variable slot, specific software is required which can identify the potential lexical items which could fill this slot. An example is the use of *WordSmith Tools* (Scott, 2008) where the asterisk can replace the word *something*. Applying this principle to the author corpus enables *sort it out* and *sort things out* to be identified. However, using a reference list will not typically enable the identification of such variation in form. Therefore, whilst the formulaic expression *all of a sudden* can be identified in a corpus if it is included in the reference list, a slight variant such as *all of the sudden*, which also occurs in the author corpus, would not be identified, even though the two are clearly related.

In order to carry out automated searches for identifying formulaic sequences, computer-based approaches require a clear definition of what is to be counted. The software then identifies anything that meets the pre-determined criteria and automatically excludes anything else. Therefore, unlike intuitive methods, a computer-based approach is not appropriate for exploring the ‘grey areas’ of formulaic sequences since computer software is no better at detecting the boundaries than researchers. In order to carry out a search, the researcher has to make decisions regarding the length of the string to be identified and the minimum frequency of occurrence, with such decisions being dictated by the size of the corpus. Whilst it is convenient to think of automated methods as more objective, the researcher still needs to make post hoc decisions about which of the identified strings to include and which to exclude: “[W]hile it might seem sensible to simply count everything, it is often intuitively clear that some patterns are more important and relevant than others” (Wray, 2002: 27). Clearly, once the researcher starts making decisions, objectivity can be compromised which can undermine the value of the automated search.

Reference lists such as dictionaries and text books provide a source of established examples of formulaic sequences (Wray, 2008: 109). It is possible, using such sources, to match a given dataset against a reference list and identify those examples which occur. Wray (2008) cautions that if a researcher wishes to use a reference list, it is important for them to think about why that list was produced and what decisions were made about what to include and exclude. Published lists are sometimes used as an alternative to intuition which leads Wray to further comment that it is important to know how the list was compiled since there is little point in trading off one's own intuition in favour of somebody else's:

> An important question for any researcher to consider before using existing lists to identify formulaic sequences is whether the list has gained authority simply by virtue of being published (Wray, 2008: 109).

Automated analysis offers a more systematic approach to identifying formulaic sequences. Reliability is higher than can be achieved using intuition, since once the criteria have been specified, the software will extract all, and only, the instances that fall within the search parameters. This means that human factors such as tiredness will not affect the results and that the same results can be obtained by multiple linguists working independently, provided that clear and unambiguous details are outlined for how the search was conducted. This aspect of the corpus analysis technique makes it both reliable and feasible for use in forensic contexts.

Demonstrating the validity of some of the formulaic sequences that automated analysis will reveal can sometimes be problematic. Corpus searches are sometimes used to extract sequences of words that occur frequently over a pre-determined threshold which can then be classed as 'formulaic'. This is on the basis that "the more often a string is needed, the more likely it is to be stored in prefabricated form to save processing effort, and once it is stored, the more likely it is to be the preferred choice when that message needs to be expressed" (Wray, 2002: 25). The problem, as described by Wray (2002) is that "some patterns are more important and relevant than others" and so some subjective human assessment is required, which can undermine the value of the objective automated search (p. 27). On this point, Schmitt, Grandage and Adolphs (2004), as described in Section 3.4.2, draw a distinction between corpus-derived recurrent strings which are not psycholinguistically valid and formulaic sequences which are psycholinguistically valid. Therefore, if automated methods are used, the researcher needs to be able to justify on what basis the material identified is actually formulaic and if it is frequency alone, then this will need to be clearly stated.

3.5.3 Structural analysis

Formal criteria can also be used to identify formulaic sequences and the two most widely recognised are i) non-compositionality (that a literal interpretation is not possible), and (ii) fixedness (the degree to which the word order can be changed and lexical insertions, inflections and replacements are possible). Such criteria are particularly suited to the identification of idioms. The main problem with using these criteria for identification is that they generally lie on a continuum with some formulaic sequences being more fixed than others (Read & Nation, 2004: 32). Since this form of analysis focuses on non-compositionality and fixedness, it would be possible to construct a reference list of idioms and use automated software to highlight those strings of words which match those entries found in the list (e.g. Moon, 1998a), which, as with the corpus methods described above, would enable reliability since the process would be automated. The contention may arise regarding how the list is formed and what it contains (Wray, 2008). Some form of authoritative list, preferably derived from large corpora may ensure better reliability. However, automated reference list-matching methods would not guarantee that only non-compositional formulaic sequences are identified; in other words, a computer cannot decide whether a string of words used in context is idiomatic or whether a literal meaning is intended. This would be easy to establish based on context, but a linguist would be required for such a purpose.

Establishing the validity of word strings highlighted using structural analysis is relatively straightforward since it is idioms that are identified. Since idioms are non-compositional, they have to be processed holistically in order for their meaning to be derived.

In terms of feasibility, again, this technique seems promising, particularly if the process is automated. The problem may arise in relation to how much material will actually be identified. Moon (1998a) argues that idioms which we know well and assume to be common in language, do not actually occur with any great frequency. In fact, in her research, some idioms occurred with zero frequency in the Oxford Hector Pilot Corpus (OHPC) consisting of 18 million words. Some examples of idioms that did not occur at all include: *kick the bucket*, *one man's meat is another man's poison*, and *when the cat's away, the mice will play* (p. 60). This is of even greater concern given that many forensic texts are rather short, so the likelihood of sufficient examples being identified is low.

3.5.4 Pragmatic and functional analysis

Some types of formulaic sequence are linked to specific functions. Pragmatic/functional analysis recognises this characteristic. If data from specific social settings have been collected, it is possible to identify formulaic sequences that are fully fixed and which exhibit a lack of transparency:

> Idioms are said to lack *semantic* transparency because their meaning is not interpretable from knowledge of the individual lexical components. To this we can add *pragmatic* transparency, which refers to the need for knowledge of the social context in which particular formulaic expressions are used in order to be able to understand their role in discourse (Read & Nation, 2004: 33, original emphasis).

This approach is appropriate for identifying, for example, routine formulae (e.g., *don't mention it* and *my pleasure*, as defined in Section 3.3.1, p. 53). Identifying formulaic sequences based on the contexts in which they are used and the functions that they perform can be argued to be valid, since, by their routine nature, they are likely to be processed holistically to reduce cognitive burden. Less clear is the reliability of this approach to identification. A researcher is required to identify formulaic sequences in this way, since computers are not able to decode contextual and pragmatic cues. This approach is therefore less reliable than automated methods. An additional problem for this technique is feasibility. Since the defining characteristic of this approach to identification is that data are collected from specific social settings (e.g. Kuiper, 2009, Chapter 6), there is unlikely to be wide enough appeal to other forms of data that are not tied to clearly definable situations (such as diary entries, personal letters etc.). In addition, the researcher would need a great deal of insight into each particular context in order to identify and understand the pragmatic effect of the formulaic sequence. Therefore, given the limited context, a pragmatic/functional approach to identifying formulaic sequences is unlikely to be robust enough for the forensic context. Each of the approaches described can be summarised as in Table 3.3:

Table 3-3 Summary of appropriateness of identification techniques

Approach	Reliable	Valid	Feasible
Intuition and shared knowledge	X	✓	X
Corpus analysis and reference lists	✓	X / ✓	✓
Structural analysis	✓	✓	X
Pragmatic and functional analysis	X	✓	X

Read and Nation (2004) argue that none of the approaches described above are independently adequate for the identification of formulaic sequences and they argue, as also does Wray (2002), that valid results can only be obtained through using more than one form of analysis. To this end, triangulation is likely to produce findings which are more reliable and valid for the researcher and more robust for the courts. From the evaluation of the above approaches to identification, it seems that using intuition and pragmatic/functional analysis can be ruled out on the basis that they lack the reliability required of the courts and that they are not likely to be feasible for the forensic linguist. Corpus methods and methods which draw on reference lists, provided that their authority can be attested, will carry the most evidential value in terms of reliability, validity and feasibility. The limitation of this approach is that only fixed forms can be identified. However, since it has been noted throughout this chapter that there are many grey areas with no clear boundaries between what is formulaic and what is not, it makes sense to remain focussed on only a narrower set of formulaic sequences rather than attempting to identify everything that can be argued to be formulaic. As such, automated approaches are the most appropriate to adopt for this research.

3.6 Summary

At the end of Chapter 2, the argument was made that irrespective of whether idiolect is related to the specific choices or habits that an author makes, using a theory grounded in psycholinguistics and sociolinguistics will carry more validity when accounting for any findings. The argument in this chapter is that formulaic sequences hold this potential and to this end, the literature which explores how formulaic sequences are processed and stored and also the available evidence into whether formulaic sequences can act as a marker of authorship has been reviewed. Having established that formulaic sequences do hold the potential to distinguish between authors, methods for identification have been presented and it was argued that automated methods are the most appropriate. Having established this, the next stage is to determine which formulaic sequences should be identified through automated methods—in other words, what the computer will actually find. This is dealt with in the next chapter.

It is now possible to begin the empirical work to determine the extent to which formulaic sequences can actually differentiate authors in the forensic context. The next chapter takes into account the discussion reported in this chapter and establishes clear hypotheses and methods for how to determine whether idiolectal formulaicity exists. To do this, the author corpus, that is, the data to be used for the empirical investigation into this marker of authorship, will also be introduced.

Chapter 4

'Failing to plan is planning to fail': research design

In Chapter 3, several claims were made about the nature of formulaic sequences in relation to the individual:

1) Individuals make preferred choices, or have habits, even when considerable variation is possible (Erman, 2007; Erman & Warren, 2000; Johnstone, 1996; Kuiper, 2004; Mollin, 2009; Peters, 1983; Waltman, 1973).
2) Individual lexicons contain formulaic sequences which each person has found to be useful based on direct experience and language input. Therefore, no person has the same set of formulaic sequences as another since we all have different experiences (Hoey, 2005; Schmitt, Grandage, & Adolphs, 2004).
3) The lexicon directly reflects the way that language operates for a person and is determined by individual needs when handling language input (Wray, 2002).

These three key findings provide a solid foundation for the belief that formulaic sequences should be effective as a marker of authorship and as such it is now possible to formulate a research question which will guide the empirical analyses carried out in Chapters 5—7.

4.1 Central research question

As a result of the issues arising from Chapters 2 and 3, and the three claims outlined above, the question guiding the empirical analysis is:

> *Given that all individuals have a different store of formulaic sequences acquired through a different range of life experiences, can formulaic sequences be used as a marker of authorship to the extent that a Questioned Document can be correctly attributed to its author from a relatively disparate sample of candidate authors?*

Upon completion of the analytical work, it will be possible to offer an answer to this question (Chapter 8). In order to find answers, a series of methods are required so that a range of different aspects of formulaic sequences can be investigated. It should be noted at this stage that as identified in Chapter 3, there is a variety of approaches available for the identification of formulaic language. However, since the aim of this research is to develop a method suitable for forensic application, it stands to reason that the approaches adopted should be the ones that hold the most potential for success. It is impossible to test every approach to the identification of formulaic language, so instead, informed by the discussion in Chapter 3, three of the most promising approaches will now be pursued.

The first method aims to identify only a small quantity of formulaic sequences. With each successive analysis, a wider variety of formulaic sequences will be identified. Drawing upon the results from these analyses, a variety of information based on different approaches and different aspects of formulaic sequences will be available so that the findings can be triangulated. This will enable a strong evaluation of the evidence and consequently, a robust answer to the central research question. The three approaches to analysis are outlined below.

4.2 Methods of analysis

4.2.1 Formulaic clusters

As an initial starting point, it makes sense to identify only a small subset of formulaic sequences because this way, it will be possible to more closely interrogate the data whilst being reasonably confident that the sequences identified are actually formulaic with fewer examples of 'grey areas'.

In order to achieve this, a quantitative approach (e.g. Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004; Hoover, 2002; Stubbs, 2002; Stubbs & Barth, 2003) will be used in order to identify recurrent sequences—clusters of words—in the corpus, which may be indicative of authorship. It is the number of occurrences and consistency with which individual clusters occur that qualifies them as formulaic (as discussed in Section 3.5.2, p. 70) and which separates the research presented here from other investigations which have used clusters as a marker of authorship.

At this juncture, it may be useful to consider whether the research presented here adopts a corpus-based or corpus-driven approach to the data analysis, which Römer (2005) describes as being "two different opposing disciplines within corpus linguistics" (p. 22). The corpus-driven approach is more prone to the alteration and development of theory leading to new theoretical insights. Corpus-based linguists, alternatively, "do not put the corpus at the centre of their research but see it as a welcome tool which provides them with frequency data, attested illustrative examples, or with answers to questions of grammaticality or acceptability" (p. 23). Although the author corpus is not annotated (a preference of corpus-driven linguists who avoid relying on other researchers' views of language), there are pre-formulated ideas and hypotheses in mind (p. 23) which the corpus evidence is then used to either support or refute. On this basis, it is more accurate to describe the present research as being corpus-based.

Genre can be an important feature of some cluster based investigations (notably, lexical bundles e.g. Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2004, cf. Section 3.3.1 for definition). Since the present research is interested in a more universally applicable approach, a robust method for authorship attribution needs to be independent of genre or context. As such, it is necessary to develop a definition of what exactly will be identified as formulaic. The term *formulaic cluster* has been coined here for this purpose, and should be understood to mean:

> Sequences of three words or more which are not necessarily complete meaningful units and which are not overtly related to context. Formulaic clusters occur in the majority of texts produced by an individual author and can be argued to be idiolectal based on the recurrence of form across separate texts and to be formulaic in terms of their frequency.

The fact that formulaic clusters are found in the majority of texts demonstrates that they are a strong and, crucially, recurring part of that author's lexical repertoire (as opposed to clusters which might be very frequent in one text but not across a series; these are also likely to be idiolectal but less consistent and therefore less reliable). Repetition across texts also reduces the likelihood of clusters being content specific or chance occurrences. The threshold for determining what 'majority' means will be dependent on the data available in terms of quantity of texts and the length of texts. In a later section (cf. Section 4.4, p. 84), the author corpus is described, in which each author produced a total of five texts. As a guide, occurrence in three of the available texts (60%) is justified as the minimum since this equates to over half of the texts produced by an author (and obviously, formulaic clusters which occur in 80% or 100% of texts should be more characteristic of idiolect). Other researchers wishing to draw on this definition would be required to justify their own thresholds based on their own data.

The definition specifies that formulaic clusters must consist of at least three words, since two word clusters will typically consist of grammatical items (e.g. Biber, Conrad, & Cortes, 2004). Although the diagnostic potential of grammatical items has been claimed (e.g. Mosteller & Wallace, 2007), it may be less convincing to argue that they will be useful in this context. After all, grammatical items are required for the organisation of text whereas lexical items allow for more variability. Although grammatical items may well be stored formulaically, being a smaller set of words means that there is more limited variation in how authors can use them compared to lexical words. A cline will naturally be generated between clusters which occur more frequently across fewer texts and those which occur less frequently over more texts.

Finally, focusing on the recurrence of form means that variability cannot be tolerated; in other words, authors must produce the identical forms over three of their texts. The limitation of this approach is that clusters which naturally allow for some variability (e.g. *it's his choice* and *it's her choice* where the pronominal choice is content dependent) will not be identified as formulaic clusters in this research. However, the method will enable an initial automated analysis, contributing to the requirement that a method based on formulaic language should be robust (cf. Section 3.5).

If individual lexicons do contain preferred formulaic sequences, differences between authors' formulaic clusters should manifest. Using this definition, formulaic clusters will be identified in the data in Chapter 5. It will then be possible to determine whether their occurrence in texts can differentiate authors and enable the correct attribution of a Questioned Document.

4.2.2 Core word

Having determined the extent to which a small and circumscribed sub-set of formulaic sequences are employed by some authors, it will be possible to increase the range of sequences to determine whether the authors employ sequences from different sets—or at least, whether the choices made in the texts differ significantly from author to author.

It stands to reason that if one particular word can be isolated which occurs predominantly and frequently in formulaic sequences—a core word—then a reasonable sub-set of sequences, the majority of which could be expected to be formulaic, will also be identified. 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 concept 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)

In this example, *way* should be frequent in large corpora because it is central to "numerous expressions". It follows that identifying all instances of *way* in a corpus should provide a direct path to a range of formulaic sequences.

So what candidates are available as core words? Of the content words, two in particular stand out as worthy candidates. The first is *thing*. Willis (1990) observes that *thing* is very common in the English language, it has a clear meaning and its grammatical behaviour is known (p. 39). These are important factors for ensuring that sufficient data are extracted from the corpus and that the marked and unmarked uses of the word are understood. However, what makes *thing* especially suitable as a core word is that it is incorporated into a variety of formulaic sequences e.g. *one thing after another, the shape of things to come* (Willis, 1990: 39). *Thing* (and its plural form, *things*) occurs 168 times in the author corpus (described in Section 4.4, p. 84) so it generates plentiful data.

The second potential candidate is of course the noun *way* described above. *Way* is also singled out by Willis (1990) as a key content vocabulary item which occurs in numerous formulaic sequences and enjoys the same advantages as *thing* described above. In comparison to *thing* though, *way* (and its plural, *ways*) occurs less frequently in the author corpus, a total of only 105 times. Even so, *way* does hold certain other advantages. There is existing literature to support the range of meanings conveyed by *way* along with its specific uses (e.g. Goldberg, 1996; Sinclair, 1999; Willis, 1990). More importantly though, there are more entries and definitions in the *Oxford Online Reference* tool (2010) for expressions containing *way*. The importance of this will become clear later (cf. Chapter 5) when glosses for the sequences identified through the core word are required. It stands to reason that for an exploratory piece of research there is inherent value in utilizing the findings of existing research and resources to inform the methods adopted. Overall then, *way* is judged to be the most appropriate core word on which to concentrate.

The first task will be to establish whether individual authors use a different set of *way*-phrases. Then it will be possible to determine whether other authors use the same *way*-phrases (i.e. the distinctiveness of those formulaic sequences), and, for comparative purposes, whether similar meanings are expressed in different forms (i.e. using expressions that do not include *way*). This is the approach adopted in Chapter 6.

4.2.3 Reference list of formulaic sequences

The final investigation into formulaic sequences as a marker of authorship takes an increasingly inclusive approach in order to establish how many formulaic sequences occur in the texts, whether authors use different proportions of formulaic sequences, and crucially, whether texts can be attributed to their authors on this variable. In Chapter 3, the difference in the size of individuals' store of formulaic sequences was highlighted. If Wray's (2002) assertion is correct that the range of formulaic sequences available to an individual varies, then each author in the corpus should rely on formulaic sequences to differing degrees. Some authors may use a higher proportion of formulaic sequences whilst others may use a higher proportion of novel language, depending on what each author has stored as communicatively more useful and the complexity of their individual needs.

In Section 3.5, it became clear that from the researcher's perspective there were different advantages to each of the methods, but which ultimately raised problems from a forensic-orientated perspective. For example, using intuitions about formulaic sequences may be particularly suited to exploratory investigations such as this, but the level of objectivity and reliability renders the findings too problematic to be used as forensic evidence. The solution, drawing on the shared knowledge of a panel of judges, can lead to consensus regarding whether a given example can be considered formulaic. However, forensic case work does not typically lend itself to analysis by a panel of judges due to issues of confidentiality, restrictions over access and limitations on time.

On the other hand, reference lists can overcome such problems and they afford the linguist an opportunity to analyse substantially more data, on their own, relatively quickly. The key consideration is which items are contained in such a list, and what decisions are made at the compilation stage. The ideal basis on which to proceed is to create a unique reference list of formulaic sequences based on the shared knowledge of a considerable number of judges. Such an approach draws on the strengths of both approaches: i) There is an element of consensus derived from a large panel of judges and ii) The list can be applied to data reliably without having to actually involve individual judges. The result will be a list of formulaic sequences that can be applied to individual forensic cases. Through the marriage of these two approaches a greater level of reliability, validity and feasibility can be achieved.

The proposed method is to develop a reference list based on examples of formulaic sequences obtained from the internet. The internet represents language as it used by a huge range of language communities. If there is consensus amongst internet users over what is acceptable as a formulaic sequence (what Peters (1983: 11) calls *community-wide formulas*), it is a reasonable assumption that such items will actually be formulaic. Numerous lists created on the basis of different aspects of formulaic sequences are available on the internet (e.g. clichés, idioms etc.), usually created as a reference tool for non-native speakers of English. In many cases, such lists are amended and added to following suggestions from readers. This satisfies the requirement of using a panel of judges.

The empirical research in the final analytical chapter sets out to test whether in fact such a list can be created and usefully applied to data, and more importantly, whether identifying formulaic sequences in this way produces results which firstly actually do differentiate authors (legitimising formulaic sequences as a marker of authorship) and secondly whether the method could be presented in evidence (legitimising the method as a forensically robust approach to identifying formulaic language).

Before the analytical work can begin, appropriate data are required. The data that will be used in this research are described in the following section.

4.3 Statistical testing

At the end of Chapter 2 (Sections 2.2.1—2.2.2), the advantages and disadvantages of qualitative and quantitative analyses were discussed. It was argued that whilst qualitative analyses may be more amenable to identifying idiosyncratic features of language use, the results of which may be more comprehensible to jurors, it is becoming harder for authorship attribution evidence based on qualitative analyses to be admitted into courtrooms due to the potential lack of scientific rigour. Quantification, when conducted properly, is more scientific since the analyses are typically objective and replicable with the potential for subjective judgement from the forensic linguist being reduced. In light of these considerations, a quantitative approach to the data analysis will be adopted in this research, taking into account Grant's (2007) recommendation to integrate statistical testing into the research design at early stage of research planning. Therefore, in the empirical chapters (Chapters 5—7), a range of statistical tests will be used to quantify similarities and differences between texts. Whilst the statistics employed in this research are commonly used (with the possible exception of Jaccard's Coefficient which is dealt with in more detail in Chapter 5, p. 108), a rudimentary knowledge of statistics is assumed in the reporting of the analyses and results. For those who are confident interpreting statistical results, this section may be disregarded. However, those without grounding in statistics may benefit from consulting Table 4.1 below to better understand why the statistics used in the empirical chapters have been selected and how results have been calculated.

Table 4-1 Statistical tests used in empirical analysis

Statistic	Purpose	Formula
t-test	For comparing two sets of numbers to determine, on average, whether those numbers are different, even when the numbers are small. The t-test assumes that scores more or less follow a normal distribution, that they are interval or ratio data and that there aren't any outliers in the data.	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ $\bar{x}_1$ = mean score for group 1 $\bar{x}_2$ = mean score for group 2 s_1 = standard deviation for group 1 s_2 = standard deviation for group 2 n_1 = number of people in group 1 n_2 = number of people in group 2
Mann-Whitney *U*	Non-parametric equivalent of the t-test, so appropriate when data don't meet the assumptions (e.g., data are ordinal or they contain outliers). Uses ranks rather than scores. The lowest score will have the lowest rank so Mann-Whitney *U* can be calculated to tell whether one group has higher ranks than the other.	The lowest value of either U_A or U_B is the *U* value used as the result of the test. $U_A = \sum R_A - \frac{n_{A\,(n_A+1)}}{2}$ $U_B = \sum R_B - \frac{n_{B\,(n_B+1)}}{2}$ n_A = number of people in group A

		n_B = number of people in group B $\sum R_A$ = total of ranks for group A $\sum R_B$ = total of ranks for group B
Kruskal-Wallis	For comparing multiple groups. Non-parametric rank based alternative to one-way between-subjects analysis of variance.	For each group, the mean rank is calculated. The sum of the ranks in each group is squared and then divided by the number of individuals in that group. $H = \left[\frac{12}{N\ (N+1)}\right]\left(\sum \frac{R_j^2}{n_j}\right) - 3\ (N+1)$ N = total sample size R_j = sum of the rank scores for group *j* n_j = sample size for group j
Log-Linear Analysis (saturated model)	For the simultaneous analysis of three or more between-subjects variables. A saturated model is constructed with all component effects present which perfectly predicts cell frequencies. The highest-order interaction is then removed to determine what effect it has—how well the model predicts cell frequencies. Removing this interaction may not affect how well the model predicts target frequencies, so the next highest-order interaction is then removed and the process is repeated until the best fit model is found. The likelihood ratio test is used to assess the goodness-to-fit at each stage.	$Ln(F_{ij}) = \mu + \lambda_i{}^A + \lambda_j{}^B + \lambda_{ij}{}^{AB}$ $Ln(Fij)$ = the log of expected cell frequency of the cases for cell *ij* in the contingency table μ = overall mean of natural log of expected frequencies λ = effects of the variables on cell frequencies *A, B* = variables *i, j* = categories within variables Likelihood ratio test: $D = -2\ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right)$
Kolmogorov-Smirnov	Goodness-to-fit test. Has the sample of measurements been drawn from a normal population or are there outliers? The cumulative probabilities of values in the data set are compared with the cumulative probabilities of the same values in a specified theoretical distribution. If the discrepancy is sufficiently great, the test indicates that the data are not well fitted by the theoretical distribution.	First, the empirical cumulative distribution function is calculated: $F_n(x) = \frac{1}{n}\ [\text{Number of observations } \leq x]$ After which, the Kolmogorov-Smirnov (*D*) statistic can be calculated: $D_n = \underset{x}{\sup} \lvert F_n\ (x) - F\ (x) \rvert$ sup_x= supremum of the set of distances

The formulae for calculating the statistics have been provided in Table 4.1 for reference only since in actuality all of the tests reported in this research were performed using *PASW Statistics* (version 18, formerly known as *SPSS*), a commercially available software package for the statistical analysis of data. There are many resources available which explain statistics in general (e.g., Hinton, 2004; Walker, 2010), the application of statistics to language data (e.g., Hatch & Farhady, 1982; Oakes, 1986; Woods, Fletcher & Hughes, 1986) and using *SPSS/PASW* to calculate statistics (e.g., Kinnear & Gray, 2000) to which the interested reader may usefully refer.

4.4 Data collection: the author corpus

As established in the previous chapter, no formal research exists which specifically explores the potential application of formulaic language in cases of forensic authorship attribution. Therefore, the research questions posed in this thesis are novel and need to be tested on a carefully controlled corpus (e.g. Hänlein, 1999). Arising from the issues raised in Chapter 2, four criteria need to be considered when selecting data:

i. Known authorship – are the authors of the texts verifiable?
ii. Composition dates – were all the texts written during similar periods?
iii. Length – do the texts contain a similar number of words?
iv. Genre – is like being compared with like?

Research which tests a new marker of authorship should use data which has been controlled for these four characteristics. That way, the marker under investigation (in this case formulaic sequences), is the only variable under investigation and any differences between authors cannot be attributed to the fact that, for example, they composed texts of different lengths. In other words, if differences in the use of formulaic sequences are found, they should be explainable as differences between authors and not differences between the texts. The best way to control these variables is to use texts that are written specifically for the research purpose through a structured writing task.

4.4.1 Method

Participants for the study were recruited through a snowball sampling technique. In this way, an initial group of participants was identified and from them the names of other participants were solicited (Heiman, 1999: 289). Specifically, an initial e-mail was sent to friends, relations and colleagues as well as to a university undergraduate mailing list which provided information about the research and invited people to participate. Recipients were then asked to forward the e-mail to any of their contacts who they thought might also be interested in participating.

A structured writing task was designed which required participants to write personal narratives in response to a choice of questions. The decision to elicit narratives, rather than, for example, diary entries or e-mails, was motivated by ethical (cf. Section 4.4.2, p. 87) and practical considerations. Asking participants to keep a diary may have been too invasive (i.e. requiring participants to recount what they did on a particular day) and too hard since if the participants did not do anything noteworthy on a particular day, they may have struggled to reach a set word limit. Collecting e-mails would have been another alternative, although this again may have been problematic. Whilst it may be possible to control the known authors and composition dates, controlling the length of an e-mail, the type of addressee, the topic etc., may be more difficult. A participant who ordinarily sends e-mails consisting of two or three sentences to set up meetings may struggle to reach an imposed word count. Furthermore, it may be hard to control the variable of genre since e-mails are likely to be produced for a variety of purposes.

Eliciting narratives overcomes these particular problems since participants can be provided with a choice of topics to write about and by virtue of being structured, there is a set word length target to work towards. Finally, since the texts are produced specifically for the task, they all serve the same purpose so they should be comparable in terms of genre. An additional benefit is that “telling about personal experiences seems to be something all humans do” (Johnstone, 1997: 316), so the likelihood that a participant will fail to understand the task is reduced.

It was established in Chapter 3 that formulaic sequences are produced automatically. Therefore, to inform participants that this particular aspect of their authorial style is important would be to foreground an otherwise automatic behaviour which could affect the reliability of the formulaic sequences elicited as a marker of authorship. For this reason, participants were not told the real aim of the research at the outset, although they were fully debriefed at the end of the task and were provided with the opportunity to withdraw their data. Labov (1970, 1972a) and Labov and Waletsky (1997) propose an additional measure for reducing the experimental effect. They propose that through describing past events—*producing narratives of personal experience*—participants focus less on their writing style. The questions posed to participants as part of the structured writing task were therefore open-ended and designed to engage participants with their personal experiences.

Participants were asked over a period of five days to write one text each day. 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 well within this range and ensures that at a rate of producing one text per day, participants could complete their task in less than a week and within the same time period. Of course, though, there is no consensus in the literature regarding the dates that “the same time period” spans. In his research into quantitative approaches to literary authorship studies, Grieve (2007) pointed out that author-based corpora should contain Questioned Documents and Known Documents produced “around the same point in time” (p. 255) and in his corpus, he selected texts produced over a five year period. Hänlein (1999) collected texts produced over a period of one year. Therefore, for the purposes of this research, texts produced over five days can justifiably be considered to have similar composition dates. Although all participants wrote their texts within their five day slot, not all participants started on the same date. The data were in fact collected over a six week period, in late 2006.

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. Some authorship attribution research has been conducted on shorter 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. Similarly, although estimates exist regarding how much formulaic language material may be found in written texts (cf. Section 3.1, p. 44), there is as yet no consensus regarding the ideal text length for establishing individual patterns. 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), whilst others have used texts of similar lengths in forensic investigations (Chaski, 2001; Hänlein, 1999; Winter, 1996), this is a reasonable length of text on which to establish whether patterns of formulaic sequences can reliably differentiate authors.

Each morning participants were sent two narrative-eliciting questions and were asked to answer just one of them. They were given a choice so that they could answer whichever question they preferred. The list of questions appears below in Table 4.2. A list of reserve substitute questions was also available and participants were invited to request one of these if they could not answer either of a particular day’s questions. The substitute questions were designed to be hypothetical so that they could be answered more easily and are reproduced as Table 4.3.

Table 4-2 Text generating questions

	Questions (Participants answer only one each 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?

Table 4-3 Substitute text generating questions

	Questions (Participants answer one as required)
Substitute A	If you could change anything in the world, what would it be and why?
Substitute B	Who do you admire and why?
Substitute C	If you could be invisible for a day, what would you do?
Substitute D	What would you do if you won £1,000,000?
Substitute E	Would you like to be a housemate on *Big Brother* and what are your reasons?

Participants were sent an electronic *Microsoft Word* template by e-mail into which they could type their answer. The template was locked so that they could not change the size of the font which restricted participants to writing a maximum of approximately 850 words. The template was big enough to hold more words than required so that if a participant could not fill the template, they may still have written enough for the text to be useful and so that their answers did not have to end abruptly if they reached the word limit, hopefully encouraging a less-experimental feel to the task.

After writing an answer to the question, participants saved the file and returned it, via e-mail, on the same day. This process was repeated for each of the five days. When five answers had been received, participants were thanked for their time and were paid £10. Upon completion of the tasks, participants were debriefed.

4.4.2 Ethical considerations

Full approval for this research was granted by the departmental ethics committee. All potential participants were sent an information sheet and declaration form (Appendix A). The information sheet explained that texts were required for an investigation into how people write about their personal experiences. They were told this half-truth so that they would not focus directly on their authorial style as outlined in Section 4.4.1. Participants were given sufficient time to consider participating and to ask any questions that they might have. All participants were required to sign the declaration form and indicate that they consented to take part in the research.

The declaration form asked participants to confirm that they were native speakers of British English to ensure that an additional variable of 'native language' was not introduced. They were also asked to confirm that they understood that they alone should write the texts and that they could withdraw from the research at any time. A small amount of personal information was collected from the participants. Each participant was asked for their name and address to ensure that payment could be sent. They were additionally asked to provide their e-mail addresses to enable the writing task to be administered and their gender, age and highest level of education (cf. Table 4.4) to enable the possibility of identifying trends along these variables.

Upon completion of their task, participants were told the real purpose of the research. They were invited to ask questions and discuss any concerns. They were also told that the results of all of the analyses would be available to them. Finally, they were given sufficient time to re-consider their participation and withdraw their data, which none of them did.

A potential ethical concern may be the nature of the topics that participants were asked to write about, since questions which elicit personal narratives may invoke sensitive and emotive memories, potentially causing distress. To this end, before agreeing to take part in the study, participants were provided with two example questions (which were not used in the actual task) to help them orientate to the sorts of questions they would be asked. Of the two questions that participants were sent daily, a potentially more emotive question was off-set against a less emotive question (e.g. Day 3: 'What is the biggest decision you have ever made and did you make the right one?' and 'How did you find out that Santa Claus doesn't exist?') so that participants could avoid writing about a sensitive issue (and of course, the list of substitute questions was available for anyone who felt uncomfortable or unable to answer the set questions).

Full anonymity was guaranteed to participants and all identifying material was altered. Such material typically included names, places and dates and all were replaced with fictitious information which was inserted into the text following the original format used by the author.

4.4.3 Participants

A total of 21 people took part in the research. One participant withdrew from the study on the third day (in line with standard ethical procedures, she was not asked for a reason). Table 4.4 shows the gender, age and highest level of education of the remaining 20 participants. Since a snowball sampling technique was used to recruit participants, it was not possible to control for these variables. Therefore, nine males and 11 females participated with an age range of 18—48 and all participants possessed a post-secondary school academic qualification, ranging from college level qualifications (A-Level and AS-Level) to a doctorate.

