Article

A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship

Larner, Samuel

Larner, Samuel (2014) A preliminary investigation into the use of fixed formulaic sequences as a marker of authorship. International Journal of Speech Language and the Law, 21 (1). ISSN 1748-8885

This is a post-refereed version of an article which has been published by Equinox Publishing:

Larner, S. (2014) A Preliminary Investigation into the Use of Fixed Formulaic Sequences as a Marker of Authorship, *International Journal of Speech, Language and the Law*, 21(1), 1—22.

A Preliminary Investigation into the Use of Fixed Formulaic Sequences as a Marker of Authorship[1]

Introduction

Whilst there now exist many accepted markers of authorship in the area of forensic authorship attribution (e.g. McMenamin, 2002), these markers often rely on authors doing something marked with language (e.g. punctuation, (mis)spelling, syntax etc.), with linguists relying on their expertise and judgement to determine which features are salient. However, sometimes a text can be somewhat unremarkable. An author who adheres to the standard form of a language and uses grammar and spelling auto-correct features in a word processing package may yield less distinctive features for analysis. The holy grail of authorship markers would be one which occurs simply by virtue of producing language and which occurs frequently in all texts. The aim of this paper is to explore the potential development of one such marker based on formulaic sequences—prefabricated sequences of words believed to be stored as holistic units.

The Nature of Formulaic Language

Evidence from psycholinguistics (Wray, 2002), sociolinguistics (Coulmas, 1979), corpus linguistics (Moon, 1998) and language acquisition (Pawley & Syder, 1983; Peters, 1983; Vihman, 1982) shows that when communicating, we rely on patterns in language and have 'preferred formulations' for expressing ideas (Wray, 2006:591). Theoretical viewpoints such as Construction Grammar also make this point (Goldberg, 2003). Such routine language can be termed formulaic which Wray (2002) defines as '[w]ords and word strings which appear to be processed without recourse to their lowest level of composition' (p.4). *Formulaic language* is an umbrella term and a survey of the literature soon reveals that many labels exist to describe different characteristics of formulaicity. These include Collocations, Idioms, Multi-word Items, and Recurrent Phrases to name just a few. In fact, Wray (2002:9) found 57 different terms for describing sequences of words which can be characterised as formulaic. The definition adopted in this research is that of the 'formulaic sequence'—a term 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:

> [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. (Wray, 2002:9)

In addition to this psycholinguistic definition, Wray (2002) also adds a sociolinguistic dimension. 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). 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' (p.267). As such, there is potential for a relatively disparate group of authors to have different inventories of formulaic sequences resulting from having differing needs and interactions in various speech communities. Providing that there is an effective way to identify them, formulaic sequences have potential to mark out an individual author.

Estimates vary regarding how much of everyday language use is formulaic. Erman and Warren (2000) claim that 55% of spoken and written language may be formulaic whilst Chenoweth

(1995) found 77% of written answers to essay style exam questions contained formulaic expressions (p.292). 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 context is wholly or largely fixed' (p.191). A lack of consensus over the occurrence of formulaic 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 in language (Wray, 2002) and so should occur frequently in texts.

Formulaic Sequences as a Marker of Authorship

Although not for the forensic context, some researchers have empirically investigated the idiolectal potential of formulaic sequences. Kuiper (2009) offers insight into individual variation in the use of formulaic sequences during routine interactions, specifically, ritual talk at the supermarket checkout. Kuiper was interested 'in the formulaic inventory and discourse structure used by checkout operators' (p.99—100) but not with sociolinguistic variation. Therefore, he did not control for social variables such as age, socioeconomic status or gender of the customers, nor were an equal number of interactions recorded for each checkout operator. The data were collected in 1991 from two supermarkets in different suburbs of Christchurch, New Zealand. Over a period of one month, 200 interactions were recorded from nine checkout operators. He found that the entire 'checkout' interaction consisted of stages and each stage had a set of formulae (in his terms). At the 'opening of monetary exchange with customer' stage, operators used formulae such as *will that be all?, there we are,* and *there we go.* At the 'farewell to customers' stage, operators used formulae such as *have a nice day, enjoy the rest of your day,* and *have a good weekend now won't you,* indicating marked points in the discourse where formulae were used and restrictions on where in a discourse sequence particular formulae were permissible (p.109). Given the highly ritualised nature of this interaction, Kuiper argued 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. (p.109)

