Bridging the Gap between Stylistic and Cognitive Approaches to Authorship Analysis using Systemic Functional Linguistics and Multidimensional Analysis

Abstract

In the present state of the art of authorship attribution there seems to be an opposition between two approaches: the cognitive and the stylistic methodologies. It is proposed in this paper that these two approaches are complementary and that the apparent gap between them can be bridged using Systemic Functional Linguistics (SFL) and in particular some of its theoretical constructions, such as *codal variation*. This paper deals with the theoretical explanation of why such a theory would solve the debate between the two approaches and shows how these two views of authorship attribution are indeed complementary. Although the paper is fundamentally theoretical, two examples of experimental trials are reported to show how this theory can be developed into a workable methodology of doing authorship attribution. In Trial 1, a SFL analysis was carried out of a small dataset consisting of three 300 word texts collected from three different authors whose socio-demographic background matched across a number of parameters. This trial led to some conclusions about developing a methodology based on SFL and suggested the development of another trial, which might hint at a more accurate and useful methodology. In Trial 2, Biber's (1988) multidimensional framework is employed and a final methodology of authorship analysis based on this kind of analysis is proposed for future research.

Keywords: authorship attribution, systemic functional linguistics, multidimensional analysis, forensic stylistics

Current issues in authorship analysis: stylistic versus cognitive approaches

The distinction between the stylistic and cognitive methods of authorship analysis is perhaps the most significant of the current debates in forensic authorship analysis. The debate has developed from Chaski's (2001) paper and the subsequent responses of Grant & Baker (2001) and McMenamin (2001), with more recent contributions by Howald (2009) and Grant (2010). The significance of the debate for the field of forensic linguistics lies not only in the

issue of finding reliable ways to determine the authorship of anonymous writings but also in the claims to scientific validity of the methods used. Solving the methodological problem would be a significant step forward towards the acceptance of a set of standard, scientific methods of authorship analysis. Here we first report the key points of the discussion, which mainly revolve around the concept of idiolect in two approaches that have been defined in the previous literature as the *stylistic* and the *cognitive* (Grant, 2010).

Stylistic methodology

On the stylistic side, we have forensic linguists claiming that (a) a theory of idiolect is important but not necessary; (b) what makes an individual unique and, therefore, identifiable, is their style, which, as pictured by McMenamin (2002: 109), is 'the variable element of human behaviour' that includes every aspect of human life that involve a choice; (c) the style of an individual can be found in language through the analysis of the array of *style markers* that an individual habitually selects from their repertoire; (d) the repertoire of style markers that an individual has at their disposal is generated by the variability inherent in the historical-sociological differences that naturally distinguish human beings. A *style marker*, thus, as weakly defined by McMenamin (2002), seems to roughly correspond to Labov's sociolinguistic variant of a variable (Labov, 1966: 15). McMenamin's forensic stylistics style markers could be formalised as, for example, alternations like: (cannot):[cannot] *vs* (cannot):[can not]. In the same way as Labov's research showed that, for example, (r):[r] *vs* (r):[ø] can distinguish certain varieties of English, McMenamin's analysis of the style markers like the one presented above is believed to distinguish certain authors. Whenever in language there is the possibility of a choice, at any level, the choice that the author makes may be revealing of their background and therefore lead to an identification.

Forensic linguists working within this stylistic paradigm tend to believe that each human being has a different repertoire of linguistic variables that manifest themselves in their writings and that these can be accounted for by socio-historical differences. The socio-historic differences would include static differences such as gender and ethnicity but also dynamic differences such as age, job, geography (dialect), second language influences and many other factors. The differing socio-historic background leads to differences in repertoires of choice and this can make discrimination between individuals feasible, especially so in cases where the background of the possible writers of the anonymous text varies.

Cognitive methodology

On the cognitive side of the debate, we have forensic linguists claiming (a) that it is fundamental to operate with a theory of idiolect, that is, a theory that describes and predicts a person's 'own distinct and individual version of the language they speak and write' (Coulthard & Johnson, 2007: 161); (b) that what makes an individual unique is the structure of their cognition and that this is reflected in the language structures that they produce; (c) that it is by analysing syntactic patterns that it is possible to distinguish individuals; and they also claim, importantly, (d) that a methodology of authorship analysis has to be grounded in a cognitive scientific linguistic theory and has to be reliably and quantitatively tested.