Table 4-4 Biographical information for each participant

Author	Age	Education
Alan	19	College
Carla	25	Undergraduate
David	28	Doctorate
Elaine	24	Postgraduate
Greg	25	Undergraduate
Hannah	25	Postgraduate
Jenny	23	Undergraduate
John	24	Postgraduate
Judy	24	Undergraduate
June	24	Undergraduate
Keith	25	Undergraduate
Mark	19	College
Michael	20	College
Melanie	48	Undergraduate
Nicola	20	College
Rick	28	Undergraduate
Rose	21	Undergraduate
Sarah	24	Undergraduate
Sue	18	College
Thomas	25	College

4.5 Description of the data

Since 20 participants completed the task by producing five texts, a total of 100 texts were collected, 15 of which are appended. Appendix B contains all five texts produced by Melanie. Appendix C contains Thomas' texts. Appendix D contains the first text from each of John, Jenny, Greg, Judy and Alan who all answered the same first question as Thomas and Melanie (when did you last cry and what made you cry?) in order to show how different authors answered the same question.

The total number of words in the corpus was 65,113 with the longest text containing 822 words and the shortest text containing 485 words. The average number of words per text was 651 and each author produced an average of 3,256 words ranging from 2,844 to 3,916 words. In an early case of forensic authorship attribution, Eagleson (1994) compiled two corpora, one containing 3,725 words and one containing 3,294 words for comparison with a Questioned Document containing 2,551 words. Eagleson explained that although the size of each corpus was not large, they were comparable in size and formality. Therefore, if formulaic sequences can be demonstrated to be a robust marker of authorship in these data, the potential application in the forensic context based on the lengths of such texts may be justified.

There is an argument that in a forensic investigation, if the linguist is convinced that the Known Documents available for analysis are composed by the same author, they can be amalgamated and be treated for the purposes of analysis as one or more longer texts, rather than as a series of independent shorter texts. Since the authorship of the data in this research has been controlled, the five texts could feasibly be treated as one. However, the central aim of this research is to determine whether formulaic sequences are a potential marker of authorship, and to do this, consistency and distinctiveness across a series of texts will be the main determinant of any patterns. In other words, it is not simply the fact that formulaic sequences occur in one text that is important in testing this marker, but whether they occur repeatedly over several texts. Therefore, throughout this research, the texts will be treated independently.

Figure 4.1 shows the distribution of questions answered on each day. For days 1, 3 and 5, both questions were answered fairly evenly. However, there was a clear preference for one question on days 2 and 4 ('What has been the worst moment of your life?' and 'What is the most life-threatening situation you have ever been in?' respectively). No inferences can be drawn from these differences based on gender since the frequencies are so low (however, see Section 4.5.1, p. 92, for discussion of these differences in relation to the quality of data). Only two people requested a substitute question, one on day 4 and the other on day 5.

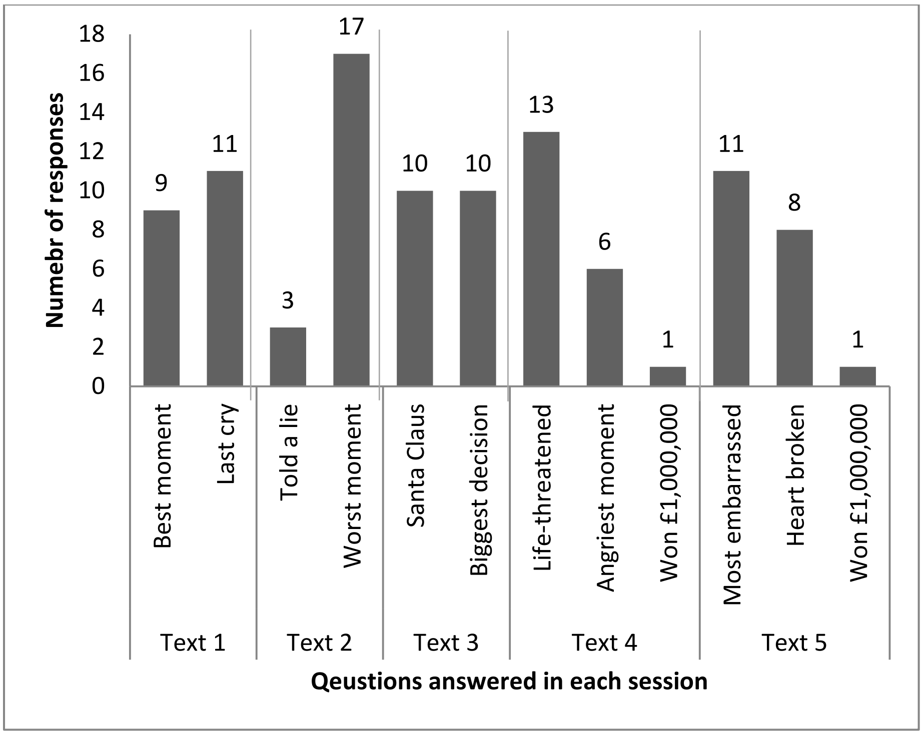

Figure 4-1 Number of participants answering each question

The task was designed so that all answers would be comparable regardless of which questions were answered. It is arguable though that topic might affect the level of formulaicity. Indeed, this point was made by Wood (2009) who found that the quantity of formulaic sequences increased over two samples produced by one person with the first being a narrative about attending a concert and the second being a narrative of childhood summer vacations. Although this was largely attributable to the formulas that his subject, Sachie, had learned through a series of workshops, Wood argued;

> The topic Sachie chose for the second sample may have been of more immediate relevance to her life, and she may have engaged more with the themes she found herself conceptualizing and formulating into speech. In a sense she took flight in the second sample and produced speech which was delivered at a faster rate, was more complex in terms of formulaic sequences, and displayed a greater range of emotion and depth than her first sample (p. 53).

If Wood is correct, then the fact that topic varies across the five texts by each author should be secondary to the formulaic sequences produced as a result of being a personal and emotive narrative, of which all five texts were. Furthermore, as noted above, by treating each text separately, any variation attributable to content should be minor compared to the overall style across all five texts.

A further important consideration is the variation produced from the authors' perception of audience. Some of the authors wrote formally whilst others wrote very informally as evidenced through lexical features such as phonological spelling (e.g. kinda, helluva, grr), frequent use of contractions (e.g. couldn’t, wouldn’t, hadn’t) and colloquial lexis (e.g. ended the relationship and was feel [sic.] *grotty* about it, most of the other *kids* had already gone, mum rang me *like* three times). This is something that will be considered in later discussion (cf. Chapter 8).

4.5.1 Quality of the data

It is now necessary to assess these data and their suitability for investigating formulaic sequences as a marker of authorship. The first question is whether the four criteria of known authorship, similar composition dates, similar length and similar genre have been fulfilled. The importance of each participant writing their own answer was highlighted throughout the data collection period. Participants were told this beforehand and were required to sign a declaration indicating that they would abide by this rule. They were also reminded on a daily basis with the accompanying questions, that they should produce their answer without assistance. Therefore, whilst it is not possible to categorically state that all participants authored the whole of their own texts, all reasonable measures were taken to ensure that this was in fact the case. In terms of similar composition dates, the texts were produced within the same time frame and every participant returned their texts daily within their individual five day slot. The entire data collection period lasted only six weeks. Therefore, the data that have been collected reflect language in use for this group of individuals at that particular period in time. Also, since all participants answered similar questions, all of which required personal narratives, the criterion of similar genre was achieved.

Potentially more problematic is variation in the lengths of the texts. It was anticipated that texts would be approximately 500 words long. However, the average word length was 651 with a range of 485—822 words. The significance therefore needs to be evaluated. As previously mentioned, research which explores markers of authorship in ‘short’ texts lacks consensus over what exactly constitutes a short text. In the research literature, Chaski (2001) used three texts produced by each of four authors as her data, with the shortest text containing just 93 words and the longest 372 words. The texts taken in total for each author only came to 531, 998, 900 and 345 words for comparison with a document containing 341 words. Winter (1996) analysed three short texts containing 616, 805 and 481 words. Hänlein (1999) used essays from *Time* magazine, the shortest of which contained 771 words whilst the longest contained 1054 words. In all of these cases, the researchers consider the variation in length to be minimal and therefore unproblematic.

Alternatively, it is possible to adopt the same procedure as Johnson (1997) who coped with variation by taking the first 500 words and ending at the next sentence boundary. It is worth acknowledging that Johnson was interested in plagiarism (rather than attribution as with the previous research cited) so to demonstrate similarity of text against only the first 500 words was more than adequate for her purposes. However, since the proportion of formulaic sequences in text is under focus in Chapter 7, to arbitrarily cut the text before the end may result in the loss of important information. Therefore, although perhaps not ideal, as with Chaski, Hänlein and Winter, the variation in texts was accepted as an unavoidable result of collecting naturally occurring data and, where comparable text lengths are required, data will be converted to a normalised frequency.

As noted above the real purpose of the study was initially withheld from participants so that they did not scrutinise their writing style and so it is necessary to assess whether this objective was in fact achieved. Many spelling errors, reduplications, missing words and errors in flow occurred throughout the texts (cf. Section 4.6). Participants therefore did not appear to edit their texts before returning them so it can reasonably be inferred that participants did not focus unduly on the actual language they were using. This is perhaps also evidenced through the fact that many of the texts were a lot longer than requested. Participants, therefore, were more likely to be engaged in telling their stories rather than counting how many words they had written and whether they had filled the minimum daily quotient. Furthermore, there is some evidence that participants became emotionally involved with the task, as intended through asking emotive questions:

> Ah, this is the most painful one yet. (Carla-5)

> Right as i [sic.] write the closing on this final chapter in the short book of five, I realise that I have learnt a few things and done some deep searching into myself to write these papers. I have listened to my heart and written down the previously undocumented memoirs of some key parts of my life. These questions have a theme depending on which ones you pick, and they involve opening up you [sic.] heart to reveal what makes you the kind of person you are. You start to question this as [sic.] think about how you see yourself and how others would label you. This one in particular has made me think. Am I a heartbreaker? (Thomas-5)

It is clear from responses such as these that at least some of the participants invested their emotions in the task. This level of involvement with the questions may explain why there was a preference for a particular question on day 2 and day 4. On day 2, participants preferred *What has been the worst moment of your life?* over *Have you ever told a lie and what were the consequences?*, while on day 4 most of the participants chose to answer *What is the most life-threatening situation you have ever been in?* rather than *What is the angriest you have ever been?*. Those who talked about the worst moment of their life, described emotive experiences including being in car crashes, family members dying or being diagnosed with a terminal illness, being in abusive relationships and a school prank ending in a friend being hospitalised. Similarly, those who described a life-threatening situation recounted, amongst others, nearly drowning, being in car accidents, having strokes and seizures, being mugged and being accidentally electrocuted. In addition, of the 17 participants who described the worst moment of their lives 11 of them also described a life-threatening situation, yet no participant showed any overlap between their two answers, drawing instead on two entirely independent experiences. It seems, therefore, that these participants were drawn to recalling and describing emotive, oftentimes troubling memories and as such, it is reasonable to argue that the task eliciting method was successful in achieving its objective.

In conclusion, the data are suitable for an investigation into formulaic sequences as a marker of authorship. However, before the data can be interrogated, some editing of the texts was required to ensure uniformity.

4.6 Data editing

Each text was assigned a number plus the author's (anonymous) name to indicate in which sequence the text was authored (ranging from 1 to 5), so a text labelled Carla-5 indicates that it is the fifth text produced by Carla and likewise, Keith-1 denotes the text written first by Keith.

The previous chapter ended by advocating the use of automated methods for identifying formulaic sequences in the forensic context and the methods outlined in Section 4.2 utilise such an approach. However, a computer can only identify strings of words that it has been programed to find. Therefore, if the search criteria or the data involve a word which is misspelled a match will not be made. Researchers who have used automated methods for identifying authorship have often relied on published texts as their data (Hänlein, 1999; Hoover, 2001, 2002, 2003a). By virtue of being published, such texts will have already been subjected to heavy editing to ensure that spelling, punctuation and formatting are all standardised. However, research using data that has not been professionally edited raises the question of whether spelling should be corrected:

> If it is not [corrected], then a misspelled word will not be recognized as an instance of that word in its correct form, and, indeed, may be counted as a nonword, a *hapax legomenon* (single-occurrence) or as an instance of another word with which the spelling coincides. Misspellings can be precisely what separates out one writer from another, but they will be unhelpful in many analyses (Mollet, Wray, Fitzpatrick, Wray, & Wright, 2010: 434).

Furthermore, deciding to correct spelling is not straight forward, since an author can make both 'performance' mistakes—mistakes that an author knows they have produced—and 'competence' errors—where non-standard rules are broken consistently (Coulthard, 2005b: 15). Coulthard also describes the problem of working with typewritten text:

> [E]rrors and mistakes may be confused and compounded—one may not know, for any given item, particularly if it only occurs once, whether the 'wrong' form is the product of a mis-typing or a non-standard rule—for instance if a (British English) text includes the word 'color'

> is this a typing mistake or a spelling error, or even worse the result of the computer user being unable to change the spell-check to British English (Coulthard, 2005b: 16).

Since spelling is not the focus of this thesis, and since automated methods will be used for identifying formulaic sequences, the decision was made to standardise the data, using the autocorrect feature in *Microsoft Word 2010* as a guide. Such changes included:

1) Inserting spaces as need for punctuation:

	Original	**Edited**
June-5	I would learn from my mistakes,but no fear, I don’t.	I would learn from my mistakes, but no fear, I don’t.
Elaine-1	and it was beautiful-it really was exactly what I would have chosen	and it was beautiful – it really was exactly what I would have chosen
Carla-1	it was a beautiful day and isn’t right .Firstly it’s	it was a beautiful day and isn’t right. Firstly it’s

2) Inserting or deleting spaces between words:

	Original	**Edited**
John-2	Those 6months were very hard	Those 6 months were very hard
Sue-1	I had somewhat convinced myself that Iw as to get AAAB	I had somewhat convinced myself that I was to get AAAB
June-1	Duke of EdinburghAward	Duke of Edinburgh Award
Hannah-1	The blindness went on for a bout 4 minutes	The blindness went on for about 4 minutes

3) Adding/removing apostrophes, accents and extraneous punctuation:

	Original	**Edited**
John-4	I was completely at everyones mercy	I was completely at everyone’s mercy
June-2	I saw him in my minds eye	I saw him in my mind’s eye
Carla-1	cliche	cliché
Greg-2	no serious damage done,.although I wasn’t aware of it at the time	no serious damage done, although I wasn’t aware of it at the time

4) Correcting some spellings, often of homophones:

	Original	**Edited**
Keith-2	to the local boarder crossing police	to the local border crossing police
Elaine-3	when your young	when you're young
Keith-3	I had heard a bit more about the shear numbers of people	I had heard a bit more about the sheer numbers of people

Some irregularities identified by the autocorrect feature were not corrected. These included:

5) Unrecognized lexical items:

Sarah-1 I was absolutely outstanded when he told me that I had passed
Mark-1 Me and JP have had a few bumpings off head but that's just our characters really
David-3 The existence of Santa Clause has always been one of magic and intrepidation

6) Inconsistent/incorrect capitalisation:

Rose-5 it Was also quite embarrassing
Keith-2 at 4:30am i was dead tired and left everyone in the club and headed back to the Hotel
Thomas-4 back to Normal

7) Features of spoken register:

June-3 We then run to our parents room and jump on their bed and begin opening our presents, oohing and aahing about what we have got
Alan-1 It takes a helluva lot to make me cry
Thomas-3 The other thing that kinda makes you stop believing

8) Lexical reduplication for emphasis:

June-2 I'm sure there [sic.] something really really bad that's happened to me
Judy-1 I had got up very very early in the morning
Alan-2 It was just constant pain, pain, pain

In the few examples of unrecognized lexical items, they were automatically identified as spelling errors and it was possible to make an educated guess about what the target word was. However, this could not be categorically known, and so the decision was made to err on the side of caution and not to second-guess the author. For example, in the case of outstanded, it is likely that Sarah blended *outstanding* with *astounded* but we cannot know for sure which was the target word.

There was no need to correct capitalisation as this would not interfere with any automated matching. Features of spoken register and reduplication for emphasis were not standardised since these were judged to be potentially characteristic of how each author used lexis. Being the central focus of this research, it would therefore be unjustifiable to alter this aspect.

In addition to the errors identified by the autocorrect feature, the data were manually checked and a series of additional errors were found:

9) Perceived errors in flow:

June-2	Moving back up North to and going to primary school
Sue-5	I refused on the basis there WAS no more room they start pushing along it, ramming into my sitting next to me
Mark-1	So it that is, that's the last time i cried

10) Omitted words:

Rose-5	When we were all sat in the hall waiting for the presentation ø begin
Michael-1	Later in ø morning we would take a stroll
Hannah-5	We had all drunk too much and there was ø of flirting

11) Incorrect lexical choices and/or potential typing errors:

Sue-3	I was still to select my choices, yet alone start writing a personal statement
Sarah-5	There are other elements rather that money that make people happy
Greg-5	smacked me on the back and pushed me fast first onto the snow

12) Some homophones:

Mark-3	but as the years past my love for animals hasn't changed
David-2	but buy definition they were accidents
Keith-3	but he actually sailed down my road in his slay

13) Incorrect word boundaries that formed complete, recognisable words:

Mark-4	but it came down to the stupidest thing of a miss understanding of what was happening
Rose-5	The nit was my turn!

This final category is akin to metanalysis in Old and Middle English where *napron* came to be pronounced as *an apron* and *a nadder* became *an adder* (Campbell, 2004). However, the difference is that whilst *napron* and *nadder* are not recognised as standard spellings in Modern English, the examples above are and so are not instantly recognisable as misspellings. Categories 9—13 were not corrected for two reasons. Firstly, they were not identified by the autocorrect feature in *Microsoft Word 2010* and therefore the task of identifying every single example could be too cumbersome for the forensic context. Secondly, whilst in some cases it would be possible to establish the target word (i.e. homophones, incorrect word boundaries that formed complete, recognisable words), an element of second-guessing the author would be required for other categories (i.e. perceived errors in flow, omitted words, incorrect lexical choices and/or potential typing errors). Rather, they have been highlighted through this manual checking to illustrate the authenticity of these texts and to explicate potential problems with any analytical techniques, and conversely the robustness of formulaic sequences as a marker of authorship if, even in the face of these problems, evidence in favour of the marker can still be found.

In the case of homophones, it is clear that some were identified by *Microsoft Word 2010* and some were not. Those that were corrected were those that were automatically identified whilst those that were not corrected were those which required manual identification. This divide in the same category highlights a potential limitation in the use of automated methods—the research is limited by the software's level of sophistication. A summary of the changes made to the data can be found below in Table 4.5:

Table 4-5 Summary of changes made to data

	Edited	**Unedited**
Identified by automated check	• Inserting space between punctuation • Inserting or deleting space between words • Adding/removing apostrophes, accents and extraneous punctuation • Some homophones	• Unrecognised lexical items • Incorrect/inconsistent capitalization • Features of spoken register • Lexical reduplication for emphasis
Identified by manual check	• Names, dates, places and any other identifying material	• Perceived errors in flow • Omitted words • Incorrect lexical choices and/or potential typing errors • Some homophones • Incorrect word boundaries that formed complete, recognizable words

The fact that so many errors of different types occurred in the data is an unavoidable characteristic of the data collection design; clearly, asking people to type their answers relies on individual typing ability, though this does highlight that the data are authentic and that a linguist is faced with many of the same problems in a 'real' case of forensic authorship attribution. Nonetheless, it is acknowledged that editing the data in this way may be problematic for some.

4.7 Summary

In this chapter, three key claims have been stated about the nature of formulaic sequences as they relate to the individual. Three approaches have been proposed, all of which, although influenced by other approaches, are novel in their approach to how formulaic sequences may be identified and a corpus has been described on which these claims can be tested. Over the next three chapters, each of these approaches will be described with a full account of the results so that an answer to the central research question can be determined.

Chapter 5

'Seek and ye shall find': formulaic clusters as evidence of authorship

This chapter is the first in a series of three which begin to test the argument proposed in Chapter 4, that formulaic sequences hold potential to be a marker of authorship. In Chapter 4 (Section 4.2.1, p. 77) the first analytical procedure was proposed—identifying formulaic clusters as evidence of authorship. Previous research has explored clusters, word sequences and ngrams in relation to authorship attribution (e.g. Clement & Sharp, 2003; Hoover, 2001, 2002, 2003a; Smith, 1994; Stubbs, 2005) and whilst the occurrence of these lexical strings may be argued to be idiolectal (since they have been selected by an author and so must be part of that author's idiolect), there is no reason to suspect that they are necessarily formulaic. The argument presented in this chapter is that if clusters can be identified which recur across a series of texts, their occurrence may constitute evidence of idiolectal formulaicity. After a full description of the results, the efficacy of this method in the forensic context will be considered, focusing on the key issues of validity, reliability and feasibility.

5.1 Aims and hypotheses

The aim of this chapter is to explore recurrent clusters used by individual authors to determine whether they use forms which can be argued to be formulaic for them and therefore, whether there is potential to use such forms as markers of authorship. It is predicted that authors will repeat across their writing, certain formulations which they have found to be communicatively useful. Therefore, three hypotheses will be tested:

1) Authors will use distinctive patterns of clusters consistently across their texts which can be argued to be formulaic;
2) Authors will be differentiated based on the patterns of formulaic clusters found within their texts;
3) A Questioned Document can usually be correctly attributed to its author based on the occurrence of formulaic clusters.

5.2 Method

Using *Wordsmith Tools* (Scott, 2008), it is relatively straightforward to generate a list of clusters for each author's group of five texts by firstly creating an index file of all the words contained in each author sub-corpus using the 'Wordlist' function and then computing clusters. All clusters of between three and six words which occurred at least twice were extracted from each author sub-corpus. Requiring each cluster to occur minimally only twice in the five texts was a deliberately low threshold set to generate as many potentially formulaic clusters as possible and 1,424 clusters were identified (98 types). Table 5.1 shows the total number of clusters per author while Table 5.2 shows how many types and tokens of each length of cluster were identified, along with some representative examples.

Table 5-1 Number of clusters per author

Author	Number of Clusters
Rose	166
Elaine	101
Rick	=93
Jenny	=93
Mark	83
Hannah	77
Sue	76
John	75
Alan	72
Nicola	=66
Keith	=66
Sarah	=66
Judy	61
Thomas	60
Carla	59
David	49
Melanie	46
Greg	45
June	41
Michael	29
Total	1424

Table 5-2 Examples of clusters found in the author corpus

Length of clusters	Types/Tokens	Examples
3 words	85/1294	A couple of All the time At the time Down the road In a way The same time What had happened
4 words	11/116	And as a result At the same time For the rest of I was going to In a way I
5 words	2/14	Enjoying each other's company Moment of my life was
6+ words	0/0	
Total	98/1424	

As would be expected, there are many more of the shorter (three word) clusters, both types and tokens, than the four word clusters and likewise, the frequency of types and tokens drops dramatically with an increase in size to five word clusters and no clusters of six words or greater being identified at all. The authors vary significantly in their use of clusters ranging from 29 to 166.

At this stage, although many clusters have been identified, there was no certainty that there would be anything necessarily formulaic about them. Therefore, the next stage was to refine the list. To do this all those clusters which occurred in at least three texts produced by a single author were selected in line with the definition of formulaic clusters presented in Chapter 4. This created for each author a range of clusters which could be argued to be formulaic on the basis of recurrence. These clusters are presented as Table 5.3. Column 2, 'Formulaic clusters', lists all of the clusters for each author. The third column indicates in how many of each author's five texts each cluster occurred. This figure merely indicates the number of texts in which a feature occurred so the totals range from a minimum of three to a maximum of five. The actual frequency of occurrence for each author is indicated in column four, 'Total occurrences of formulaic cluster across all five texts'. The fifth column, 'Total occurrences in entire corpus', shows how many tokens of the formulaic cluster type occurred across the entire 20-author corpus and the final column indicates how many of the 20 authors used a particular formulaic cluster. These two columns are discussed further below.

Table 5-3 Formulaic clusters identified for each author and in comparison to all other authors

Author	Formulaic cluster	In *N* files	Total occurrences of formulaic cluster across all five texts	Total occurrences in entire corpus	Used by *N* authors
Alan	BY THE TIME*	3	3	19	9
	FOR A WHILE	3	4	9	5
	I DON'T KNOW*	4	5	13	5
	I REALLY DON'T	4	4	5	2
Carla	AT THE TIME*	5	6	34	15
	IT WAS A*	5	6	41	17
	THE WHOLE THING*	3	3	11	5
David	IN MY LIFE	3	8	28	13
	IT IS THE	3	4	5	2
	THE NEXT DAY	3	3	17	11
	WHEN I WAS*	3	3	41	18
Elaine	AND I JUST	3	4	8	5
	AND WE WERE	3	4	12	9
	IT WAS TIME	3	4	9	6
	MOMENT OF MY LIFE*	3	6	32	13
	MOMENT OF MY LIFE WAS	3	3	9	5
	MY LIFE WAS	3	3	14	10
	OF MY LIFE*	3	6	49	16
	OF MY LIFE WAS	3	3	9	5
	THAT I WAS*	3	3	45	17
	THAT IT WAS	4	5	14	9
	TO GET OUT	3	4	9	6
	TO GO TO	3	3	26	15
Greg	OUT OF THE	3	3	21	14
	THAT THERE WAS	3	3	12	8
	THE FEELING OF	3	3	5	3
	THERE WAS NO	3	3	12	8
Hannah	AT THE TIME*	4	5	34	15
	GOING TO BE	3	3	15	8
	HAVE EVER BEEN	3	3	8	6
	I HAVE EVER	3	3	15	11
	I HAVE EVER BEEN	3	3	8	6
	I REMEMBER THINKING	3	3	3	1
	I WAS SO*	3	3	23	10
	WAS GOING TO*	3	3	24	9
	WAS GOING TO BE	3	3	11	6
Jenny	AND AS A RESULT	3	3	3	1

	AND I WAS*	4	4	33	16
	AS A RESULT	3	4	6	3
	AS I WAS*	3	4	21	12
	I WAS REALLY	3	3	7	3
	IN THE END*	4	4	20	9
	SOME OF THE	3	3	9	7
	THAT I WAS*	3	4	45	17
	THOUGHT IT WOULD	3	3	3	1
	WAS WHEN I	3	3	9	6
	WHEN I WAS*	3	3	41	18
John	I WAS GOING*	3	4	16	10
	I WAS GOING TO	3	4	13	8
	I WAS IN	3	4	30	14
	I WAS THE	3	3	8	5
	IT WAS A*	3	4	41	18
	WAS IN THE	3	4	12	9
	WAS GOING TO*	3	5	24	9
Judy	A COUPLE OF*	4	6	23	11
	I DON'T KNOW*	3	3	13	5
	I WENT TO*	3	4	24	12
	IN THE END*	3	4	20	9
	THAT I WAS*	3	4	45	17
June	ALL THE TIME	3	3	7	5
	END OF THE	3	3	11	6
	ONE OF THE*	4	4	27	13
	WHEN I WAS*	3	3	41	18
Keith	BE ABLE TO	3	3	14	9
	I HAD BEEN*	4	6	34	16
	PEOPLE IN THE	3	3	4	2
	WAS GOING TO*	3	3	24	9
Mark	AND I WAS*	3	3	33	16
	AT THE SAME TIME	3	3	9	5
	FOR THE REST OF	3	3	6	4
	IN THE END*	3	4	20	9
	IN THE SAME	3	3	8	6
	ME AND MY	3	3	11	6
	THE REST OF*	3	3	25	12
	THE SAME TIME	3	3	11	7
	WENT TO MY	3	3	3	1
Melanie	A COUPLE OF*	3	3	23	11
	I WAS ABOUT	3	3	6	3
	ONE OF THE*	3	4	27	13
	SHE SAID THAT	3	3	7	4
	THAT I HAD*	3	5	42	16

Michael	ENJOYING EACH OTHER	3	3	3	3
Nicola	I HAD BEEN*	4	5	34	16
	I WENT TO*	4	5	24	12
	THAT I HAD*	3	3	42	16
	WHEN I WAS*	3	4	41	18
Rick	DOWN THE ROAD	3	7	8	2
	I HAD BEEN*	4	5	34	16
	I HAD TO *	3	5	36	16
	I WAS SO*	3	4	23	10
	IT WAS A*	3	3	41	18
	OF MY LIFE*	3	7	49	16
	THAT I WAS*	4	6	45	17
	THE REST OF*	3	3	25	12
	THE TIME I	3	3	18	10
	WHAT HAD HAPPENED	3	4	18	6
Rose	A COUPLE OF*	3	4	23	11
	A LOT OF	3	3	14	8
	A WAY I	3	3	8	5
	AND I WAS*	3	3	33	16
	AS I WAS*	3	3	21	12
	BUT I KNEW	3	3	3	1
	BY THE TIME*	3	3	19	9
	I KNEW THAT	4	5	18	10
	I REALLY FELT	3	4	4	1
	I THINK THE	3	5	11	5
	I WAS GLAD	3	3	4	2
	I WAS GOING*	3	4	16	10
	I WAS SO*	3	5	23	10
	IN A WAY	5	10	19	8
	IN A WAY I	3	3	7	4
	IT WAS A*	3	4	41	18
	LOOKING FORWARD TO	3	3	7	5
	MADE ME FEEL	3	4	8	5
	ME IN A	3	3	3	1
	MOMENT OF MY LIFE*	3	4	32	13
	OF MY LIFE*	4	6	49	16
	THAT I WAS*	3	3	45	17
	THE WHOLE THING*	3	4	11	5
	WAS GOING TO*	3	6	24	9
	WHEN I WAS*	3	3	41	18
	WHICH I WAS	4	5	6	2
Sarah	FRIENDS AND FAMILY	3	4	6	3
	HE TOLD ME	3	4	12	6
	THAT I HAD*	5	7	42	16

	TOLD ME THAT	3	3	7	5
Sue	AT THE TIME*	3	5	34	15
	BACK INTO THE	3	3	4	2
	I COULD NOT	3	3	9	6
	I DID NOT	3	4	13	8
	I WAS NOT	3	3	11	7
	IT WAS NOT	3	4	12	6
	THAT I WAS*	3	3	45	17
	TO BE THE	3	3	8	6
Thomas	AND IN THE	3	3	7	5
	I HAD TO*	3	5	36	16
	IT WAS LIKE	3	3	7	3
	THAT I WAS*	3	3	45	17
	THE END OF	3	4	23	11
	THE FACT THAT	3	7	30	15

From Table 5.3, it can be seen that all twenty authors have a least one formulaic cluster (i.e. a cluster which they use across at least three of their five texts) although no single cluster is used by all 20 authors: the clusters most shared are *when I was* and *it was a* which are used by 18 authors. It is also apparent that more formulaic clusters have been identified for some authors than others. This difference is perhaps most evident between Michael who has only one formulaic cluster, and Rose, for whom 26 formulaic clusters were identified.

The majority of the formulaic clusters occur in only three texts of the five texts written by any given author, although there are a very few formulaic clusters which occur at least once in all five texts: *at the time, it was a* (Carla), *in a way* (Rose) and *that I had* (Sarah). Some formulaic clusters are particularly noteworthy because of their frequency. For example, Carla uses both *at the time* and *it was a* a total of six times across all five of her texts. Rose uses *in a way* ten times across all her five texts and Sarah uses *that I had* a total of seven times across all her texts. It is also important to acknowledge that five clusters in particular are directly primed by the data-eliciting questions: *moment of my life, moment of my life was, my life was, of my life,* and *of my life was* all of which are in response to the three questions: *what has been the best moment of your life, what has been the worst moment of your life* and *what has been the most embarrassing moment of your life?* As such, to comply with the context-free nature of formulaic clusters (as discussed in Section 4.2.1, p. 77) these clusters were excluded from further analysis, leaving 93 formulaic clusters.

A set of formulaic clusters have been isolated—that is, clusters that occur at least once, and often more, across a series of at least three texts for each author. However, what is not known is the significance of the formulaic clusters for an individual author—whether they are commonplace items of little significance or whether they are potentially diagnostic of authorship. The entire corpus was therefore searched and all the instances of formulaic clusters identified in Table 5.3 were counted (indicated in the fifth column). A total of 1,311 tokens were identified for the 93 cluster types of which 22 types were shared with another author, as indicated by an asterisk following the cluster in the second column. The sixth column shows how many authors across the entire corpus used the formulaic cluster. By examining these two columns, it is possible to determine how distinctive each formulaic cluster is for each author e.g. Rose's use of *I really felt* four times across three texts appears to be more prominent in her lexicon since she is the only author to use the cluster, whereas another cluster such as *to go to* occurs 26 times across the author corpus and is used by 15 authors, so the fact that Elaine uses this cluster three times across three texts is not sufficient to claim this cluster to be distinctive for her.

Of particular interest in this regard are clusters produced by only one author and produced in at least three of their texts. For example, Hannah's use of *I remember thinking*, Jenny's use of *and as a result* and *thought it would*, Mark's use of *went to my* and Rose's use of *but I knew, I really felt* and *me in a*, none of which occur in the rest of the corpus (in other words, each author's uses of these formulaic clusters accounts for 100% of their occurrences in the whole corpus). In fact, Rose's use of *I really felt* occurs in three separate texts, a total of four times (so in one text she uses this formulaic cluster twice) and these four occurrences are the only occurrences in the corpus. This is in contrast to other formulaic clusters which occur relatively frequently for each author and for other authors in the corpus. Such examples include Carla's use of *at the time* which occurs in all her five texts, and a total of six times but a total of 34 times across the whole corpus and Sarah's use of *that I had* which occurs seven times across all five of her texts, against a total of 42 occurrences across the whole corpus. The results of Table 5.3 add support to the first hypothesis, that authors use different patterns of clusters with some consistency across their texts. It is now possible to determine whether formulaic clusters can be used as a marker of authorship.

5.2.1 Statistics and short texts

Although formulaic sequences have been argued to be pervasive (cf. Section 3.1, p. 43), the fact that shorter texts are under investigation means that only low occurrences of formulaic clusters are available for analysis. As such, the statistical testing of data may be problematic since statistics require minimum thresholds before the tests carry validity. A solution is to draw on the types of statistical tests devised for small data sets. Grant (2010, 2011), when establishing the authorship of mobile telephone text messages, drew upon statistics used in psychology to establish case-linkage in crimes. One statistical test in particular, Jaccard's coefficient, appears suited to short texts.