By analysing the formulae used during the greetings phase, Kuiper found that some formulae were used frequently by the majority of operators (e.g. *How are you?; How are you today?*); some greetings formulae were used rarely and only by one operator (e.g. *Alright?*; *How are you going?*); some greetings formulae were the preferred formulae for operators (e.g. four operators used *Hi, how are you today?* more frequently than any other formula in the data); and some greetings formulae were used more regularly by only one operator (e.g. only one operator used *Good morning. How are you?* and *Morning*). Kuiper argued that some operators used formulae much more flexibly than others and that all operators used particular formulae, equivalent to a 'signature' (p.113), leading to the conclusion 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 suggest that it should be possible to identify an individual operator based on the formulae they use. However, the formulae described by Kuiper are limited to only one context—the supermarket checkout— and his discussion of individual preferences is confined to only the 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. 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. In addition, 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.

Although not the primary focus of their investigation, Schmitt, Grandage and Adolphs (2004) provide a small amount of evidence that formulaic sequences may be linked to idiolect. The aim of their study was to determine whether recurrent clusters were psycholinguistically valid—that is, stored holistically (p.128). They drew a distinction between word strings identified through corpus analysis which may or may not be stored holistically (e.g. *in a variety of*), termed *recurrent clusters*, and word strings that are stored holistically, which, following Wray (2002), they refer to as *formulaic sequences.*

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 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.

Of particular interest 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. Some speakers appeared to store 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. (p.138)

Like Wray (2002), they argue that individual formulaic inventories include 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' (p.138) which may also be linked to topic and discourse situation. They conclude 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' (p.138). This is clearly useful for the claim that formulaic sequences hold potential as a marker of authorship, particularly when we take into account that other aspects of linguistic history such as exposure, experience, and participation in various communities of practice are also likely to influence individual formulaic inventories (Wray, 2002). 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) but it is interesting that in a more general context, idiolectal differences were still found lending further support to Kuiper's (2009) context-specific research.

More specific evidence that formulaic sequences may be a marker of authorship comes from Waltman (1973) who 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). The poem is typically split into two parts so that they can be compared with each other. Literary differences 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, and a higher frequency of assonance in the second part (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). Despite this focus on performance context, Waltman investigated the use of formulaic expressions in the *Poema de Mío Cid* by claiming that patterns of formulaic expressions would reveal something about the poem's authorship: 'A variance in formulaic expressions in the poem would tend to point toward two composers' (p.571).

Defining *formulaic expressions* as 'a group of words with similarity of vocabulary under somewhat the same metrical conditions', Waltman used concordance software and found 26 formulaic expressions (p.571—2). Such phrases included: *El de Biuar*, *Moros & christianos*, and *Vala el Criado* and 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 and selected sections which dealt with the same topic: a parting, farewell scene. He found that both segments contained the same formulaic expressions. Furthermore, 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).

In assessing this evidence, it should be borne in mind that Waltman was careful to select segments which were comparable in topic. Therefore, it is possible that recurrence throughout the *Poema de Mío Cid* was directly linked to this rather than authorial style. 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, Waltman's focus on 'the same metrical conditions' would appear to be redundant outside the literary context. Furthermore, the focus on metrical conditions relates to the oral

nature of the poem which may have led to a greater incidence of formulaic sequences since these serve to aid memory during public performances (cf. Rubin, 1998), rather than being a feature of authorship.

This review of the limited empirical research into formulaic sequences as a marker of authorship has lent support to the assumption that formulaic language holds potential as a marker of authorship. However, 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 limited. More detailed empirical investigation into the relationship between formulaic sequences and authorship is therefore required 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.

Identifying formulaic sequences for forensic purposes

There are two problems in the identification of formulaic sequences. The first is that the task is not an easy one; so difficult in fact that Wray (2008) comments it 'can be rather like trying to find black cats in a dark room: you know they're there but you just can't pick them out from everything else' (p.101). Erman and Warren (2000) highlight the problem that formulaic sequences ('prefabs' in their terms) can be easily overlooked in a text and caution that 'the identification of "all and only" the prefabs in a text is in practice impossible' (p.33). The second is that if formulaic sequences do hold potential as a marker of authorship, any method developed needs to satisfy the three tenets of forensic authorship analysis: reliability (analyses which can be repeated to produce the same results), validity (whether what has been identified is actually what was intended), and feasibility (how well the method can be applied to forensic data).