Chaski (2001) tried to show empirically what she claims theoretically by trying to demonstrate that the cognitive methodology, exemplified by her analysis of syntactic structures, performs better than other kinds of methodologies. According to Chaski, forensic stylistics is not on her terms scientific, it fails to replicate its results in her test and, furthermore, it 'rests on erroneous assumptions about individuality in linguistic performance and violates theoretical principles of modern linguistics' (Chaski, 2001: 3).

Some critiques of both approaches

Both sides of this debate claim the high ground of scientific status. More than once McMenamin (2002) collocates the noun *science* with the word *stylistic*. Chaski, however, claims that McMenamin's approach is not empirical and she fails to replicate a version of stylistic analysis in her experiment. She in contrast declares her approach to be scientific and she argues that her syntactic markers are perhaps easier to replicate and evaluate than some of the stylistic markers used in forensic stylistics. This is because Chaski's syntactic markers can be selected *a priori* to approaching a text and they are amenable to clear objective definition and grounded in a linguistic theory. In contrast, a stylistic approach will not define the specific markers to be used in advance of approaching a text. When stylistic contrasts are drawn, these are sometimes loosely defined and can be harder to measure and evaluate. This makes it hard to replicate the analysis and therefore to claim objectivity and universality.

A problem then for forensic stylistics is that it does not provide the linguist with an array of markers that has to be used each time and in every case. This lack leads to a significant difficulty. The method starts with the analysis of texts and the search for style markers which will distinguish between texts but it could be the case that there are always arrays of markers that distinguish *texts*, even when those texts are produced by the same *author*. The

generalisation from observations of *texts* to conclusions about *authorship* may be weakly evidenced. Where this occurs the weighting of evidence and judgements will then depend upon a more subjective expertise of the analyst. One danger for the forensic analyst is that the result of a forensic stylistic analysis will always be subject to the analyst's bias. In other apparently more objective areas of forensic science such as forensic finger print examination it has been shown that analysts' decision making is always susceptible to bias. It is also shown that such biases can significantly affect outcome decisions by forensic scientists (Dror *et al*., 2011; Dror *et al*., 2006). The issue of subjectivity might be reduced if the markers are defined and consistent across cases. The *a priori* definition of markers is conversely the strength of those who take the cognitive approach.

Chaski (2001) develops a set of features described as syntactically classified punctuation, for example, distinguishing between commas used as clause boundaries and commas used in lists. These provide a good example of markers which are defined clearly in advance of any forensic problem. The most important difficulty with their application is that Chaski does not explain why and how her maker should be able to distinguish authors (Grant & Baker, 2001; Grant, 2010). The justification that Chaski claims for her syntactic markers is that syntactic analysis is grounded in linguistic theory and that syntactic behaviour is 'automatized, unconscious behaviour and therefore is difficult either to disguise or imitate' (Chaski, 2001: 8). She does not state, however, why there should be difference in syntactic behaviour between individuals. This weakness in the cognitive method reflects a strength in the more stylistic methods of authorship analysis. Stylistic methods do provide a justification on why their markers distinguish authors. The differences in spelling, word forms, grammatical constructions and so on originate from the different sociolinguistic background that each individual presents.

A proposal for complementarity

The question that this paper deals with is therefore: is it possible to bridge the gap between these two stances and find a method that has the advantages of both?

A possible step towards this goal is the method that Grant & Baker (2001) nod towards at the end of their paper. They propose to use a statistical technique such as Principal Component Analysis (PCA) to find which of a large set of authorship markers is more distinctive for each author in the pool of suspects. PCA is just one approach of a set of statistical and computational approaches which allow the analyst to find which marker is able to

differentiate the authors in an individual case. By using the resulting variable or variables, it is then possible to assign the questioned text to its author. Since 2001 there have been considerable advances in data driven text analytic approaches for both authorship analysis and other classification tasks. Reviews include Grieve (2007) and Stamatatos (2009).

This collection of methods might be used to solve the major problem of McMenamin's original approach. This computational data driven analysis is an objective method that can be applied in every case and that includes a clear definition of markers to be used each time, avoiding therefore the problem of analyst's bias. However, and this is a point that Chaski (2001) and Howald (2009) raised, the methods lack a theory of idiolect. Grant & Baker explain that 'In PCA strange combinations of markers [...] might prove to be effective discriminatory components. The explicit lack of validity, however, prevents any generalization beyond the data set analysed' (Grant & Baker, 2001: 77).

If a linguistic theory of idiolect could be added to a multivariate data driven method, then the picture would be complete and it could be possible to develop a method that is:

1. Grounded in modern linguistic theory;
2. Justified by a theory of idiolect that explains why individuals should vary, both in cognitive and in sociolinguistic terms;
3. Objective, in the sense that it is always applied in the same way for every case using always the same markers.

