Proceedings of The International Association of Forensic Linguists' Tenth Biennial Conference

Edited by
Samuel Tomblin, Nicci MacLeod,
Rui Sousa-Silva and Malcolm Coulthard

Editorial Assistant
Andrea Nini

IAFL 10

Centre for Forensic Linguistics

ISBN: 978 1 85449 432 0

Published in 2012 by the Centre for Forensic Linguistics

Centre for Forensic Linguistics
Aston University
Aston Triangle
Birmingham
B4 7ET, UK.

Investigating formulaic language as a marker of authorship

Samuel Tomblin
Centre for Forensic Linguistics, Aston University, UK
tomblisd@aston.ac.uk

Abstract

This research unites the psycholinguistic theory of formulaic language, that is, prefabricated sequences of words believed to be stored as holistic units, and the practice of forensic authorship attribution with a view to developing a new marker of authorship. Since formulaic sequences are holistically processed, it stands to reason that they are likely to elude a writer's attempts to disguise their style. It follows that research into formulaic language usage may therefore assist in the development of new tools for authorship attribution. In order to test this hypothesis, a reference list containing 13,412 examples of formulaic sequences was compiled from multiple online sources (e.g., lists of clichés, idioms) which was then used to identify formulaic language in a 20 author corpus containing 100 personal narratives. A series of statistical tests were used to determine whether the proportion of formulaic language compared to novel language was sufficient to differentiate authors and to attribute a Questioned Text to its author. The results are discussed with reference to the reliability and validity of the method.

Keywords: FORMULAIC LANGUAGE; FORMULAIC SEQUENCES; AUTHORSHIP ANALYSIS; AUTHORSHIP ATTRIBUTION; MARKER OF AUTHORSHIP

Introduction: formulaic sequences

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).

However, whilst the potential for novel utterances is limitless, speakers appear 'to renounce the great freedom that the language offers'. 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., Wray, 2002), sociolinguistics (e.g., Coulmas, 1979), corpus linguistics (e.g., Moon, 1997; 1998a; 1998b) and both L1 and L2 language acquisition (e.g., Pawley and Syder, 1983; Peters, 1983; Peters, 2009; Peters, 1977; Vihman, 1982) shows that when communicating, we 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 (Ellis, 1996; Fillmore, 1979; Coulmas, 1981; Pawley and Syder, 1983; Howarth, 1998).

Such routine language can in a global sense 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). 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 analysing it into its component parts' (2002: 3). In this way, 'Rice Krispies' seems 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 and Lieven, 2009; Ellis, 1996; Erman and Warren, 2000; Pawley and Syder, 1983; Wray, 2000; 2002; 2008).

Formulaic language is an umbrella term and a survey of the literature soon reveals that many terms exist to describe different characteristics of formulaic language. These include Collocations (Herbst, 1996; Gledhill, 2000; Stubbs, 1995), Idioms (Grant and Bauer, 2004; Simpson and Mendis, 2003), Fixed Expressions including Idioms (Moon, 1998a) Formulaic Sequences (Wray, 2002; Schmitt and Carter, 2004), Multi-word Items (Moon, 1997), Phrasal Lexemes (Moon, 1998b), Recurrent phrases (Stubbs and Barth, 2003) and Situation Bound Utterances (Kecskés, 2000), to name just a few. In fact, Wray (2002: 9) found 57 different terms each describing what can be characterised as formulaic. Though related, these terms 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 amount of distance between words (Hoover, 2003) or whether sequences of words should be contiguous (Hoover, 2002; Stubbs, 2002; Stubbs and Barth, 2003). The definition adopted in this research is that of 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 (Wray, 2002: 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).

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 found 77% of written answers to essay style exam questions contained formulaic expressions regardless of length (1995: 292). Bannard and Lieven (2009) found that between 86% and 97% of utterances spoken by toddlers were derived from recurring strings and 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 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).