Formulaic sequences can be identified in several ways, depending on whether the language is spoken or written, but two common approaches include using intuition and using automated approaches (see Wray, 2002 for a comprehensive critical review of additional techniques). The intuition approach is based on the belief that native speakers recognize formulaic sequences as having special status (Van Lancker-Sidtis & Rallon, 2004:208). This is clearly open to subjectivity and the 'status of the intuition of an individual investigator is dubious from a modern "scientific" perspective' (Read & Nation, 2004:29) due to the lack of reliability—what one researcher may judge to be formulaic may not be so for another, leading to variation between judges. To carry more reliability, at least a second rater should be used (Read & Nation, 2004:29), and better still, panels of independent judges can be used to reach consensus about whether a string of words is indeed formulaic (e.g. Foster, 2001). However, whilst using a panel of judges may increase reliability, the majority of forensic linguists work in isolation due to confidentiality. Furthermore, intuitive analysis is often restricted to smaller datasets given that each text has to be read carefully and more than once which can make it a slow and laborious process. Given the time pressures often involved in producing authorship evidence (Shuy, 2006), using intuition as a technique for identifying formulaic sequences in forensic texts may be problematic.

Reference lists such as dictionaries and textbooks 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. Reliability is higher than can be achieved using one person's intuition and marking texts manually, since once the

criteria have been specified, the software will extract all, and only, the instances that fall within the search parameters. Automating the approach guards against human error, and this technique is also feasible for use with larger sets of data, such as when a relatively small set of individual forensic texts is compared against a larger reference corpus (e.g. Wright, 2013), or when a larger corpus of many smaller forensic texts is compiled (e.g. Grant, 2010) since analysis can be fast and reliable. However, Wray (2008) cautions that if using a reference list, it is important to consider why that list was produced and what decisions were made about what to include/exclude and 'whether the list has gained authority simply by virtue of being published' (p.109). Furthermore, identifying formulaic sequences in this way will come at a cost—this method is limited to identifying only fixed formulaic sequences (rather than semi-fixed or formulaic frames). For the present research, a compromise is proposed: using the internet to build a reference list. Drawing on a multitude of different sources compiled by numerous members of various speech communities—and therefore lots of individual intuitions— should ensure that the list is as representative of formulaic sequences as possible. This is especially valuable given that formulaic sequences are difficult to identify quantitatively or objectively. Furthermore, using a reference list compiled of multiple intuitions enables an automated approach which satisfies the requirements of a forensic investigation.

The Present Experiment

The aim of this research is to begin to answer the question: do formulaic sequences hold potential as a marker of authorship, as the surveyed literature suggests? To most effectively do this, variation must firstly be established: inter-author variation needs to be high whilst intra-author variation needs to be low. In other words, the use of formulaic sequences needs to remain similar for one author whilst being used differently by another. Secondly, if variation of this kind can be demonstrated, the next stage will be to determine whether a Questioned Document (a document of unknown authorship, henceforth QD) can be correctly attributed to a corpus of documents with known authorship.

Data

The data comprise 100 texts written by 20 authors, with each author producing five texts. Authors were sent a daily structured writing task consisting of two questions over a five day period. Authors were required to answer whichever question they preferred. Open-ended questions which elicited personal narratives were used. By asking emotionally-charged questions, it is hoped that the likelihood of participants focussing on their language use was reduced (Labov, 1972; Labov & Waletsky, 1997). The 100 texts totalled 65,113 words, with each author producing an average of 3,325 words across their five texts. The average text length was 651 words with the shortest being 485 words and the longest being 822 words.

A snowball sampling technique was used to identify participants (where potential participants invite others to take part, who in turn invite their contacts and so on), so it was not possible to control for social variables. It is therefore beyond the scope of this study to consider the effect of sociolinguistic variables on formulaic sequence usage. Nonetheless, the sample contained nine males and 11 females, with an age range of 18—48 (mean age=24). In terms of education, six participants reported Further Education (e.g. A-Levels) as their highest qualification, ten participants were undergraduate-level educated, three were postgraduate level, and one had a doctorate. Sociolinguistic information specific to each author is provided as an appendix.