Jaccard's coefficient establishes the correlation between whether a series of particular features are present in a sample, rather than the frequency with which particular features occur. A particular advantage to using Jaccard's coefficient is that the absence of a feature does not increase or decrease the similarity between two texts or crimes (Grant, 2011; Woodhams, Grant, & Price, 2007; Woodhams & Toye, 2007). In other words, the fact that an author does not use a particular feature in the data is not conflated to suggest that the author would never use that feature in any other texts. In the example of case-linkage, features may include the presence or absence of offender behaviours such as whether the perpetrator used a weapon or blindfolded the victim, and how the perpetrator left the scene of the crime and avoided detection (e.g. by wearing a mask and destroying semen) (Woodhams, Grant, & Price, 2007; Woodhams & Toye, 2007). In the authorship of text messages, this may include the presence or absence of specific features such as particular spellings, abbreviations and formatting conventions (Grant, 2010, 2011). Jaccard's coefficient score is calculated between linked pairs (a text by the same author compared to another text by the same author) and unlinked pairs (a text by one author and a text by another author) resulting in a distance measure of between zero and one where zero indicates that two texts are completely different and one indicates that they are identical. Decimals between zero and one indicate variation between these two extremes. The formula used to calculate Jaccard is:

$$J = \frac{a}{(a + b + c)}$$

a = total number of features

b = total number of features in text 1

c = total number of features in text 2

The statistical significance of the resulting distance measure is then calculated using an appropriate test (as will be seen in Section 5.3.1, the non-parametric Mann Whitney *U* is the most appropriate with these data).

5.3 Results

With relation to the data presented in the current investigation, each formulaic cluster (e.g. *by the time, made me feel, the fact that, the same time, what had happened*) constituted a feature, resulting in 93 features. All 100 texts as described in Chapter 4 were used in the analysis resulting in 4,950 pairs of texts.

5.3.1 Establishing variation

The Jaccard's coefficient for each of the two groups of linked and unlinked pairs was tested to see if the coefficients were normally distributed. Although Jaccard values for linked pairs showed no significant difference from normal (KSZ=0.768, N=200, p=0.597) the unlinked pairs were significantly different from normal (KSZ=7.661, N=4750, p<0.001). Therefore, the non-parametric Mann-Whitney *U* test was carried out to test whether Jaccard was significantly lower in unlinked pairs. The Mann-Whitney *U* test showed a significant difference in mean ranks between linked and unlinked pairs (Z=11.3, N=4950, p<0.001) where unlinked pairs were lower. This means that texts produced by the same author are more similar in their use of specific formulaic clusters than texts by different authors and provides support for Hypothesis 2, that authors can be differentiated on the basis of the occurrence of formulaic clusters.

It is now necessary to determine whether a Questioned Document can be successfully attributed to its author. However, the point of Grant's (2011) approach using Jaccard's coefficient is that it is not an authorship attribution in the traditional sense (e.g. attributing a single Questioned Document to one of two candidate authors). Rather, it is a statistical method for describing consistency and distinctiveness (Grant, personal communication). Therefore, whilst idiolectal analysis may focus on specific features of the language which appear to be characteristic of an author, Jaccard's coefficient cannot tell the linguist whether a feature is unique or not, only whether it is consistently used across the data. As a result, in order to attribute a Questioned Document to its author, it is necessary to use qualitative analysis to describe the consistent and distinctive features between writers. Having established in Section 5.3.1 that formulaic clusters can be shown to be more consistent between texts produced by the same author than by different authors, a descriptive approach can be used to attribute a text. Such an approach is in keeping with Grant (2011) who, in the aforementioned SMS authorship research, established variation between the authors through the use of Jaccard's coefficient and then attributed Questioned Documents through qualitative analysis based on the occurrence of features shared between the texts.

5.3.2 Attributing a questioned document: two candidate authors

Using the random case selection function in *PASW Statistics*, two authors were selected for the analysis: Rose and Mark. Of the ten texts produced by these two authors, *PASW Statistics* was again used to randomly select one text to act as the Questioned Document: the first text produced by Mark.

Selecting one of the documents as a Questioned Document means that there will be a 5-text to 4-text comparison and although the majority of clusters occur in only three texts, this uneven comparison may skew the results. Whilst the argument can be made that in a forensic investigation it is less likely that exactly the same number of texts will be available for analysis, in an exploratory study such as this, limits must be established where possible. Therefore, the first part of the analysis will proceed with the 5-text to 4-text comparison, before reducing Rose's texts by one to see how the results are affected by a 4-text to 4-text comparison.

The results of this analysis are presented in Table 5.4. Column 1 shows the formulaic clusters used by Rose. The third column lists all of the formulaic clusters identified in the four 'Known Documents' produced by Mark (i.e. those that occurred in at least three texts). The Questioned Document was then searched for each of Rose and Mark's formulaic clusters and those which were present are shown in the second column. It is important to point out that those items in the second column are only 'candidate formulaic clusters', since by definition a formulaic cluster would need to occur in three texts whereas only one Questioned Document is available for analysis. Therefore, this column represents the occurrence of a cluster which has been claimed to be formulaic for another author (either Rose or Mark), and it is predicted that more clusters in the Questioned Document should be shared with its author (Mark) than with the other candidate author (Rose). The fourth column is discussed further below.

Table 5-4 Formulaic clusters used by Rose, Mark and QD in comparison to all other authors

Formulaic clusters used by Rose	*Clusters occurring in QD*	Formulaic clusters used by Mark	Total authors using formulaic cluster
A COUPLE OF			11
A LOT OF			8
A WAY I			5
AND I WAS	*AND I WAS*		16
AS I WAS			12
		AT THE SAME TIME	5
BUT I KNEW			1
BY THE TIME	*BY THE TIME*		9
I KNEW THAT			10
I REALLY FELT			1
I THINK THE			5
I WAS GLAD			2
I WAS GOING			10
I WAS SO	*I WAS SO*		10
IN A WAY			8
IN A WAY I			4
	IN THE END	IN THE END	9
		IN THE SAME	6
IT WAS A	*IT WAS A*		18
LOOKING FORWARD TO			5
MADE ME FEEL			5
		ME AND MY	6
ME IN A			1
THAT I WAS			17
		THE SAME TIME	7
THE WHOLE THING			5
WAS GOING TO			9
		WENT TO MY	1
WHEN I WAS			18
WHICH I WAS			2

As can be seen from Table 5.4, 24 formulaic clusters were identified in Rose's texts, whilst only six were identified in Mark's texts and five formulaic clusters were identified in the Questioned Document. The first thing to notice is that Rose and Mark do not share any of the same formulaic clusters. This adds some weight to the argument that there is inter-author variation in the use of formulaic clusters. Secondly far fewer formulaic clusters were identified for Mark than for Rose. This is partly as a result of research design and is clearly related to the texts analysed (cf. below and Sections 5.3.3—5.3.4 for further testing on different texts). Referring back to Table 5.3, it is evident that nine formulaic clusters were originally identified for Mark, based on five texts. Here, since one of Mark's texts has been selected as a Questioned Document, only four 'Known Documents' were available for analysis, which explains why fewer formulaic clusters were identified than previously.

Given that only five clusters were identified in the Questioned Document and that four are formulaic for Rose and one is formulaic for Mark, it is unlikely that persuasive evidence can be found for authorship. However, the fact that they are formulaic clusters for an author only means that they are used frequently (at least once in three texts) for that author, not that they are used exclusively by that author. In other words, in line with Solan and Tiersma (2005: 156), the distinctiveness of a feature needs to be assessed in relation to other authors. This is shown in the fourth column in Table 5.4. With the benefit of 18 other authors with whom to compare the texts, it is possible to show how many of the 20 authors also used the identified formulaic clusters in their texts. Note, though, that the occurrence could be as low as once across all five texts produced by an individual author, so the claim is not necessarily that the cluster is also distinctive, or even formulaic, for them; rather, that it is also available in their lexical repertoire. A summary of the salient points is shown in Table 5.5.

Table 5-5 Relative significance of formulaic clusters in comparison to other authors

Formulaic cluster	**Significance**
AND I WAS	Used by 16 authors
BY THE TIME	Used by 9 authors
I WAS SO	Used by 10 authors
IN THE END	Used by 9 authors
IT WAS A	Used by 18 authors

Viewed in this light, it can be seen that whilst Rose shares the majority of the formulaic clusters isolated in the Questioned Document (rather than Mark), they do not seem to offer any discriminatory power since all of the formulaic clusters are used by several other authors—almost 50% in each case with *and I was* and *it was a* being used by 80% and 90% of the authors respectively. Therefore, no attribution is possible, and nor is it possible to exclude either author as a potential author of the Questioned Document. It is important to acknowledge though that if an attribution had been based purely on the quantity of 'matched' formulaic clusters, the wrong attribution would have been made with Rose looking like the more likely author.

At this stage, it is necessary to consider the fact that five texts produced by Rose have been compared against four texts produced by Mark and that the extra text available for analysis in Rose's set of Known Documents may well have skewed the results. The point was made above that using fewer texts reduced the quantity of formulaic clusters identified for Mark. Therefore reducing the number of Known Documents written by Rose should also affect the outcome of the qualitative analysis and forms the next part of testing the method. *PASW Statistics* was instructed to randomly select one text from Rose. Her second text was selected and was removed from the pool of Known

Documents resulting in four texts by Rose, four by Mark and one Questioned Document. The formulaic cluster analysis based on these texts is presented as Table 5.6.

Table 5-6 Formulaic clusters used by Mark and Rose in comparison to QD (4 Known Documents each)

Formulaic clusters used by Rose	*Clusters occurring in QD*	Formulaic clusters used by Mark	Total authors using formulaic cluster
A COUPLE OF			11
AND I WAS	*AND I WAS*		16
		AT THE SAME TIME	5
BY THE TIME	*BY THE TIME*		9
I REALLY FELT			1
I THINK THE			5
I WAS GLAD			2
I WAS GOING			10
	IN THE END	IN THE END	9
		IN THE SAME	6
LOOKING FORWARD TO			5
		ME AND MY	6
THAT I WAS			17
		THE SAME TIME	7
THE WHOLE THING			5
WAS GOING TO			9
		WENT TO MY	1
WHEN I WAS			18

As predicted, the number of Rose's formulaic clusters was significantly reduced from 24 to 12 and as a consequence, two of the clusters which occurred in the Questioned Document are discounted. The result is that there are now only two of Rose's formulaic clusters to place against the one for Mark. This in no way clarifies or otherwise strengthens/weakens the conclusions reached above but simply reduces the data on which conclusions can be based. This reinforces the position of forensic linguists that more data (i.e. more and longer texts) enable stronger conclusions and, more importantly for this method, it appears that data sets should be similar in size to enable more valid comparisons.

So far, formulaic clusters which occur in five texts and four texts have been identified and no attribution was possible. It may be the case that formulaic clusters do still hold potential to be diagnostic of authorship, but that a larger set of candidate authors is required to make differences more apparent. The next investigation tests this assertion.

5.3.3 Attributing a questioned document: five candidate authors

Using the random case selection function in *PASW Statistics*, five authors were now selected for the analysis: Keith, Jenny, Sue, Michael and Judy. Of the 25 texts they produced, *PASW Statistics* was again used to randomly select one text to act as the Questioned Document. The first text produced by Jenny was selected. Since this left Jenny with only four texts for comparison, and taking into account the findings from Section 5.3.2, the first text for all of the other authors was also removed from the analysis so that four Known Documents were available for each author.

In Section 4.2.1 (p. 77), the definition of the formulaic cluster was provided which stated that clusters need to occur in the majority of texts and that just how many texts this equates to would vary depending on how many are available for analysis. In this investigation, four texts for each author are available for analysis and so the threshold could be lowered to clusters which occur at least once in two texts which would certainly generate more formulaic clusters. However, this would lead to the identification of a range of clusters which occur at least once in only 50% of an already small range of texts, so the decision was made to firstly test the method with a threshold of occurrence set to at least once in three texts. A smaller range of formulaic clusters will be identified, but stronger evidence of formulaicity based on recurrence can also be argued as a result of this decision.

The set of formulaic clusters for each of the authors, as identified in Table 5.3, was refined to remove any clusters that were identified on the basis of their occurrence in Text 1. This left a subset of formulaic clusters which occurred in at least three of the four texts, as shown in Table 5.7, organised alphabetically by author (the significance of these clusters in relation to the author corpus is shown in Table 5.10):

Table 5-7 Formulaic clusters in four known documents (five candidate authors)

Author	Formulaic clusters in at least 3 files	In *N* files	Excluding Text 1
Jenny	AND AS A RESULT	3	2
	AND I WAS	4	4
	AS A RESULT	3	2
	AS I WAS	3	2
	I WAS REALLY	3	2
	IN THE END	4	3
	SOME OF THE	3	2
	THAT I WAS	3	2
	THOUGHT IT WOULD	3	2
	WAS WHEN I	3	3
	WHEN I WAS	3	3
Judy	A COUPLE OF	4	4
	I DON'T KNOW	3	3
	I WENT TO	3	3
	IN THE END	3	2
	THAT I WAS	3	2
Keith	BE ABLE TO	3	2
	I HAD BEEN	4	3
	PEOPLE IN THE	3	2
	WAS GOING TO	3	2
Michael	ENJOYING EACH OTHER	3	2
Sue	AT THE TIME	3	3
	BACK INTO THE	3	3
	I COULD NOT	3	3
	I DID NOT	3	3
	I WAS NOT	3	2
	IT WAS NOT	3	2
	THAT I WAS	3	2
	TO BE THE	3	2

From Table 5.7, it can be seen that with the exception of Michael, at least one formulaic cluster was identified for all authors. Since the threshold was set at occurrence in three texts, once Text 1 was removed, the following 12 formulaic clusters were available for analysis: *I had been, and I was, in the end, was when I, when I was, at the time, back into the, I could not, I did not, a couple of, I don't know,* and *I went to.*

The Questioned Document was searched for each of these clusters, but only one cluster was found: *in the end*, which is a formulaic cluster for Jenny. Whilst it is true that Jenny is the author of the Questioned Document, the occurrence of this one formulaic cluster is certainly less than persuasive as evidence of authorship, although only two other authors in the corpus actually used this cluster.

Therefore, whether or not *in the end* is formulaic (discussed below, Section 5.4.1, p. 122), this cluster does show how rarity may be used as a feature in authorship analysis, particularly since it is used by only three authors.

This invites the question of how effective the method will be if the threshold *is* lowered and formulaic clusters that occur in just two texts out of four are identified, taking into account the possibility that more clusters which can less readily be argued to be formulaic may be identified. The results are displayed in Table 5.8, which is organised alphabetically by author. The third column shows how many of each authors' texts contained a specific formulaic cluster, which is typically two texts. The fourth column shows how many times each author used a particular formulaic cluster across their four texts and crucially, the fifth column shows whether the formulaic clusters also occur in the Questioned Document, and if so, how often (in shaded cells).

Table 5-8 Formulaic clusters in four known documents (five candidate authors, lowered threshold)

Author	**Formulaic clusters in at least two files**	**In *N* files (excluding text 1)**	**Total occurrences across four texts**	**In Questioned Document**
Keith	AND I WAS	2	2	No
	BE ABLE TO	2	2	
	GOING TO BE	2	2	
	I COULD NOT	2	2	
	I HAD BEEN	3	4	
	I WENT TO	2	3	
	PEOPLE IN THE	2	2	
	THE END OF	2	3	
	THERE WAS NO	2	2	
	WAS GOING TO	2	2	
	WAS GOING TO BE	2	2	
Jenny	AND AS A	2	2	1
	AND AS A RESULT	2	2	1
	AND I WAS	4	4	0
	AS A RESULT	2	3	1
	AS I WAS	2	2	2
	BE ABLE TO	2	2	0
	I HAVE EVER	2	2	0
	I WAS IN	2	3	0
	I WAS REALLY	2	2	1
	IN THE END	3	3	1
	ONE OF THE	2	3	0
	SOME OF THE	2	2	1
	THAT I WAS	2	2	2
	THE FACT THAT	2	2	0
	THOUGHT IT WOULD	2	2	1
	WAS WHEN I	3	3	0
	WHEN I WAS	3	3	0

Judy	A COUPLE OF	4	6	0
	AND I WAS	2	2	0
	I DON’T KNOW	3	3	0
	I HAD TO	2	3	1
	I WAS SO	2	2	0
	I WENT TO	3	4	0
	IN THE END	2	2	1
	OUT OF THE	2	3	0
	THAT I HAD	2	2	2
	THAT I WAS	2	2	2
	THE NEXT DAY	2	2	0
	THE REST OF	2	2	0
	WHEN I WAS	2	3	0
Michael	A COUPLE OF	2	2	No
	ENJOYING EACH OTHER	2	2	
	IT WAS A	2	2	
	ONE OF THE	2	2	
Sue	AT THE TIME	2	4	0
	BACK INTO THE	2	2	0
	I COULD NOT	3	3	0
	I DID NOT	2	2	0
	I HAD TO	2	3	1
	I WAS NOT	2	2	0
	IT WAS NOT	2	3	0
	ONE OF THE	2	3	0
	THAT I WAS	2	2	2
	THE TIME I	2	4	0
	TO BE THE	2	2	0

As can be seen from Table 5.8, a total of 56 formulaic clusters were identified. As would be expected, this is obviously a greater number now that the threshold has been lowered, although it is perhaps noteworthy that the increase in formulaic clusters does not remain stable, as shown in Table 5.9:

Table 5-9 Difference in formulaic clusters identified depending on thresholds

Author	Formulaic clusters (occurrences in three texts)	Formulaic clusters (occurrences in two texts)
Michael	1	4
Keith	4	11
Judy	5	13
Sue	8	11
Jenny	11	17

Table 5.9 starts with Michael, who uses the fewest formulaic clusters and shows that when the threshold is set at occurrences in three texts, only one formulaic cluster is identified, and when the threshold is lowered to two texts, four formulaic clusters are identified. The greatest user of

formulaic clusters, Jenny, uses 11 when the threshold is three texts and 17 when the threshold is lowered. Sue is the second highest user of formulaic clusters when the threshold is three texts, but when the threshold is lowered, she becomes the joint second lowest user. The increase is also not stable across these authors, with Michael and Sue having an extra three formulaic clusters identified with the lower threshold, whilst Keith has an additional seven identified, and Judy has an extra eight. The key point then is that altering the threshold will have a significant impact on any analysis.

Returning to the attribution problem, of the 56 formulaic clusters, 11 occurred in the Questioned Document. Two of the authors: Keith and Michael shared none of the clusters with the Questioned Document. This reduces the closed-set of candidate authors from five to three. The next stage is to assess the significance of the clusters for the remaining three authors, summarised in Table 5.10 below:

Table 5-10 Relative significance of formulaic clusters shared between three authors and QD

Author	***N* formulaic clusters shared with QD**	**Shared formulaic clusters**	***N* occurrences in QD**	**Total occurrences across author corpus**	**Used by *N* authors**
Jenny	9	*And as a*	1	3	1
		And as a result	1	3	1
		As a result	1	6	3
		As I was	2	21	12
		I was really	1	7	3
		In the end	1	20	9
		Some of the	1	9	7
		That I was	2	45	17
		Thought it would	1	3	1
Judy	4	*I had to*	1	36	16
		In the end	1	20	9
		That I had	2	43	16
		That I was	2	45	17
Sue	2	*I had to*	1	36	16
		That I was	2	45	17

Table 5.10 is organised alphabetically by author. The second column shows how many formulaic clusters are shared with the Questioned Document. Jenny shares the most with a total of nine, whilst Sue only shares two. The third column lists the actual formulaic clusters that occur in the Questioned Document and the figure in the fourth column shows how many times each formulaic cluster occurs in the Questioned Document. On the right hand side of the table, the fifth column shows how many times each particular formulaic cluster occurs across the entire author corpus (for example, *that I was* occurs 45 times whilst *and as a* occurs only three times) and the final column shows how many

of the 20 authors use that formulaic cluster (although, obviously, these are not formulaic for other authors and this column is instead useful for gaining a sense of distinctiveness).

It seems that Jenny shares more formulaic clusters with the Questioned Document. However, the first three of Jenny's clusters are essentially derived from the same phrase "and as a result". If those three formulaic clusters are counted as one, this gives Jenny a score of seven, which is still higher than the other authors. Next to consider is the formulaic clusters in relation to the author corpus. All of the formulaic clusters used by Judy and Sue occur frequently across the corpus. For example, *that I was,* a formulaic cluster for all three authors, occurs 45 times and is used by 17 of the authors. Likewise, *that I had*, a formulaic cluster for Judy, occurs 43 times and is used by 16 of the authors. On their own then, these formulaic clusters are likely to reveal very little about authorship—they have been demonstrated to be consistent through this process, but demonstrating distinctiveness has failed. However, in addition to Jenny sharing the most formulaic clusters, three of those look particularly distinctive: *and as a, and as a result* and *thought it would*. Of course, counting *and as a* and *and as a result* as one formulaic cluster, still leaves these two, which are used only three times across the entire corpus and are only used by one author: Jenny. This makes these two formulaic clusters consistent and distinctive. This may provide some grounds for claiming Jenny to be the author, from an initial closed set of five candidate authors.

5.3.4 Reversing the process: identifying clusters in the Questioned Document

One final approach may still enable a Questioned Document to be attributed to its author and that is by reversing the process—in other words, identifying clusters in the Questioned Document and then searching the candidate authors' texts for similarities. To this end, the same authors and texts were used as in Section 5.3.3. All clusters of between 3 and 6 words which occurred at least twice in the Questioned Document were identified, of which there were 35. The texts by the candidate authors were then searched for the same clusters, of which 16 were matched as shown in Table 5.11:

Table 5-11 Clusters shared between QD and Known Documents

QD Clusters	Jenny	Judy	Keith	Michael	Sue
WAS TRYING TO	1				
WHAT TO DO	1				2
I WAS STILL		3			1
AS I WAS	2	1			1
AS I WAS STILL		1			
I HAVE BEEN		1		1	
I KNEW I			2		1
I KNEW THAT			1		1
I WAS TRYING	1				
I WAS TRYING TO	1				
IT WAS PROBABLY		1			
KNEW I WAS	1				
ON THE PHONE			1		
THAT I HAD	2	2	1		2
THAT I WAS	2	2		1	2
WHAT I WAS					2
Tokens in common with QD	11	11	5	2	12
Types in common with QD	8	7	4	2	8

As Table 5.11 shows, the four Known Documents produced by Jenny and Sue shared eight clusters in common with the Questioned Document with Judy sharing seven. Keith shared four and Michael shared the least, with only two clusters in common. On this basis, it may be possible to exclude Keith and Michael as the authors of the Questioned Document, thereby reducing the pool of candidate authors to three, although it would not be possible to eliminate Jenny, Judy or Sue. The same conclusions can be reached by examining how many cluster tokens are shared between the Known Documents and the Questioned Document. Of course, the point needs to be made that whilst this may be helpful, it cannot be claimed that these clusters are formulaic—certainly not in accordance with the definition of formulaic clusters or idiolectal formulaicity used in this research—they are instead ngrams that actually do not occur that frequently (i.e. *I was still* occurs three times in Judy's four Known Documents whilst the remaining clusters occur only once or twice). Ngrams have received attention as a lexical feature of authorship in their own right (as established in Section 5, p. 100) and pursuing this particular line of investigation falls outside the scope of this research.

In assessing the final hypothesis, very limited, tentative support can be provided in determining whether a Questioned Document can be correctly attributed to its author based on the occurrence of formulaic clusters and it should be acknowledged that through this process, the method has been tested with the aim of trying to obtain positive results in order to test the procedure. It may be more cautiously stated that rather than safely attributing a text to its author, the method may enable a larger pool of candidate authors to be narrowed. What must be borne in mind, when evaluating this research, is that no forensic linguist would attribute a Questioned Document to an author with any certainty, based on the occurrence or absence of just one feature in isolation. This explains why the results of this stage of the analysis are so tempered, and a stronger attribution to an author would likely be more possible if other established markers of authorship were also taken into consideration (for example, see Eagleson, 1994).

5.4 Discussion

The method reported in this chapter attempts to do something slightly different from previous investigations which explore the relationship between clusters and authorial style. Rather than simply identifying clusters, a decision was made to focus only on those clusters which can be argued to be formulaic for an author because of their recurrence across a minimum threshold of texts and these formulaic clusters were assessed for distinctiveness in comparison to other authors. Using the Jaccard's coefficient statistical test proved successful in demonstrating that formulaic clusters were consistent and distinctive for authors, leading to the conclusion that authors can be differentiated based on their use of formulaic clusters. In this regard, support could be provided for the first and second hypotheses. However, it became more difficult to actually attribute a Questioned Document to its correct author through the ensuing descriptive approach; a situation which became further compounded when fewer texts were available for analysis. Therefore, whilst it would be wrong to suggest that support was definitely provided for the third hypothesis, some areas of commonality between the Questioned Document and the texts produced by the candidate authors could be identified. It is this potential in the method which may prove to be a foundation for future research into this special type of cluster.

As expected, reducing the number of texts available for analysis means that fewer formulaic clusters are identified. The significance of this is that the method outlined in this chapter will clearly carry more investigative value if larger data sets are available for analysis and it is perhaps not a suitable approach for those investigations where fewer/shorter texts are available. Whilst it may not be possible to speculate about the ideal number of texts that would be required to make the method more robust, it is important to note that no reliable predictions could be made about which particular clusters might occur in another random text, since no formulaic cluster was used sufficiently frequently or regularly. If any of the formulaic clusters had occurred more than once in all five of the texts for an author, there may be grounds to predict that the cluster would also occur in a sixth, seventh or *n*th text also by that author. But since this situation did not occur, the fact that the next best formulaic clusters, those occurring in four out of five texts, already means that 20% of the texts produced by an author will not contain that cluster.

Likewise, it is likely that the length of the texts available for an investigation will affect analysis based on this method. Whilst Grant (2011) had considerably more texts available during his analysis (407 texts produced by known authors), it is less likely that a feature such as formulaic clusters would have had sufficient opportunity to manifest in short-form communications such as SMS text messages. A forensic linguist may therefore first need to appraise their data and then decide whether a method such as this is appropriate to use. The method cannot therefore claim to be applicable to all types of texts.

One final issue that has not received attention in this chapter since it falls outside the scope of the research is the actual number of formulaic clusters that were identified for each author—only those formulaic clusters that occurred in a Questioned Document have been explored in detail. However, should any significance be attached to the fact that 26 formulaic clusters (based on at least one occurrence in three out of five texts) were identified for Rose, whilst only one was identified for Michael, or 12 for Elaine but only four for Sarah (see Table 5.3)? It is likely that this level of recurrence would create the sense of a repetitive style for Rose and presumably more novel language and less repetition for Michael. This in itself may provide a useful avenue for future investigation.

5.4.1 Is the method valid?

The case has been made in this chapter that formulaic clusters are valid as evidence of formulaic sequences since they recur across a series of texts; they therefore hold potential to be pre-fabricated in these particular forms, ready for use when required. In other words, the authors have found them to be communicatively useful. Whilst some of the formulaic clusters may appear to be quite acceptable as evidence of formulaic sequences (e.g. *the whole thing, the next day, as a result, in the end, all the time*), others, due to their semantic incompleteness, appear less so (e.g. *it was a, and I just, to go to, out of the, me and my*). There are certainly features in common with previous research into formulaic language. Notably, Wray (2002) and those who use the *formulaic sequence* as their definition of choice, do not see the lack of meaning (in other words, the fact that the units are incomplete) as a problem. Therefore, the fact that formulaic clusters such as *it is a*, *and I was*, and *I was really* are incomplete does not preclude them from being formulaic. They are, though, certainly less intuitively satisfying. A stronger argument for the classification of these clusters as formulaic is based on the frequency approach to formulaic language (cf. Section 3.5.2, p. 70). That is, they occur over a certain threshold for a particular author and can therefore be argued to be formulaic for a particular individual based on their recurrence in texts. In other words, the individual has found a particular formulaic cluster which enables them to express their meaning, or produce cohesive discourse, in a way which operates best for them. In this way, formulaic clusters can be argued to be formulaic sequences. However, claiming that a formulaic cluster is a formulaic sequence just because it occurs in at least three texts is clearly arbitrary and entirely different results will be obtained if this threshold is changed. The key point perhaps lies less in whether the formulaic clusters identified in this research can be argued to be formulaic sequences and more in their diagnostic potential as a tool for the forensic linguist.

5.4.2 Is the method reliable?

Quite whether the method outlined in this chapter is reliable is not clear. There are elements of the analysis which are automated, and as such, are completely replicable leading to high reliability. However, an element of linguistic decision making is also introduced which can only be performed by a linguist as opposed to a computer, and is therefore subject to the usual fallibilities (cf. Section 3.5.1). The most obvious example is the decision to exclude formulaic clusters which *appear* to be context-specific. Furthermore, whilst demonstrating distinctiveness could be achieved through an automated approach, a forensic linguist is required to assess the output in order to make an attribution. Whilst a skilled forensic linguist may well reach the same conclusions time and time again, there is nothing to mitigate against the possibility that other forensic linguists might reach different conclusions (e.g. Finegan, 1990).

5.4.3 Is the method feasible for forensic purposes?

It has already been established that the method outlined here is unlikely to be universally applicable. Therefore, selecting formulaic clusters as a marker of authorship will not be appropriate for every type of forensic investigation (but then, neither is selecting other text-specific features such as initialisms in text messages which would be inappropriate for academic essays). It is also important to point out that the specific formulaic clusters identified here will not be applicable to other cases. In other words, the fact that 93 formulaic clusters were identified and used in the analysis does not mean that the same 93 formulaic clusters can be used in other forensic investigations. Clearly, the selection of formulaic clusters will rest entirely on the texts available for analysis, and the clusters occurring within.

In light of this, the bigger issue raised is whether the formulaic clusters identified based on the five texts available for each author would also be identified if a different set of texts by the same authors had been used. Since it is claimed that these clusters are formulaic, it is hoped that they would, but this is unlikely since any text will only exemplify a very small number of formulaic clusters. In reality, the fact that many of the clusters occurred in no more than two texts by the same authors leaves three texts remaining where the formulaic cluster was not used. If the Questioned Document happens to be one of those texts where the author does not use a particular (set of) formulaic clusters, then the analysis is no longer applicable (speaking of course only about the qualitative attribution phase, since, as explained earlier, the justification for using Jaccard's co-efficient is that non-occurrence is not taken into consideration). It is therefore unlikely that this method, in its current state, is forensically robust. It is also not an insignificant fact that the texts used in this investigation are short and so the range of formulaic clusters which have had the opportunity to manifest may be limited. Just as lowering the threshold increased the number of formulaic clusters that were identified, presumably, increasing the lengths of the texts would also lead to a larger quantity of formulaic clusters.

In conclusion, the method has shown some promise at demonstrating consistency and distinctiveness between authors, but far more research into the area would be required before the method is forensically robust and, as an initial suggestion, longer texts may need to be the focus of future research.

5.5 Conclusion

The analysis presented in this chapter adopted a corpus-driven approach and it was argued that clusters which recur in separate texts by one author may be argued to be formulaic. However, this method only explores inter-author variability and does not accommodate intra-author variability. That is, the identification of formulaic clusters relies on authors producing the same fixed forms across their texts, so any variability that an author demonstrates has to remain both consistent and with a frequency of occurrence high enough for it to be identified. The next chapter therefore adopts a completely different approach to the identification of formulaic sequences in texts, namely, a qualitative approach which allows for variation to be captured in a systematic and principled way, rather than a quantitative approach such as that adopted in this chapter which relies on the consistency and rigidity of fixed forms.

Chapter 6

‘Dig *way* down deep’: using a core word to identify formulaic sequences

In Section 4.2.2 (p. 79), the second analytical procedure was proposed—using a core word (i.e. a word frequently embedded in formulaic sequences) to directly identify a range of formulaic sequences. The word *way* was argued to be the most appropriate core word to investigate. It was also argued that to fully appreciate the significance of how often specific formulaic sequences are used by a given author, it is essential to also know how frequently other authors use the same formulaic sequences, and crucially, whether they use any alternatives. After all, it cannot be argued that one author does not use a particular formulaic sequence if they simply do not have reason to express that particular meaning. Therefore, two approaches are outlined in this chapter. The first approach assesses whether any of the authors appear to have preferences for formulaic sequences based around the core word *way*. The second approach then attempts to establish whether, on the occasions that the authors have reason to express the same meaning, they use the *way*-phrases, or alternatives that do not include this core word; however, it is important to stress that the aim of this chapter is to test a method, rather than to be exhaustive. Through using these two approaches, a contrast can be made with the previous chapter, where rigidity of form was essential to argue formulaicity; by contrast the method presented in this chapter enables more variation. After a description of the methods and results for both of the approaches, the efficacy of these methods in the forensic context will be considered, focussing on the key issues of validity, reliability and feasibility.

6.1 Approach 1: authorial preferences for specific *way*-phrases

As established in Section 4.2.2, since *way* is expected to form part of numerous formulaic sequences, the first approach seeks to establish if this is in fact the case and, if so, whether authors demonstrate a preference for certain formulaic *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 idiolectal. Three hypotheses inform this approach:

1) It will be possible to identify a range of formulaic sequences which have the word *way* at their core;
2) Authors will have consistent preferences in the *way*-phrases that they choose from the available set; and
3) As a consequence, it will be possible to distinguish texts produced by one author from those produced by another based on the occurrence of *way*-phrases.

6.1.1 Method

Using *WordSmith Tools* (Scott, 2008), *way** was entered as the *node* (that is, the term used in corpus analysis for the word being investigated which is usually displayed in the centre of the concordance line (Hoey, 2005: 4—5)). From the 100 text author corpus 105 occurrences were extracted (94 instances of *way* and 11 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:

1	good food and my father singing 'My	**way**	' on the karaoke. It was a typical party
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 either *way* or *ways* that would need to be removed if an alternative formulation were to be used instead. Five examples are provided for illustrative purposes:

3	chronic diarrhoea and I drove all the	**way**	down to Oxford (where he lived at the time)
4	of my masters, it is linked in several	**ways**	, and the experience and life experience gained
5	that Santa doesn't exist. I suppose in a	**way**	I must have done, as when I was younger
6	120 miles north of Liverpool a long	**way**	from Deeside and when John got a job in
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 entirely 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	I knew that I was standing right in the	**way**	. What I didn't know was that the driver was

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 to each other than the word *right*, which is more likely to be an optional addition, although admittedly an important one included for rhetorical effect. All the 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 times per 1,000 words). The frequency of *way* across each author sub-corpus per 1,000 words is shown in Table 6.1:

Table 6-1 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. John and Sue) while some use *way* considerably more (e.g. Alan, Jenny, June, 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 than in the BNC.

6.1.2 Results

The 103 instances were made up of 55 different phrases. The range of phrases used is presented in Table 6.2 (organised from most frequent to least frequent), alongside their total frequency across the corpus and the number of authors who used a particular phrase. All 20 authors used at least one phrase.