A function of formulaic sequences: reducing cognitive burden

The exact size of the mental lexicon is not known, although de Bot (1992) estimates there to be about 30,000 words in the active lexicon. De Bot calculates that with an average rate of speech of 150 words per minute, peaking at approximately 300 words per minute, the average speaker has 200—400 milliseconds to select the words they wish to use. Expressed another way:

> 2 to 5 times a second we have to make the right choice from those 30,000 words. And usually we are successful; it is estimated that the probability of making the wrong choice is one in a thousand (p. 11)

Producing language is clearly a cumbersome task, albeit one which humans manage with relative ease. However, if sequences of words are stored as single lexical items (as the theory of formulaic language suggests), then accessing that sequence of words from the lexicon, rather than constructing a novel sequence from individual words, plausibly reduces the amount of cognitive processing required:

> Formulae make the business of speaking (and that of hearing) easier. I assume that when a speaker uses a formula he or she needs only to retrieve it from the dictionary instead of building it up from its constituent parts. In other words, such expressions likely exist as whole or part utterances within the speaker's dictionary and need not be built up from scratch on every new occasion (Kuiper, 1996: 3).

Therefore, there is consensus that reducing cognitive burden is a function of formulaic language (Kuiper, 2000; Peters, 1983; Wray, 2002; Wray and Perkins, 2000;). By decreasing cognitive load, fluency can be increased (Fillmore, 1979; Kuiper, 1996).

Since formulaic sequences are stored in this pre-packaged, holistic form they are likely to escape conscious regulation by authors—in other words, authors will produce sequences of words without necessarily thinking about each individual word. 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)

It stands to reason that if evidence can be found of formulaic sequence usage varying between authors it should make an excellent candidate for a new marker of authorship.

Aims and hypotheses

The aim of this paper is to determine whether the overall proportion of formulaic sequences in texts is sufficient to differentiate authors. By concentrating on the proportion of text that is formulaic, it will be possible to make claims about whether the language used by one author is more or less formulaic than that of another. If this is the case, the consistency in levels of formulaic sequences across a series of 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 analysis. To carry out this investigation, a series of hypotheses must firstly be proposed.

Individuals are socialised differently and this affects their repertoires of formulaic sequences (Wray, 2002). Therefore, since each of the authors that contributed data to this research will have a different set of life experiences, they should have a different range of formulaic sequences to draw upon; some with larger repertoires and some with smaller. Taking into account each author's potential formulaic repertoire and the range of cognitive demands placed on them in producing language, it is hypothesized that authors will use differing proportions of formulaic sequences. Secondly, if the first hypothesis is correct, there should be a significant difference in the proportion of formulaic sequences compared to novel language used by an author and, based on this variable it should be possible to differentiate authors. This is important for the forensic context in demonstrating that the variation between authors is significant. Thirdly, if support is found for the first and second hypotheses, using a corpus of texts which have been carefully controlled for genre and length, which have also been composed in the same time period and on similar topics, it should be possible to attribute a text to its author.

In summary, the following hypotheses will be tested:

i) Variation in proportion of formulaic sequences will be greater between authors than within authors;

ii) Authors will be potentially differentiable from each other based on the proportion of formulaic sequences in their texts;

iii) A randomly selected Questioned Text will be correctly attributed to its author based on the proportion of formulaic sequences in the Questioned Text and in the author's other texts.

Method

The task of identifying formulaic sequences in texts is not an easy one; so difficult in fact that 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). Formulaic sequences can be identified in several ways, depending on whether the language is spoken or written. In spoken language, formulaicity can be identified through phonological analysis which focuses on stress, articulation, fluency and pausing (e.g., Pawley and Syder, 1983; Peters, 1983; Wray, 2002 for a comprehensive review). Whether formulaic input or output is under investigation is also relevant. Eye-gaze studies, for example, can be used to monitor how participants read formulaic sequences if input is the focus (Underwood, Schmitt and Galpin, 2004).

In written language, 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, however, 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:

> 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).

Whilst a reference list of formulaic sequences may be an excellent resource in practical terms (e.g., analysis on 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 for the list to be nothing more than the intuitions of one individual. For the present research, a compromise is proposed: using the internet to build a reference list. By drawing on a multitude of different sources compiled by many members of various speech communities should ensure that the list is as representative of formulaic sequences as possible.

Compiling a reference list

Terms commonly accepted as 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 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 into the database regardless of whether or not they were intuitively pleasing as examples of formulaic sequences.

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 formulaic sequences was made when it was evident that relatively few new examples were actually being added to the list. The list at this stage contained 17,973 entries.

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

Table 1: Proportion of different types of formulaic language 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 over half of the entire list of formulaic sequences. The category 'Everyday Expressions and Sayings' highlights the problem of relying on selfreports for categorisation purposes: the dividing line between a cliché, idiom and everyday saying is in no way clear cut.