Method

Terms commonly accepted as names for various types 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*. For each search string, all of the links from the first five pages of results were explored. There did not appear to be any benefit in 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 onto a list regardless of whether or not they were intuitively pleasing as examples of formulaic sequences. This process was repeated until no new websites were identified. The list contained 17,973 entries.

It is difficult to account for the contents of the list in terms of how each formulaic sequence can be classified (e.g. idiom, collocation, metaphor etc.) since formulaic sequences can often be classified into several categories (Moon, 1998). Based on how the websites self-identified themselves, the list appears to be composed of the following proportions:

Table 1: Categories of formulaic sequences included in the reference list

Type of formulaic language	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 %
Totals	17,973	100%

Clichés and idioms account for just under 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 everyday saying is in no way clear cut.

Whilst many of the webpages visited in compiling the list of formulaic sequences ended with the domain .co.uk, it is impossible to definitively account for the origin of the entries and it is notable that some contained North American spelling variants. 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. This may be problematic since using a reference list of formulaic sequences for one speech community against data produced by authors from another may be incongruous. However, since it was argued above that people's store of formulaic sequences will be shaped by the speech communities in which they interact, it may be precisely in the identification of these types of sequences that an author may be identified. If authors do not use them, then they cause no harm since they simply will not be identified in the data. Finally, there were many duplicates in the list which were removed resulting in a reference list containing 13,412 entries.

Bespoke software (Menacere et al., 2008) was used in this research. The software was firstly used to compile the entries of formulaic sequences into a machine-readable, editable reference list, and secondly to identify matches between the reference list and the narratives data. A specific feature was built into the software which enabled it to cross-reference to a separate list of 86 pronouns 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*) whenever it encountered an asterisk in the formulaic sequences reference list. Therefore, to prepare the reference list for comparison with the database of narratives, pronouns in the list were replaced with an asterisk. By changing the entry *his bark is bigger than his bite* to * *bark is bigger than* * *bite* the examples 1—3 could theoretically be matched:

(1) *her* bark is bigger than *her* bite
(2) *my* bark is bigger than *my* bite
(3) *your* bark is bigger than *your* bite.

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 its bite*. However, since it is unlikely that an author would produce these strings under normal circumstances, the advantages of allowing substitution outweigh the disadvantages of having only fully fixed forms in the list.

Individual researchers may find some entries more problematic and less prototypical than others (e.g. *Jiminy Christmas* would typically be considered to be more formulaic than *date rape)*. However, the aim of the list is not to reach universal agreement about what actually constitutes a formulaic sequence; rather, the aim is to collate as many *potentially* formulaic sequences as possible. There are limitations to the list, both in terms of what it contains and how well it can match examples of formulaic sequences in real text since the list cannot claim to identify every single instance of formulaic sequences in text, nor will it identify variants of items contained in the list (with the exception of pronoun substitution). Furthermore, the list does not account for the possibility that items are not fully representative of what authors use in their writing. It cannot even guarantee that every instance it identifies will be formulaic. In this regard, the results that follow are limited to testing based on those items included in the list. 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 initial exploration into the relationship between formulaic sequences and authorship.

Results

A total of 604 formulaic sequence tokens were identified in the data, of which there were 301 types. Table 2 shows the ten most frequently occurring formulaic sequences whilst a selection of ten formulaic sequences that were used only once across the corpus include: *under the influence, under the weather, vice versa, what on earth, what will be will be, wide awake, with flying colours, with the exception of, worst nightmare,* and *x factor.* It is interesting to notice from Table 2 that the most commonly occurring formulaic sequences contain just two or three words (with the exception of *at the same time*, containing four). Some five word formulaic sequences did occur, but only once (including *in my heart of hearts, in the middle of nowhere, take the bull by the horns* and *everything but the kitchen sink*). Additionally, the longest formulaic sequence *at the end of the day* occurred just three times across the corpus.