Table 6-2 55 *way*-phrases identified in the 100 text author corpus

Formulaic Phrase	Frequency across entire corpus	*N* authors using phrase
in a way	19	8
the way	6	4
way	6	4
all the way	4	4
on the way	3	1
the only way	3	2
way of Xing	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
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
a certain way	1	1
a long way	1	1
along the way	1	1
any other way	1	1
any way	1	1
by the way	1	1
either way	1	1
for ways to	1	1
gave way	1	1
get out of the way	1	1
go out of my way to	1	1
half way	1	1
in a different way	1	1
in a roundabout way	1	1
in any serious way	1	1

in any sordid way	1	1
in many other ways	1	1
in many ways	1	1
in several ways	1	1
in some way	1	1
in such a kind way	1	1
in such a way	1	1
in the way	1	1
let's put it that way	1	1
make their way	1	1
making his way	1	1
on my way	1	3
one way or the other	1	1
some ways	1	1
the exact way	1	1
the only way to	1	1
the other way around	1	1
the rest of the way	1	1
the ways	1	1
the whole way	1	1
ways	1	1
worked my way	1	1

The first important observation is that none of the phrases was used by every author. From Table 6.2, it can also be seen that not only are the majority of the phrases used by only one author but they also occur only once (e.g. *in a roundabout way, some ways, the other way around, the whole way*). By contrast the phrase *in a way* is used by eight authors and occurs 19 times. As such, this phrase requires further investigation. Table 6.3 below shows which authors use this phrase, how frequently, and in how many of their texts:

Table 6-3 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

As Table 6.3 shows, *in a way* is used only by Rose 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, this phrase may have significance as a marker of authorship for Rose. In the BNC there are only 2,751 occurrences of *in a way*. As such, the formulaic sequence *in a way* actually appears to be relatively rare which adds more significance to the fact that Rose uses it consistently and frequently in comparison both with the other authors and with 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 idiolect.

As an additional measure, phrases were grouped and reduced to their underlying structures (e.g. *in a/an 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*). Again, no patterns emerged across the entire corpus or for any individual author sub-corpus.

6.1.3 Discussion

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 20 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. This provides support for the first hypothesis, that it is possible to identify a range of formulaic sequences by using the core word *way*. This also suggests that the selection of *way* as core word is justified and bodes well for the second approach.

With reference to the second hypothesis, that authors will have consistent patterns in the *way*-phrases that they use, there is some, but very limited, support. One author out of 20 used a *way*-phrase across all five of her texts. Since Rose used the formulaic sequence *in a way* in all five texts a total of ten times, it is possible to argue support for the second hypothesis, if only for this author. Following on from this, the third hypothesis also receives very limited support—the collection of five 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 the forensic context. Therefore, these findings need to be confined to this particular research context; that is, additional research on longer and many more texts would certainly be desirable.

Given that with the exception of *in a way*, no other phrase and no other author comes close to exhibiting any kind of pattern, the second approach seeks to look beneath the forms that these particular *way*-phrases take, and instead focuses on the meanings that are conveyed.

6.2 Approach 2: identifying alternatives to *way*-phrases

It was established through Approach 1 that focussing only on *way*-phrases may be limited since alternative realisations of the same semantic content will not be identified. 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 meanings similar to those encoded in individual way sequences should be the next focus. The rationale for this approach is described by Wray (2002):

> [R]aw frequency is not an adequate measure of formulaicity. 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—1).

That authors can prefer different formulations to convey similar meanings is well-recognised and in addition to the quotation from Wray, above, Cortes (2004) too raises the question of different forms being used: “It would be interesting to focus on the language forms students use instead of these expressions [academic lexical bundles] to convey these functions” (p. 414). Outside of the formulaic language literature, the same question has been raised, 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. Approach 2 therefore aims to contribute to this research gap by assessing how robust formulaic sequences extracted using the core word *way* are for differentiating texts produced by different authors. By ‘tagging’ sequences for their meaning, 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 attribute texts to their authors based on this method. Three hypotheses guide the investigation:

1) Specific meanings will be expressed in different forms, some based around the core word *way* and others around a different set of words;
2) Authors will have preferences for the forms they use to express specific meanings; and
3) Such form selections will be consistent across a series of texts and may therefore be idiolectal.

6.2.1 Method

As a product of Approach 1, there existed a list of the *way*-phrases that each author used. The next stage was to produce a gloss for each *way*-phrase in order to determine the exact meaning being conveyed. A selection of nine *way*-phrases (underlined) which occurred a total of 26 times in the author corpus is presented below organised under four clearly discernible glosses:

=do more than necessary/expected (1 occurrence)

9	caring 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	said that 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	outright but this made it worse in many	**ways**	as he was searching for an excuse and
13	field of my masters, it is linked in several	**ways**	and the experience and life experience
14	to fill this gap and while this is some	**ways**	positive it may mean significant changes

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

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

In total, 29 different glosses were derived from the 103 *way*-phrases. For each of the glosses, a series of synonyms were extracted from the dictionary and thesaurus components available through *Oxford Reference Online* (2010). Drawing again on the examples provided above (lines 9—17), Table 6.4 shows the synonyms that were identified (quite whether these are in fact synonyms, or even near-synonyms is discussed in Section 6.3):

Table 6-4 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 6.4, based on these synonyms, a series of nodes with which to search the corpus were created. Many of the nodes recurred throughout the process. For example, in the first row of Table 6.4, the node *required* occurs twice. Duplicates were therefore removed. Through this process, 242 search nodes were identified that could potentially convey the same meaning as any one of the identified *way*-phrases. Using *WordSmith Tools*, a total of 2,458 concordance lines were extracted based on this list of nodes. 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 a similar meaning, the phrase was retained. The process for determining which words constituted the phrase was the same as that outlined in Section 6.1.1 (p. 126) 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.

6.2.2 Worked example

As a product of Approach 1, 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 6.3 for discussion), as indicated in the examples below:

18	situation, well at least never in a	way	that would ordinarily be thought of
19	girls could easily be behaving in such a	way	. After that incident it wasn't quite as
20	that 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 her fourth text (line 18) in the sense of '=in a certain manner, fashion'. 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 four phrases as being indicative of authorship—no other author uses these phrases to convey this meaning. However, before such a claim can be confidently made, the following need to be established: 1) whether any other authors expressed this meaning differently, since it cannot be argued that they did not use these phrases if they had no need to express the meaning, and 2) if they did express this meaning, what phrase did they use? The gloss for this meaning ('=in a certain manner, fashion') can be used to derive synonyms as the basis for identifying other phrases that express the same meaning in the entire author corpus. A selection of 25 concordances, organised alphabetically by node, are presented below. The near synonymous expressions which convey this meaning are underlined:

21	. I was the last cast member to arrive	as I did	not need any make up. I pulled on
22	Society and still went out as much	as I did	in the first two years (it's a wonder I
23	interested in. The more creative	aspects	of my life I decided to keep as
24a	most afford to drop, as was the	convention	in my school - I had decided that
25	feelings known to him or anything	like that	! Luckily, i think some people have
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 to
28a	on him but obviously I didn't see it	like that	. Unfortunately I wasn't
29a	I never knew I could betray someone	like that	. The next day I went over to Andy
30	Being lanky	means	there have been many
31a	I had achieved AABC - by no	means	bad results, but over the last few
32	bad lies. I will tell a lie if that	means	I won't hurt somebody's feelings
33	results 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 hours
35	.went in and saw her, because of the	nature	of the operation she was lying in
36	and used to call me names as they	regarded	me as one of the 'clever' people.
37	on my spine, I did cry, but carried on	regardless	. The kindness of the girls who
38	fact that he didn't have the decency,	respect	, courtesy or balls to tell me. This
39	fault." He was joking about it. I lost	respect	for him then. I texted him a while
40	of her mother's and my teacher's	respect	. I also argued with my friend who
41a	and 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 own
43a	me. I had no longer felt quite the	same	about the relationship for several
44	and was blurred at first. I was in a	state	of shock, I sat down and was unable

45a	back in. At this point I was in such a	state	that my sister ran out to save me

From the selection of 25 concordances presented in lines 21—45a, 16 can be discarded since they do not convey the same meaning as '=in a certain manner, fashion'. The remaining 9 concordances do express this meaning and can replaced with the following *way*-phrases, whilst still retaining a similar meaning:

24b	most afford to drop, as was the	*way*	in my school - I had decided that
27b	Josh wouldn't have wanted to exist	*in that way*	; to have been such a burden to
28b	on him but obviously I didn't see it	*in that way*	. Unfortunately I wasn't
29b	I never knew I could betray someone	*in that way*	. The next day I went over to Andy
31b	I had achieved AABC –	*In no way*	bad results, but over the last few
34b	Suddenly thankful for my hands-on	*way*	I took over and after two hours
41b	and 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 several
45b	back in. At this point I was in such a	*way*	that my sister ran out to save me

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 of this analysis are presented below.

6.2.3 Results

From the 2,458 concordance lines generated from 242 nodes, a total of 141 concordances contained words or expressions which were considered to be alternatives or near-synonyms for one of the *way*-phrases identified through Approach 1. When these 141 alternatives are added to the 103 *way*-phrases, 29 different meanings were expressed a total of 244 times across the entire 100 text author corpus. All of the *way*-phrases and alternative expressions were plotted on a grid to enable clear cross-referencing. The grid is reproduced as Appendix E. Table 6.5 below, organised according to frequency of occurrence, summarises how many times each meaning occurred in the corpus, along with how many authors expressed that meaning:

Table 6-5 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 6.5 that the meaning '=to some extent, in some respects' occurs the most frequently, a total of 35 times, and is used by 11 of the 20 authors. The second most frequently occurring meaning, '=method, how to achieve an objective', occurs 31 times and is used by 14 authors. The third most frequent category, '=emphasis', occurs slightly fewer times, 29, but is used by slightly more authors, 15. At the bottom end of the table is a selection of meanings which are expressed only once in the entire 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 authorship 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 6.6:

Table 6-6 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 Xing

=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

As previously stated, by comparing all of the expressions used to convey all of the meanings identified by each of the twenty authors, it should be possible to determine whether 1) authors have a preference, and 2) how distinctive that preference is in comparison to other authors. In fact no preferences were found—indeed only two authors expressed the same meaning 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)* so there is no evidence of an authorial preference for him when expressing this meaning. 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 convincing pattern for her. However, in her fifth text, Rose also used the expressions *kind of, in that respect,* and *in the other sense*, along with *in a way* three times—in other words, although she does have some variation in the forms she uses to express this meaning, there is a predominant form, *in a way*.

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 6.5 and Appendix A): '=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 authorship 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 7 and 5 times respectively, so it may just be that there was insufficient data to explore alternatives.

In addition, some authors do seem to remain faithful to just one expression to express a specific meaning e.g. Rose only used *much* to express '=emphasis' and Thomas used *how* to express '=in a certain manner, how', exclusively across three of their five texts. Whilst they do not occur with particularly high frequencies (Rose used *much* once in each of three texts and Thomas used *how* six times across three texts), it may be this lack of variability which is marked for authorship, given that all other authors use at least two expressions. Again, more data would be required in order to argue this fact more convincingly and it should be remembered that this is very much an exploratory study.

6.2.4 Discussion

For forensic purposes, 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. With these results, it is now possible to return to the three hypotheses that informed this approach, restated for convenience below:

1) Specific meanings will be expressed in different forms, some based around the core word *way* and others around a different set of words;
2) Authors will have preferences for the forms they use to express specific meanings; and
3) Such form selections will be consistent across a series of texts and may therefore be idiolectal.

Hypothesis 1 is confirmed—as can be seen from Table 6.5 and Appendix A, there is 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. Hypotheses 2 and 3 receive the same level of very limited support as for Approach 1. The expression *in a way* again seems to be characteristic of Rose's idiolect by being both a preferential choice and a consistent choice.

In this chapter, a case study of *way* has been presented with two approaches to identifying potentially formulaic sequences being described. Both approaches have achieved the same very limited results—that is, for one author, patterns for how *way* is used have been demonstrated, but for all other authors, there is no evidence of authorial preferences.

6.3 Evaluation of the approaches

There are three key issues that need to be addressed in evaluating the approaches presented in this chapter. Firstly, is *way* one word with a range of different meanings, or are there numerous words which happen to be homonyms? Secondly, a set of alternative phrases were identified for Approach 2—are these alternative phrases really synonymous? Finally, what would be the effect of working with a larger, or indeed smaller, set of data? Each of these issues will be dealt with in turn.

Wray (2002) writes:

> In a standard dictionary, dozens of entries may be needed to capture all the different aspects of a word's meaning, and it is often difficult to judge just where to draw the line between one word having multiple, related meanings and there actually being two (or more) words which happen to be spelled and pronounced the same way. (p. 29)

This is certainly relevant in the current context, given that *way* has been assumed for these purposes to be one word which conveys many different meanings—29 in fact. Intuition alone suggests that 29 senses for a single word is considerable and it is more likely that *way* could be a homonym. After all, is *way* in the sense of conveying a location as in *I was stood half way between the door and the table* really the same word as *way* conveying that something has collapsed as in *my leg gave way and I fell over*? For the present purposes, they have been assumed to be so, but arguing the case in a court of law would require a far stronger conviction. Proposing that they are in fact two separate words which happen to look and sound alike immediately opens the door to the question of where the dividing line occurs, as Wray comments above, and is certainly a question that reaches far beyond the scope of this research. In a sense, however, even if *way* is a homonym, is it really problematic? Whilst such a question may provoke academic debate in linguistics circles, it is perhaps questionable whether the average person authoring a document would be sensitive to such differences, and if this proposition is accepted, does it matter where the dividing line is? Approach 1 revealed that *way* is part of numerous formulaic sequences and if they *are* formulaic, it follows that the constituent parts in those sequences have not been analysed individually and the authors themselves are unlikely to be aware of the senses of *way* being conveyed. To this end, the fact that Rose utilises *in a way* frequently and consistently is perhaps a more salient point than whether her use of *way* as part of that expression is the same as, for example, *in any sordid way* used by Sue since, by virtue of being formulaic, neither of these authors should have broken these sequences down into their constituent words—whatever those words may be. Nonetheless, a far larger corpus with far more diverse data would be required to reach a suitable answer to this question.

The second issue relates to synonymy. In Approach 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 comprehensive dictionary and thesaurus tools in *Oxford Reference Online* (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, was it possible to replace the *way* formulaic sequence with an alternative whilst still conveying a similar meaning? In this vein, is it really appropriate to argue that the formulaic sequence *in a way* to mean '=to some extent, in some respects' is 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 that "we may hypothesise that synonyms differ in the 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 surely 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. Again, these issues go beyond what is possible in the current research, but a more informed and principled analysis would be required before attempting to move these methods into a forensic context.

The final issue—that of the corpus itself—also warrants attention. The formulaic sequences 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 20 authors. Would using only four texts or as many as ten texts have made a difference? It is possible to speculate that this would be a very important factor, but where to put the cut off point for having the appropriate quantity of texts, per author and overall, seems entirely arbitrary. One way to establish this is to ascertain the frequencies of *way* on fewer texts in each author sub-corpus. This will determine whether having fewer data will significantly alter the results, as shown in Table 6.7 which 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 6-7 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 6.7 shows, as would be expected, that with fewer texts, so too are there fewer occurrences of *way*. More important 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 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 in this research, though, is whether using more texts would create the same effect. There is the possibility that authors' use of *way* stabilises over five texts, but there is no real reason to believe that this should be the case.

In spite of these important considerations, it should be borne in mind that this is a piece of exploratory research and so whilst the answers provided in these pages raise more questions than they answer, the important point is that at the time of writing, there is a lack of research which investigates this particular aspect of formulaic sequences in the forensic authorship context and so highlighting just a few of these issues is essential groundwork. And whilst such important questions cannot be merely pushed to one side and ignored as a result of these two approaches, the same result has repeatedly manifested itself—Rose does use *way* differently to the 19 other authors. How valid, reliable and forensically useful, therefore, are these results?

6.3.1 Are the methods valid?

Can the expressions highlighted through Approach 1 reasonably be termed 'formulaic sequences'? Are they valid as examples of formulaic language? In Section 3.3.1 (p. 54), the definition of the *formulaic sequence* was provided, a key criterion of which is that phrases are "processed holistically". 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 in Section 6.3 above, quite where the dividing line between the literal and the non-literal occurs is not clear.

Approach 2 is valid as far as frequency is accepted as a technique for identifying formulaic sequences (see Section 3.5.2, p. 70). If a phrase is used consistently and frequently, then there is potential for it to be formulaic. By identifying the range of possible expressions for specific meanings, it was possible to determine which expressions, if any, were favoured, and therefore, likely to be formulaic, for each author.

6.3.2 Are the methods reliable?

Reliability relates to replication—achieving the same results each time the approaches are attempted. The two approaches vary in their degree of reliability. The first approach is reliable because it is automated, that is, all concordances containing the node *way* were extracted. However, determining where the boundary between the formulaic sequence and the non-formulaic part of the message is less reliable and more open to subjective judgement. However, adopting a replacement principle (that the formulaic sequence comprised all the words that would need to be removed to maintain grammaticality) at least ensured consistency. The second approach, like the first, combined reliability with subjectivity. The synonyms were acquired from a reputable and reliable source, and again, their extraction from the data was automated, so subjective issues such as tiredness in the researcher were irrelevant and the same concordance lines would be extracted on all occasions. However, the decision regarding whether a phrase was in fact synonymous with the *way*-phrase was subjective, and therefore the results are open to debate.

6.3.3 Are the methods feasible for forensic purposes?

Taking into account the issues discussed above, it is unlikely that Approach 2 meets the standards required for a forensic investigation, or for producing evidence for a court of law. However, it is important to reiterate that this case study of *way* is itself investigative by offering research in a previously unexplored field. Therefore, although the method is not yet developed sufficiently for application in the forensic context, the results may provide a good foundation for future research which might generate analyses which are forensically applicable. It is certainly encouraging that both approaches achieved the same results indicating a level of support through triangulation. Approach 2 also offers a method which may be persuasive in the forensic context—by demonstrating not just how a particular author expresses a meaning, but also how all other authors in a sample express the same meaning adds more weight to the findings by drawing on robust comparative data. Therefore, with considerably more research, the combination of both approaches may be a useful indicator of authorship.

6.4 Conclusion

This chapter has described two approaches to identifying formulaic sequences in the author corpus. Each approach has its limitations, but it is intriguing that in each case, the same result was found for

Rose, that the phrase *in a way* may well be a distinctive marker for her. However, this chapter has presented *way* as a case study, and it is important to remember that regardless of the results, it would not be prudent to view *way* as a magic bullet—that is, there are other potential candidate core words (as described in Section 4.2.2) and 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 a full-scale investigation of a variety of different core words in order to determine whether other combinations of formulaic sequences provide more intriguing results. In line with Grant's (2010) view of idiolect (as described in Section 2.1.5, p. 29), it may be the combination of a variety of features that is more indicative of authorship than the patterns of usage for any one word, or one marker of authorship. Consistent combinations of formulaic sequences would certainly provide stronger evidence of authorship.

The methods so far investigated in this and the previous chapter have taken a very narrow view of formulaic sequences by focussing only on a small and limited subset of the potential formulaic sequence pool. The next chapter develops the lessons learned from this analysis by focussing on a far wider variety of potential formulaic sequences, identified in a different way, in order that the overall use of formulaic sequences in comparison to novel language may be investigated as being more convincing as evidence of authorship in a more valid, reliable and forensically feasible way.

Chapter 7

‘More clichés than you can shake a stick at’: adopting a reference list approach

In this final analytical chapter, rather than individual formulaic sequences being the unit of analysis, it is the quantity of formulaic sequences compared to novel language that forms the basis of the investigation. In Section 3.4.1, it became apparent that individuals are socialised differently and have a wide range of different life experiences and that this, in combination with their specific needs in handling language, affects their repertoires of, and indeed reliance upon, formulaic sequences. Therefore, since each of the authors that contributed data to this research will have a different range of cognitive abilities in handling language, some authors should have larger formulaic repertoires than others.

It is useful to consider at this stage that whilst each of the authors' texts were roughly the same size, there was a small amount of variation in length, so to cope with these differences, the measure used is a normalised count of the number of words which make up a formulaic sequence, per 100 words (henceforth ‘count’ for brevity). The advantage of using a single count is that it makes fewer assumptions about the nature of the data and so allows for a wider range of appropriate statistical tests to be used. By concentrating on the count of text that is formulaic, it will be possible to make claims about whether the language used by one individual author is more or less formulaic than that of another. If this is the case, the consistency in levels of formulaic sequences across all five of each authors' texts can be investigated. Finally, it will be possible to determine whether a given text can be successfully attributed to its author, as would be necessary in a case of forensic authorship attribution. To carry out this investigation, a series of hypotheses must firstly be proposed.

7.1 Aim and hypotheses

In this chapter, the central aim is to examine whether the count of formulaic sequences is sufficient to enable the correct attribution of a text to its author. Taking into account each author's individual abilities to produce language—in other words, their cognitive abilities—it is hypothesized that authors will use different counts of formulaic sequences from each other. Some authors may rely on more formulaic sequences to reduce cognitive burden (cf. Section 3.4.1., p. 56). Since each of the five personal narratives were collected over a five period day period, there should not have been any significant change in authors' formulaic sequence repertories, nor their encoding strategies and so by taking all five texts as a set, it can be predicted that each of the five texts should contain similar counts of formulaic sequences.

Secondly, if the first hypothesis is correct, it should be possible to differentiate authors based on the count of formulaic sequences that occur across the totality of their five texts. This is important for the forensic context in demonstrating that the variation between authors is significant.

Thirdly, the authors are expected to use differing counts of formulaic sequences and since those counts are hypothesized to be similar across a series of texts authored in the same period and in the same genre, any randomly selected text should be attributable to its author in a mock forensic authorship attribution case. In other words, finding support for the second hypothesis will not be sufficient for using the count of formulaic sequences as a marker of authorship on its own. It will also be important to demonstrate that a Questioned Document can be successfully attributed.

In summary, the following hypotheses will be tested:

i. Variation in the count of formulaic sequences in texts will be greater between authors than within authors;

ii. Authors will be potentially differentiable from each other based on the count of formulaic sequence usage;

iii. A randomly selected Questioned Document can usually be correctly attributed to its author based on the closeness between the count of formulaic sequences in the text and in the author's other four texts

7.2 Method

In Section 3.5.2 it became apparent that whilst a reference list of formulaic sequences may be an excellent resource in practical terms (e.g. analysis of large sets of data can be fast and reliable), caution needs to be expressed over which items are included in the list since without clear justification, there is the possibility that the list may be nothing more than the intuitions of one individual. A compromise was proposed that satisfies the needs of exploratory research and which holds the potential to conform to criteria governing the admissibility of expert evidence. The compromise was to use the internet to build the reference list by drawing on a multitude of different sources to ensure that it would be as representative of formulaic sequences as possible. This section describes how the reference list was created.

7.2.1 Creating a reference list of formulaic sequences

Potential search terms for lists of examples of formulaic sequences were entered into the online search engine *Google*. These included, for example, *list of proverbs, list of clichés, list of common phrases, list of similes,* and *list of popular sayings*. The search term *list of regular expressions* could not be used since *regular expression* is a specific technical term from the field of computer science and so returned too many irrelevant results. Similarly, the search string *list of formulaic language* did not provide any useful lists (mainly links to online books and articles related to formulaic language) since no such list has been widely publicised. For each search string, all of the links from the first five pages were explored. There did not appear to be any benefit to exploring beyond the fifth page since these typically included irrelevant links, or links that had already been explored. Every time a link led to a website which contained examples of formulaic sequences, those examples were entered into the database regardless of whether or not they were intuitively convincing.

This process was repeated until no new websites were identified. It became clear that several of the websites were sharing examples of formulaic sequences between themselves and so the decision to discontinue adding examples was made when it was evident that relatively few new ones were actually being added to the list. The final list contained 17,973 entries. Examples of individual formulaic sequences included in the reference list are provided in Section 7.2.2 (p. 149) and in Tables 7.2, 7.4 and 7.5. Furthermore, it will have been noticed that not only the title of this thesis but also each chapter heading incorporates a formulaic sequence which has been taken from the reference list.

It is difficult to account for the contents of the list in terms of classification (e.g. idiom, collocation, metaphor etc.) since formulaic sequences can often be members of more than one category (e.g. Moon, 1998a). However, based on how the websites self-identified themselves, the list appears to be composed of the following proportions:

Table 7-1 Proportion of different types of formulaic sequence included in the reference list

Type of formulaic sequence	**Number of entries**	**Percentage of entries**
Clichés	5131	28.6%
Idioms	3772	21%
Everyday expressions and sayings	3497	19.5%
Proverbs	2539	14.1%
Similes	1992	11.1%
Other (including prepositional phrases, collocations, Latin phrases and phrasal verbs)	1042	5.8 %
Total	17,973	100%

Clichés and idioms account for half of the entire list. The category ‘Everyday expressions and sayings’ highlights the problem of relying on self-reports for categorisation purposes: the dividing line between a cliché, idiom and an everyday saying is in no way transparent.

7.2.2 Editing the list

Having created a list of formulaic sequences, some editing was required to ensure consistency across the entries and to improve reliability (that is, the correct identification of formulaic sequences) as far as possible. The following procedures were undertaken:

i) All pronouns were replaced with an asterisk. The software used to identify matches in the data (cf. Section 7.2.3, p. 150) was capable of cross-referencing to a separate list of pronouns and the asterisk indicated the place where any item from the pronoun list was permissible. The pronoun list contained 86 entries including personal pronouns (e.g. *me, you, her, it*), possessive pronouns (e.g. *mine, yours, hers, its*) and possessive determiners (e.g. *my, your, her*). For example, by changing the entry *his bark is bigger than his bite* to ** bark is bigger than * bite* would allow matches in the data including *her bark is bigger than her bite, its bark is bigger than its bite, my bark is bigger than my bite, your bark is bigger than your bite* etc. A problem with this substitution approach is that there is potential for a nonsense string to be identified e.g. *her bark is bigger than his bite, your bark is bigger than its bite* etc. However, since it is unlikely that an author would produce these strings, and if they did, it would be for creative purposes rather than formulaic, the advantages of allowing substitution outweigh the potential disadvantage of having only fully fixed forms in the list. The only pronouns that remained fixed in the list were those where substitution would affect the meaning e.g. *get thee behind <u>me</u> Satan, love that dare not speak <u>its</u> name, one small step for <u>man</u>, cry <u>me</u> a river* etc.

ii) Some entries contained the word *something*, indicating that a free choice was available from the lexicon, i.e. they were semi-fixed phrases. Such entries included *cut something down, cut something off, do something over,* and *drop something off*. Theoretically, it would be possible to include a wild card for a lexical item along the lines for the pronoun substitution. However, since there are many more lexical possibilities than pronouns, far more nonsense strings, and more importantly, far more non-formulaic strings would be identified which would compromise the usefulness of the approach. These entries with *something* were removed from the reference list.

iii) Bearing in mind that the software only matches identical strings (with the exception of pronoun variation), the decision was made to split longer phrases into shorter stretches of text. This was useful for two reasons. Firstly, as Wray (2002) observes, sometimes simply starting an idiom can be sufficient for it to be recognised, rendering the need for writing the complete phrase obsolete (p. 24). Secondly, longer stretches of text are less likely to be matched in their exact form (including punctuation), so the compromise is to match a shorter stretch of that phrase rather than to miss the match altogether. Such changes included *a rose by any other name would smell as sweet* being shortened to *a rose by any other name* and added alongside the longer entry on the list. Likewise, the entry *build a better mousetrap and the world will beat a path to your door* was edited to include the separate entry *beat a path to * door* (including shortening and adding the asterisk) which in this case was judged to be the more fixed part of the longer phrase.

iv) Punctuation was generally removed from the list, except from some shorter phrases. For example, *look out!* can be understood to mean "be aware of an imminent danger". However, removing the exclamation mark would identify non-formulaic senses of *look out* such as in the third text authored by Keith: *So I went to look out the window*. Other examples of formulaic sequences which did not have their punctuation removed include *can it!, come again?, do birds fly?, do you feel me?* and *Duck!*

v) Many of the entries were obtained from American websites. Since the data to be analysed were produced by native English speakers living in England, UK spelling variants were added to the list alongside the original American spellings. Examples include *good fences make good <u>neighbours</u>, horse of a different <u>colour</u>, in <u>honour</u> of,* and *in self-<u>defence</u>*.

vi) As noted above (Section 7.2.1), there were many duplicates in the list. The last editing procedure removed these duplicates.

The final reference list contained 13,412 entries.

7.2.3 A note on software

The software required for this analysis needed to be capable of doing three things:

1) Assist in the construction of a computer-readable reference list of formulaic sequences;
2) Match a large reference list against an indefinitely innumerable set of texts;

3) Allow for a small amount of pronoun variation between entries in the reference list and examples in the texts.

Since no freeware software could be found which successfully and satisfactorily met all these criteria, it was deemed necessary to use bespoke software. The software, *Linguistic Analysis Suite v.1.3* (Menacere, Taylor, & Tomblin, 2008), was originally commissioned to enable the automated analysis of written texts for a research project which explored linguistic indicators of deception. As a researcher on this project, I was in a position to advise on the software design from an early stage, ensuring that it was designed to work specifically for the needs of a linguist. Additionally, although not a formal criterion, having an in-depth understanding of the software's design, capabilities and testing procedures would be a potential asset to a forensic linguistics expert during cross-examination, rather than simply trusting the software to do an accurate job. Unfortunately, the software is not currently commercially available.

Using this software, it was possible to create a large reference list which was automatically converted into a machine-readable format. The software was programmed to recognise the asterisk as a wildcard, which instructed it to consult an additional reference list of pronouns to enable pronoun matching in the texts. The software then compared the reference list to any texts that were input and produced an output file of the original texts with all matched examples highlighted and a calculation of how many individual entries were matched.

7.3 Evaluating the list

Before identifying formulaic sequences in the data, it was necessary to establish the reliability of the list on unrelated data. For this purpose, stories published on the Cancer Research UK website[3] were collected. Such stories offer advice, support and inspiration for people diagnosed with cancer, written by people who have themselves been diagnosed. As such, these data were judged to be similar to the personal narratives in the author corpus, i.e. they were short narratives written about emotive and personal topics, and therefore made excellent comparison data with which to evaluate the list. The first 25 stories published on this website were collected and generated 13,539 words. Each story started with an introduction and ended with a 'fact file'. The introduction was typically a one or two sentence summary of the author's background. The fact file included relevant information and statistics on the particular form of cancer under discussion. These did not form part of the main narrative and were therefore discarded.

[3] http://cancerhelp.org.uk/coping-with-cancer/ [Accessed: 17/01/2011]

Using the *Linguistic Analysis Suite* software, the list of formulaic sequences was matched against the cancer survival stories data. A total of 124 matches were made (cf. Table 7.2 for examples). This equated to 328 of the 13,539 words being classed as items in formulaic sequences. Based on this data, a word in a formulaic sequence occurred once for every 41 overall words or rather, per 100 words, 2.42 words were parts of formulaic sequences (cf. Section 7.5, p. 162 for comparison).

In order to assess the reliability and validity of the list, manual checking of the data was required. All of the data were read closely several times and phrases which were potentially formulaic but which were not identified and those which were identified but were less likely to be formulaic were highlighted and cross-checked against the list to account for any discrepancies. Additionally, the phrases that were identified by the software were also manually checked to ensure that the formulaic sense was captured and not some more literal, non-formulaic usage of the phrase. The confusion matrix presented as Table 7.2 highlights those sequences that:

i. Are potentially formulaic and were identified as such (correct matches);
ii. Are less likely to be formulaic but were still identified as such (false positives);
iii. Are potentially formulaic but were not identified as such (false negatives); and
iv. Are less likely to be formulaic and were not identified.

This last cell is shaded since language which was not identified as formulaic and seems unlikely to actually be formulaic would be novel text illustrated by any number of examples.

Table 7-2 Formulaic sequences confusion matrix

	Potentially formulaic sequences	Less likely to be formulaic sequences
Identified as formulaic sequences	i) at the same time watch and wait now or never how on earth over the moon burst into tears at this stage believe it or not face to face fingers crossed go hand in hand keep an eye on it light at the end of the tunnel doom and gloom without a doubt life goes on my heart sank back and forth from strength to strength emotional roller coaster	ii) get pregnant not even for life
Not identified as formulaic sequences	iii) ok all the above round the corner no one is perfect to be quite honest as fit as a flea and my heart goes out take the wind out of my sails shout it from the roof tops my wife and I goes with the territory clutching at straws day and night dropped the bomb shell	iv)

Since the objective was to correctly identify formulaic sequences and to not identify novel language, of all the entries in Table 7.2, it is those entries which were identified that are less likely to be formulaic sequences, and those sequences that were not identified which were potentially formulaic that are the most interesting and require further discussion.

As can be seen from Table 7.2, 14 potentially formulaic sequences were not identified by the software. They are potentially formulaic because they were variants of entries in the formulaic list; however, they were not identified because they were not identical matches (as shown in Table 7.3).

Table 7-3 Variants of formulaic sequences

Potentially formulaic sequences	**Related entries in the list of formulaic sequences**
ok	okay
all the above	all of the above
round the corner	just around the corner
no one is perfect	nobody is perfect
to be quite honest	to be honest with you
as fit as a flea	as fit as a butcher's dog as fit as a fiddle
my heart goes out to them	* heart goes out to *
blew the wind out of my sails	take the wind out of * sails
shout it from the roof tops	proclaim it from the rooftops
my wife and I	my husband and I
goes with the territory	come with the territory comes with the territory
clutching at straws	clutch at straws
day and night	night and day
dropped the bomb shell	the bomb

My heart goes out to them was not matched because the pronoun *them* was mistakenly omitted from the pronoun list. As a legitimate pronoun, this omission was corrected in the pronoun list. The remaining examples were not identified because their form was not identical to the list entry, even if the meaning they conveyed was the same which highlights a limitation of identifying fixed forms. The question of how variation between forms should be handled warrants further discussion.

It would be possible, upon reviewing potentially formulaic sequences that were not identified, to simply add those entries to the list. However, since the list was created to represent what is potentially formulaic for the wider language community, adding entries identified by one person would introduce an element of bias which would compromise objectivity; something that has been a fundamental consideration during the construction of the reference list. Further bias in the list of entries would also arise since only those potentially formulaic sequences identified in the Cancer Research UK data would be added. Admitting these 14 entries to the list when there are undoubtedly many more variants in other types of data could not be justified.