Pronouns in the reference list were replaced with an asterisk. The software used to identify matches in the data 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*). Through this process, by changing the entry *his bark is bigger than his bite* to ** bark is bigger than * bite* enabled 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. (although, of course, some of these might not be nonsense and may be deliberate playing with words). 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. The only pronouns that remained fixed in the list were those where substitution would affect the meaning e.g., *get thee behind me Satan, love that dare not speak its name, one small step for man, cry me a river* etc.

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>.* Finally, there were many duplicates in the list, as noted above, which were removed. The final reference list contained 13,412 entries.

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 (e.g., *Jiminy Christmas, date rape)*. The aim of the list is not really to reach universal agreement about what constitutes a formulaic sequence; rather, the aim is to collate as many potentially formulaic sequences as possible in order to investigate whether evidence can be found that some individual authors use formulaic sequences more than others. Just as the list cannot claim to be representative of formulaic sequences for each individual, questions must also be asked about its authority. That is to say that the entries of formulaic sequences have not been verified by independent means, other than by their inclusion on public websites as opposed to being included, for example, on the basis of corpus frequency counts. 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 language in texts (e.g., Foster, 2001; Van Lancker-Sidtis and 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 examples of formulaic sequences in real text. However, it does hold certain advantages which are particularly favourable for the forensic context. By using an automated approach, large volumes of data can be analysed almost instantaneously. It offers reliability; formulaic sequences included in the list will be matched in any data on any occasion. However, 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). It cannot even guarantee that every instance it identifies will 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 initial exploration into the relationship between formulaic language 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 begin our investigation into whether formulaic language has potential as a marker of authorship.

Data

The data comprise 100 texts written by 20 authors, each author producing five texts. Authors were provided over a five day period with a daily structured writing task. Authors were sent two essay-style questions each morning and were required to answer whichever one they felt most comfortable writing about. Open-ended questions which elicited personal narratives were asked. By asking emotionally-charged questions, it is hoped that the likelihood of participants focussing on their language use was reduced (Labov, 1970; Labov, 1972; Labov and Waletsky, 1997).

The 100 texts contained 65,113 words with each author producing an average of 3,325 words across their five texts. The average text length was 651 words with the shortest being 485 words and the longest being 822 words. The software compared the data with the reference list and highlighted all instances of exact matches.

Results

A total of 604 formulaic sequence tokens were identified in the data, of which there were 300 types. Table 2 shows the ten most frequently occurring formulaic sequences whilst Table 3 shows a selection of ten formulaic sequences that were used only once across the whole data set.

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

Table 3: Least frequently occurring formulaic sequences across the data

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

Table 4 shows how many words each author produced over the total of their fives 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 whilst *at the same time* counts as four and so on). The authors are listed in numerical order from the author using the lowest frequency of formulaic sequences to the one using the highest. To facilitate comparison between the authors, a normalised frequency of formulaic language per 100 words is also provided.

Table 4: Proportion of formulaic language across the data

Author	Total Words	Total Formulaic Words	Normalised Frequency of Formulaic Language per 100 words
MELANIE	2879	34	5.97
SARAH	2957	46	7.57
ROSE	3820	66	8.63
JOHN	3119	55	8.81
CARLA	3217	59	8.99
JUNE	3151	59	9.28
MARK	2844	56	9.92
DAVID	3058	63	10.05
NICOLA	3021	62	10.24
GREG	2980	70	11.62
ALAN	3916	92	11.67
MICHAEL	2516	61	12.12
SUE	3716	94	12.63
RICK	3583	93	12.90
JENNY	3518	103	14.82
JUDY	3427	104	15.26
HANNAH	3559	111	15.54
KEITH	3067	95	15.74
ELAINE	2941	94	16.03
THOMAS	3824	130	17.02

Establishing variation between authors

Frequency of use of formulaic sequences was tested between 20 authors providing five texts each. A Kruskal-Wallis test showed significantly more variation 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 proportion 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 proportion of formulaic sequence usage for gender, age or education.

With regard to the second hypothesis, that authors will be differentiated from each other based on the normalised frequency of formulaic language usage, it is evident from Table 4 that these authors do in fact use different normalised frequencies of formulaic sequences in their texts. However, the statistical significance of these differences is not clear. In what follows, the second hypothesis is statistically tested.