Table 2: Most frequently occurring formulaic sequences across the data

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

Although formulaic sequences have been argued to be pervasive, the fact that shorter texts are under investigation may be problematic since minimum thresholds must be met before statistical tests carry validity. For establishing the authorship of very short mobile telephone text messages, Grant (2010) drew upon Jaccard's coefficient (typically used in psychology to establish case-linkage). Jaccard's coefficient establishes the correlation between whether a series of particular features are present in a sample, rather than the frequency with which they occur. A particular advantage to using Jaccard's coefficient is that the absence of a feature does not increase or decrease the similarity measure between two texts or crimes (Grant, 2010) so 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. 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. In the current investigation, every formulaic sequence highlighted in the corpus constituted a feature, resulting in 301 features. All 100 texts were tested against each other text, resulting in 4,950 pairs of texts.

The Jaccard's coefficient for each of the two groups of linked and unlinked pairs was tested using the Kolmogorov-Smirnov Goodness-of-fit test to see if the coefficients were normally distributed. The distribution was significantly different from normal for both the linked ($D=0.452$, $N=200$, $p=<0.001$) and the unlinked pairs ($D=0.494$, $N=4750$, $p=<0.001$). 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 ($Z=2.67$, $N=4750$, $p=0.008$) where unlinked pairs were lower. This means that texts produced by the same author are more similar in their use of formulaic sequence types than texts by different authors.

Having established that inter-author variation is greater than intra-author variation, it is next necessary to determine whether a QD can be successfully attributed to its author. However, the point of Grant's (2010) approach using Jaccard's coefficient is that it is not an authorship attribution technique in itself. Rather, it is a statistical method for describing consistency and distinctiveness. Having established that the use of specific formulaic sequences is more consistent between texts produced by the same author than by different authors, it is necessary to use qualitative analysis in order to attribute a QD to its author.

For this purpose, the 26 most frequently used formulaic sequences in the corpus were selected, which in combination were used 228 times across the corpus. These 26 were selected since they occurred a minimum of five times in the corpus. Theoretically, any one of these 26

formulaic sequences could be used across all five texts by a single author which would demonstrate consistency, but a formulaic sequence that occurs only four times in the corpus could not. However, no author did in fact use a single formulaic sequence across all five of their texts. This is problematic because out of five texts, any one of those could be a QD and so could not be attributed on this feature alone. The formulaic sequence used most consistently by an author was *in the end*—the most frequently used formulaic sequence in the corpus—used by Jenny, a 23 year old female educated to undergraduate level, in four of her five texts, but, crucially, only once on each occasion. Even if *in the end* did occur in her fifth text, an occurrence of only once in each text would be too low to be persuasive as evidence of authorship, particularly since this formulaic sequence was also used by eight other authors at least once with 20 total occurrences across the corpus, so it cannot even be claimed to be distinctive. These formulaic sequences clearly do not occur frequently enough, or with enough consistency to enable an attribution. Furthermore, none of these formulaic sequences occur in clusters together—idiolectal co-selection (Coulthard, 2004)—so neither the type of formulaic sequence, nor the clustering of specific formulaic sequences follow any patterns in these data that would be useful in attributing a QD.

The normalised count of formulaic sequences

Whilst the specific formulaic sequences used by authors do not appear to hold potential as a marker of authorship, some authors did use more formulaic sequences than others and so overall usage of formulaic sequences may still hold potential as a marker of authorship. In each of the texts there was a small amount of variation in length. To cope with these differences, the measure used was the normalised count of formulaic words (rather than sequences) per 100 words (henceforth 'count' for brevity). By calculating the count, it will be possible to make claims about whether the language used by one author is more or less formulaic than that of another. It may be problematic to use the count of formulaic words since some formulaic sequences may be longer than others and an author who uses fewer formulaic sequences consisting of more words may not be directly comparable to an author who uses many formulaic sequences comprising fewer words. Taking the measure of 'hits' (i.e. the total number of formulaic sequences) is justifiable from a theoretical perspective: if formulaic sequences are stored holistically, all words that comprise a formulaic sequence should be treated as one. However, in practice this is equally problematic since the resulting scores from short texts may be so low that they prohibit meaningful analysis. Therefore, the decision to use the count of formulaic words was taken on the basis that although there can be great variation in the length of formulaic sequences, this appears not to be the case in the dataset (mean=2.6, median=2.6, mode=2.5). This is coupled with a low standard deviation (the variability of distribution of average formulaic words is 0.459). Therefore, for the present purposes, it is acceptable to use the count of formulaic words as the measure, although clearly this may not be appropriate for other types of data.