Aside from this practical issue of adding new entries to the list, there is a theoretical consideration about what actually counts as a variant of a formulaic sequence already contained in the list. Wray (2002: 28) cites an example from Altenberg (1990) which shows that a simple formulaic sequence like *thank you* does not necessarily occur alone and can be found with other strings such as *thank you very much, thank you very much indeed* and *thank you bye*. The question of whether these are different formulaic sequences, variants of the same formulaic sequence, or formulaic sequences with other formulaic sequences embedded in them falls to the individual researcher to decide: "These questions cannot be answered without the application of common sense and a clear idea of the direction of one's research: the latter automatically creates bias in the interpretation of the raw data" (Wray, 2002: 28). Given the need for forensic evidence to be objective and reliable, the decision was therefore made not to add new entries to the list. However, if the method can be shown to work, it will be possible to treat this research as a first approximation which can be developed with a basic strategy for adding all new items in an attempt to make the list more complete.

Next to consider are those sequences which were identified, but are less likely to be formulaic and the issue of how these extraneous examples should be handled. In the Cancer Research UK data, three such examples were identified: *for life*, *get pregnant* and *not even*. In order to understand why these sequences may or may not be formulaic, it is necessary to read them in their original contexts, as shown below:

1 and we have just taken part in The Race For Life and loved every minute of it and look
2 and was honoured to be a part of Relay for Life . Thank you for reading his story, we are

3 that I had to make sure that I did not get pregnant . The problem was, my partner and I

4 on the same ward as me die having not even lived their lives yet. The hardest part
5 falling out and was I going to die? I'm not even going to go there about how hard it is
6 , no one ever really asked about it – not even about the red wig! Although it was

Based on (biased) intuition, none of these examples appear to be necessarily formulaic—although, of course, that is not to say that another researcher may not consider them to be perfectly satisfactory examples. The argument that these examples are less formulaic can most strongly be waged against *for life*, which occurs twice across the Cancer Research UK Data, and in both instances as part of noun phrase for a specific event (rather than the perhaps less controversial, fabricated example, *a dog is for life* which may more convincingly be called formulaic). The key point here is that even though the validity of these few examples as formulaic sequences is questionable, to make that judgement call would draw entirely on individual intuition which would introduce a level of bias which, as far as possible, has been minimised in the creation of the list as discussed above. They therefore remained unaltered in the reference list.

7.3.1 Efficacy of the list

Through applying the list of formulaic sequences to unrelated data, several claims about its efficacy can now be made. Firstly, not every entry in the list will be acceptable to everyone as an example of a formulaic sequence. Some people will find some entries more problematic and less prototypical than others. The aim of the list is not really to reach universal agreement about what constitutes formulaic sequences; rather, the aim is to collate as many potentially formulaic sequences as possible in order to investigate whether evidence can be found that individual authors use some more so than others.

As the list cannot claim to be representative of formulaic sequences for each individual, questions must be asked about its authority. That is to say that the entries have not been verified by independent means, other than by their inclusion on websites as opposed to being included, for example, on the basis of corpus frequency counts (cf. Section 3.5.2, p. 70). The result is that a broader, more inclusive list has been created. However, the trade-off has been a lack of authority in as much as entries are those that other people have decided are special in some way (be it as a cliché, idiom, common expression or collocation etc.) which in turn can be considered to be 'formulaic' rather than being independently identified in corpora. Whilst the authority of the list may be called into question, the counter argument is that it is in fact representative of the language community—that is, people identified and recognised these examples as holding special status. Therefore, whilst the data collection method differs significantly, the end product equates to asking members of the same speech community to identify formulaic sequences in texts (e.g. P. Foster, 2001; Van Lancker-Sidtis & Rallon, 2004) and therefore a level of resilience and authority can be claimed through consensus.

In conclusion, there are limitations to the list, both in terms of what it contains and how well it can match formulaic sequences in real text. However, at the same time it does hold certain advantages which are particularly favourable in the forensic context. By using an automated approach, it enables large volumes of data to be analysed almost instantaneously. It offers reliability; items included in the list will be identified in any data on any occasion. However, the list cannot claim to identify every instance of a formulaic sequence; nor will it identify variants of items which are contained in the list, except when variants are already part of the list like 'as fit as a butcher's dog' and 'as fit as a fiddle'. It cannot even guarantee that every instance it identifies will actually be formulaic. However, the list is large and varied, so the crucial point is that it contains items which have the potential to be formulaic. It is this potential that makes the list a satisfactory tool for the initial exploration of the relationship between formulaic sequences and authorship. With a full understanding of the benefits and limitations of the list, it is now possible to apply it to the authorship data in order to continue the investigation of formulaic sequences as a marker of authorship.

7.4 Identifying formulaic sequences in the data: results

All 100 texts described in the author corpus were used in the analysis. The software compared the data with the formulaic sequences reference list and highlighted all instances of exact matches. A total of 604 formulaic sequence tokens were identified in the data, of which there were 300 types. Table 7.4 shows the ten most frequently occurring sequences whilst Table 7.5 shows a selection of ten formulaic sequences that were used only once across the whole data set.

Table 7-4 Most frequently occurring formulaic sequences across the author corpus

Formulaic sequence	**Frequency of occurrence across all data**
In the end	20
At least	17
Go back	14
At the end	12
In front of	12
In fact	11
On the phone	11
At home	9
At the same time	9
As if	8

Table 7-5 Least frequently occurring formulaic sequences across the author corpus

Formulaic sequence	**Frequency of occurrence across all data**
Under the influence	1
Under the weather	1
Vice versa	1
What on earth	1
What will be will be	1
Wide awake	1
With flying colours	1
With the exception of	1
Worst nightmare	1
X Factor	1

7.4.1 To what extent are the texts formulaic?

By establishing how many words there are in each text, and how many of those words form part of an identified formulaic sequence, it is possible to establish a count of formulaic sequences—in other words, how much of the text is formulaic and how much is novel.

Table 7-6 Count per 100 words of formulaic words as part of formulaic sequences across the author corpus

Author	Total words	Total words as part of formulaic sequences	Count of formulaic words per 100 words
MELANIE	2879	34	1.18
SARAH	2957	46	1.56
ROSE	3820	66	1.73
JOHN	3119	55	1.76
CARLA	3217	59	1.83
JUNE	3151	59	1.87
MARK	2844	56	1.97
NICOLA	3021	62	2.05
DAVID	3058	63	2.06
GREG	2980	70	=2.35
ALAN	3916	92	=2.35
MICHAEL	2516	61	2.42
SUE	3716	94	2.53
RICK	3583	93	2.60
JENNY	3518	103	2.93
JUDY	3427	104	3.03
KEITH	3067	95	3.10
HANNAH	3559	111	3.12
ELAINE	2941	94	3.20
THOMAS	3824	130	3.40

Table 7-6 shows the count of words identified as formulaic, broken down by author. The table is ranked from the author who uses the lowest count of formulaic sequences over the total of five texts, Melanie with 1.18 to the author who uses the greatest, Thomas, with a total count of 3.40. The mean average count of formulaic sequences in these texts is 2.35 (σ = 0.63).

7.4.2 Can a count of formulaic sequences differentiate authors?

7.4.2.1 Establishing variation between authors

A Kruskal-Wallis test showed significantly more variation between authors than within texts by the same author (χ^2 = 35, df = 19, p = 0.013)—in other words, the five texts produced by a single author are more alike in the count of formulaic sequences contained therein, compared to the texts produced by other authors. The first hypothesis, that variation between authors will be greater than within authors, is therefore supported.

A log linear analysis was carried out to determine any interactions between the factors gender (male/female), age (below 25/above 25) and education (Pre-university/ Undergraduate/Postgraduate). Analysis showed that no significant interactions could be separated out from the saturated model indicating that there were no significant patterns in the count of formulaic sequence usage for gender, age or education.

7.4.2.2 Differentiating authors

To determine whether authors can be differentiated taking the count of formulaic sequences as a marker of authorship, a series of statistical tests was performed, summarised below as Table 7.7.

As a test case the highest and lowest mean ranked authors were compared (Thomas and Melanie respectively). Taking the total count of words in formulaic sequences across all five texts for each author, it was possible to differentiate these two authors. Although this result shows that the method works in ideal circumstances, taking the highest and lowest mean ranked authors significantly improves the likelihood of reliably establishing difference since these authors were at the extreme ends of formulaic sequence usage. Therefore, to further test the method, the two authors with the most similar count of formulaic sequences (excluding Greg and Alan who had exactly the same) were compared (Nicola and David). In this test, it was not possible to differentiate the texts produced by these two authors.

In light of these findings, two sets of authors were selected to explore the limits of the method. Carla and Judy (1.83 and 3.03 words in formulaic sequences per 100 respectively) were firstly selected and it was possible to statistically differentiate these two authors. Rick and Mark were then compared with 2.60 and 1.97 words in formulaic sequences per 100 respectively and this time it was not possible to statistically differentiate the five texts produced by these two authors. It can be seen that the count of formulaic sequences was too close for Rick and Mark, whereas the texts produced by Carla and Judy did enable differentiation.

Table 7-7 Differentiating authors: summary of statistical testing

Purpose of test	Authors compared	Result	Outcome
Test case	Thomas and Melanie	Mann-Whitney *U* = 1, N = 10, p = 0.016	Possible to differentiate authors
Harder case	Nicola and David	Mann-Whitney *U* = 9.5, N = 10, p = 0.548	Not possible to differentiate authors
Exploring the limits	Carla and Judy	Mann-Whitney *U* = 23, N = 10, p = 0.032	Possible to differentiate authors
	Rick and Mark	Mann-Whitney *U* = 6, N = 10, p = 0.222	Not possible to differentiate authors

Taking these results into account, only partial support can be claimed for the second hypothesis, since the method only appears to work when the difference in counts between the authors is larger (although this is a relative term and future statistical testing would be required in order to accurately establish the boundaries of this distance). Therefore, we can more safely say that based on the twenty authors investigated in this study, some authors exhibit measurably different counts of formulaic sequences in their texts from some other authors (the limitation of this is discussed in Section 7.5). It may, additionally, be worth considering that although for some pairs of authors there was a non-significant tendency, in a genuine forensic case the count of formulaic sequences may add weight to other markers of authorship for the total to become significant. (cf. Section 8.4, p. 174 for discussion of this point in relation to Bayes' theorem).

7.4.2.3 Assessing forensic potential

These results show that in some cases, it is possible to differentiate authors. The next stage in the testing process is therefore to simulate a forensic scenario to see if, in addition to differentiating between texts produced by different authors, it is also possible to successfully attribute a Questioned Document. For this purpose, five texts by each of two authors were randomly selected for analysis: Nicola and Hannah. The two groups of texts were tested to see if they were normally distributed; that is, whether the count of formulaic sequences across all five texts in each sub-corpus were equivalent. Both groups showed no significant difference from normal meaning that no single text had an uncharacteristically high or low count of formulaic sequences (Nicola: KSZ = 0.913, N = 5, p = 0.376; Hannah: KSZ = 0.445, N = 5, p = 0.989). The second text by Nicola was randomly selected by *PASW Statistics* to act as the Questioned Document.

A two-tailed one-sample t-test showed no significant difference between the count of formulaic sequences in the four texts by Nicola compared to the Questioned Document, also by Nicola (t(3) = 0.601, p = 0.590). In real terms, we can say that there is a 95% chance that Nicola wrote the Questioned Document which is arguably an acceptable level of confidence for forensic linguistics. As the prediction from the means was that Nicola's scores would be lower than those of Hannah, a uni-directional hypothesis was tested. A one-tailed one-sample t-test showed a significantly higher count of formulaic sequences in the five texts by Hannah compared to the Questioned Document (t(4) = 2.157, p = 0.0485). Since there was no significant difference between the texts produced by Nicola and the Questioned Document and there was a significant difference between the texts produced by Hannah and the Questioned Document, it follows that Nicola authored the Questioned Document, which we know to be a correct attribution. The method appears to work.

This is clearly a positive result although it should be noted that only four texts by Nicola have been compared to the Questioned Document whilst all of Hannah’s five texts have been compared which automatically raises the question of whether the additional text has helped to improve the results. Therefore, Hannah’s second text was removed from the analysis. A one-tailed one sample t-test did not show a significantly higher count of formulaic sequences in the four texts by Hannah compared to the Questioned Document (t(3) = 2.037, p = 0.067). In terms of the final hypothesis then, that a randomly selected Questioned Document will be correctly assigned to its author based on the count of formulaic sequences, the results demonstrate that when a Questioned Document is compared to nine Known Documents produced by a closed set of two authors, it is possible to correctly attribute the Questioned Document to its correct author. However, it also appears to be the case that when a Questioned Document is compared to eight Known Documents produced by a closed set of two authors, it is not possible to correctly attribute the text. The final hypothesis therefore receives partial support.

The analysis carried out here relies on pairwise distinctions (e.g. Grant, 2010)—comparisons between pairs of authors—as opposed to population wide distinctions (e.g. Chaski, 2001)—comparisons between more than two authors in a sample. There are of course 190 unique pairs of authors that could be tested when exploring the limits, but to do so would be to introduce the potential for a Type I error in statistical terms (a false positive error). In other words, continuing to test each possible pair to determine where differentiation occurs would introduce a higher level of error than is acceptable. It can therefore be said that the variable ‘count of formulaic sequences’ holds potential to differentiate some pairs of authors but not all pairs of authors. In this regard, it is analogous to using the visual description of height as a variable on which to differentiate people. Some people will be taller, some will be shorter, and some will be the same height and it would not be possible to establish a threshold at which differentiation between people becomes possible. The same is true of using the count of formulaic sequences in a text as a marker of authorship. In a closed sample, some authors can be differentiated whilst some others cannot. Therefore, it is not possible to claim the method described here as a universal method that will be applicable in all cases.

7.5 Discussion

These results provide evidence that taking the count of formulaic sequence usage as a marker of authorship does have potential to differentiate some authors and, more importantly, to attribute a Questioned Document correctly to its author. The focus in this approach has been the count of formulaic sequences rather than specific sequences that may be used consistently and/or distinctively by the authors. It is nonetheless important to acknowledge that no patterns were evident—that is, no author showed a preference for any single formulaic sequence. The overall count of formulaic sequences is therefore more indicative of authorship than any single formulaic sequence. In order to fully contextualise the success and effectiveness of the method, it is necessary to discuss whether the method is valid, whether the method is reliable, and whether the method is forensically robust, that is, whether the method holds value as an investigative and/or evidential tool. Before doing so, some more general observations can be made.

Firstly, the count of formulaic sequences compared to novel text that was identified using the reference list method is low compared to other qualitative approaches such as 24.8% identified by Van Lancker-Sidtis and Rallon (2004) in the screenplay *Some Like it Hot* and 77% in essay-style exam answers identified by Chenoweth (1995). However, this is not necessarily surprising. In Section 3.5.2, the point was made that using automated methods to identify formulaic sequences yields lower results than alternative methods. Even though the evidence suggests that formulaic language is ubiquitous and plentiful, there was never an expectation that large quantities of formulaic sequences would be identified using this method, which relies on largely fixed forms (with the exception of pronouns). Given that the difference between formulaic sequences identified using this method is enormous compared to other qualitative approaches, it is interesting to consider how these counts compare to other, similarly obtained results.

Moon (1998a) used a reference list containing 6,776 items "of the commonest FEIs [Fixed Expressions and Idioms] in current British English, together with some commoner FEIs in American English" (1998a: 44) in the c.18 million word Oxford Hector Pilot Corpus. She found that very few of the FEIs occurred with a frequency greater than one per million words. Simpson and Mendis (2003) created a reference list from three ESL textbooks to identify idioms in the Michigan Corpus of Academic Spoken English (MICASE) corpus which contained 1.7 million words. They identified 238 different idiom types with 562 tokens. Of the 238 types, 123 occurred only once and only 23 occurred more than four times (p. 425). Simpson and Mendis found that the overall frequency of idioms in MICASE was approximately 330 tokens per million words (p. 427). In the present research, as reported in Section 7.4 (p. 157), 604 tokens (300 types) were identified in the 65,113 word author corpus, which can be alternatively expressed as 1.08 tokens per million words.

As might be expected, owing to the different definitions for what constitutes formulaic language, which, and how many items were included in the reference list, the findings of Moon (1998a) and Simpson and Mendis (2003) differ from each other with the findings of this research being closer to those of Moon. For these reasons, quite whether the count of formulaic sequences in text is similar or different to the findings of other research cannot be taken as a comparative indicator of the effectiveness or inefficiency of the method.

Secondly, although not a formal hypothesis, the fact that there were no interactions between gender, age and education level with the count of formulaic sequences is an important finding since the relationship between these factors and overall use of formulaic sequences have not been reported elsewhere. These traditional sociolinguistic variables do not appear to have an impact on how much of a person’s lexicon is formulaic compared to novel in these data. Rather more precisely, these variables do not appear to correlate with how many formulaic sequences these authors actually use compared to novel language. The implication for forensic authorship attribution is that author profiling along these lines does not appear possible for these particular authors. However, author profiling has not been a central concern and considerably more research in this area would be required in order to make the claim bolder and more substantiated—and of course, testing on other data would be a requirement before it could be claimed that there is no relationship between formulaic sequence usage and age, gender or education level. The implication for formulaic language theory is that, if, as Wray (2002) argues, the history of our socialisation affects our formulaic sequence repertoires, either i) our socialisation (certainly in relation to education, gender and age) does not appear to affect the overall count of language that is formulaic for these authors in these data, or ii) other aspects of socialisation, excluding these variables, may impact formulaic sequence repertoires.

7.5.1 Is the method valid?

To assess the validity of the method, it is necessary to critically examine whether formulaic sequences have actually been identified through this process. There are two considerations in this regard: i) the entries that were included in the reference list, and ii) the entries that were actually identified in the data. Dealing firstly with the reference list, as has been argued throughout this chapter, it is highly unlikely that everybody will agree that what is included in the reference list is a formulaic sequence. The key point, as has also been emphasised, is not that any one individual agrees with every item on the list; rather, that each item on the list holds an equal opportunity to be formulaic for any author. Furthermore, the list cannot claim to be exhaustive and there are undoubtedly other entries that could have been included. However, to compensate for these unavoidable shortfalls, the list is as deliberately large and inclusive as possible, covering a multitude of different types of formulaic sequence. As long as researchers accept collocations, idioms, similes, everyday sayings and so on to be formulaic, the list is valid.

Next to consider is whether those formulaic sequences identified in the data are valid in terms of being evidence of authorship, or whether they are indicative of something else. The theoretical claim underlying this research is that authors will have different repertoires of formulaic sequences to draw upon based on their exposure to language through their socialisation. The reality is that several other factors may have had an impact on any author's use of formulaic sequences. Such factors may include how well rehearsed or edited their particular narrative was and whether they were concentrating fully and solely on the task (or whether they were concurrently preparing a meal, chatting on a social networking website, watching television etc.) i.e. their cognitive load. However, it is hoped that by collecting five texts from each author over a series of five days, such additional cognitive pressures may have been mitigated by texts produced on days when there were perhaps less cognitive pressure to give a representative account of each individual author's average cognitive load when producing language. (Although, clearly, producing a threat letter, suicide note, or ransom demand will carry additional cognitive pressures that go far beyond the scope of this research.)

7.5.2 Is the method reliable?

In order for the method to be reliable, it would need to be proven that the same results would be achieved each time the analysis is replicated. We know this to be true since the method is automated and so is unaffected by factors which normally affect reliability (cf. Section 3.5.2). However, establishing that the method is reliable each time the analysis is carried out is only useful if the method can be applied to any type of data. The research described in this chapter has focussed only on the count of formulaic sequences as they occur in a very restricted type of data—short personal narratives. Just as there may be grounds to question the validity of the method on the basis that it may not actually be authorship that is detected, so too may the reliability of the method be criticised on the basis that the data used are in some way special. Did the questions asked to elicit the narrative data encourage a higher count of formulaic sequences in the responses? To assess this, all of the questions were matched against the reference list. No incidences of formulaic sequences were identified. It is therefore unlikely that the authors were primed in their use of formulaic sequences and the data can be argued to have occurred naturally. However, a potential criticism may be that the narratives themselves are not representative of normal, everyday language. After all, the narratives were deliberately intended to encapsulate the authors (cf. Section 4.4.1). As entertaining personal narratives, it is conceivable, perhaps even probable, that the authors will have told these narratives in various ways on various occasions, and they may therefore be rehearsed, revised and may contain hyperbole. As such, it may be hard to argue them to be naturally-occurring (in the same way that traditional oral stories contain higher occurrences of formulaic sequences to aid memory during public performances, cf. Rubin (1998)). This is not to dismiss the method though, since, in this sense of naturally-occurring, a suicide note or threat letter could also be the product of several revisions.

There is also a temporal issue to consider in relation to the reliability of the method. Some formulaic sequences can be considered “15 minutes of fame” expressions (Wray, 2002: 27) or “dynamic vocabulary items” (Moon, 1998a: 51). The reference list does not appear to contain such examples, with the possible exceptions of *X Factor* and *Big Brother*, which could be argued to be relatively contemporary formulaic sequences in British English, based on the popularity of the same titled television shows. Incidentally, both *X Factor* and *Big Brother* were identified on one occasion each in the authorship data and both were actually used in relation to the television programmes. However, the lack of fleeting expressions in the reference list and their relative scarcity in the authorship data suggests that such dynamic formulaic sequences are unproblematic for the analysis. Likewise, if formulaic sequences can be demonstrated to remain stable over a longer period of time, the reference list will remain unproblematic. Future additions to the list may affect reliability since if new entries are included, a retrospective analysis using the new list may reveal different results although it is also arguable that a more complete reference list would simply produce a greater count of formulaic sequences for all authors. Therefore, this particular aspect would need to be investigated to ensure reliability and long-term evidential status.

A final consideration is the range of speech communities represented by the list. A wide variety of UK and USA variants have been included. In principle, the reference list can therefore be used to identify formulaic sequences in texts produced by speakers of British or American variants of English. However, it can only be applied to texts which follow the standard conventions of English and may be less applicable to non-standard varieties of English (such as text message language, computer-mediated communication etc.). It therefore may be unreliable as a universally-applicable method for authorship attribution.

7.5.3 Is the method feasible for forensic purposes?

Based on the analysis carried out in this chapter, it would be improper to claim that the methodology outlined has been tested extensively enough to warrant use in a forensic investigation, let alone its use in court. The results in Section 7.4.2.3 (p. 160) showed that a Questioned Document could be correctly attributed to its author with a 95% level of confidence but there was also clearly an effect depending on how many texts are included in the analysis. In practical terms, the method is forensically viable, since a single researcher can apply the reference list to innumerable texts relatively quickly and reliably. It must of course be remembered that if the 'styles' of authors (i.e. counts of formulaic sequences) are too close, or if there are too few texts available for analysis it will not be possible to use the method—clearly more testing is needed to establish these thresholds. However, of the three analytical approaches to formulaic sequences as a marker of authorship, the formulaic sequence reference list method appears to hold the most potential both as an analytical tool and as a forensically robust method.

7.6 Conclusion

This chapter has outlined a method of authorship attribution which takes the count of formulaic sequences compared to novel language as a marker of authorship. Using just five texts totalling approximately 3,000 words from each of 20 authors, it was established that there is more variation between authors than within, that two authors with markedly different counts of formulaic sequences can be successfully differentiated and that when two authors are randomly selected, a Questioned Document can sometimes be correctly attributed to its author. The method has also been argued to be valid, although far more testing than is possible in this initial exploration will be required in order to demonstrate reliability. Despite the positive conclusions that can be drawn from this method, it is important to consider that although the two authors with the greatest and the least count of formulaic sequences in their texts could be differentiated, it was not possible to differentiate the two authors with the most similar count of formulaic sequences. This is not necessarily problematic in itself since not all markers of authorship work on all sets of data. The method would only be problematic if the method produced false positives, which, through the testing carried out in this chapter, it has not. However, as established in Chapter 2, there is currently no unified method to authorship attribution, and instead the linguist must select the most appropriate methods from a rich toolkit. With further testing, the method described in this chapter could conceivably be added to that toolkit as another variable on which some authors have been demonstrated to vary from others and may add further evidence in some cases of authorship attribution.

Chapter 8

'Come to think of it': a consideration of the issues

Over the course of the past three chapters, a series of experiments have been devised to test whether formulaic sequences are a potential marker of authorship. A series of issues have arisen which warrant full discussion and an answer must be provided to for the central research question. However, before the research can be assessed, it will be helpful at this stage to have a summary of all of the approaches and their corresponding results.

8.1 Summary of results

The research carried out in Chapters 5—7 is summarised as Table 8.1 which shows each of the approaches adopted and the results achieved. Table 8.1 should be read in conjunction with Table 8.2 which lists each of the hypotheses tested over the course of the research, and the level of support received.

Table 8-1 Summary of empirical research

Chapter	Approach	Test	Summary of results
5	Formulaic clusters	1) Establishing variation between authors	Texts linked by the same author are more similar in use of specific formulaic clusters than texts by different authors
		2) Attributing a QD: two candidate authors	Not possible to exclude either author as potential author of QD
		3) Attributing a QD: five candidate authors	Two formulaic clusters identified which were consistent and distinctive for one author. May provide limited grounds for attributing a QD
6	Core word (*way*-phrases)	1) Authorial preferences for specific *way*-phrases	Very limited results—for one author, consistent and distinctive preference for *in a way*. No patterns for any other authors
		2) Identifying alternatives to *way*-phrases	No evidence of authorial preferences
7	Formulaic sequences reference list	1) Establishing variation between authors	Significantly more variation between authors than within authors
		2) Differentiating authors: a test case	Possible to differentiate texts produced by highest and lowest mean ranked authors based on count of formulaic sequences
		3) Differentiating authors: a harder case	Not possible to differentiate texts by two closest authors based on count of formulaic sequences
		4) Differentiating authors: exploring the limits	Possible to differentiate texts by one pair of authors but not the other

		5) Assessing forensic potential: attributing a QD (nine known documents, two candidate authors)	Correct attribution based on significantly higher count of formulaic sequences for one author compared to the other
		6) Assessing forensic potential: attributing a QD (eight known documents, two candidate authors)	Unsuccessful. No significant difference between texts produced by two authors

Table 8-2 Summary of hypothesis testing

Chapter	Approach	Hypotheses	Level of support
5	Formulaic clusters	1) Authors will use distinctive patterns of clusters consistently across their texts which can be argued to be formulaic	Full support—authors used different patterns of clusters with some consistency across their texts
		2) Authors will be differentiated based on the patterns of formulaic clusters found within their texts	Full support—authors were differentiated based on the occurrence of formulaic clusters in their texts
		3) A QD can usually be correctly attributed to its author based on the occurrence of formulaic clusters	Only very limited, tentative support—QD shared two formulaic clusters with one author from closed set of five candidate authors. No attribution with only two candidate authors
6	Formulaic core word (*way*-phrases)	1) It will be possible to identify a range of formulaic sequences which have the word *way* at their core	Full support—55 *way*-phrases identified
		2) Authors will have consistent patterns in the *way*-phrases that they choose from the available set	Very limited support—one author out of 20 used a *way*-phrase consistently
		3) As a consequence, it will be possible to distinguish texts produced by one author from those produced by another based on *way*-phrases	Very limited support—texts produced by only one author appear to be marked as different from all other texts in the corpus due to the frequency and consistency with which *in a way* occurs
		4) Specific meanings will be expressed in different forms, some based around the core word *way* and others around a different set of words	Full support—141 phrases identified which express similar meanings to the *way*-phrases
		5) Authors will have preferences for the selections of words they use to express specific meanings	Very limited support—expression *in a way* is characteristic of one author by being a distinctive and consistent choice across all five texts. No preferences or consistent patterns evident for any other author
		6) Such word selections will be consistent across a series of texts and may therefore be idiolectal	

7	Formulaic sequences reference list	1) Variation in the count of formulaic sequences in texts will be greater between authors than within authors	Full support—significantly more variation between authors than within authors
		2) Authors will be potentially differentiable from each other based on the count of formulaic sequence usage	Partial support—method only appears to work when difference in counts between the authors is larger but breaks down when counts are more similar
		3) A randomly selected QD can usually be correctly attributed to its author based on the closeness between the count of formulaic sequences in the text and in the author's other four texts	Partial support—successful attribution when QD is compared to nine Known Documents produced by a closed set of two authors. Unsuccessful attribution when QD is compared to eight Known Documents produced by a closed set of two authors

It can be seen from Tables 8.1—2 that the three approaches adopted in this research have received differing levels of support. The salient points are these:

i) VARIATION BETWEEN AUTHORS: Formulaic sequences identified through the core word *way* are not sufficient to universally detect authorial patterns, although the phrase *in a way* was identified as being a consistent and distinctive feature for one author out of 20. Formulaic clusters are sufficient for demonstrating variation between the authors with pairs of texts produced by the same author being more similar than mixed authorship pairs. Using a reference list to identify formulaic sequences was also successful in showing variation between sets of texts with some authors producing significantly higher counts of formulaic sequences than others.

ii) ATTRIBUTING A QUESTIONED DOCUMENT: Formulaic sequences identified through the core word *way* do not enable a Questioned Document to be attributed to its author since, with the exception of one author, no authorial patterns were apparent. Formulaic clusters seem to hold a small amount of potential since there were commonalities between the Questioned Document and one author, although this approach works better with greater numbers of candidate authors. A formulaic sequences reference list holds the most potential to enable the correct attribution of a text.

iii) MOST PROMISING METHOD: Only one method, the formulaic sequences reference list, convincingly enabled a Questioned Document to be attributed to its author, but was not successful for every pair of authors tested and varied according to how many texts were available for analysis, the number of candidate authors, and which particular candidate authors were included in the analysis. None of the approaches investigated in this research therefore offer a universal method with 100% accuracy.

This information now enables the central research question proposed in Chapter 4 to be answered.

> *Given that all individuals have a different store of formulaic sequences acquired through a different range of life experiences, can formulaic sequences be used as a marker of authorship to the extent that a Questioned Document can be correctly attributed to its author from a relatively disparate sample of candidate authors?*

8.2 Does formulaic language hold potential as a marker of authorship?

The answer to this question appears to be that in some cases, formulaic sequences can indeed be used as a marker of authorship since some of the Questioned Documents were successfully attributed to their authors. The approach with the least potential is formulaic sequences identified through the use of the core word *way* since only one author out of 20 exhibited a consistent and distinctive feature. However, as acknowledged in Section 6.3 (p. 140), this is not to say that other core words would not meet with more success. Formulaic clusters also had very limited success as a marker of authorship although on its own, this approach is unlikely to be persuasive as evidence of authorship. Taking the count of words which make up formulaic sequences compared to novel words through the use of a formulaic sequence reference list seems to hold the most potential—both in robustly demonstrating variation between authors and by reliably attributing a Questioned Document, depending on which pairs of authors are selected for comparison.

Despite the limited level of success for attributing a Questioned Document, it should be acknowledged that both the formulaic clusters method and the formulaic sequences reference list approach enabled statistically reliable differentiation between authors—that is, there was significant variation between the authors on these two variables and this should be taken as evidence that authors do use formulaic sequences differently.

8.3 Is there a link between formulaic language and idiolect?

The next question that naturally arises is whether there is an inherent connection between formulaic sequences and idiolect—whether evidence of idiolectal formulaicity exists. Each author produced their own narratives; therefore, all of the language that occurs *is* part of their idiolect. But this approach can be a little too simplistic—or at least less useful for forensic purposes—since the strength of the argument accounts to no more than noticing that the words *strength, of, the* and *argument* are part of my own idiolect, but shared with countless other authors: in the BNC alone, *strength* occurs 6,951 times, *of* occurs 3,887,705 times, *the* occurs 6,047,031 times and *argument* occurs 8,201 times. These words therefore hold very little diagnostic power of idiolect. However, the combination of these words does become interesting, with *strength of the argument* occurring only once in the BNC—what Coulthard (2004) would refer to as idiolectal co-selection. Of course, it is not clear whether *strength of the argument* is formulaic for me—it certainly does not occur anywhere else in these pages and nor is it knowable whether I will ever use this sequence of words again—but it is likely that analysing sequences of words may provide stronger evidence of idiolect rather than single words in isolation and therefore may be more appropriate for authorship attribution. Similarly, each of the authors in the corpus used the words *in, a* and *way*, so each of these single words were available as part of their idiolects but only eight of the authors used *in a way* as a contiguous sequence. Furthermore, only one author, Rose, used this sequence in every text she produced, indeed more than once in each text. This phrase therefore must be characteristic of Rose's idiolect in a way that it is not for other authors. Even though this phrase is not distinctive in comparison to other authors, its frequency of occurrence does appear to be idiosyncratic to Rose.

The fact that two of the approaches adopted in this research were able to detect significant variation between authors shows that formulaic sequences also appear to be useful in illustrating idiolectal differences and characterising the styles of different authors. And what of the fact that other sequences of words, formulaic clusters, were identified based on recurrence across a series of texts? Again, the repetition and consistency must characterise an author's idiolect in some way and the fact that 26 formulaic clusters were identified for Rose is surely significant in relation to the fact that only one was identified for Michael. Finally, the count of words which make up formulaic sequences compared to the overall words used might also characterise idiolect in some way. Is it a characteristic of idiolect that Melanie uses the lowest count amount of formulaic language whilst Thomas uses the most? Can it be said that Thomas is *more* formulaic than Melanie or any other author in the corpus? The fact that variation was demonstrated between authors and that on some occasions a Questioned Document could be attributed to its author would indicate that this is in fact the case. Formulaic sequences are a characteristic of idiolect, and, according to the evidence, are also a useful marker of authorship.

As discussed in Chapter 2, accepting that formulaic sequences are evidence of idiolectal variation rests on the assumption that such a notion actually exists and is not instead only the random or calculated combination of choices made by the language-user. In this research, a direct connection between formulaic sequence usage and authorship has been proposed since there is a strong relationship between formulaic sequences and reality—a reason why this marker should vary between authors as opposed to more objective measures (e.g. the number of words starting with a vowel as used in the Cusum technique, for which no reasonable linguistic explanation exists for why inter-author variation should occur). That is, the way that people have been socialised, the contact with, and interest in language that they have, and the priorities they face when producing language all have an impact on the individual stores of formulaic sequences contained in each author's mental lexicon. To this end, there is a good sociolinguistic and psycholinguistic theory to support the notion that idiolectal differences between authors exist and that authors use formulaic sequences differently from each other.