Differentiating authors: a test case

As a test case the highest and lowest mean ranked authors were compared (Thomas and Melanie respectively). With just five texts each (equivalent to a total of 3,824 words and 2,879 words) it was possible to differentiate these two authors based on the normalised frequencies of occurrences of formulaic language (Mann-Whitney U = 1, N = 10, p = 0.016). This provides evidence that using the normalised frequency of formulaic language as a marker of authorship works. The question that remains is how well it works for authors whose normalised frequency of formulaic language is more similar.

Differentiating authors: a harder case

Taking the highest and lowest mean ranked authors somewhat improves the likelihood of reaching significance, since these authors were at the extreme ends of normalised frequency of formulaic language usage. Therefore, to further test the method, the two authors with the closest normalised frequency of formulaic language were compared (Greg and Alan, with 11.62 and 11.67 formulaic words per 100 respectively). With just five texts each (equivalent to a total of 2,980 words and 3,916 words) it was not possible to differentiate these two authors (Mann-Whitney U = 11, N = 10, p = 0.841). This result clearly invites the question of how effective the method is when two authors whose normalised frequency of formulaic language is neither very similar nor very different.

Differentiating authors: exploring the limits

Two sets of authors were selected to explore the limits of the method. The 5th ranked author, Carla, and the 16th ranked author, Judy, were selected for the analysis (8.99 and 15.26 formulaic words per 100 respectively). With just five texts each (equivalent to a total of 3,217 words and 3,427 words) it was possible to differentiate these two authors (Mann-Whitney U = 23, N = 10, p = 0.032). Secondly, the 7th ranked author, Rick, and the 14th ranked author, Mark, with 12.90 and 9.92 formulaic words per 100 respectively were compared. With five texts each (equivalent to a total of 3,583 words and 2,844 words) it was not possible to differentiate these two authors (Mann-Whitney U = 6, N = 10, p = 0.222). It can be seen that the normalised frequency of formulaic language was too close for Rick and Mark, whereas the texts produced by Carla and Judy enabled differentiation.

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 normalised frequency of formulaic language 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 different proportions of formulaic sequences in their texts from some other authors.

The analysis carried out here relies on pairwise distinctions (e.g., Grant, 2010) as opposed to population wide distinctions (e.g., Chaski, 2001). Therefore, the variable, normalised frequency of formulaic language, 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 quantity of formulaic sequences in a text as a marker of authorship. In a closed sample, some authors can be differentiated whilst others cannot. Therefore, it is not possible to claim the method described here as a universal method that will be applicable in all cases.

Assessing forensic potential

In an attempt to replicate a realistic forensic case, five texts by each of two authors were randomly selected for analysis: Nicola and Greg. The two groups of texts were tested to see if they were normally distributed. Both groups showed no significant difference from normal (Nicola: KSZ = 0.913, N = 5, p = 0.376; Greg: KSZ = 0.445, N = 5, p = 0.989). The second text by Nicola was randomly selected by SPSS to act as the Questioned Text.

A two-tailed one-sample t-test showed no significant difference between the normalised frequency of formulaic language in the four texts by Nicola compared to the Questioned Text, also by Nicola (t(3) = 0.601, p = 0.590). As the prediction from the means was that Nicola's scores would be lower than those of Greg, a uni-directional hypothesis was tested. A one-tailed one-sample t-test showed a significantly higher normalised frequency of formulaic language in the five texts by Greg compared to the Questioned Text (t(4) = 2.157, p = 0.0485). In real terms, we can say that there is a 95% chance that Nicola wrote the Questioned Text which is arguably an acceptable level of confidence for forensic linguistics evidence and which we know to be a correct attribution. In terms of the final hypothesis, that a randomly selected Questioned Text will be correctly assigned to its author based on the normalised frequency of formulaic language in that author’s other four texts, the results demonstrate that when a Questioned Text is compared to nine Known Texts produced by a closed set of two authors, it is possible to correctly attribute the Questioned Text to its correct author. The final hypothesis is therefore supported.

Discussion

These results provide evidence that taking the normalised frequency of formulaic language usage as a marker of authorship does have the potential to differentiate authors and, more importantly, to attribute a Questioned Text correctly to its author. In line with the aims of this research, the focus has only been on the proportion of formulaic language usage compared to novel language, in other words, whether authors or more or less ‘formulaic’ than others. However, it is important to acknowledge that no individual formulaic sequences emerged as being characteristic of authorship, that is, no individual formulaic sequence appears to be related to idiolect.

In order to fully contextualise the success and effectiveness of the method, it is necessary to discuss the validity and reliability of the method.

Is the method valid?