Table 3, shows the total words produced by each author over their five texts and how many of those words were identified as being formulaic, that is, part of a formulaic sequence (e.g. *in fact* counts as two formulaic words, *in the end* counts as three formulaic words and so on). Table 3 is ranked from the author who uses the lowest count 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 ($\sigma = 0.63$).

Table 3 Count of formulaic words per 100 words across author corpus

Author	Total words	Total formulaic words	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

A Kruskal-Wallis test showed significantly more variation in the count of words forming formulaic sequences in texts between authors than within texts by the same author ($\chi^2 = 35$, 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: inter-author variation is greater than intra-author variation.

The next stage is to determine if it is possible to successfully attribute a QD. This requires a series of known texts and questioned documents for comparison. Using the narrative data, four texts from every author were selected as the known texts. A fifth, randomly selected text from only one of the authors in each pair was used as the QD with the equivalent text from the other author being discounted from the analysis. To test as extensively as possible, every author in the corpus was compared against every other author. Using the One Sample Kolmogorov-Smirnov Goodness-of-Fit test the twenty groups of texts were individually tested to determine if they were normally distributed. Each group of texts showed no significant difference from the norm, meaning that no single text had an uncharacteristically high or low count of formulaic sequences (see Table 4).

Table 4 One Sample Kolmogorov-Smirnov Goodness-of-Fit results

Author	Result
John	D=0.342, p=1.0
Rose	D=0.450, p=0.988
June	D=0.547, p=0.925
Keith	D=0.393, p=0.998

Jenny	D=0.509, p=0.958
Sue	D=0.706, p=0.701
Michael	D=0.593, p=0.873
Carla	D=0.507, p=0.960
Nicola	D=0.913, p=0.376
Hannah	D=0.445, p=0.989
Melanie	D=0.668, p=0.764
Sarah	D=0.630, p=0.823
Rick	D=0.537, p=0.935
Greg	D=0.538, p=0.934
Judy	D=0.522, p=0.948
Elaine	D=0.482, p=0.974
Thomas	D=0.384, p=0.998
Mark	D=0.486, p=0.972
David	D=0.592, p=0.875
Alan	D=0.372, p=0.999

Since none of the 20 authors' texts contained counts which were significantly different from normal, a one-sample two-tailed t-test was used for each comparison between a candidate author's texts and the QD. With a total of 20 authors, 190 different pairs could be formed. A sample of the results are produced as Table 5. In each instance, the first listed author is the actual author of the QD.

Table 5 Sample of t-test results

Author Pair	**Result**	**Outcome**
Melanie and Sarah	Melanie: t(3)=1.823, p=0.166 Sarah: t(3)=4.089, p=0.026*	Correct attribution—Sarah's four texts differ significantly from the QD whilst Melanie's do not.
Sarah and Nicola	Sarah: t(3)=2.381, p=0.097 Nicola: t(3)=14.400, p=0.001**	Correct attribution—Nicola's four texts differ significantly from the QD whilst Sarah's do not.
Rose and Sue	Rose: t(3)=-3.217, p=0.049* Sue: t(3)=-2.719, p=0.073	Misattribution—Rose's four texts differ significantly from the QD whilst Sue's do not.
June and Judy	June: t(3)=-7.326, p=0.005** Judy: t(3)=-0.673, p=0.549	Misattribution—June's four texts differ significantly from the QD whilst Judy's do not.
Carla and Mark	Carla: t(3)=-0.403, p=0.714 Mark: t(3)—0.235, p=0.829	No attribution—no statistical differences between either candidate author and QD.
David and Thomas	David: t(3)-0.606, p=0.588 Thomas: t(3)=2.900, p=0.062	No attribution—no statistical differences between either candidate author and QD.
Hannah and Elaine	Hannah: t(3)=4.115, p=0.026* Elaine: t(3)=3.858, p=0.031*	No attribution—statistical differences between both candidate authors and QD.
David and Alan	David: t(3)=-9.077, p=0.003** Alan: t(3)=-5.218, p=0.014*	No attribution—statistical differences between both candidate authors and QD.

*p<0.05, **p<0.01

Table 5 exemplifies four possible outcomes: 1) a correct attribution (i.e. no statistically significant difference between the QD and the author of the QD's texts, but a significant difference between the QD and the other author); 2) an incorrect attribution (i.e. a statistically significant difference between the QD and the correct author's texts); 3) no attribution possible where neither author's four known texts were significantly different from the QD; and 4) no attribution because both of the candidate authors' four texts were significantly different from the QD.