This research has made no direct attempt to find evidence of idiolect and was only ever searching for markers of authorship. It is important to keep the distinction between these two endeavours separate, since, in light of the discussion in Chapter 2, far more data from many more authors produced over far longer periods would be required before any strong claim of the existence of idiolect could be substantiated (in other words, a closer approximation to the "totality of speech habits" than can be achieved in five short narratives). However, it should be apparent that there is an underlying theory behind this marker of authorship—authors have a different store of formulaic sequences built up over a lifetime of differing experiences—so in Grant's (2010) terms, this research is not idiolect-free authorship attribution. Therefore, although further testing along the lines suggested in Section 99.1 (p. 186) will be required before any attempt to demonstrate the existence of idiolect, the following comments on idiolect in general may be offered.

This research contributes to the theory of idiolect in as much as it demonstrates that formulaic language is a component. But the crucial point is that it only provides a very small snapshot of idiolect, namely, of each individual author's idiolect as it relates to writing short personal narratives over a five day period. This would differ, one assumes, from the features of their idiolect that would manifest whilst, for example, discussing an important matter with colleagues in a meeting over the course of one hour. The point is that whilst any form of language an author produces can legitimately be argued to be an aspect of their idiolect (as it has been defined in this research), there are as yet, far too many unknown variables affecting which aspects of an idiolect are invoked on one occasion compared to another. In light of this, it seems more appropriate, specifically for forensic purposes, to talk about the stability of linguistic features (Barlow, 2010). Barlow (2010) explains that "the language of an individual changes depending on the interlocutor and the general context" (p. 2)—a point which has been repeatedly made throughout this research. In his investigation of the language of White House Press Secretaries, Barlow argued that some features of idiolect, such as bigrams, could be shown to be stable over a one year period since intra-author variation across different speech samples was low, whilst inter-author variation was higher.

Rather than looking for features of idiolect, then, the forensic linguist may benefit from searching for features of idiolectal stability—those features which seem to characterise an individual's language use regardless of content and despite text length, genre, composition date and medium—the Holy Grail of authorship markers. Such features will likely be rare (if they exist at all) and it is argued here that deeper level features which lie beyond the individual's conscious language decisions will be the most fruitful avenues for investigation. Of course, formulaic language is one such possibility and so whilst it has been argued that formulaic sequences are part of these twenty authors' idiolects, and that this characteristic appears to enable differentiation between some authors, what we cannot know at this stage is whether this feature is evidence of idiolectal stability.

Can we therefore adjust the existing theories of idiolect (as described in Chapter 2) on the basis of the empirical work carried out here? It would be tempting to say yes given that formulaic language, particularly the count of formulaic sequences, appears to be a characteristic of language which remains stable over a five day period and which differs enough for a Questioned Document to be attributed to its author. However, my more cautious side prevents such a conclusion since this research project was not designed as an investigation into idiolect and some of the criticisms I have levelled against other researchers in Chapter 2 could, I suggest, to differing extents apply to my own. To reach such a conclusion would be to judge my research in a different light to others'. It is therefore more cautious to conclude that through this research an aspect of language use has been identified that appears to differ between individuals. However, the scope for generalisation is very small since the analysis has focussed on one very small aspect of language use (namely, the production of short written narratives over a five day period). For now, I would prefer to conclude that formulaic language appears to be a consistent and distinctive authorship marker (Grant, 2010) which warrants further investigation to determine its limits. How that feeds into the idiolect debate, if at all, will be for other researchers to assess.

8.4 Are formulaic sequences a forensically robust marker of authorship?

Having established that formulaic sequences are a potential marker of authorship, it is necessary to assess how this tool may be used and what its future may hold taking into account its forensic value. Table 8.3 summarises the three methods outlined in Chapters 5—7 and indicates whether each method is valid, reliable and feasible for forensic application, summarising the discussion at the end of each of these chapters.

Table 8-3 Summary of methods in terms of validity, reliability and forensic feasibility

Method	Valid	Reliable	Feasible
Formulaic clusters	✓	X	X
Core word	✓	X	X / ✓
Formulaic sequences reference list	✓	✓	✓

Of the three methods presented in this research, only one, the formulaic sequences reference list, successfully meets the criteria of being valid and reliable as an analytical tool, whilst holding the most potential to meet evidential standards. The core word method is valid but not reliable, with half of the approach being forensically robust and the other half being less so. Based on this research, then, an approach based on the formulaic sequences reference list appears to be the most suitable method to adopt for analysing texts for authorship. However, again, reference must be made to the fact that only short texts have been the focus of this investigation and it is quite possible that with a greater number of texts and/or texts of greater length, the other two methods may become more suitable. It stands to reason that with more data, a higher number of formulaic clusters will likely be identified. Likewise, there is potential for more *way*-phrases to be identified along with a wider variety of meanings and therefore alternatives when more data is available for analysis. In such a situation, stronger patterns may be established than was possible here. However, both of these approaches will still be limited in terms of their reliability and forensic feasibility.

The most obvious point to make is that using formulaic sequences as a marker of authorship is clearly not developed enough to hold any evidential value, taking into account the discussion of the *Daubert* criteria in Chapter 2. More testing is required in order to establish known rates of error and the exact limitations of the method. Peer review and acceptance or rejection by the community can then follow. In light of this, assuming these stages are followed, will there be any evidential value to using formulaic sequences as a marker of authorship? This is certainly more questionable since predicting which formulaic sequences authors are likely to use does not seem possible, with the possible exception of Rose for whom *in a way* could be predicted to occur in additional texts. However, the fact that texts produced by Rose could be compared to texts by other authors with *in a way* being far more distinctive may carry evidential value and, of course, formulaic sequences would never be used as evidence alone—they would always be combined with other markers of authorship. Therefore, as part of the forensic linguists' basket of authorship attribution tools, formulaic sequences may indeed turn out to have significant evidential value.

It should also not be forgotten that the most successful of the methods was based on the normalised count of words which are part of formulaic sequences in comparison to the overall words and this appears to be a far more robust marker of authorship with some authors being more formulaic than others. This method may be especially persuasive for jurors who are likely to recognise that some people use more clichés, for example, than others. It should also not be forgotten that the approaches outlined in this research are lexically based, and since other authorship attribution evidence has been admitted in UK and USA court cases based on lexis and notably strings of words (Coulthard, 2004; Fitzgerald, 2004; McMenamin, 2002), there is no reason to automatically assume that this tool would not be admissible. In reality, after further testing, whether the evidence is admissible will come down to the judgement of the courts.

What then of investigative value? Although Questioned Documents could not always be successfully attributed to their authors, there was a measure of success which may be useful at the investigative stage. Furthermore, it was possible to narrow down a larger pool of candidate authors using formulaic clusters. This would certainly be helpful in an investigation with multiple suspects. And it should be borne in mind that these results were achieved based on limited data so an investigation fortunate enough to yield longer and more texts would likely benefit from incorporating these analyses in combination with other markers of authorship. Through such an approach, the culmination of smaller similarities between texts may produce stronger evidence of authorship as in Bayesian analysis (e.g. Mosteller & Wallace, 2007).

A key point about the approaches outlined in this research is that some people are more similar than others, such as twins (e.g. Künzel, 2010). So too can it be expected that some idiolects will be harder to differentiate than others as was found by using the count of formulaic sequences as a marker of authorship: some authors could be differentiated whilst others could not. Notably, there appears to be a threshold at which the method no longer works whilst the most formulaic and least formulaic authors can successfully be differentiated. In real terms, this means that this method, like any other marker of authorship that relies on pairwise distinctions, cannot guarantee to work in every case. However, the point should be made that no such marker of authorship exists that provides 100% success and the forensic linguist will always have to appraise the data to establish which markers are the most useful to apply based on the available data.

In addition to the forensic resilience of formulaic sequences, comments regarding their general nature can now be made.

8.5 The nature of formulaic sequences

One of the main arguments for supposing that formulaic sequences would make an ideal marker of authorship was based on the fact that they are ubiquitous. This claim should therefore be assessed. In Chapter 7, some comparison was made between the findings from the present research and other research, and the point was made that the count of formulaic sequences in the author corpus was lower than other researchers had found in their data. This is unsurprising given that the approaches to identification differed, along with definitions of what was counted. Therefore, drawing comparisons with incompatible research is less insightful. Instead, three claims based on the present research can be made:

1) Authors used an average of seven formulaic clusters in each of five texts. The average sub-corpus length is 3,256 words and the average formulaic cluster consists of 3.1 words equating to 6.6 formulaically-used words per 1,000 overall words;
2) Formulaic sequences using the core word *way* occur 103 times in the total corpus of 65,113 words. The average length of a *way*-phrase is 3.2 words, so *way*-phrases occur 1.58 times per 1,000 words with 5.06 formulaically-used words per 1,000 overall words; and
3) Items Identified using the formulaic sequences reference list had an average length of 2.7 words. A total of 604 formulaic sequences were identified, equating to approximately 25.05 words per 1,000.

Viewed in this light, it cannot be claimed that these particular formulaic sequences are ubiquitous in short written narratives. It can barely be argued that these formulaic sequences are even frequent given that the most prominent measure calculates that a word which is part of a formulaic sequence occurs roughly 25 times per 1,000 words. The reason for these low frequencies has been acknowledged throughout this research—an automated approach was always expected to yield less data than the more intuitive approaches used by other researchers and described in Chapter 3. It also should be considered that identifying formulaic sequences which can be classed as formulaic clusters and identifying formulaic sequences through the use of just one core word are two very limited and very narrow approaches. These frequency scores are therefore understandably low. The frequency score for the formulaic sequences reference list is perhaps more surprising since by virtue of being a very large, inclusive list of formulaic sequences, more occurrences should have been identified. As previously stated, the automated nature of the approach bears the brunt of the blame. However, another contributory factor is likely to be that found by Moon (1998a)—quite simply, some formulaic sequences, particularly idioms and fixed expressions (e.g. *kick the bucket*), just do not occur as frequently in the English language as intuition might suggest. Basing a reference list on such items will therefore have limitations. Nonetheless, the fact that this approach in particular was successful at establishing variation between authors and attributing Questioned Documents, the fact that some authors have a higher count of formulaic sequences than others adds further support to the rationale behind this research; that is, if intuition suggests such formulaic sequences to be common in English, the fact that they are not for some authors, and the fact that they are higher than average for other authors indicates that this is a useful marker of authorship of which authors are unaware.

It is interesting that no author showed a preference for any one formulaic sequence, again with the exception of Rose and *in a way*. None of the three approaches revealed that an author has a favourite phrase which they use consistently, whether distinctive or not. The conclusion to reach from this is that some phrases are distinctive (e.g. the Unabomber's use of *cool-headed logicians* and *you can't eat your cake and have it too*) but rare, whilst others are consistently used but not distinctive (e.g. Rose's use of *in a way*). This is an important point, since the three methods outlined in this research cannot identify distinctive and rare formulaic sequences and qualitative analysis would be necessary to identify such admittedly important phrases. The approaches outlined in this research cannot, and should not, replace the traditional qualitative approach of close-reading a text to identify distinctiveness. It is also possible that focussing on specific patterns of formulaic sequences is not the only way to approach the data, and drawing on other lexical richness measures such as authorial pace, "the frequency with which he [the author] generates new words and allows them to enter his manuscript" (Baker, 1988: 36), may be developed. This is particularly appealing since authorial pace has been argued to be characteristic of authorial style regardless of text length or genre (Baker, 1988). Where authorial pace is expressed as *Pace = 1/Type token ratio*, a new measure which calculates the rate at which new *sequences* of words enter text may offer additional avenues to explore, particularly since formulaic sequences have been argued to constitute one lexical choice in this research.

The fact that no other author appeared to have a set of preferred formulaic sequences is in keeping with Wray (2002) who argues that:

> [E]ach person, in each unique situation, will apply slightly different selection criteria to a slightly different set of options, from those available to anyone else. Certainly there will be very many similarities between individuals, insofar as they share, within a given environment or speech community, an inventory of idiomatic forms and certain interactional expectations of, and towards, each other. But, just as it will be possible, through such similarities in formulaic speech patterns, to spot people who come from the same place, are the same age or share the same interests or beliefs, so it will rarely be possible fully to predict which formulaic sequences a given speaker will select, since the balance of priorities is constantly shifting, and with it, the relative usefulness of the stored sequences (Wray, 2002: 101—2).

Wray's assertion that predictions about which formulaic sequences an author will select are not possible accounts for why the authors in this research did not appear to prefer certain formulaic sequences, since, if the priorities on the language user are constantly shifting, then texts composed on different days at different times will inevitably not be fully comparable. This adds further support to the notion that the overall count of formulaic sequences may be more indicative of authorship than any single formulaic sequence—since the actual forms of formulaic sequences that authors use may vary, the degree of overall reliance on formulaic sequences may not.

But are the word sequences identified in this research really *formulaic* sequences?

In Chapter 3, Hoey's (2005) theory of lexical priming was introduced. Hoey proposed that words are primed for other words, and so too are word sequences ('nesting'), and crucially that primings can vary for individuals, depending on their experiences of the contexts and cotexts in which those words are used. Wray (2008), talking of her *needs-only analysis* model (also described in Chapter 3), highlights the difference between lexical priming and her model. Both offer a psychological explanation for how words co-occur to reduce processing effort, making the models "highly compatible" (Wray, 2008: 67). However, the crucial difference lies in the fact that for Hoey single words are primed first with word sequences being later primed whereas Wray places the emphasis on whichever lexical unit "constitutes the largest form-meaning mapping so far found adequate to handle the effective manipulation of input and output" (2008: 67). This is in many cases the word, but can also be larger or smaller components—*morpheme equivalent units* (as described in Chapter 3)—which provide "a layer of wrapping that protects the components from analysis under normal circumstances" (p. 67). Therefore, according to Wray's model, words that occur in a morpheme equivalent unit can be considered to be collocation associations if they occur adjacently in text, but they "are not really 'associates' in his [Hoey's] sense at all. Rather, they are sub-parts of a single large unit, much as 'im-' often occurs adjacent to 'possible'" (p. 67).

What this means for the present research is that there are two potentially valid theories for why some authors use particular sequences of words, whereas others use different combinations. The present research focuses only on written output and so it is not possible to argue support for either theory. It might therefore be observed that Alan-3 used the word sequence *besides the point* which occurred only once in the entire author corpus and there were no occurrences of *beside the point*, yet in the BNC, *beside the point* occurs 74 times with *besides the point* occurring only once. This is therefore a rare word sequence which appears to be distinctive for Alan. What cannot be claimed is whether this is holistically processed as a single lexical item, for which Alan has no need to process the constituent words (as in Wray's needs-only analysis model), or whether it was a low-level writing mistake, or whether Alan has a specific nested priming which was appropriate in this particular context at that particular moment of text creation, just as Hoey (2005) found *around the* *world* to be five times more frequent in his corpus of newspaper articles than *round the world* (discussed in Section 6.3, p. 141). On this point, Hoey (2005) claims:

> [I]t could be that for such co-occurrences one speaker has *round the world* (and not *around the world*) primed while a second speaker is primed to use *around the world* (and not *round the world*) (p. 74).

In short, it has only been possible to speculate on what *should* be formulaic, or what *appears* to be formulaic, either for an individual author or for the group. It has been helpful to characterise the items identified as formulaic since there is a rationale for why those sequences occur and because they are a low-level feature. But what evidence exists that they actually are formulaic? For each of the methods, the case has been made that the items are formulaic (albeit with a small measure of non-formulaic material such as the single words *way* and *ways* for example). However, in strict evidential terms, there is no way of knowing whether these items actually are formulaic for these authors. Such theories are based on proposals for how the mental lexical may operate, and not on how the mental lexicon *actually* operates for each of the 20 authors under investigation in this study. In reality, this means that whilst evidence can be claimed of formulaic sequences being processed differently to novel language (Conklin & Schmitt, 2008; Erman, 2007; Erman & Warren, 2000; Underwood, Schmitt, & Galpin, 2004), this evidence focuses on groups of people, rather than on individuals. This is clearly at odds with the current research which rests on the notion of individual uses of formulaic sequences. Therefore, even if some find the label 'formulaic' in this context to be problematic, the value of the results still stand—authors do use different counts of items included in the reference list and they do use different forms of clusters, both of which have been shown to statistically vary between authors and in some cases enable the successful attribution of a Questioned Document. In short, the label may be wrong, but this marker of authorship still shows promise.

8.6 Suitability of the data for forensic linguistics research: limitations

In Section 7.5.2 (p. 164), the possibility that the data collection method may have primed or otherwise influenced the narratives produced by the authors was raised. It was demonstrated for that approach that there was no relationship between any formulaic sequences in the data-eliciting questions and the narratives. Nonetheless, it is now necessary to focus more closely on the data in order to assess the impact of this research and, consequently, the level of generalizability that can be afforded to the results. The suitability of the data for this research can therefore be organised under three key questions:

1. Are narratives appropriate data for forensic linguistics research?
2. What are the limitations of the data?
3. How generalizable are the results derived from the data?

To engage with the first question, clearly, if narratives are deemed inappropriate data on which to investigate the potential of formulaic sequences as a marker of authorship, the generalizability of the results will be severely limited. In Section 4.4.1, the argument was made that narratives were appropriate data since the participants needed a structured writing task which would be familiar to them. Clearly, however, this is not the only consideration and other issues that must be borne in mind include the nature of narratives, their relationship with occurrences of formulaic sequences and, ultimately, their appropriateness for the forensic context.

There is certainly some agreement that narratives are 'special': that they are not necessarily spontaneous and they are inherently linked to the identity of their authors. Toolan (2001) argues that one of the characteristics of a narrative is the "degree of artificial fabrication or constructedness not usually apparent in spontaneous conversation" because narratives are planned, revised and refined (p. 4). Furthermore, narratives contain a "degree of *pre*fabrication":

> In other words, narratives often seem to have bits we have seen or heard, or think we have seen or heard, before (recurrent chunks far larger than the recurrent chunks we call words). One Mills and Boon heroine or hero seems much like another—and some degree of typicality seems to apply to heroes and heroines in more elevated fictions too, such as nineteenth-century British novels (p. 4)

Of course, in the latter half of the quotation Toolan appears to be describing formulaic genres (Kuiper, 2009) but the reference to 'recurrent chunks' no doubt can be subsumed under the definition of the formulaic sequence. It may be argued that narratives inherently attract or demand a higher count of formulaic sequences than other types of text such as a letter of complaint or an application form for employment might. Furthermore, narratives may be more closely connected to identity than other texts:

> Because narratives are, relative to ordinary turns of talk, long texts and personalized or evaluated texts, there is a way in which, while your conversational remarks *reflect* who you are (your identity and values), in the course of any narrative the narrator's text *describes* that narrator. In brief snatches of conversation, a person may be able, through accent-mimicry for example, to 'pass' for someone of a different class or gender or ethnic identity; but to take on another's identity in a sustained fashion, across a number of personal narratives, is ordinarily very difficult, and may even imply disabling confusion or a personality disorder (Toolan, 2001: 3)

This is in keeping with Johnstone (1996), as discussed in Section 2.1.5, who argued that linguistic differences between people are especially evident in narratives and that "it is precisely in narrative that people's individuality is expressed most obviously, because the purpose of narrating is precisely the creation of an autonomous, unique self in discourse" (1996: 56). This creates something of a duality: narratives may well contain more prefabrication and increased opportunity for the use of formulaic sequences than other types of texts, but at the same time, since narratives are so linked to identity, the ground may well be fertile for individual style to manifest. Also, the fact that narratives may invite a higher occurrence of formulaic sequences is not sufficient grounds to dismiss the data. Ong (1982) argues:

> Human memory and language grow out of the unconscious into consciousness. Writing and print and electronic devices are produced by conscious planning—though of course their use, like all human activities, involves the unconscious as well as consciousness (p. 22).

Therefore, whilst it may well be true that the narratives analysed in this research were produced with conscious planning, there will also have been elements of unconscious planning, and since formulaic sequences are argued to be produced at a lower, less conscious level, their occurrences may still be more linked to authorship rather than being determined by genre indicating that narratives are no less suitable data than any other potential text. Clearly, finding an answer to this puzzle goes far beyond the scope of this research.

What this research cannot establish is whether the 20 authors investigated would use the same formulaic sequences or the same count of formulaic sequences in any other texts that they authored, and to this effect, there is a valid argument that the data are special. However, since this research is only in the initial stages of testing formulaic sequences as a marker of authorship, it is argued here that selecting narratives as data was reasonable and even though this particular decision limits the results, those results may still be useful for the forensic context:

> Some texts, through their content, are clearly of interest to police investigators and the wider judicial process. These texts might include, for example, threatening or abusive letters, ransom notes or sexually explicit internet conversations between middle-aged men and under-aged girls. Many texts, however, which are analysed as part of forensic casework, are not inherently criminal; they may be more mundane including for instance, personal letters and diaries. Such texts may provide an alibi or their content may assist an investigation in a less direct way (Grant, 2008: 216).

In this way, the selection of narratives as data for this study is just as valid (and equally, as limited) as any other type of text that could have been selected and clearly, the next stage in the research process will be to explore the relationship between formulaic sequences and individual language use on a wider variety of genres. Having established that narratives are as valid as any type of data for an investigation of this kind, it is next necessary to determine whether other choices made as part of the research design might have limited the generalizability of the data.

The first issue to consider is the length of the texts, and in loose terms, whether the data should have been longer or shorter than the 500 words decided upon. The decision to use shorter texts (and the arbitrariness of this term was discussed in Section 4.4.1) was motivated by the desire to make the results relevant to the forensic context, where texts are typically short. Just as Kredens (2001) argued that demonstrating differences in idiolect between two very similar speakers should make it more likely that it would be possible to find differences between dissimilar speakers, so too it can be claimed of this research that finding evidence of author specific uses of formulaic sequences in shorter texts should increase the likelihood of finding differences between authors when longer texts are analysed. Certainly, this logic would be in keeping with the notion that more data makes conclusions more solid, or as D. Foster (2001) claims: "[g]ive anonymous offenders enough verbal rope and column inches and they will hang themselves for you, every time" (p. 12). Since two of the three approaches outlined in this research met with relative success, it is likely that using longer texts would generate stronger results. Likewise, since the results were limited in some cases, it may be wise to concede that the approaches will be unlikely to work on even shorter texts (i.e. less than 500 words).

A further potential limitation of the data is the context in which they were produced. Whilst the point was repeatedly reinforced to participants that they alone should author the texts, beyond that stipulation there was no control over how the texts were produced. It is not known, for instance, whether each author dedicated a reasonable amount of time to the composition of each text or whether they rushed the task. It is not known whether there were any distractions (such as watching television or talking on the telephone) whilst composing these texts. To this end, there may be an argument for further testing of the method using data produced under laboratory conditions. However, since the results from this research indicate a level of success for two of the methods, it is unlikely that there is a need to control the data so tightly. This is not an insignificant fact given that in a forensic investigation, no data would be available which had been produced under laboratory conditions (with the possible exception of encouraging a suspect to complete a writing task under police supervision). Therefore, to suggest that the research should be replicated on more strictly controlled language would seem to be a step backward in the testing process since it would carry less forensic applicability.

The fact that the data were collected over a five day period may also be problematic. This was necessary to ensure that participants were not faced with an unmanageable task (e.g. producing 2,500 words in one day) and also to capture each author's use of formulaic sequences over a period of time, rather than in the same sitting where tiredness could have increased (or conversely decreased) the use of formulaic sequences towards the end of the task. What the data have therefore captured is each author's use of formulaic sequences on five separate occasions over five days. No claims can be made about whether the same authors would use the same formulaic sequences, or indeed count of formulaic sequences, if the research was repeated. There is clearly a need for future research to investigate changes in individual repertoires of formulaic sequences over a longer and intermittent period. One potential solution may be to adopt a similar research design to Barlow (2010). In collecting his data from the White House Press Conferences he deliberately avoided using speech produced on consecutive days, typically leaving an interval of three days, in order to reduce the likelihood of priming from one day to the next although the obvious point of departure is that those data were produced for non-experimental purposes.

It is also important to consider that in Section 4.6, several procedures for editing the data were outlined. How much effect did these procedures have on the analysis? Is editing data in this way practical under typical forensic circumstances? To deal with the second question first, editing the data was relatively straightforward since the grammar and spelling features available in *Word 2010* were available to automate the process. Therefore, editing the 65,113 word corpus was both reliable and achievable in a short period of time. Therefore, the practical issue is of minor concern. Of greater concern is whether editing the data had an effect on the analysis and whether standardising the data (correcting spellings, apostrophe misuse etc.) actually covered up idiosyncrasies or other aspects of author style. To assess this, all of the changes that were made were individually reviewed and only two of them potentially affected the results. In Keith-4, *vicer versa* was corrected to *vice versa* and *besides the point* was standardised to *beside the point* in Alan-3, both of which were subsequently identified as formulaic sequences using the reference list approach. None of the other changes made affected the analysis and the occurrence of these two formulaic sequences across the entire corpus would have been insufficient to alter the results. Therefore, editing the data to enable automated approaches was justified.

Finally then, assessing the overall generalizability of this research, the following points can be made:

- The results and conclusions derived are limited to short narratives and generalizability to other data sets is not yet possible;

- The texts were composed over a five day period so claims about consistency in formulaic sequence use over a longer period cannot be made; and
- The effects of editing the data are minimal, so analyses using texts which have potentially been auto-corrected by word-processor spelling and grammar features may be possible using the approaches outlined in this research.

Chapter 9
'All good things must come to an end': final thoughts

The research presented in these pages has been undertaken in pursuit of one objective: to determine whether formulaic sequences hold potential as a new marker of authorship. To this end, three approaches to the identification of formulaic sequences have been described along with a series of experiments designed to test whether texts produced by different authors can be differentiated, and ultimately whether a Questioned Document can be correctly attributed in a mock forensic context. The results show mixed success with one approach in particular—utilizing an all-inclusive reference list and measuring the overall count of formulaic sequences—providing compelling evidence that authors do vary in the level of formulaicity in their texts. As such, some aspects of formulaic sequences (namely, formulaic clusters and count of formulaic sequences) warrant further attention since this initial investigation indicates that there *is* evidence of authorial differences in the use of formulaic sequences.

The thesis opened with brief details from the FBI UNABOM investigation and it was proposed that the phrases *cool-headed logicians* and *you can't eat your cake and have it too* were likely to be formulaic for Theodore Kaczynski. If this proposition is accepted, it neatly accounts for why Kaczynski would include them in his writings even though he was actively evading detection—he would have been unaware that these phrases were so distinctive and characteristic of his authorial style because they were likely to be holistically stored and communicatively useful for him in a way that they would not be for other people. And because they were so communicatively useful for him, it probably never even occurred to him *not* to use them. Such sub-conscious, low level features of language use will always provide promising avenues to explore as markers of authorship.

9.1 Formulaic sequences as a marker of authorship: the next steps

The description of the limitations of the data described in the previous chapter, combined with the general discussion provided throughout this thesis suggests several areas for further research in the area of formulaic sequences as a marker of authorship. Given that this research is the first to investigate formulaic sequences as a marker of authorship for forensic application, the range of potential further research is enormous, but the most salient issues arising may be addressed:

- The effect of the authors' perception of audience—as acknowledged in Section 4.5, some of the authors wrote their narratives formally whilst others wrote far less formally. This is likely to be an effect of not being given a clear audience for whom to write. This is a strength as far as the present research is concerned, particularly since differences in the formality of documents available for comparison are likely to exist in forensic investigations. If the differential use of formulaic sequences can differentiate authors and enable Questioned Documents to be attributed to their authors, then it seems less likely that the effect of formality will be an important variable. Nonetheless, research into the wider effect of audience may be necessary to ensure reliability in the forensic context. Is the effect of writing for an unknown audience likely to affect the count of formulaic sequences used by an author?

- The effect of genre—related to the issue of audience is of course the effect of genre. Do authors remain consistent in their use of formulaic sequences across different text types? Are narratives comparable with diary entries and formal letters? The unknown effect of genre is a recurrent theme in forensic linguistic analysis and so clearly more research into this area in general, and of course with specific reference to formulaic sequences, is required. Wray (2009) argues (specifically in the L2 context, but presumably equally appropriate for the L1 context) that "[p]atterns of formulaic language, like those of vocabulary more generally and also of grammar, will vary according to genre and medium" (p. 10). Likewise, Johnstone (1996: 128—9) acknowledges that speakers undoubtedly vary their styles to take into account audience, topic, context etc. Therefore, since genre is strictly controlled in the present research, there is a need to establish whether differences in authors' use in both the type and count of formulaic sequences remains the same across different texts in order to be more confident about forensic potential.

- The effect of forensic testing—a necessarily simplistic approach to forensic testing has been adopted in this research; namely, attributing one Questioned Document to a small subset of candidate authors. Is this a useful way to test new and existing markers of authorship? Such an approach clearly does not account for all of the permutations of a forensic investigation (e.g. investigations with multiple Questioned Documents where common authorship between the Questioned set needs to be established before comparison to Known Documents and which opens the door to more than one candidate author, to give just one example). However, as a starting position, it is reasonable to investigate markers of authorship on more straight forward scenarios. It will be a challenge for future and additional research to test the limits.

- The effect of cognitive load—cognitive load is likely to have an impact on the use of formulaic sequences (e.g. Kuiper, 1996, 2000; Peters, 1983; Wray, 2002; Wray & Perkins, 2000), particularly the count of formulaic sequences as identified in Chapter 7. It stands to reason that since the data were not produced under laboratory conditions, the individual cognitive loads of each author during the creation of each text is unknown and may well have had an effect. Future testing would therefore be required to establish whether authors produce the same type and count of formulaic sequences under a range of different cognitive pressures.

- The relative occurrence of formulaic sequences in non-native speakers of English—this research has focussed only on the L1 user of English. An important consideration is that L2 users of English often have non-native speaker formulae in their formulaic repertoires (Pawley & Syder, 1983; Wray, 2002) which may in themselves be useful markers of authorship, being another example of a lexical feature which may be distinctive and consistent. Therefore, assessing how groups of language users other than L1 English speakers use formulaic sequences may provide fruitful avenues for additional research into formulaic sequences as a marker of authorship.

- Changing formulaic repertoires—how dynamic are formulaic sequences and does an individual's store of formulaic sequences change over the course of their lifetime? Will Rose still be using *in a way* so frequently in five years' time? What about people who regularly talk to Rose—will they increase their use of *in a way* as a consequence? What about people who regularly communicate with Thomas who had the highest count of formulaically-used words: will his interlocutors be sensitive to this and consequently will their own linguistic styles converge or diverge? Clearly, if formulaic repertories are not static over a lifetime or even stable across interlocutors and contexts, their reliability as a marker of authorship will be compromised.

- Deliberate disguise—it has been argued in this research that formulaic language should be a strong marker of authorship because it occurs at the subconscious level, so someone attempting to disguise their style may concentrate on increasing (or decreasing) the quantity of misspellings in their writing (e.g. Eagleson, 1994) but they may be less aware of the quantity or type of formulaic sequences they use. Empirical testing would be valuable in determining whether this is actually the case.

- Other aspects of authorship analysis—the focus in this research has been on the ability to correctly attribute a Questioned Document to one of a small pool of candidate authors. It has been widely acknowledged throughout this research that there is a large sociolinguistic component which affects formulaic language usage. Where you have grown up and lived, and who you have interacted with should shape your store of formulaic sequences. It stands to reason that if stronger connections can be made, then authorship profiling should become a realistic next step. If a relationship between age, gender and education, for example, and formulaic language usage can be established, then such evidence would be extremely useful for profiling an unknown author. Whilst the results in this research were unable to establish any sociolinguistic variation along these variables, this is not evidence that it does not exist at all—just that there is insufficient evidence in this set of data.

Naturally, it has been possible to only begin to scratch the surface and far more research is required to establish the parameters in which this marker of authorship can be applied, but as an initial investigation, the future is certainly promising. As outlined above, more research is clearly needed both for the identification of formulaic sequences in text, and for their applications in forensic investigations, extending potentially, to authorship profiling if clear sociolectal formulas can be identified. There is also a wealth of literature which deals with formulaic language usage by linguistically-defined sub-groups (such as bilinguals, second language users and those with communication disorders) and whilst case studies of individual users of formulaic language exist, these appear to be limited to 'special' language users such as infants acquiring their first language (e.g. Peters, 1977; cf. Vihman, 1982 for overview) rather than normally developed adult native speakers. Importantly, therefore, it is proposed here that future research into formulaic language should focus on the individual. What is it that motivates the use of formulaic language? How dynamic or static are formulaic sequences in the mental lexicon? How does formulaic language usage change, if at all, over a lifetime? How much impact does interaction with another interlocutor have on our individual stores of formulaic sequences? By finding answers to these key questions, not only will formulaic language receive the attention it deserves, clearer evidence of idiolect may also be established.

In conclusion, while this research describes only a very small investigation into the application of formulaic sequences as a marker of authorship it is distinctive because it places the individual author at the centre of the investigation. Although a Questioned Document cannot always be correctly attributed to its author, the existence of variation in the use of formulaic sequences in one text genre has become an empirically supported fact. Authors do use formulaic sequences distinctively and distinguishably.

REFERENCES

Bagavandas, M., & Manimannon, G. (2008). Style consistency and authorship attribution: a statistical investigation. *Journal of Quantitative Linguistics, 15*(1), 100—110.

Baker, J. (1988). Pace: a test of authorship based on the rate at which new words enter an author's text. *Literary and Linguistic Computing, 3*(1), 36—39.

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.

Barlow, M. (2010). 'Individual Usage: A corpus-based study of idiolects'. Paper presented at L*AUD Conference*, Landau. [WWW] http://michaelbarlow.com/barlowLAUD/pdf [Accessed: August, 2012.

Bel, N., Queralt Estevez, S., Spassova, M. S., & Turell, M. T. (2012). The use of sequences of linguistic categories in forensic written text comparison revisited. In S. Tomblin, N. MacLeod, R. Sousa-Silva & M. Coulthard (eds), *Proceedings of the International Association of Forensic Linguists' Tenth Biennial Conference*. Aston University, Birmingham, UK: The Centre for Forensic Linguistics. Retrieved April 2012 from www.forensiclinguistics.net. 192—209.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics, 14*(3), 275—311.

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*. Amsterdam: Rodopi. 181—190.

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

Bloch, B. (1948). A set of postulates for phonemic analysis. *Language, 24*, 3—46.

Bond, G., & Lee, A. (2005). Language of lies in prison: linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19*, 313—329.

Burrows, J. (2002). 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267—287.

Campbell, L. (2004). *Historical Linguistics: An introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

Carter, R. (1999). Common language: corpus, creativity and cognition. *Language and Literature, 8*(3), 195—216.