The position of this paper is that a theoretical base for a linguistically valid, more objective method of authorship attribution may be bridged using Systemic Functional Linguistics.

Systemic Functional Linguistics, Codal Variation and a theory of idiolect

Systemic functional linguistics (SFL) is a holistic theory of language developed by M. A. K. Halliday from 1961 (for a general review of SFL see Halliday & Matthiessen (2004), Eggins (2004) and Matthiessen (1995)). With the term 'holistic' it is meant that it is a linguistic theory that describes every level of language within one single model, starting from phonology to pragmatics up to sociolinguistic variation. That being so, it is proposed in this paper that SFL is a good candidate for a valid theory applicable to authorship attribution, since this holistic approach provides consistent terminology and coherence between the analytic tools. In addition to this, SFL is a sociolinguistic theory of language that expects a variation at group level and a variation at the individual level. It is, therefore, exactly the theory that McMenamin (2002) invokes when he suggests that what is needed is a theory that conceives variation as

inherent in language. In addition, SFL has also found support of its basic principles at the cognitive level thanks to the work of Lamb (2013), who developed correlations between neurocognitive linguistics and SFL.

SFL presupposes three kinds of variation which are inherent in language. This means that it is expected that each stretch of language produced is a function of three sources of variation. These kinds of variation are: *registerial* (from the *register* of a text), *codal* (from the *code* of an individual or group of individuals) and *dialectal* (from the *dialect* of a group of individuals or individual) (Matthiessen, 1995; 2007).

Registerial variation and dialectal variation are well established in more general linguistic theory and analysis. However, in SFL these two have a precise and specialist definition. Registerial variation is formally the skewing of probability of selection of options in the systems of a language that depend on contextual variables. For example, if we take a general reference corpus of English, we might find that the frequency of occurrence of a certain variable, say, 'past tense', is a number X per thousand words. In SFL terms, this translates into a certain probability X for 'past tense' to be selected in the system TENSE, whose other options would include 'present tense' and 'future tense'. However, if we randomly extract from that general reference corpus a sub-corpus of narrative texts, then we might find out that the frequency of occurrence (or, again, the probability of selection) for the variable 'past tense' has changed into the number Y, which is higher than X. If we analyse the texts closely, we perhaps find out that this skewing of probability is due to the contextual or generic nature of the texts, that is, the fact that those texts have to narrate past events more often than the average. This variation in the semantics (as 'past tense' is part of the 'interpersonal meaning' of a clause) can only be explained by reference to the context and it is defined in SFL as *registerial variation*.

Dialectal variation, on the other hand, is the realisations of the options of the systems. What we normally call 'dialect' in SFL is the realisation of semantic options like, as in the previous example, 'past tense'. These semantic options are shaped by a phonic or graphic substance. Therefore, for instance, two possible realisations of the past tense of the verb *burn* are *burned* or *burnt*. This is defined as dialectal variation, since the semantic option selected by the speaker is the same (the meaning of 'past tense') and what varies is the realisation of this (*burned* or *burnt*). Since SFL explains that the realisations of semantic features are norms established by a network of people that interact with each other, dialectal variation depends on the individual's social-geographical background, as it is the version of the language that they know (Matthiessen, 2007: 538).

In addition to this, SFL attempts to answer a further question: How do we explain the variation that depends on the social background of the speaker/writer but that involves meanings rather than form? This kind of variation can be explained by referencing one of the most cited cases in the forensic linguistics literature: the Derek Bentley case (Coulthard & Johnson, 2007). In this case, Coulthard showed that the position of *then* in respect of the pronoun *I* (therefore *then I* versus *I then*) varied significantly in relation to what group of speakers produced the report: lay people or police officers. It is important to note here that what varies in this case is the semantics. According to SFL, the difference in the position of the pronoun in this case changes the Theme of the clause, which is *then* in the first case (focus on the sequence of events) and *I* in the second case (focus on the person who is acting in the event). That being so, this variation cannot be defined as dialectal, since neither group is using words or pronunciations that are not used by the other group. However, this variation does not seem registerial either, since in this case both groups operated in a similar context (a report). SFL, therefore, lists another form of variation that explains cases like this: *codal variation*. Formally speaking, codal variation is a variation in the skewing of probability of selection of options in the systems of a language (like in the registerial variation) but that is caused by the social background of an individual or group of individuals or, more specifically, by how different social groupings interpret the context.