Of the 190 pairs tested, 38 (20%) were correct attributions whilst 28 (15%) were incorrect attributions. In 124 cases (65%) no attribution was possible: in 107 cases there was no significant difference between either of the candidate authors' four texts and the QD (i.e. neither could be ruled out as the author) and in 16 cases both of the candidate authors' four texts differed significantly from the QD (i.e. both were ruled out as the author). In some respects, the high level of no attributions is not problematic—the marker could not be applied in the same way that spelling errors would not be useful in a text which adhered to standard spelling. However, the proportion of misattributions is too high for the forensic context. The success rate of 20% must also be considered alongside the potential for a Type 1 statistical error—it is entirely possible that after carrying out 380 t-tests, the chance of obtaining a statistically significant difference could occur by chance alone.

Discussion

At the outset, it was claimed that the holy grail of authorship markers would be one which occurs simply by virtue of producing language and which occurs frequently in texts, and formulaic sequences were argued to hold potential as one such marker. It has been demonstrated that there whilst there is variation in formulaic sequence usage between the group of 20 authors investigated—with inter-author variation being greater than intra-author variation for both type and count of formulaic sequences—neither the type of formulaic sequences used, nor the overall count of formulaic words enables the attribution of a text to its author to a satisfactory level. Some observations about the nature of formulaic sequence usage can nonetheless be made.

The results support Kuiper's (2009) research; just as individual variation was found in the checkout data, individual variation was found in the author corpus, although not with the same 'signature' potential—that is, no formulaic sequence was used rarely overall but consistently by only one author. It is likely that the routinized nature of operating a checkout made individual differences more marked than was possible in the less routine act of writing narratives. Likewise, whilst this research cannot claim to explore the distinction between recurrent clusters and formulaic sequences, it does seem to lend support to Schmitt et al.'s (2004) claim that formulaic sequences are linked to idiolect—these authors do appear to use different types and, with the exception of two authors, Greg and Alan, different counts, so individual differences are apparent. Finally, Waltman's findings are not supported since there was no consistency in the type of formulaic sequences used across the five texts produced by each author. This is likely to be down to both Waltman's literary definition of formulaic sequences and the oral nature of the poem which may have led to a greater incidence of formulaic sequences.

Formulaic sequences identified using the method outlined in this paper do not seem to hold the potential as a marker of authorship, so it is now necessary to discuss why this may be the case.

The overriding claim about formulaic sequences is their ubiquity in language yet in these data their occurrence appears to be rather low with a mean average of 2.35 formulaic words per 100 overall words. It is likely that the method itself limited how much formulaic material could be identified since Moon (1998) argues that well-known idioms assumed to be common in language do not actually occur with any great frequency. In fact, in the Oxford Hector Pilot Corpus consisting of 18 million words, some idioms did not occur at all (e.g. *kick the bucket*, *one man's meat is another man's poison*, and *when the cat's away, the mice will play)* (p.60). Given that the reference list used in this research contained a large proportion of these types of items, it is perhaps not surprising that they were not identified in the data in higher numbers. The point, however, is that whether formulaic sequences do hold potential as a marker of authorship or not is irrelevant if a forensically robust method of identification cannot be identified and since, as Erman and Warren (2000) explain, it is impossible to identify all and only formulaic sequences, this research has at least ruled out the viability of fixed formulaic sequences as a marker of authorship.

At the same time, though, it can be argued that the data used in this study are not representative of 'natural' or 'everyday' language, since they are narratives. It is conceivable, perhaps even probable, that the authors will have told these narratives in various ways on various occasions, and as Wray (2002) comments, 'if you are called upon to tell the story of a recent incident several times in quick succession to different people, then the words you use will soon become relatively fixed' (p.84). Hence, there may be a higher incidence of formulaic sequences resulting from repetition in narratives than in 'natural' speech; however, narratives may also contain a potentially more individualistic set of formulaic sequences than could be captured using the quite general list compiled in this research.