To assess the validity of the method, it is necessary to critically examine whether formulaic language has 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 in this paper, it is highly unlikely that everybody will agree that what is included in the reference list is a formulaic sequence and therefore is an example of formulaic language. 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 basis for this paper has been that different authors will have different cognitive abilities which will be evidenced through their reliance on formulaic sequences. The reality is that several other factors may have had an impact on an 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.). 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 fewer cognitive pressures 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.)

Is the method reliable?

In order for the method to be reliable, it would need to be proven that the same examples would be identified each time the analysis is replicated. This is true since the method is automated and so is unaffected by factors which commonly affect reliability (e.g., tiredness of the researcher, unprincipled analysis of large quantities of data, etc,.). 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 paper has focussed only on formulaic sequences occurring in a very restricted type of data—short personal narratives.

The reliability of the method may 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 normalised frequency of formulaic language in the responses? To assess this, all of the narrative eliciting 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. 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 language to aid memory during public performances, cf. Rubin (1998)).

Finally the range of speech communities represented by the list should also be considered. A wide variety of UK and USA variants have been included. In principle, the reference list can therefore be used to identify formulaic language in texts produced by speakers of British or American variants of English. However, it could 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 is therefore unreliable as a universally-applicable method for authorship analysis.

Conclusions

This paper has outlined a method of authorship attribution which takes the normalised frequency of formulaic language 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 some pairs of authors with different normalised frequencies of formulaic language usage can be statistically differentiated and that when two authors are randomly selected, a Questioned Text can 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 is required in order to demonstrate reliability. Despite the positive conclusions that can be drawn from this pilot investigation method, it is important to stress that although the two authors with the highest and lowest normalised frequencies of formulaic language in their texts could be differentiated, it was not possible to differentiate the two authors with the most similar normalised frequencies of formulaic language. However, there is currently no unified method for authorship analysis, and instead the linguist must select the most appropriate methods from a rich toolkit. With further testing, the method described here 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.

References

Bannard, C. and Lieven, E. (2009) Repetition and reuse in child language learning. In R. Corrigan, E. Moravcsick, H. Ouali and K. Wheatley (eds) *Formulaic Language: Acquisition, Loss, Psychological Reality, and Functional Explanations* Vol. 2. Amsterdam: John Benjamins Publishing Co., 299—321.

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—97.

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—66.

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.

de Bot, K. (1992) A bilingual production model: Levelt's 'Speaking' model adapted. *Applied Linguistics* 13(1): 1—24.

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

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

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

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.* London: Longman, 75—94.

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

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

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

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

Hoover, D. L. (2002) Frequent word sequences and statistical stylistics. *Literary and Linguistic Computing* 17(2): 157—80.

Hoover, D. L. (2003) Frequent collocations and authorial style. *Literary and Linguistic Computing* 18(3): 261—86.

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

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

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.

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

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—4): 3—38.

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

Moon, R. (1997) Vocabulary connections: multi-word items in English. In N. Schmitt and 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.

Nattinger, J. R. and DeCarrico, J. S. (1992) *Lexical Phrases and Language Teaching.* Oxford: Oxford University 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.* New York: Longman, 191—226.

Peters, A. (1977) Language learning strategies: does the whole equal the sum of the parts? *Language* 53(3): 560—73.

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. Moravcsick, H. Ouali and K. Wheatley (eds) *Formulaic Language: Acquisition, Loss, Psychological Reality, and Functional Explanations* Vol. 2. Amsterdam: John Benjamins Publishing Co., 387—404.

Rubin, D. C. (1998) *Memory in Oral Traditions: the Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes.* Oxford: Oxford University Press.

Schmitt, N. and 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.

Simpson, R. and Mendis, D. (2003) A corpus-based study of idioms in academic speech. *TESOL Quarterly* 37(3): 419—441.

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—44.

Stubbs, M. and Barth, I. (2003) Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language* 10(1): 61—104.

Underwood, G., Schmitt, N. and 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. and Rallon, G. (2004) Tracking the incidence of formulaic expressions in everyday speech: methods for classification and verification. *Language and Communication* 24: 207—240.

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

Wray, A. (2000) Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics* 21(4): 463—89.

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—7.

Wray, A. (2008) *Formulaic Language: Pushing the Boundaries.* Oxford: Oxford University Press.

Wray, A. and Perkins, M. (2000) The functions of formulaic language: An integrated model. *Language and Communication* 20: 1—28.