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), 1—65.

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

Clement, R., & Sharp, D. (2003). Ngram and Baysian classification of documents for topic and authorship. *Literary and Linguistic Computing, 18*(4), 423—447.

Clemit, P., & Woolls, D. (2001). Two new pamphlets by William Godwin: a case of computer-assisted authorship attribution. *Studies in Bibliography, 54*, 265—284.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: are they processed more quickly than non-formulaic language by native and non-native speakers? *Applied Linguistics, 29*(1), 72—89.

Corrigan, R., Moravcsik, E., Ouali, H., & Wheatley, K. (2009). Introduction. Approaches to the study of formuale. In R. Corrigan, E. Moravcsik, H. Ouali & K. Wheatley (eds), *Formulaic Language: Distribution and historical change* Vol. 1. Amsterdam: John Benjamins Publishing Co. xi—xxiv.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes, 23*, 397—423.

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

Coulmas, F. (1981). Introduction: conversational routine. In F. Coulmas (ed.), *Conversational Routine: Explorations in standardized communication situations and prepatterned speech*. The Hague, Netherlands: Mouton Publishers. 1—17.

Coulthard, M. (1994). On the use of corpora in the analysis of forensic texts. *Forensic Linguistics, 1*(1), 27—43.

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

Coulthard, M. (2005a). The linguist as expert witness. *Linguistics and the Human Sciences, 1*(1), 39—58.

Coulthard, M. (2005b). Some forensic applications of descriptive linguistics. *VEREDAS - Revista de Estudos Linguísticos, 9*(1—2), 9—28.

Coulthard, M. (2010). Experts and opinions: in my opinion. In M. Coulthard & A. Johnson (eds), *The Routledge Handbook of Forensic Linguistics*. Abingdon, Oxford: Routledge. 473—486.

Coulthard, M., & Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in evidence*. London/New York: Routledge.

Covella, F. D. (1976). Grammatical evidence of multiple authorship in *Piers Plowman*. *Language and Style, 9*(1), 3—16.

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), 109—127.

de Vel, O., Anderson, A., Corney, M., & Mohay, G. (2001). Multi-topic e-mail authorship attribution forensics, *ACM Conference on Computer Security: Workshop on data mining for security applications*. Philadelphia, PA, USA.

Eagleson, R. (1994). Forensic analysis of personal written texts: a case study. In J. Gibbons (ed.), *Language and the Law*. London: Longman. 362—373.

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

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics, 12*(1), 25—53.

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

Farringdon, J. (1996). *Analysing for Authorship: A guide to the Cusum technique*. Cardiff: University of Wales Press.

Feiguina, O., & Hirst, G. (2007). Authorship attribution for small texts: literary and forensic experiments, *SIGIR '07: Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*. Amsterdam.

Fillmore, C. (1979). On fluency. In C. Fillmore, D. Kempler & W. S.-Y. Wang (eds), *Individual Differences in Language Ability and Language Behavior*. London: Academic Press. 85—101.

Finegan, E. (1990). Variation in linguists' analyses of author identification. *American Speech, 65*(4), 334—340.

Fitzgerald, J. R. (2004). Using a forensic linguistic approach to track the Unabomber. In J. H. Campbell & D. Denivi (eds), *Profilers*. New York: Prometheus Books. 193—221.

Fokkema, D., & Ibsch, E. (1987). *Modernist Conjectures: A mainstream in European literature 1910—1940*. London: Hurst.

Foster, D. (2001). *Author Unknown: On the trail of anonymous*. Basingstoke: Macmillan.

Foster, P. (2001). Rules and routines: a consideration of their role in the task-based production of native and non-native speakers. In M. Bygate, P. Skehan & M. Swain (eds), *Researching Pedagogic Tasks: Second language learning, teaching and testing*. London: Longman. 75—94.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes, 19*, 115—135.

Goldberg, A. (1996). Making one's way through the data. In M. Shibatani & S. Thompson (eds), *Grammatical Constructions: Their form and meaning*. Oxford: Oxford University Press. 29—53.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: are we barking up the wrong tree? *Applied Linguistics, 25*(1), 38—61.

Grant, T. (1992). *Evaluation of the Cusum Method for the Attribution of Authorship.* Unpublished MSc., University of Birmingham, Birmingham.

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.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons & M. T. Turell (eds), *Dimensions in Forensic Lingusitics*. Amsterdam: John Benjamins Publishing Co. 215—229.

Grant, T. (2010). Text messaging forensics: txt 4n6: idiolect free authorship analysis? In M. Coulthard & A. Johnson (eds), *The Routledge Handbook of Forensic Linguistics*. Abingdon, Oxford: Routledge. 508—522.

Grant, T. (2011). TXT 4N6: Methods for the forensic authorship analysis of SMS text messages. *Unpublished Paper*.

Grant, T., & Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics: The International Journal of speech, Language and the Law, 8*(1), 66—79.

Grieve, J. (2007). Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251—270.

Hancock, J., Curry, L., Goorha, S., & Woodworth, M. (2004). *Lies in Conversation: An examination of deception using automated linguistic analysis.* Paper presented at the Procedings of the Annual Conference of the Cognitive Science Society.

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

Hardcastle, R. A. (1997). CUSUM: a credible method for the determination of authorship? *Science and Justice, 37*(2), 129—138.

Hatch, E., & Farhady, H. (1982). *Research Design and Statistics for Applied Linguistics*. London: Newbury House Publishers, Inc.

Heiman, G. (1999). *Research Methods in Psychology* (2nd ed.). New York: Hoighton Mifflin Company.

Herbst, T. (1996). What are collocations: sandy beaches or false teeth? *English Studies, 4*, 379—393.

Hinton, P.R. (2004), *Statistics Explained* (2nd ed.). London: Routledge.

Hockett, C. (1958). *A Course in Modern Linguistics*. New York: The Macmillan Company.

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

Holmes, D., & Forsyth, R. (1995). The *Federalist* revisited: new directions in authorship attribution. *Literary and Linguistic Computing, 10*(2), 111—127.

Hoover, D. L. (2001). Statistical stylistics and authorship attribution: an empirical investigation. *Literary and Linguistic Computing, 16*(4), 421—444.

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

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

Hoover, D. L. (2003b). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing, 18*(4), 341—360.

Hoover, D. L. (2004a). Delta prime? *Literary and Linguistic Computing, 19*(4), 477—495.

Hoover, D. L. (2004b). Testing Burrows's delta. *Literary and Linguistic Computing, 19*(4), 453—475.

Howald, B. (2008). Authorship attribution under the rules of evidence: empirical approaches-a layperson's legal system. *The International Journal of Speech, Language and the Law, 15*(2), 219—247.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24—44.

Johnson, A. (1997). Textual kidnapping—a case of plagiarism among three student texts? *Forensic Linguistics, 4*(2), 210—225.

Johnstone, B. (1996). *The Linguistic Individual: Self-expression in language and linguistics*. Oxford: Oxford University Press.

Johnstone, B. (1997). Social charactersitics and self-expression in narrative. *Journal of Narrative and Life History, 7*(1—4), 315—320.

Juola, P., Sofko, J., & Brennan, P. (2006). A prototype for authorship attribution studies. *Literary and Linguistic Computing, 21*(2), 169—178.

Kecskés, I. (2000). A cognitive-pragmatic approach to situation-bound utterances. *Journal of Pragmatics, 32*, 605—625.

Kinnear, P. & Gray, C. (2000). *SPSS for Windows Made Simple: Release 10*. Hove, East Sussex: Psychology Press Ltd.

Kniffka, H. (2007). *Working in Language and Law: A German perspective*. 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*. Frankfurt: Peter Lang. 405—446.

Kredens, K. (2002). Idiolect in forensic authorship attribution. In P. Stalmaszczyk (ed.), *Folia Linguistica Anglica* Vol. 4. Lodz: Lodz University Press. 191—212.

Kredens, K. (2006). On the status of linguistic evidence in litigation. In P. Nowak & P. Nowakowski (eds), *Language, Communication, Information 1(1)*. Poznan: Sorus Publishers.

Kuhl, J. (2003). *The Idiolect, Chaos, and Language Custom Far From Equilibrium: Conversations in Morocco*. Unpublised PhD, The University of Georgia, USA. [WWW] http://athenaeum.libs.uga.edu/bitstream/handle/10724/6745/kuhl_joe_w_200308_phd.pdf ?sequence=1 [Accessed: July, 2012].

Kuiper, K. (1996). *Smooth Talkers: The linguistic performance of auctioneers and sportscasters*. New Jersey: Lawrence Erlbaum.

Kuiper, K. (2000). On the linguistic properties of formulaic speech. *Oral Tradition, 15*(2), 279—305.

Kuiper, K. (2004). Formulaic performance in conventionalised varieties of speech. In N. Schmitt (ed.), *Formulaic Sequences: Acquisition, processing and use*. Amsterdam: John Benjamins Publishing Company. 37—54.

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

Künzel, H. J. (2010). Automatic speaker recognition of identical twins. *The International Journal of Speech, Language and the Law, 17*(2), 251—277.

Labov, W. (1970). The study of language in its social context. In J. B. Pride & J. Holmes (eds), *Sociolinguistics: Selected Readings*. Harmondsworth: Penguin. 180—202.

Labov, W. (1972a). *Language in the Inner City: Studies in the Black English vernacular*. Oxford: Basil Blackwell.

Labov, W. (1972b). *Sociolinguistic Patterns*. Oxford: Basil Blackwell.

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

Lancashire, I. (1998). Paradigms of authorship. *Shakespeare Studies, 26*, 296—301.

Leech, G., & Short, M. (2007). *Style in Fiction: A linguistic introduction to English Fictional Prose*. Harlow: Pearson Education Limited.

Loakes, D. (2006). A forensic phonetic investigation into the speech patterns of identical and non-identical twins. *The International Journal of Speech, Language and the Law, 15*(1), 97—100.

Louwerse, M. (2004). Semantic variation in idiolect and sociolect: corpus linguistic evidence from literary texts. *Computers and the Humanities, 38*, 207—221.

Love, H. (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Mannion, D., & Dixon, P. (2004). Sentence-length and authorship attribution: the case of Oliver Goldsmith. *Literary and Linguistic Computing, 19*(4), 497—508.

McMenamin, G. (1993). Appendix 1: descriptive markers of style. *Forensic Science International, 58*(1—2), 183—188.

McMenamin, G. (2002). *Forensic Linguistics: Advances in forensic stylistics*. London: CRC Press.

McMenamin, G. (2004). Disputed authorship in US law. *The International Journal of Speech, Language and the Law, 11*(1), 73—82.

McMenamin, G. (2010). Forensic stylistics: theory and pratice of forensic stylistics. In M. Coulthard & A. Johnson (eds), *The Routledge Handbook of Forensic Linguistics*. Abingdon, Oxford: Routledge. 487—507.

Menacere, T., Taylor, P. J., & Tomblin, S. (2008). Linguistic Analysis Suite (Version 1.3). Lancaster University: Department of Psychology.

Mollet, E., Wray, A., Fitzpatrick, T., Wray, N. R., & Wright, M. J. (2010). Choosing the best tools for comparative analyses of texts. *International Journal of Corpus Linguistics, 15*(4), 429—473.

Mollin, S. (2009). "I entirely understand" is a Blairism: the methodology of identifying idiolectal collocations. *International Journal of Corpus Linguistics, 14*(3), 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. (1998a). *Fixed Expressions and Idioms in English*. Oxford: Clarendon Press.

Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (ed.), *Phraseology: Theory, analysis and applications*. Oxford: Oxford University Press. 79—100.

Mosteller, F., & Wallace, D. (1963). Inference in an authorship problem: a comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers. *Journal of the American Statistical Association, 302*(58 (June)), 275—309.

Mosteller, F., & Wallace, D. (2007). *Inference and Disputed Authorship: The Federalist*. Leland Stanford Junior University, USA: CSLI Publications.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

Oakes, M. (1986). *Statistical Inference: A commentary for the social and behavioural sciences*. Chichester: John Wiley & Sons.

Olsson, J. (2004). *Forensic Linguistics: An introduction to language, crime and the law*. London: Continuum.

Ong, W., J. (1982). Oral remembering and narrative structures. In D. Tannen (ed.), *Analyzing Discourse: Text and Talk*. USA: Georgetown University Press. 12—24.

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

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.

Pennebaker, J., & Lay, T. (2002). Language use and personality during crises: analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36*, 271—282.

Peters, A. (1977). Language learning strategies: does the whole equal the sum of the parts? *Language, 53*(3), 560—573.

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.

Putter, A. (2004). The language and metre of *Pater Noster* and *Three Dead Kings*. *The Review of English Studies, 55*(221), 498—526.

Read, J., & Nation, P. (2004). Measurement of formulaic sequences. In N. Schmitt (ed.), *Formulaic Sequences*. Amsterdam: John Benjamins Publishing Co. 23—35.

Römer, U. (2005). *Progressives, Patterns, Pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins Publishing Company.

Rubin, D. C. (1998). *Memory in Oral Traditions: The cognitive psychology of epic, ballads, and counting-out rhymes*. Oxford: Oxford University Press.

Sanford, A., Aked, J., Moxey, L., & Mullin, J. (1994). A critical examination of assumptions underlying the Cusum technique of forensic linguistics. *Forensic Linguistics: The International Journal of speech, Language and the Law, 1*(2), 151—167.

Sapir, E. (1927). Speech as a personality trait. *American Journal of Sociology*, 32(6), 892—905.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: an introduction. In N. Schmitt (ed.), *Formulaic Sequences: Acquisition, processing and use*. Amsterdam: John Benjamins Publishing Company. 1—22.

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

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.

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

Simpson, R., & Mendis, D. (2003). A corpus-based study of idioms in academic speech. *TESOL Quarterly, 37*(3), 419—441.

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

Sinclair, J. (1999). A way with common words. In H. Hasselgård & S. Oksefjell (eds), *"Out of Corpora": Studies in honour of Stig Johansson*. Amsterdam: Rodopi. 157—179.

Sinclair, J. (2004). *Trust the Text: Language. corpus and discourse*. London: Routledge.

Smith, M. W. A. (1986). A critical review of word-links as a method for investigating Shakespearean chronology and authorship. *Literary and Linguistic Computing, 1*(4), 202—206.

Smith, W. (1994). Computers, statistics and disputed authorship. In J. Gibbons (ed.), *Language and the Law*. London: Longman. 374—413.

Solan, L. (2010). The forensic linguist: the expert linguist meets the adversarial system. In M. Coulthard & A. Johnson (eds), *The Routledge Handbook of Forensic Linguistics*. Abingdon, Oxford: Routledge. 395—407.

Solan, L., & Tiersma, P. (2004). Author identification in American courts. *Applied Linguistics, 25*(4), 448—465.

Solan, L., & Tiersma, P. (2005). *Speaking of Crime: The language of criminal justice*. London: The University of Chicago Press.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language, 2*(1), 23—55.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics, 7*(2), 215—244.

Stubbs, M. (2005). Conrad in the computer: examples of quantitative stylistics methods. *Language and Literature, 14*(1), 5—24.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language, 10*(1), 61—104.

Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior, 18*, 523—534.

Tiersma, P., & Solan, L. (2002). The linguist on the witness stand: forensic linguistics in American courts. *Language, 78*(2), 221—239.

Toolan, M. (2001). *Narrative: A critical linguistic introduction* (2nd ed.). Abingdon: Routledge.

Trudgill, P. (1974). *Sociolinguistics: An introduction to language and society*. London: Penguin Books Ltd.

Trudgill, P. (2003). *A Glossary of Sociolinguistics*. Edinburgh: Edinburgh University Press.

Turell, M.T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law*, 17(2), 211—250.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: an eye-movement study into the processing of formulaic sequences. In N. Schmitt (ed.), *Formulaic Sequences*. Amsterdam: John Benjamins Publishing Co. 153—172.

Van Lancker-Sidtis, D., & Rallon, G. (2004). Tracking the incidence of formulaic expressions in everyday speech: methods for classification and verification. *Language & Communication, 24*, 207—240.

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.

Walker, I. (2010). *Research Methods and Statistics*. Basingstoke: Palgrave Macmillan.

Waltman, F. W. (1973). *Formulaic* expression and unity of authorship in the "Poema de Mío Cid". *Hispania, 56*(3), 569—578.

Wardhaugh, R. (2006). *An Introduction to Sociolinguistics* (5th ed.). Oxford: Blackwell Publishing.

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*. Frankfurt am Main: Peter Lang. 141—179.

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: a case study. *Canadian Journal of Applied Linguistics, 12*(1), 39—57.

Woodhams, J., & Grant, T. (2006). Developing a categorisation system for rapists' speech. *Psychology, Crime and Law, 12*, 245—260.

Woodhams, J., Grant, T., & Price, A. (2007). From Marine Ecology to Crime Analysis: Improving the detection of serial sexual offences using a taxonomic similarity measure. *Journal of Investigative Psychology and Offender Profiling, 4*, 17—27.

Woodhams, J., & Toye, K. (2007). An empirical test of the assumptions of case linkage and offender profiling with serial commerical robberies. *Psychology, Public Policy, and Law, 13*(1), 59—85.

Woods, A., Fletcher, P. & Hughes, A. (1986). *Statistics in Language Studies*. Cambridge: Cambridge University Press.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics, 21*(4), 463—489.

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.

Wray, A. (2009). Future directions in formulaic language research. *Journal of Foreign Languages, 32*(6), 2—17.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language & Communication, 20*, 1—28.

APPENDIX A – Information sheet and declaration form sent to potential participants

Information Sheet and Declaration Form

You are being invited to take part in a research investigation that is being conducted by a Doctoral Candidate in The Centre for Language and Communication Research at Cardiff University.

The research aims to learn more about how people write about personal experiences. The only stipulations are that you are a native speaker of English and have daily access to your e-mail account.

If you agree to take part, each morning you will be sent two questions which will elicit a short story and you will be required to answer only one of them. Two example questions include: *What is the most frightened you have ever been? What is the most vivid dream you have ever had?*

You will be required to make your answer one page long (approximately 500 words) and you will be given an electronic document in which to type your answer. You will be given guidance on how you should fill in the electronic document. You must return your answer each day, by 9pm at the latest. You should therefore make sure that you have daily access to your e-mail account. The investigation will last for five days so at the end you will have answered just five questions.

It is essential that you are the only person who answers these questions and you must not use materials from any other source. However, you will be allowed to change any dates, names or other information that you wish and you are allowed to elaborate your stories if you need in order to fill the page.

Once the fifth answer has been received, you will receive payment for your participation. You will be sent a cheque for £10.

If you encounter any problems which prevent you from returning your answer by the daily deadline, you must contact the Principle Investigator, Samuel Tomblin, as soon as possible. You may do this by e-mail (TomblinSD@cardiff.ac.uk) or by Telephone (07845 705648). It may be possible to make alternative arrangements, although payment will only be made when all five questions have been fully completed.

On the next page is an example of the electronic document that you will be sent each day into which you will be expected to type your answer. It should give you an indication of how much you will be required to type in order to fill the box. You will not be allowed to change the size of the box so you must finish when you reach the end. It is formatted so that text is entered in Arial 11pt font which you also cannot change.

Name:

Question Answered: What is the most vivid dream you have ever had?

Please begin typing your answer in the box below. Stop when you reach the end of the box.

SAMPLE

STOP TYPING HERE AND RETURN THE FORM

If you would like to take part in this research, please continue reading and fill in the information on the next page.

If you do not think that this research will suit you, you do not need to take any further action. Thank you for considering this study.

If you are undecided about whether this research will suit you, please feel free to contact Samuel Tomblin on the above e-mail address or telephone number for an informal, non-committal discussion.

Please read the following six statements and check the boxes to indicate that you agree to the terms of this research (place your mouse over the box and click once).

I am a native speaker of English □

I understand that I will be required to write one answer of approximately 500 words every day for five days. I must return my answers by 9pm each day unless I have made alternative arrangements with the Principal Investigator □

I understand that I will only be paid £10.00 when I have returned five fully completed answers, each of approximately 500 words □

I understand that my answers may be used for publication and presentation purposes. I stipulate that if my data are to be used in this way, they must be made anonymous □

I understand that I alone must write the answers to the questions and must not use materials from anywhere else □

I understand that I may withdraw from this study at any time without reason, although I will not be paid for my time □

When you have agreed to all six statements above, please complete the information required below.

Full Name:

E-Mail Address:

Telephone Number:

Address (to which payment should be sent):

Age:

Gender: Male

Highest Qualification Gained:

Earliest Date When the Research Can Begin:

For Current Students Only:

Degree Scheme Registered for:

Year of Study:

Now save this form using the following format: *Surname* Declaration e.g. Smith Declaration. You should save it as a *Microsoft Word* Document by using the *.doc* extension. Please save this entire document as it is – do not copy and paste the sections into a new blank file.

APPENDIX B – Five texts authored by Melanie

MELANIE – 1 – When did you last cry and what made you cry?

Last week was the last time that I cried. We met Amanda and Josh when Phil began coaching their son Craig in football. However, it turned out that we had known each other years ago when we were in our late teens and twenties.

They had been invited to Phil's 50th birthday party and then were invited to our New Year party. Two years ago Josh had become ill. It had started out with hangover or flu-like symptoms and we all thought that he would be out of action for a couple of weeks, the doctors would find a cure and everything would go back to normal. After all we were still all young and so obviously nothing too bad could happen. We put plans on hold and made promises about what we would do once Josh was well again but Josh only got worse.

From being tired and feeling hungover his symptoms progressed to dizziness and vomiting. The hospital gave him blood tests for everything from CJD to AIDS; they said it was post-viral; could be ME and eventually held their hands up and admitted that they didn't know what was wrong with him. They then left Amanda alone with a man who could do little or nothing for himself.

With some urging from friends Amanda persisted and fought with the Local Health Authority to find help for Josh. She called on the services of her MP who did help and the Health Authority gave funding for Josh to go to a International Specialist Centre in London. They were very helpful and adjusted his medication but even with the most sophisticated machines they could find no trace of the suspected brain tumour. Josh came home.

Two years after the onset of his symptoms Josh died with no definite diagnosis. Amanda was absolutely devastated. She had gone out with Josh since she was 14 and wanted nothing but to live out her life with Josh no matter what his condition was. Amanda's grief was extremely raw and I cried as she told me how she would have Josh any way as long as she had him. I felt so helpless, I would have given anything to make her feel better but all I could say were crass statements such as that Josh wouldn't have wanted to exist like that; to have been such a burden to her. All this to a woman would have given anything to have him back with her.

The funeral was where I last cried. How could you not cry? The church was packed to the seams with many people standing; one man even fainted. I cried because Josh was young, 47, and a waste of a young life with so much still to do: he was so very badly missed by his family and I also probably cried from fear that something like that might happen to my family.

I often think that at these times we might feel better if we were more visible or audible in our grief. In the church everyone was surreptitiously wiping away tears, no one was blaming God or the Doctors for Josh not being there. The Rector had told Amanda that she should shout at God for depriving her of Josh, I hope she did, I would have done.

MELANIE – 2 – What were the consequences of a lie you have told?

I do not tell lies. Well I do not tell bad lies. I will tell a lie if that means I won't hurt somebody's feelings but because I have a very poor memory about what I have told various people I would soon get caught out if I told lies.

As a child I remember a very bad lie that I told and I still feel guilt even today and wonder if I shouldn't 'do' something about making reparation. I was always the child that could be trusted; was responsible and trustworthy. As a consequence of this I was made a biscuit monitor. I was about 9 or 10 years old and at playtime the school sold biscuits which I suppose were to raise money. I was made a biscuit monitor and we were the people who had to go and sell the biscuits each playtime. Sandra Moore was another monitor and I had recently become her friend along with another couple of girls whose names I cannot remember.

My parents had never had much money, my father, who worked at various manual labouring jobs, was often out of work and there were three younger children to care for too. This meant that there was never enough money to go around let alone spare money for luxuries such as biscuits! So, I rarely had any money to buy a biscuit. If I got the chance of doing a small job for one of the neighbours, then they might give me a penny, threepence or sometimes even sixpence (all old money) to spend.

Sandra Moore never seemed to have this problem. She always had some biscuits at playtime and yet I kind of knew that she was from a similar background to me. I really don't know how she came to tell me how to steal the biscuits but she did. Coincidentally she must have told me about there not being a Father Christmas because I remember her showing me where all her Christmas presents were well before Christmas and then going home and looking and finding mine! I remember that she said that she would take the biscuits and flick the money in the dish to make it seem that she had put money in for them. I must have thought that I would try this and did.

This went on for some time and the Mr Bleck (the Headmaster) called me into his office and asked me if I was stealing biscuits. Apparently Sandra Moore had told on me! I denied that I had stolen any biscuits but my face MUST have given me away because I have always been a terrible blusher. I probably saw my life flash before my eyes. I remember feeling terrible shame and then he said that he didn't think that I would have done something like that. I must have felt grateful that he believed my lie but also must have felt shame because I had stolen the biscuits.

It seems an inadequate reason that I never had money to buy a biscuit and I suppose I was not the only child at that school to be in that situation, the area was hardly wealthy. It has remained a major lesson in my life that if you lie you will be found out and even if your lies are believed they are still lies. I have tried to make sure that my son doesn't lie, I have always told him that I must always be able to believe him. I read a book recently where one of the characters had been told that the worst thing you can do to a person is to lie because you then steal something from that person. I probably would agree with this.

MELANIE – 3 – How did you find out that Santa doesn’t exist?

Of course Santa Claus exists, everybody knows that. How else can everybody get presents at Christmas. Actually I don't remember the exact time and place that I discovered Santa didn't exist. I know that a school friend showed me her presents, that her parents had hidden and she had found, and I certainly knew after that.

I had probably had my doubts about Santa Claus for a long time. I remember when I was about 6 or 7 I was into Sindy and Tressy. Tressy was a similar kind of doll to Sindy but she actually grew her hair. This worked with a key in her tummy that you wound up or down to make the hair grow or shorten. I must have thought this was wonderful and desperately wanted one but not only did I want a Tressy I wanted all her clothes as well. Not an unreasonable request when you remember that Santa Claus brings children toys; he doesn't go to the shops and buy them; parents don't have to pay because it is nothing to do with them. On Christmas Day I was very disappointed, I had got Tressy but no clothes. I can still remember the disappointment.

One year we were lucky enough to be taken to Lewis's in Manchester to actually see Father Christmas, as we called him then. There was a huge, long queue, through this dim hall which had little tableaux showing scenes from Father Christmas's Grotto; The elves working away on the toys and reindeer flying through the air pulling the sleigh. Eventually we got to Father Christmas and he asked us what did we want him to bring us. I completely accepted that this was Father Christmas so I was a little tongue tied but I did manage to tell him what wanted. Then because he must have been on commission he asked both my brother and myself did we want a box of chocolates as well. Obviously we said yes please. Years later, whilst talking about this to my mother, she said that she could have hit him because she then had to go and buy us a box of chocolates each, money was never plentiful in our house. I did ask myself why then didn't she buy me all of Tressy's clothes!

When the positions were reversed and it became my turn to be Santa Claus I kept these events in mind when it was time to buy presents for my son. I only remember one sticky moment. He was about three and I asked him what he wanted Santa to bring him and he said a ban harster. Even after having it repeated a couple of times I still couldn't think what he was trying to say. Eventually it clicked. He wanted a combine harvester. He was in the tractor phase and thank god for Tonka trucks, they actually made a big combine harvester. He never played with it much though, I played with Tressy lots more.

All things considered I think Santa is the best thing. How could you not. Only if you don't have the money to be able to make your child’s dreams come true and that is an awful lot of people these days. That is the problem though. On one hand children are told that this wonderful old man brings children their hearts desire and on the other hand some children are fed this dream then on Christmas Day their reality is not quite the same dream.

MELANIE – 4 – Describe a life-threatening situation you've had

I have never been in a life threatening situation, well at least never in a way that would ordinarily be thought of as life threatening. I have never been in a car accident or in bank robbery. I have had a few close brushes with serious illness that could have been life threatening, but thankfully I am still here to tell the tale.

When I was about seventeen I was at a local nightclub with a friend. I started to feel unwell and lost my sight. I managed to walk to somewhere to sit down and my friend sat with me until my sight came back and I think I carried on with then night out. I probably had a couple more drinks, the usual night out thing, and I just assumed that I had nearly fainted, thinking no more about it.

Shortly after I was married I had to call the doctor out one night as I had come home from work with a sore leg, I then began to get pains in my chest. Not heart attack pains, just pains in my chest. The doctor diagnosed constipation and a sprained ankle and left after prescribing laxatives and a bandage. Eventually, after a hot curry, much better than suppositories!, and some lager, I eventually felt better. I thought no more about this.

After a miscarriage when I had suffered a couple of deep vein thrombosis and my body had begun to destroy my own blood and a successful pregnancy where I had also suffered more deep vein thrombosis and then the start of the failure of my placenta the doctors at the hospital obviously thought something was amiss and sent a sample of my blood to London to see what might be happening.

The doctors told me that I had a rare auto-immune disease that could be controlled but it might be difficult to have another successful pregnancy. I accepted this and life went on. I was told that the international specialist centre was in London and if I wanted to go they could arrange it but never bothered, thinking that they couldn't do any more than my doctors were already doing.

After a major stroke and other blood problems I eventually decided the time had come to go to London and see what they had to say about things. London asked that I have a MRI head scan and other tests before I went down, then they would have the results and draw their own conclusions.

When I saw Dr Phelps she said that I had had four strokes and my medication needed adjusting. This was the surprise. I knew I had had a stroke when I was twenty nine but the other three were a surprise. It wasn't until I sat and thought about it that I realised that the incident in the night club had been a stroke. I suppose that it was good that I was drinking because this would have had the effect of thinning my blood which is one of the complications of my disease. This also could be the case for the supposed constipation which is suspected as being a blood clot on my lungs.

These are the nearest I have ever come to being in a life threatening situation and as near as I ever want to be, well, I suppose that I could have left some accidents in my wake whilst driving but……..

MELANIE – 5 – What has been your most embarrassing moment?

Well that depends on the degree of embarrassment. It depends on how well you know the people you have embarrassed yourself in front of. It depends on so many different factors. For instance in Asda on day, quite unexpectedly, I farted. I was in the aisle with one other person that I did not know. I did not make eye contact, in fact I pretended nothing had happened. But I was so embarrassed.

I often embarrass myself by saying something before I think about what I am saying. The other day Phil said that one of his footballers had rung him and I had been speaking about big willies. I have no idea what I had said or in what context but when I see him I will be embarrassed but will have to bluff it out.

As I was walking past church one day there were some ladies in the Church yard and I called out that I knew what they were doing. (I did know them quite well). Unfortunately I didn't, because it was an anniversary of the death of one of their loved ones and they were paying their respects. Ground and open up come to mind.

I have done some things where I wished I was somewhere else but it is still the same kind of embarrassment and ninety nine times out of a hundred it is something I have said rather than done.

I really wish that I had a good singing voice or a talent for dancing or acting. I haven't. It is only the fear of embarrassment that prevents me from proving this in front of an audience. Programmes such as The X Factor amaze me, there are always so many people who have god-awful voices yet they never believe the judges. They never seem to suffer embarrassment yet they should do.

I can't think of any other seriously bad occasions where I have embarrassed myself. I suppose that I forget these types of things very quickly; thankfully. However, I also am getting to the age where I am beginning to think 'who gives a shit?' Twenty two plus years ago I was embarrassed when my new born son sneezed out a huge bogey onto his clean cot sheets that the nurse had just changed for me. I was mortified. Today I wouldn't think twice about it which is infinitely the better option.

I have tried very hard to try and think of embarrassing situations and I really can't think of many at all. This could be one of the few benefits of getting older and poor memory or I have never really been that embarrassed which could definitely be the case. As a child and teenager I was more than happy to sit in the background, which lessens the chances of a faux pas and consequent embarrassment and as I get older I can't really say that I care that much if I am embarrassed. It could also be that as you get older you are more sure of your boundaries; of what you can and can't do.

Considering that I know that if I embarrass myself then it is usually something I say, often speaking before fully thinking, why do I still manage to do this?

APPENDIX C – Five texts authored by Thomas

THOMAS – 1 – When did you last cry and what made you cry?

The last time I cried was at the end of August this year when I had to tell a girl I love very much that I no longer wanted to be in a relationship with her, as I needed my own space and freedom to go out and just be free.. I felt so confused I didn't know if I was making the right decision, part of me wanted to stay and console her after making the announcement, part of me wasn't sure if the decision in my mind was final, part of me thought it was better to go home and leave her with time to adjust to things. We sat talking with me doing lots of crying and apologising. I could not stop crying I didn't know what to do, it was like I knew it had to be done, but I can't bear to hurt people. I was the only one crying the she was just taking everything in and being very calm and rational although I could see the hurt in her eyes. After about an hour I managed to compose myself put on a brave face and get myself back to the station to come home, then something stopped me and I began to get emotional again, I couldn’t go through the gates to get on the train and instead rung her saying I didn't know what to do, by this time I was crying again, I felt i needed to go back and try and make everything ok, but at the same time knew that really there wasn't really a way I could do this. After about another 10 minutes on the phone I went and got on the train and sat down wiping the tears from my eyes aware that i must have looked like i had been crying as people were looking at me. I then got off the train and then got back on again, clearly my mind was not settled. I resolved myself that I was going home, and tried to compose myself. What helped take my mind off things was a little girl that was on the train with her mum who was playing with her toys. I just couldn't help watching her as she was playing. Her sweet simple life and her amusement and joy allowed me to escape for an hour on the journey home. I then had time to myself to go back over things which made it hurt again. I have seen her several times since and we have talked and talked about things both managing to stay composed. Now it is almost like my head is back in control and normality has been restored, it worries me slightly that my head and my heart have different levels of control over what I do, and that maybe the balance isn't what it should be. I have not cried since and almost feel numb when looking at the situation now, I don't feel a need for any more emotional outbursts, I think it is true that you do get all cried out. Crying is such a basic human reaction and I don't know why there is such a stigma attached to it about Men crying. It is an important way of releasing emotions that build up inside. I have never experienced tears of joy, only of sadness. People keep telling me that it is inevitable in life that you will hurt people, but it just kills me every time. I get past it and am able to look back and feel sure that i did things because I had to. I am so strong headed but when it comes to the heart it is a whole different story I get emotional very quickly. Some people cry when they watch things on TV or Film, sometimes I get a little bit emotional, but unless it is an amazingly brilliant piece of script writing or a painful in depth look at tragedy i find it hard to get emotional, over something that is distanced from me or is fictitious. Music however has a different effect as I relate songs to events or people and as such once memories are invoked they can cause me to get emotional. Some songs make me very happy, some songs make me very sad, some songs make me cry, some make me reflective. It is most odd how different things stir up our emotions, different people, different places, different events in your life. Whatever the emotion it is better to feel it and express it, than to keep it locked up inside, we are human, and the fact that we have emotions and react on them is at the very core of what makes us.