Further, the authority of the list of formulaic sequences compiled for this study must be questioned, since it is unverified by independent means and consists solely of sequences deemed by others to be special in some way (e.g. as clichés, idioms, etc.). At the same time, though, the list is broad and inclusive and representative of numerous different language communities, and the end product equates to asking members of a range of speech communities to identify formulaic sequences by intuition. Hence, a level of resiliency may be claimed (e.g. Foster, 2001; Van Lancker-Sidtis & Rallon, 2004) that is higher than relying on only one or a few people's intuitions. It should be pointed out, though, that some entries may be 'recurrent clusters' (Schmitt et al., 2004) rather than formulaic sequences stored as unanlysed units and so may lack psycholinguistic validity. A more cautious position may be to refer to items on the reference list as 'allegedly formulaic sequences'.

Since variation between authors was observed, there is perhaps enough evidence here to justify exploring the sociolinguistic variables that may account for individual differences in formulaic sequences further since at present, adult native L1 speakers have been somewhat neglected by the research literature. In addition, if formulaic sequences can be linked to sociolect, the potential may exist for authorship profiling. In conclusion, this study shows that there seems to be potential for formulaic sequence usage to differ between individuals, but the method outlined in this paper has not been able to capture those differences sufficiently, nor are we yet able to use type or amount of formulaic language as a reliable marker of authorship.

Appendix

Biographical information collected about each participant

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

Endnote

[1]I am extremely grateful to the anonymous reviewers for their insightful and valuable comments.

References

Chenoweth, N. A. (1995) Formulaicity in essay exam answers. *Language Sciences* 17.3: 283—97.

Coulmas, F. (1979) On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics* 3: 239—66.

Coulthard, M. (2004) Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics* 25.4: 431—447.

Erman, B. and Warren, B. (2000) The idiom principle and the open choice principle. *Text* 20.1: 29—62.

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, and M. Swain (eds) *Researching Pedagogic Tasks: second language learning, teaching and testing* 75—94. London: Longman.

Goldberg, A. (2003) Constructions: a new theoretical approach to language. TRENDS in Cognitive Sciences 7.5: 219—224.

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

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

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

Labov, W. and Waletsky, J. (1997) Narrative analysis: oral versions of personal experience. *Journal of Narrative and Life History* 7.1: 3—38.

McMenamin, G. (2002) *Forensic Linguistics: advances in forensic stylistics*. London, CRC Press.
Menacere, T., Taylor, P.J., & Tomblin, S. (2008) *Linguistic Analysis Suite* (v1.3) [computer program]. Department of Psychology, Lancaster University.
Moon, R. (1998) *Fixed Expressions and Idioms in English.* Oxford: Clarendon Press.
Pawley, A. and Syder, F. (1983) Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. Richards and R. Schmidt (eds) *Language and Communication* 191—226. New York: Longman.
Peters, A. (1983) *The Units of Language Acquisition.* Cambridge: Cambridge University Press.
Read, J. and Nation, P. (2004). Measurement of formulaic sequences. In N. Schmitt (ed.), *Formulaic Sequences* 23—35. Amsterdam: John Benjamins Publishing Co.
Rubin, D. C. (1998) *Memory in Oral Traditions: the cognitive psychology of epic, ballads, and counting-out rhymes.* Oxford: Oxford University Press.
Schmitt, N., Grandage, S., and Adolphs, S. (2004) Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (eds) *Formulaic Sequences: acquisition, processing and use* 127—151. Amsterdam, John Benjamins Publishing Company
Shuy, R. (2006) *Linguistics in the Courtroom: a practical guide*. Oxford, Oxford University Press.
Van Lancker-Sidtis, D. and Rallon, G. (2004) Tracking the incidence of formulaic expressions in everyday speech: methods for classification and verification. *Language & Communication* 24: 207—40.
Vihman, M. (1982) Formulas in first and second language acquisition. In L. Obler and L. Menn (eds) *Exceptional Language and Linguistics* 261—84. London: Academic Press Ltd.
Waltman, F. (1973) Formulaic expression and unity of authorship in the "Poema de Mío Cid. *Hispania* 56.3: 569—578.
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* 590—7. Oxford: Elsevier.
Wray, A. (2008) *Formulaic Language: pushing the boundaries.* Oxford: Oxford University Press.
Wright, D. (2013) Stylistic variation within genre conventions in the Enron email corpus: developing a text-sensitive methodology for authorship research. *International Journal of Speech, Language and the Law* 20.1: 45—75.