Codal variation is a concept elaborated by Hasan (1990). In her work, Hasan shows that there is a significant difference in how mothers interact with their children in terms of the probability of occurrence of certain linguistic features and that this difference can be accounted for only in relation to the social class of the mothers. She explains that this is so because a *code* is the style that people adopt in a certain context as a way to respond to that context in relation to what interpretation of that context they learned. The interpretation of context is given by the social background of the individual or, in extreme cases, by the individual themselves and becomes part of their cognition, as Vygotsky suggested (quoted in Hasan, 2009). In the Bentley case, two social groupings interpreted the context in different ways. It is reasonable to hypothesise that the lay people believed that in a context of police report the important piece of information is the temporal sequence of events. On the other hand, the police officer believed that, in this same context, the focus of the clauses should be the person who is doing the action. What shapes the code of an individual is therefore the experience of a certain context that the individual has gathered, which is in turn shaped by their social background.

Idiolect

The consequence of applying this theory to authorship profiling is obvious; social groupings such as age, sex, ethnicity, social class (but also job, activities, ideologies and anything that can be used to classify social groupings) are revealed through language choices reflecting codal variation. In the authorship attribution task, codal variation, if pushed to the extreme, may also explain phenomena of the single cognition of an individual which are a function of the personality and habits of that individual in a certain context. Even two individuals that have a compatible social positioning (same age, sex, social class, education and so on) are nonetheless likely to differ because of the variation in the different interpretation of the context given by the different experience of the context that they had in their life.

Matthiessen (2007) describes the sum of a person's interpretations of all the possible contexts (that is, the sum of their codes) as their *personalised meaning potential* or, in other words, their *idiolect*. In SFL, the idiolect is therefore the path through the network of systems of a language that an individual habitually selects in a certain context because of their unique sociolinguistic background. An idiolect is therefore the sum of the codal variation and dialectal variation that an individual presents in every context. It is only measurable, however, by keeping the context constant, avoiding in this way the interference of registerial variation.

SFL, linguistic variation and the multidimensional analysis of language

At this point it is necessary to introduce another tool for calculating variation in language that is theoretically consistent with theories of SFL and codal variation: Biber's (1988) multidimensional approach.

The multidimensional approach is a methodology based on multivariate statistics used by Biber (1988) to study registerial variation in English. Biber (1988) and, subsequently, other corpus linguists, applied this methodology to study variation in the English language, especially to the study of registerial variation. Biber in particular pioneered the use of factor analysis to examine a number of linguistic variables extrapolated from a general corpus of English. The result he obtained was a set of six dimensions of variations that significantly separated the 23 genres of the English language considered, which included a vast sample of the English language, from conversation and speeches to newspapers, fiction and non-fiction books, letters, academic prose and official documents. Those dimensions create a six-dimensional space where texts can be located and for which the distribution of a number of

variables is known. It is therefore possible, having a text, to determine the scores for this text for each dimension and thus locate the text in the six-dimensional space. This will verify in which genre the text belongs or, in case we already know the genre of the text, it will allow an evaluation of how different a text is from the normalised score for its genre. Biber's results were replicated by other studies, even in other languages (e.g. Grieve *et al.* 2011; Biber 1995; Biber 2003).

The possibilities offered by Biber's multidimensional analysis are extremely powerful for forensic purposes. Knowing how a genre locates in a multidimensional space is equivalent to knowing the population base-rate of linguistic features for that genre. It is a step forward in solving an issue that forensic linguistics has had since the beginning: the issue of population-level distinctiveness (Grant, 2010; Grant, 2007: 22-23). Working on this method could lead to the possibility of calculating objectively the distinctiveness of a certain author for a certain genre by using a base-rate knowledge of what is typical for the text's genre.

The relevance of the multidimensional analysis in this paper resides in the fact that this method's theoretical assumptions are compatible with our discussion of *codal variation*. Finegan & Biber (1994) reach the same conclusions as Hasan when they claim that social groups differ in the kinds of communicative tasks that they can be competent with and that therefore they vary in the employment of different kinds of registers and in how they realise them with language. In fact, Finegan & Biber (2001)'s *register axiom* is indeed another way of formalising what is claimed in Hasan's *codal variation* theory, as both authors claim essentially that the kinds of registers a speaker has access to will influence the use of linguistic features in their social dialects.

Furthermore, saying that genres vary in relation to the co-occurrence of linguistic variables (frequency of past tense, number of verbs, nouns, nominalizations and so on) means saying that it is the probability of occurrence of a combination of choices in the systems of a language that determines register variation, which is what SFL claims. In the same way, if we keep this variation constant by keeping the genre constant, any other significant variation in this distribution of probabilities of occurrences or frequencies will be due to the social characteristics of the author, or, in SFL, to their code.