THOMAS – 2 – What has been the worst moment of your life?

The worst moment I can think of has just changed, not literally but I started typing this about one thing and then realised that 'Worst' I could think of something else. So here we go again. The worst thing that has happened to me was back in 2002. I had been going out with someone for about 6 months, and in the end things hadn't worked out and we had parted company. I had not heard anything from them, so assumed the dust had settled. One Sunday morning I was getting ready to go to work as I worked at a craft centre on weekends. I left the house just after 10 and on going down the drive to where my car was parked I saw a note under the windscreen. I lifted the note off and read it, all it said was 'Guess Who'. I thought nothing of it passed the note to mum at the front door and off I went to work in my shiny clean little Red, VW Lupo. I never thought no more about it just thought it may have been kids or something messing about. When I got to work and parked up as I went round to the passenger side of the car I noticed that all across the front wing of the car, offensive words had been scratched into the paintwork of the car and were standing out as clear as day. I was shocked I couldn't believe that I had driven 20 miles with all these words on display for all to see, probably laughing at me as I went on my way. I went into work and rang home to tell my parents what was going on. Whilst on the phone mum said she thought the hand writing of the note looked like that of my ex. I didn't believe her as I suppose I didn't want to believe that someone I had loved could be so spiteful. When I got home we all looked at the car, this was after my boss had come out of work at the end of the day and had seen what had been written much to my embarrassment. Once we had all looked at the car, it seemed quite likely that this was the work of my ex. We didn't know what to do next so we rang the police to report it as vandalism. It was so daunting as had never had any involvement with the police before. A policeman came round and took a statement and said he would follow up the information that we had given. Due to the fact that we had matched the handwriting I had to hand over any samples of the handwriting I had got this included personal letters and cards, basically anything that was personal to me, and private. I felt like it was an invasion. Sure enough within a few days the police came back and confirmed what we had already known ourselves, my ex admitted to causing the damage. What got me was the cheek of not only doing it but doing it whilst we were all in the house and the car was on the drive, albeit overnight.. I was asked if I wanted to press charges and of course I said 'Yes'. I had to go to the police station and make another statement. At this point I was advised that once it went to court a member of the local press could be present and my story could end up in the local paper. What made it worse is that I was told you could not stop this from happening. I went home and discussed this with my parents who were angered by the fact that we could only get justice served if we were happy for our story to go public, which would cause embarrassment to us all especially as we lived in a small village community. In the end we agreed that we would have to drop the charges and see if we could persuade my ex to cough up, they were also given a warning by the police. It was the fact that it was so personal that was hurtful. When I spoke to my ex after the event they said they knew how much loved my car and thought that attacking it would somehow hurt me, they hurt me with what they wrote but had neglected the fact that I don't care about material things at the end of the day, so I was upset but not over the actual physical damage to the car as that is something can easily be rectified with money, but upset by the

THOMAS – 3 – How did you find out that Santa doesn't exist?

Kids and Santa, hmm. It's all part of the Christmas magic isn't it when you are little there is this great man who flies through the sky on reindeer delivering presents and bringing joy to every single house around the world in one night. Now that last sentence as an adult is questionable but as a child seems perfectly believable. I think it is as I was growing up that this realisation became apparent, you realise that logistically it isn't possible for Santa to do what you believe him to do in just one night, especially due to the time it takes for him to get his large frame down the chimney, drop the presents, eat the mince pies, and drink the brandy at each house. Can you imagine how he would feel after just one housing estate let alone the world. Even now my mum has always been one of these people who doesn't go to bed until really light so I never thought anything of it when she stayed up late on Christmas Eve, she would be busy pottering around, watching TV, and wrapping presents, I would try to get to bed at a reasonable time, because I knew the sooner I went to bed and went to sleep the sooner it would be Christmas day. So to my bed I would go putting my stocking either on the end of a drawer unit or in later years on the bedroom door handle. I would be fast asleep and as if by magic in the morning my stocking would be piled high with little presents for me to sit in bed and open, followed by me running through to my parents to show them what Santa had delivered. Now my mum like most was very convincing at being surprised at what I was showing her. Couldn't fault the performance at all. Now as the years went on seemingly either Santa was a bit more noisy in his deliveries or I just didn't sleep so deep, and on more than one I noticed that as santa was leaving my room he had the exact same profile from the rear as my own good mother, how spooky is that! Now I think that by this time I was not really surprised as I had realised that Santa and my mum had close connections as a child I was told that even if I had seen my mum it was because Santa was so busy he needed a little bit of help here and there because he was such a busy man. Who was I to argue with the great man himself. So as the logic kicked in over time that ho hum just maybe he isn't real, it didn't seem to matter, it wasn't like a sudden shock more like a slow acceptance. Christmas was still magical I still went to bed and in the morning gifts adorned the tree and my stocking, so did it really matter just how they got there?? Now at the age of 25 I still think Christmas is magical even without Santa, and to ensure the magic continues, on Christmas Eve I go back to my parents, spend a lovely evening getting ready for Christmas, and go to bed and yes you guessed it put my stocking out as per usual, and still to this very day Santa/Mum has never let me down I wake up in the morning and sit on my bed going through all the little gifts. Now these gifts have obviously changed over the years as I have grown up, but all the magic is still there. Christmas is what you make it, each of us has it within us to make it just a little bit magical, and for children especially to their faces light up at the thought of Santa and Reindeer and presents, isn't it worth going along with the magic. Reality isn't always that exciting and I think we all need a little bit of magic in our life, we all know (well think we know) that Santa isn't real, but then we all go along with it because we all remember what it was like as a child to have those wide eyes and the excitement of waking up Christmas morning and seeing what we believed to be the proof that Santa really does exist. The other thing that kinda makes you stop believing is when you see him sequentially in two different department stores!!! He isn't god he can't be everywhere at once!

THOMAS – 4 – Describe a life-threatening situation you've had

The most life threatening situation in my life was back in Feb 2005. My mother woke to find that she was paralysed from the waist down, and after several trips to hospital was admitted and told that she would need to undergo a 5 hour operation to address the problem with her spinal column and there was no guarantee of success and the operation itself was of high risk and sadly not optional.

Now those are the facts here are the feelings. We are but a small family especially with me being an only child, and to suddenly be told that your previously active youthful mother may be paralysed from the waist down for the rest of her life with no explanation as to why, is something that when I look back now I struggle to understand how either myself or my dad coped. On hearing the news at work I fled down to the hospital where mum was only to find them sitting there with mum in a wheelchair still waiting to be seen. Once seen she was then discharged until the results came in a few days later, and then she was admitted to a specialist hospital about 1 and 1/2 hours from my work and about 40 minutes from where my dad worked. To this day I don't know how she coped with the shock, all of this did threaten my life not directly, but threatened my life as I knew it, the stability and assumed continuity as I knew it had gone. The day of the operation came round and this was the worst day ever, as there was nothing we could do but wait, I had seen mum the night before and the day of the operation dad was with her. Work that day was almost distanced from reality as I kept looking at my watch imagining what would be happening at that particular time, I was waiting for the operation to be over and my day to be over so I could go and see her and see how it had all gone. The anticipation was immense, I knew of course that my dad would ring me as soon as he knew something to put my mind at ease. So The day passed and as soon as the clock struck I was out the door and in the car. Dad had phoned to say mum was out and back on an HDU ward just until they could make sure she was ok. I went in and saw her, because of the nature of the operation she was lying in bed under strict instruction not to move and had to have everything given to her with a straw because she could not sit up.

The months that followed seem like a lifetime ago now. Life as we know it was threatened, and was very fragile for a while. Slowly but surely things are settling although it would be very wrong to say back to Normal. The operation went ok it was not as successful as they would have hoped, as a result mum has a walking aid, we now have a stair lift in the family home, and she is reliant on a wheelchair if she goes out for any long period of time. Our lives have all changed quite a lot since, it is only when your life or that of a loved one is threatened that you stop and realise just how much you take everything for granted. It is clichéd but it really makes you think about what is important. I can't begin to imagine what it must have been like for mum, she was incredibly brave and strong and her determination to restore some kind of quality of life, rather than give up shows with the recovery she has made. As with anything that threatens one's life, the question which in this case is unanswerable by any expert, is Why and more so Why Me? Depending on how your life is threatened there is often something that contributes to the event taking place whatever it be, but with many medical ailments, there is often no reasoning, if your life is threatened as your car spins off the road there is a reason that the car did such a thing. No expert can say what happened that night that made the next morning so life changing, makes you think doesn't it……..life is precious and by living it we threaten it every day.

THOMAS – 5 – How close have you come to your heart being broken?

This is a tough one as I am one of these people that seems to break other people’s hearts before causing my own to ache. It is funny isn't it how we have such an intelligent mind, yet almost to counter act it, we have a heart which makes as impulsive and defiant against our head. Someone said to me once that I was happiest when I had a certain level of control over a situation and as much as I hate it there is some truth in it. I am one of these people that can't just rest on something, if something is on my mind or there is a problem, I have to take the bull by the horns and deal with it, I can't just plod along when things aren't right. It is difficult to know what being heartbroken feels like, I think I have come close to it once. The first person I split up with after 18 months I felt like I had to do it as my life wasn't heading in the direction I wanted but I loved the person so much and we had shared so many happy times together. What made it worse was the weekend before we parted company we had been away together visiting relatives who were very welcoming of me, which just made things worse as of course in my heart I knew what was coming, it made me feel like a liar. As in one other of these questions I mentioned the fact that it just seems so awful to hurt people in your life. For days after I wondered if I had made the right decision, the worst part of it which did hurt me inside the most is that we were unable to salvage any kind of friendship from it, and I haven't spoken to them since. That made it worse, as it was like literally through my own doing I had not only lost someone close to me, but they had gone out of my life forever. I have spoken to people and can understand that it can be easier for people to deal with if they cut themselves off completely, and I am well aware of the fact that I only have myself to blame but that is what hurts the most. Even now I look back and of course every now and then doubt creeps into your mind, and next to that is sadness when you look at old photos and such like, and remember what you had. I can't begin to imaging what it must have felt life being on the receiving end, I am sure it will happen to me and can't imagine how I will feel. As a great believer in the fact 'What is meant to be will be' this somehow reassures me that things that happen to me happen for a reason although you don't always see it at the time, after the event has passed sometimes it becomes clearer why things happen. I am not sure what it is like on the receiving end but I presume you reach a point where you move on and realise that as obscure as things are that what will be will be.

I think heartbreak is a strong term as is love, and both are often misused, heartbreak can be caused by many different things, one partner leaving another, one partner cheating on another, loss of a loved one, (not necessarily a spousal relationship), a friendship that gets severed through a person’s fault or through physical distance. Love & Hurt without being cynical are as inevitable in life as ageing, everything that happens to us helps us to learn more about ourselves and others.

Right as i write the closing on this the final chapter in the short book of five, I realise that I have learnt a few things and done some deep searching into myself to write these papers. I have listened to my heart and written down the previously undocumented memoirs of some key parts of my life. These questions have a theme depending on which ones you pick, and they involve opening up your heart to reveal what makes you the kind of person you are. You start to question this as think about how you see yourself and how others would label you. This one in particular has made me think. Am I a heartbreaker?

APPENDIX D – Text 1 from John, Jenny, Greg, Judy and Alan

JOHN – 1 – When did you last cry and what made you cry?

The last time I cried was on a very old and decrepit Russian Passenger plane waiting to depart from José Marti airport, Havana. It was the end of April and I had just spent the last month travelling around Cuba with my girlfriend, Shelia. Shelia's flight departed an hour after mine and while I was flying to Mexico she was returning to the UK.

We weren't sure when I was going to be back. I was aiming for sometime in September but due to a highly ambitious travelling plan this was all dependant on how quickly I could cover the several thousand miles that separated Cancun (where my Russian plane was now heading) to Chile's capital, Santiago. My ticket was non-refundable and though I could change the dates the airports themselves were fixed.

Though it was going to be an amazing trip I was going to have to do it without my girlfriend. This wasn't as selfish as you might think, she insisted that I get solo travelling out of my system before we start living together (she had already covered central america by herself) and she was obliged to stay in her current job to raise the funds needed to pay for the MSc she was planning on studying the next year.

But I already knew how difficult it would be to cover all this distance without her. I had actually left the UK in the last few days of January following the end of a highly stressful and unfeasibly ambitious Research Contract at a British university. During the last few weeks I was in my lab until 3am or occasionally sleeping in my office. It was a very stressful period in both my life and my girlfriend. My departure was imminent but I could barely afford to spend any time with friends or family. I was also submitting the Project report for an MSc and felt selfish that the time I was spending compiling my work should have been spent with my Shelia. She was clearly spending far too much time already experiencing what life would be like once I had left the country. Though the last time I cried was in Havana the time before that was in Heathrow as I walked into departures. I was driven to leave the country and explore the world alone, and I felt terrible about it.

When I arrived at my destination, Belize, I met with my voluntary organisation and spent two months living in isolated tropical jungle. Working on a project to stop deforestation I spent the next 6 weeks without electricity, gas or running water. Our only communication with the outside world was via an occasional visit from field base by landrover, if we were lucky this would carry post. Here separated from busy world I was used to and dependant on international mail, I realised how much I could miss my girl.

Despite the amount of effort I had put into getting to Belize (through fundraising throughout the previous 12 months) I was beginning to anticipate the end of the project. Once I was out of Belize I would be heading to Cuba and to real travelling with someone I was used to spending most of my time with. Not only did I find it hard being away from Sheila but the other volunteers were turning out to be essentially not-my-sort-of-people. While I was expecting dedicated and enthusiastic eco-orientated persons such as myself I was instead coming across spoilt and whiney rich kids who had fundraised nothing and cared nothing for the environment.

Now I am back in England and we live together in our own flat. Shelia is doing the MSc as she wanted and I completed my tour across 11 Latin American countries at a breakneck rate. I came back a few weeks earlier than I originally predicted and missed country that I looking forward to the most. One day I will see it, probably with Shelia.

JENNY – 1 – When did you last cry and what made you cry?

The last time I cried was when my sister and I had an argument. We were in the car and started to argue. I really didn't want to argue in public so I got in the car quickly and waited for her. I was trying to avoid an argument as I hate arguing but I could feeling the tears coming and my body was tightening with anger. We went along the road but then we started shouting at each other so I pulled over as I knew I was concentrating on driving and thought I would start crying. I was really upset by what she had said even though some of it was probably true if I admit it. She said that I kept patronising her and telling her what to do. I knew that I had being telling her what to do recently but that was because she was leaving everything in a mess or forgetting important things. I was a bit upset that she hadn't said anything before. I got more upset when she said I can't cope with her telling me what to do! I think she was right though I am so used to being the big sister who likes to be in charge and she was trying to tell me I wasn't always right! I think the tears were out of frustration as well as I have been told that I am bossy and in the back of my mind I was thinking great I have been doing this lots recently and I haven't even noticed her feelings or what I was doing. But at the same time I felt she was making out it was all me being nasty and that I had evil intentions and was trying to control her. I felt like I had to apologise but didn't as I was still feeling that she was making herself out to be a victim. Then we ended up arguing about loads of other things like how long she was spending with her boyfriend and the amount of time they were spending on the phone . I said I felt like I never saw her because she was always on the phone. I realised that was probably the reason I was being bossy when I did see her. She didn't think her phone calls were excessive and said that I was just comparing her relationship to my relationship with my boyfriend and I couldn't accept that everyone was different in relationships. In the end she was crying a bit as well as we both admitted that we were spending more and more time apart and then arguing instead of saying what was bothering us. I suggested she tried to make less phone calls when we were together and that we did something with our time rather than argued. I knew I was trying to tell her in a roundabout way that i was also annoyed at how her boyfriend was always at our house or that she was away at the weekend at his house. I was still crying later on that day as I was still upset by the way she had said everything but I wasn't feeling great anyway so I just told everyone else it was because I was ill to avoid their questions. I knew that other people in my family might agree and I would probably feel better if I spoke to them about it next time rather than bottle it up and have a big argument with her. But I do see that I find it hard to see things from other people's point of view. Also she admitted that living at home after being at uni was more difficult than she thought it would be and as a result she was taking being told what to do by anyone difficult as she had enjoyed the freedom and independence of her time at university. I did agree about some of the points she mentioned but said I found it easier than her and was quite used to it after a year. Later on she did apologise and I did as well but it was probably as well that we had had the argument and I felt better that at least she knew what I was feeling rather than me consciously being bossy or getting in a mood with her. She was also able to get the anger about being at home off her chest.

GREG – 1 – When did you last cry and what made you cry?

I last cried whilst watching television alone in my flat, about 3 weeks ago.

I don't consider myself to be particularly prone to tears, and I don't recall crying about the deaths of friends or relatives a great deal. That's not to say that I didn't feel sad or even upset, because I definitely did, but the physical act of crying is not something that happens to me a great deal.

When I do cry, it's not always out of sadness. Any time I feel a strong emotion welling up inside of me, I sense that tears may not be far away, especially if this emotion takes me by surprise. Of course, more often than not this does happen in response to negative situations, and in the last few months I had experienced the (admittedly very slow and measured) break up of a long term relationship, and had probably cried more in 3 weeks than I had cried in the previous ten years. Nevertheless I was happy with the situation overall, and enjoying the feeling of freedom, but I was still partial to the occasional phone call with my ex-partner, and they nearly always ended up in the pair of us blubbing. I mention this because I'm not quite sure whether it had a bearing on my crying on what happened to be most recent occasion I cried.

On that night I had been out for a couple of pints and a read of the paper (again, alone - but I've always been happy in my own company, and certainly didn't feel lonely), then got some chips, and headed home. A friend had texted me earlier that night to say that he didn't fancy a pint with me because he was watching a Johnny Cash documentary on BBC4. I made it back with my chips in time to watch it too.

It was a pretty standard biographical programme, and really nothing I didn't already know about someone who's music, politics and general attitude to life I so admired. As the years covered by the programme passed by, Johnny Cash went from looking like a skinny bloke with sticking out ears to a much bigger, older looking man. The film ended by showing the video for 'Hurt', a song (and video) recorded by Johnny a few years prior to his death. The song had always been a favourite of mine, and I'd always found listening to it an emotional experience. I'd also seen the video once before, in a pub, and while I stood watching it captivated on that occasion, I wasn't moved to tears. Perhaps it was the noise in the pub that prevented me from getting concentrating hard enough on the video then, more likely I checked myself from getting to engrossed because I knew that tears wouldn't be far away, and crying wasn't something i wanted to be doing sat around a table in a pub on a Saturday afternoon with a load of my mates and a few old-timers telling stories. But on this occasion, alone in my flat, with a few pints inside me, I couldn't keep it in.

Tears streamed down my face as the video showed Johnny Cash looking old and dignified but in pain as he spent time with his wife, and the thought that she would die only added to the emotion of the situation. I can honestly say that I don't think there was any emotion from the recent break-up of my relationship spilling over into this occasion. If anything, when the credits were rolling and I'd made myself a cup of tea and switched over to something more light-hearted, I thought how happy and re-assured I felt that even in the absence of a girlfriend, or even any company, there are a world of emotional experiences to be had that can be experienced alone. In fact - that are best experienced alone.

JUDY – 1 – When did you last cry and what made you cry?

The last time I cried was last night. I had got up very very early in the morning and had a very long busy day at work and was looking forward to an early night. After Mitchell and Webb I excused myself from the sofa and went up to bed, proud of myself and happy for going to bed so early, and then I think I'd only just fallen asleep when James came running up the stairs shouting "HELLO!! HELLO!! WAKE UP!" and bounding into the bedroom turning on lights, seemingly astonished that I was asleep, and then I grumbled at him and told him to be quiet and he started whimpering and went downstairs. I was bristling up waiting for him to come up to bed and he didn't but I could hear him turning off the lights and making up a bed on the sofa. For goodness sake! James gets up about an hour and a half later than me most mornings and then won't let me go to bed even half an hour before him at night…grr.

After about ten minutes on edge trying to decide what to do, and being thoroughly frustrated about all this time that I was spending not asleep I got up, stomped downstairs and tried to find out what was going on, why he was sleeping on the sofa, and in my frustration I started crying and was trying to be sensible and nice so he'd come to bed and we could both FINALLY go to sleep, but was in reality just spluttering and crying. James bizarrely started pretending to be a small child, probably to soften me up and stop me screaming at him, but it was just thoroughly annoying because I just wanted him to stop playing and stop pretending to be a small scared child about to ring Childline, and just come to bed and then we could both just go to sleep, so I was getting more and more frustrated, and was just stood naked in the living room wailing and stomping my feet because I wanted to go to sleep. I eventually pulled the blanket off James, started trying to drag him off the sofa - I think at this point he was actually properly frightened of me - and he conceded in the end and, starting to get a big grumpy now, he lost his sense of humour and told me to pull myself together and stomped upstairs angrily.

All I wanted to do was go to sleep but now seemingly I'd got myself into an argument by being too frustrated with James’s playfulness (although I don't really think it's too unreasonable to be angered by being woken up after a horrible tiring day by a giggling laughing 24 year old man who'd had LOTS of sleep recently and was pretending to be a small boy), so I sighed and followed him back upstairs preparing to either have a nasty argument in which he accused me of being old and boring, or to just have a long difficult silence in which both of us tried to go sleep out of spite, to show that we didn't really care we'd have an argument. I went into the bedroom sighing loudly to show that I still considered him to be the guilty party here, that I REALLY wasn't going to cave in and apologise. I got into bed next to James and he turned away pointedly, so after five tense minutes where both of us were unsuccessfully pretending to be asleep, I gave in and apologised, because maybe I shouldn't have been so snappy, he was only being young and playful (although I WAS asleep). He accepted my apology but was still being a bit distant with me, which made my eyes start to spill over again, because I had apologised and it wasn't even my fault! Ironically James went to sleep almost immediately and I lay awake feeling awful for ages even though I was surely the most tired one - I felt tempted to shake him awake shouting "HELLO! WAKE UP!" but maybe I'm a bit kinder than he is. I got to sleep in the end and both of us apologised properly this morning.

ALAN – 1 – When did you last cry and what made you cry?

Well I find this quite coincidental. The last time I cried was about three weeks ago. I honestly cannot remember the time before then that I actually cried; we're talking years. It takes a helluva lot to make me cry. I really wish I was able to do it more freely. I get the feeling that I want to cry but no tears come out. This is why I knew when I was crying a few weeks ago that my feelings for the person who had reduced me this would never be same again. I met Chris at a New Year's Eve Party last year in London, a friend of my then boyfriend. He was very flirtatious, had awful hair, not my taste at all. But he was a nice enough guy. We would chat about anything and everything once or twice a week. After a few weeks of dating Nick (the then boyfriend) we decided it wasn't really working but we'd be better off as friends instead. Then Chris and I at his house soon after, kissed, hugged..hugged with less clothes on..you get the gist. A week later after consulting Nick and getting the all clear we started dating. We dated for 8 months. We had the typical highs and lows not dissimilar to every other relationship in the world really. By the time we broke up, neither of us were happy. We were on the phone for hours every night trying to figure out a way to save the relationship. The general problem with this. We were best friends, we knew everything about each other and more. There was nothing neither of us couldn't say to one another. A few months ago he was suffering with chronic diarrhoea and I drove all the way down to Oxford (where he lived at the time) to look after him. At another time I had very bad tonsillitis and he sat up with me all night, while I was screaming in pain into my pillow. There was nothing we wouldn't done for each other. The snag was this, we didn't love each other. It just never happened. We had both acknowledged this and hoped that one day love would spring. It never did, and it had now got to the point that continuing a loveless relationship was too much for us to handle. We arranged that he was going to come up to Stafford for the weekend and we would talk about things, and if needed, give each other back belongings which we had acquired by and by. He was due to come up Saturday evening, after he had been to a friend's house. He called me Friday evening, while waiting for the train to go and see his friend. He told me he didn't see any point in coming up as in his eyes, there was no hope. He broke up with me from platform 1 of Stafford Station. The last thing he said to me before he hung up was "Oh my God, the train's really busy, I'll have to stand. I should go..I'm sorry. Bye" He slept with his friend that night. I knew he would and a few days later if he had. He dodged the question. I asked him again. And again. Finally he answers "well if I did, it was his fault." He was joking about it. I lost respect for him then. I texted him a while later saying if he had just apologised it would have been okay and I would have gotten over it. It wasn't the fact he had slept with this guy that upset me. It was the fact that he didn't have the decency, respect, courtesy or balls to tell me. This was then followed by various other examples of him saying the wrong thing until by the end of the night he had made me so angry I was white and shaking (how dramatic). I very very very very rarely get angry. I don't know how to. So when I do, I get upset as I don't know how to handle it. I stayed up all night with some friends who were wonderful. At 5.30am I went for a walk. I arrived back at my campus just after 7. I was in tears. I was so upset that I had let him get me that angry. No one has the right to do that to anyone. I cried probably around 5 or 6 tears before I dried up. That's a lot for me. We've spoken now, and all is forgiven. He apologised for a lot of things he did wrong. It's just such a shame that now I really don't see how we can ever be friends. Forgiving I can do. Forgetting is always much harder and if someone has done something to me that actually made me cry then I don't see how I'll ever be able to forget that.

APPENDIX E – Grid for all *way*-phrases and alternative expressions

	John				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions				*system*	
=do more than necessary/expected					
=embarked on a route, journey				*on their route / the only way to / worked my way*	
=embarked on a route, journey (metaphorical)					
=emphasis	*far too*				
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion				*nature / style*	
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective			*a chance*		

=method, no options/possibilities				*chance out*	*impossible*
=mid-point					
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)	*out of my system*				
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time	*all this distance / throughout*				
=to some extent, in some respects					*in a way*
=vice versa					

	Rose				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected	*make sure*				
=embarked on a route, journey	*made our way x 2*				
=embarked on a route, journey (metaphorical)					
=emphasis		*much*		*much*	*much*
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion		*in a strange way*			*in such a kind way*
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective					
=method, no options/possibilities					
=mid-point					
=move to safety, away from path of danger					

=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)		*coming*		*in a certain direction / a certain way*	*coming*
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects	*in a way*	*in a way x 2*	*kind of / in a way x 2*	*in a way x 2*	*kind of / in that respect / in the other sense / in a way x 3*
=vice versa					

	June				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey	*on the way*				
=embarked on a route, journey (metaphorical)					
=emphasis					
=from available options					
=great distance, far					*a long way*
=helped through alternative means					
=in a certain manner, fashion				*in such a state*	
=in a certain manner, how		*the way x 2*			
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective	*how*			*how*	*the only way x 2*
=method, no options/possibilities					
=mid-point				*middle of / of the middle*	
=move to safety, away from path of danger					

=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time		*all the way*	*througho ut*		
=to some extent, in some respects					
=vice versa					

	Keith				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					*any other way*
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)					
=emphasis			*way*	*much*	
=from available options					
=great distance, far					*so far*
=helped through alternative means					
=in a certain manner, fashion		*in some way*			
=in a certain manner, how	*the way*				*the way*
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective			*how*		*how*
=method, no options/possibilities					
=mid-point			*in the middle of*		
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					

=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects	*in a way*				
=vice versa			*vice versa*		

	Jenny				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions				*ways*	
=do more than necessary/expected					
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)					
=emphasis		*much*			
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion	*by the way / in a roundabout way*	*in such a way*			*like that*
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways			*in many other ways*		
=method, how to achieve an objective					
=method, no options/possibilities					
=mid-point				*in the middle of*	

=move to safety, away from path of danger		*away*			
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons			*some ways*		*in many ways*
=particular direction, towards an outcome (metaphorical)	*coming*	*a way*			
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects		*in a way x 2*			
=vice versa					

	Sue				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)					
=emphasis		*much more*		*much / rather / far more*	
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion	*in any sordid way / by no means*	*quite the same / way*	*the convention*	*way*	
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective		*how*		*how*	
=method, no options/possibilities					
=mid-point					

=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					*throughout*
=to some extent, in some respects	*somewhat*			*sort of x 2*	*somewhat*
=vice versa					

	Michael				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected		*make best efforts / go out of my way to*			
=embarked on a route, journey	*my journey*				
=embarked on a route, journey (metaphorical)					
=emphasis					
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					*like that*
=in a certain manner, how		*how*			
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective			*only one way*		
=method, no options/possibilities					

=mid-point					
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects					
=vice versa					

	Carla				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario				*the ways*	
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey				*my way*	
=embarked on a route, journey (metaphorical)					
=emphasis	*significantly*	*far too*		*much / far too*	
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion				*sense of style*	*style*
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right		*both ways*		*in the other direction*	
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective		*how*			*how / the only way*
=method, no options/possibilities					
=mid-point					
=move to safety, away from path of danger					

=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons	*in so many ways*				
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects	*somewhat / in a way*				
=vice versa					

	Nicola				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					*gave way*
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)					*approac hing*
=emphasis				*so much*	
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how				*the way*	
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective	*only one way*				
=method, no options/possibilities					
=mid-point					
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					

=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time		*through out*			
=to some extent, in some respects					
=vice versa					

	Hannah				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey				*distance*	
=embarked on a route, journey (metaphorical)					
=emphasis	*much*				
=from available options					
=great distance, far					
=helped through alternative means			*in a different way*		
=in a certain manner, fashion	*in any serious way*				
=in a certain manner, how	*how*		*how*		
=in any condition, state		*any different*			
=in each direction, left and right				*in the other direction x 2*	
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective			*my way*		
=method, no options/possibilities			*little to no*		

			chance		
=mid-point				*in the middle of*	
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons		*in so many ways*	*in several ways*		
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time	*all the way*		*throughout x 2*		*throughout x 2*
=to some extent, in some respects		*in a way*	*somewhat*		
=vice versa					

	Melanie				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected		*tried to make sure*			
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)					
=emphasis				*much*	
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion	*like that*			*in a way*	
=in a certain manner, how					
=in any condition, state	*any way*				
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective		*how*			*option*
=method, no options/possibilities					
=mid-point					

=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects		*kind of*			
=vice versa					

	Sarah				
Meaning (29)	1	2	3	4	5
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey	*made my way*	*made my way*			
=embarked on a route, journey (metaphorical)					
=emphasis					
=from available options			*one way or the other*		
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective				*way of Xing x 2*	*how*
=method, no options/possibilities					
=mid-point					

=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time	*throughout*				
=to some extent, in some respects					
=vice versa					

	Rick				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey					*on my way*
=embarked on a route, journey (metaphorical)					
=emphasis	*far too*				
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					*the old routine*
=in a certain manner, how			*how x 2* */* *the way*		
=in any condition, state					
=in each direction, left and right				*both ways*	
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective	*how*				
=method, no options/possibilities					*there is no way*

=mid-point				*half way*	
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					*the rest of the way*
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects					
=vice versa					

	Greg				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)					*make their way*
=emphasis	*much*				
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger				*in the way*	
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective		*a chance*			*my best course of action*
=method, no options/possibilities					*way*
=mid-point					

=move to safety, away from path of danger				*get out of the way*	*out of the way*
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects					
=vice versa					

	Judy				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed	*cave in*				
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)					
=emphasis					
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective					
=method, no options/possibilities					
=mid-point				*middle of the*	
=move to safety, away from path of danger					

=move to safety, away from path of danger (metaphorical)			*out of the way*		
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects			*somewhat*		
=vice versa					

	Elaine				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey					*coming*
=embarked on a route, journey (metaphorical)					
=emphasis			*much*		
=from available options			*either way*		
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion				*in a strange way*	
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective					
=method, no options/possibilities					
=mid-point					

=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)			*the exact way*		
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time				*all the way*	
=to some extent, in some respects					
=vice versa					

	Thomas				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey	*on the journey / get myself back*	*on my way*			
=embarked on a route, journey (metaphorical)					
=emphasis					
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how			*how*	*how x 2*	*how x 4*
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective	*how / way of Xing*			*how x 2*	
=method, no options/possibilities	*a way*		*it isn't*		

			possible		
=mid-point					
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					*in the direction / coming*
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects					
=vice versa					

	Mark				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey		*on my way*			
=embarked on a route, journey (metaphorical)					
=emphasis			*way*		
=from available options					
=great distance, far			*far*		
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how					
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					
=manner, in different ways					
=method, how to achieve an objective					
=method, no options/possibilities					
=mid-point					
=move to safety, away from path of danger					

=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects					
=vice versa			*the other way around*		

	David				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions					
=do more than necessary/expected					
=embarked on a route, journey			*making his way*		
=embarked on a route, journey (metaphorical)					
=emphasis	*much x 2*				
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how		*how / manner*			
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion					*similar*
=manner, in different ways					
=method, how to achieve an objective					
=method, no options/possibilities					
=mid-point					
=move to safety, away from path of danger					

=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)		*coming*		*coming*	
=remainder of the journey					
=tactfully express					
=the entire distance, journey, time					
=to some extent, in some respects					*somewhat x 2*
=vice versa					

	Alan				
Meaning (29)	**1**	**2**	**3**	**4**	**5**
=a different situation, alternative scenario					
=broke, collapsed					
=devising plans, solutions			*for ways to*		
=do more than necessary/expected					
=embarked on a route, journey					
=embarked on a route, journey (metaphorical)			*along the way*		
=emphasis	*much*	*far*	*way*	*significantly*	*far too*
=from available options					
=great distance, far					
=helped through alternative means					
=in a certain manner, fashion					
=in a certain manner, how	*how*				
=in any condition, state					
=in each direction, left and right					
=in the direct path of danger					
=like, in a similar fashion			*the same way x 2*		
=manner, in different ways					
=method, how to achieve an objective	*how x 2 / a way*				*how*
=method, no options/possibilities					*there is no way*

=mid-point					
=move to safety, away from path of danger					
=move to safety, away from path of danger (metaphorical)					
=on several levels, for different reasons					
=particular direction, towards an outcome (metaphorical)					
=remainder of the journey					
=tactfully express				*let's put it that way*	
=the entire distance, journey, time	*all the way*				*the whole way*
=to some extent, in some respects		*sort of*	*in a way*		*sort of x 2 /* *in a way*
=vice versa					