In this paper, Biber's method is interpreted as a reliable way of computationally calculating registerial and codal variation and it will be further explored in Trial 2 below.

A theory-based stylistic authorship analysis

Using this SFL understanding it is possible to develop a socio-cognitive based theory of stylistic authorship attribution situated within a modern linguistic theory. In this hypothetical method, McMenamin's fundamental concept of *choice* as distinctive of human behaviour would be taken into account, as *choice* is the fundamental concept in SFL. At the same time, this method would include and account for syntactic markers, such as those advocated by Chaski, and why these should vary between authors. In fact, we can think of the analysis of the frequency of occurrence of any of Chaski's syntactic markers as an instance of analysis of codal variation, since those syntactic markers represent different ways of creating meanings. Furthermore, the method also addresses an apparent weakness of stylistic approaches such as McMenamin (2002); the method solves the issue of replicability by providing a list of markers that has to be tested every time, thus avoiding this aspect of the analyst's bias. The analysis is firmly grounded in linguistic theory, allows for quantification and its underlying theory explains not only why there is idiolectal variation but also how this can be measured (as suggested by Howald (2009)).

Empirical exploration

The theoretical discussion above can only be transformed into a methodology after extensive testing and there is not scope for this in the current paper. We are, however, able to exemplify and extend the theoretical into the practical through the exploration of some data in brief trials. Although the goal of the present paper is to bridge the theoretical gap, it was felt that the points argued could be presented and explained more thoroughly by producing empirical trials, which can then be used in the future as pilot studies for further, more wide-scale experiments. We intend that it will be possible to see the practical dimension of the theoretical points previously elaborated and the following trials can be taken as a first step towards this elaboration of a methodology. In particular, this experimental dimension is important to show why the multidimensional analysis is needed and why just applying a SFL analysis may not be enough.

Dataset

The data for the trials was collected from three second year undergraduate students of a UK university and conducted with approval of the University ethics committee. Just as Chaski

(2001) selected demographically close participants, the participants in this study were selected from a wider pool having been matched for gender, age, education level, degree programme, and ethnicity. We were unable to also match for regional variation as Chaski did. Nonetheless, in the current study this was considered of less importance, as within the framework of SFL geographical origin is more likely to significantly affect dialectal variation rather than codal variation (Matthiessen, 2007: 538). The rationale for choosing people coming from a similar background is the same proposed by Chaski (2001: 4): an authorship attribution technique has to be tested in extreme conditions, that is, when the possible candidate authors come from very similar linguistic backgrounds.

Each participant contributed three academic assignments from which the first 300 words were chosen. As in Chaski's (2001) experiment, we wanted to test short texts and also control for their extra-linguistic contextual parameters, that is, their combination of Field, Tenor and Mode, also defined in SFL as Contextual Configuration (Halliday & Hasan, 1989). Controlling for the Contextual Configuration would guarantee that the texts are in a comparable contexts and that therefore only codal variation can be measured with little or no significant influence of registerial variation. Although the texts analysed are not texts on their own but fragments of longer texts, this is not problematic for the theory because what was chosen was the first 300 words of the "Introduction" section only. Although the Contextual Configuration of an academic assignment is relatively unexplored, it is still possible to conclude that the first 300 words of the Introduction section to an essay are produced within a comparable phase of the genre of academic assignment and can be therefore said to belong to the same Contextual Configuration.

Trial 1

Texts were analysed and tagged using *UAM CorpusTool* (O'Donnell, 2010), which allows the SF analyst to upload systems and manually tag the data to produce counts. The systems used for the analysis are the ones elaborated by the seminal work of Halliday & Matthiessen (2004) and Matthiessen (1995). A sample of the system network can be seen in Appendix 1.

A fragment of the result of the tagging appears in Table 1 below:

	T 5.txt	T 1.txt	T 9.txt	T 6.txt	T 2.txt	T 8.txt	T 7.txt	T 4.txt	T 3.txt
Feature	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**	**Mean**
Single	0.91	1.00	0.87	0.88	0.81	0.92	0.88	0.73	0.90
Multiple	0.09	0.00	0.13	0.12	0.19	0.08	0.12	0.27	0.10
Interpersonal-adjunct	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00
Textual-adjunct	1.00	0.00	1.00	1.00	1.00	1.00	0.50	0.83	1.00
Both-types	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00
Marked-theme	0.00	0.05	0.13	0.08	0.06	0.21	0.00	0.00	0.05
Unmarked-theme	1.00	0.95	0.87	0.92	0.94	0.79	1.00	1.00	0.95
As-theme-matter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
As-transitivity-role	0.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	1.00

Table 1 - A Fragment of the results for THEME

Table 1 above it is possible to see as an example nine variables taken from the system of THEME. Each variable is one option in a system and the values of those variables represent their percentage of selection. Rows for reciprocal variables such as *single* and *multiple* Theme, are grouped in the table. Therefore, for example, the option *single* has been selected 91% of the time in Text 5 (and *multiple* 9% of the time) and *single* is selected 100% of the time in Text 1. The complete analysis resulted in 506 variables. The first attempt at creating a methodology was tried by feeding these variables into a classificatory statistical model like discriminate function analysis, as suggested by Grant & Baker (2001), so to classify one text selected randomly from the ones available to the correct author. However, applying this methodology to this data set cannot give fruitful results as with this large amount of variables and little data, multivariate techniques such as PCA or classificatory models like discriminant function analysis cannot be applied. The proliferation of variables, as well as the limited data is therefore the first problem of a possible methodology based on codal variation. To side-step this problem, an ANOVA was run on the variable set to verify whether there were variables that would vary more between-authors than within-authors, thus at least confirming the hypotheses produced by the theory of codal variation exposed above.

An ANOVA was run for each system. Post-hoc tests and visual inspection of the error bars with 95% Confidence Limit determined the specific location of the between-author variation and the following results:

1. **Clause Complex**: no significant differences found;

2. **Conjunction**: no significant differences found;
3. **Modal Assessment**: no significant differences found;
4. **Mood**: no significant differences found;
5. **Nominal Group**: showed significant variation between authors under the following systems:
    - *demonstrative* vs *personal* (p = 0.049, discriminating Author1 from Author2)*;*
    - *inter-conscious* (p = 0.031, discriminating Author1 from Author2)*;*
    - *determinative* vs *interrogative* (p = 0.001, discriminating Author1 *vs* Author2 and Author3)
    - *ellipted* vs *non-ellipted* (p = 0.021, discriminating Author1 *vs* Author2 and Author3) [The couplets of variables with a *vs* come from the same system and are therefore the same variable, since the choice of, say, *ellipted* precludes the choice of *non-ellipted*].
6. **Theme:** no significant differences;
7. **Transitivity**: *relational* (p = 0.012, discriminating Author2 *vs* Author1 and Author3).

For each system where a quantitative difference was found the data can be explored in more detail.

The difference between *demonstrative* and *personal* in the nominal group resolves as the choice between using a determiner or a pronoun in the Deictic of a nominal group. In Figure 1 below examples of these features are presented. Quantitatively, A1 and A2 show differences in this respect whereas A3 cannot be discriminated from other two. A1 and A2 use a high level of demonstratives, as is typical of the genre. However, A2 has a higher use. The result is that A1's texts are less formal and more involving through a higher use of pronouns. A3, on the other hand, is not consistent in either style as it sometimes selects more demonstrative, sometimes more personal.

A1	A2
We intend to examine whether our perception of a celebrity changes when we are presented with their accent. That is to say, is Adrian Chiles better looking when he is not matched with his recognisable Brummie accent? Or is Cheryl Cole less likeable without her regional Geordie twang?	**The enforcers of the law** are effectively responsible for **the outcome of the final sentencing**, impacting heavily upon **the entire life of the defendant**, but linguistically they are also in a higher position of power; **the judge** and barristers are **the only people present in the court with the right to speak freely and to determine when others can do so**. **This** is essential to **the structure of the courtroom**, since **the content of what is said in any speech act** will always be dictated by **the person who is dominant in the questioning process**, **this** also being **the lawyers** and judge.

Figure 1 - Examples of A1 and A2 writings; 'personal' nominal groups are underlined, 'demonstrative' nominal groups are in **bold**.

Another variable that separates A1 from A2 is the system labelled *inter-conscious*. This counts the selection of non-plural conscious personal Deictics (*he, his, she, her*). Here A2 is distinctive; they never select this option, favouring other kinds of personal reference. A1, on the other hand, selects *inter-conscious* more frequently and A3, just as before, falls in between.

The variable *interrogative*, which is complementary to the variable *determinative*, counts nominal groups constituted by relative pronouns. The distinction between *determinative* or *interrogative* nominal group is critical to distinguish A1 from the other two. This is because A1 always selects *determinative* whenever possible, therefore with a significant consistency, whereas the other two authors select determinative very often but use some interrogative nominal groups too. In **Error! Reference source not found.**Figure 2 below, we can see that A2 and A3 present some interrogative nominal groups, whereas this is never the case in A1.

A2	A3
In addition to this, the language used is a variety which fulfils a different the purpose of conveying with precision the intricacies of the law, and as such this places those who are trained to fully understand this language in an immediate position of power and strength. It is the power balance, or rather imbalance, deriving from these factors **which** have been of great interest to those researching the language of the court.	However, P1 has also been living in London and Birmingham for the past 6 years and has thereby acquired a somewhat 'softened' accent. Participant 2 (P2) has lived in South Northamptonshire his whole life but has attended university within Birmingham for the past four years and has **what** could be considered an accent of Received Pronunciation.

Figure 2- Examples of A2 and A3 writings; 'determinative' nominal groups are in underlined, 'interrogative' nominal groups are in **bold**

A similar pattern is for *ellipted* vs. *non-ellipted*, that is, the variable that counts how many times an element of a nominal group or a nominal group complex has been ellipted. The analysis of this system showed that A1 selects an ellipsis of an element in nominal groups or coordinated nominal groups more often than the other two authors.

Finally, *relational*, an option of the TRANSITIVITY system, separates A1 and A3 from A2, who selects this choice significantly more often than the other possible options (*material*, *behavioural*, *mental*, *verbal* and *existential*).

A1	A2
Should this be the case, then surely the media has an effect on election results for Political Parties whether it is positive or negative? *This investigation will look at the political leanings of Britain's best-selling newspaper The Sun (BBC: 2009)* and how these are shown through photographs chosen for publication. Many studies have been carried out on election images found in U.S. newspapers *in order to establish whether they are politically biased, but, I've yet to find a comprehensive study for British newspapers*. **The closest is that carried out by William Miller (1991)** based on the 1987 General Election. *He looks at television/ press and the effect these had on the perceptions of the main political parties in the 1987 General Election (1991:11)*. **Miller chose his basis of results to be a consumer panel (1991:169)**. **This election was made up of four party leaders; Thatcher, Kinnock, Steel and Owen** and each were rated by the panel in terms of,	**A dialect is defined by the Longman Dictionary of Language Teaching and Applied Linguistics (2002) as a "variety of language, spoken in one part of a country (regional dialect) or by people belonging to a particular social class". It is possible to realise from this definition key elements of analysing a dialect; location and sociolinguistic factors, such as class. Language is "identity relevant under a social interpretation" (Hirschfield, 1998: 134)** and thereby shaped by these factors, **aspects of which are conveyed both conciously [*sic*] and subconsciously through a person's use of language. This can be exemplified by any variant of language**, although this research will primarily investigate phonology as the central feature of dialect.

Figure 3- Examples of A1 and A2 writings; Relational clauses are in **bold**, material clauses are in underlined, mental clauses are in *italics*; the other categories were left out from this example as they do not occur.

Using only the variable *relational* we can reasonably assign all of A2's texts to A2, because in all of them, at least 50% of the clauses is relational. The other authors range between 30% to 40%. However this observation about a single variable raises the wider crucial question of base rates. It is necessary to know how distinctive a feature is within a particular genre to understand how distinctive a writer is. That is, for example in this system (but the logic can be applied to any other variable): what is the percentage of relational clauses that one has to expect for academic writing? If this is about 30%, then A2 is distinctive and consistent for this feature and this can constitute evidence for the case. For this variable there is the additional problem

that transitivity analysis of a clause always presents a degree of subjectivity such that O'Donnell *et al.* (2008) conclude that it can be unreliable.

Trial1 – Discussion

The overall picture that emerges is that A1 and A2 each have a distinctive and internally consistent style. A3 is less consistent in their style across these texts which will make their texts more difficult to attribute correctly. We can also see how A1 and A2 are more different between themselves than other pairings.

The homogeneity in style across texts that has been found is as predicted by codal variation theory: the subjects share the same background and had experience with the genre, therefore they share most of the style. However, as hypothesised, even in this extreme case of homogeneity between texts and authors, there are significant differences that can be explained by idiolect variation or, in SFL terms, personalised meaning potential caused by personal codal variation. The historico-social explanation of why it is exactly those variables that vary significantly, and not others, cannot be answered here and perhaps not at all. The difference that we expect in terms of personal codal variation, according to codal variation theory, is generated by the different experiences that the individuals have of that genre, in this case, of the different readings and academic writing that they had done before, the different approaches to academic writing that they have been taught, etc.

However, it has been pointed out during the description of the trial that many problems would arise if a method based on this analysis had to be developed. First of all, we cannot determine why these particular variables vary significantly. One of the reasons might be the different topics of the essays, for example. A qualitative assessment of the variables is therefore a necessary step for the method. Equally fundamental is the possibility that quantification of how distinctive a variable be possible. This issue of the difference between the *expected variation* due to the context, compared with *observable variation* due to the code of the individual must be resolved if the method is to be applied. To achieve this, something similar to a base-rate knowledge of the variables that would form the basis of the method is needed. In addition to this problem, two other problems have arisen: the issue of reliability of subjective coding, as noted for transitivity analysis, and the sheer number of system variables. When a large number of variables has to be analysed, the issue is not just one of time or effort. The statistical methods used to classify texts can work only if the total number of variables is low.

Using few variables becomes therefore a necessity. These three issues are addressed and partially resolved by applying Biber's multidimensional analyses, as shown in Trial 2.

Trial 2

After Trial 1, it is clear that just applying a SFL analysis on forensic data of the kind similar to the ones analysed can be problematic. We encountered specifically three problems: (1) base-rate knowledge: the base-rate distribution of the SFL variables is not known and in this way it is difficult to demonstrate distinctiveness; (2) proliferation of variables: there are too many variables in a SFL analyses and typically in a forensic scenario the data is scarce. This means that the ratio variables/cases is not in favour of classificatory statistical techniques or multivariate statistics; (3) SFL analysis is at times too subjective to be reliable for forensic purposes (e.g. transitivity). If we had an extensive and comprehensive corpus of the English language automatically tagged with a SFL parser we would solve all of these problems, as we would know the base-rate distribution of all the variables by genre and at the same time this analysis would be based on computer algorithms that are reliable to reproduce. Unfortunately, such an endeavour is not available. It is for this reason that Biber's Multidimensional Analysis has been chosen (Biber, 1998).

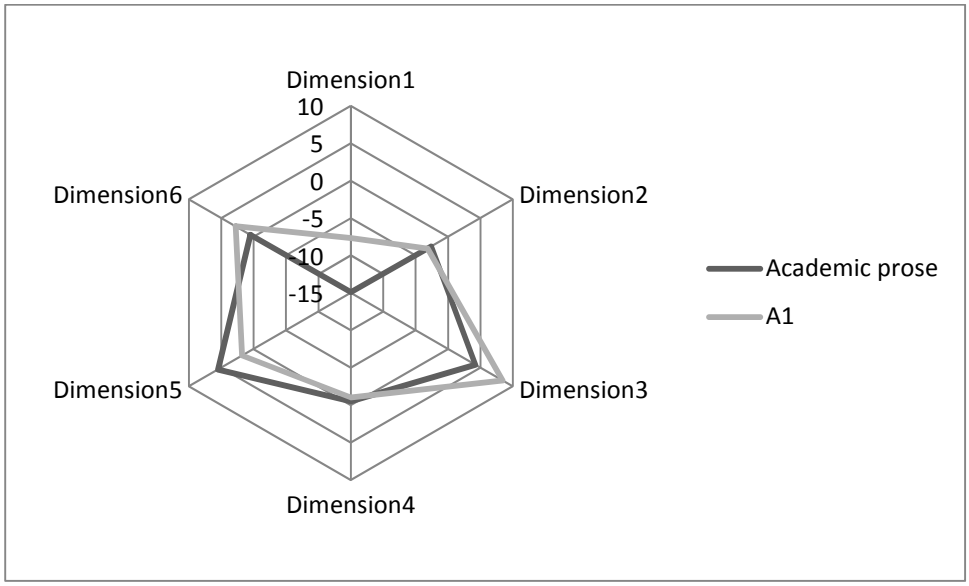
As previously introduced, Biber selected 67 linguistic variables from a survey of previous studies and fed them in a factor analysis to determine how these variables covary and can distinguish the main genres of the English language that he considered. Using his work we therefore have at our disposal a computerised set of algorithms to compute frequencies of variables and we have the distribution of these variables for a large set of genres. If we assume that these 67 variables are an approximation to a full SFL analysis then we have a way to solve the three problems encountered. First of all, the framework solves subjectivity of analysis, since Biber used algorithms that would determine automatically certain variables using a computer programme. Additionally, it solves the problem of the proliferation of variables as well, since the variables analysed are only the six factors that take into account the covariation of the 67 grammatical variables. Finally, as mentioned above, it is reasonable to assume that Biber's method is theoretically compatible with SFL and *codal variation*.

Trial 2

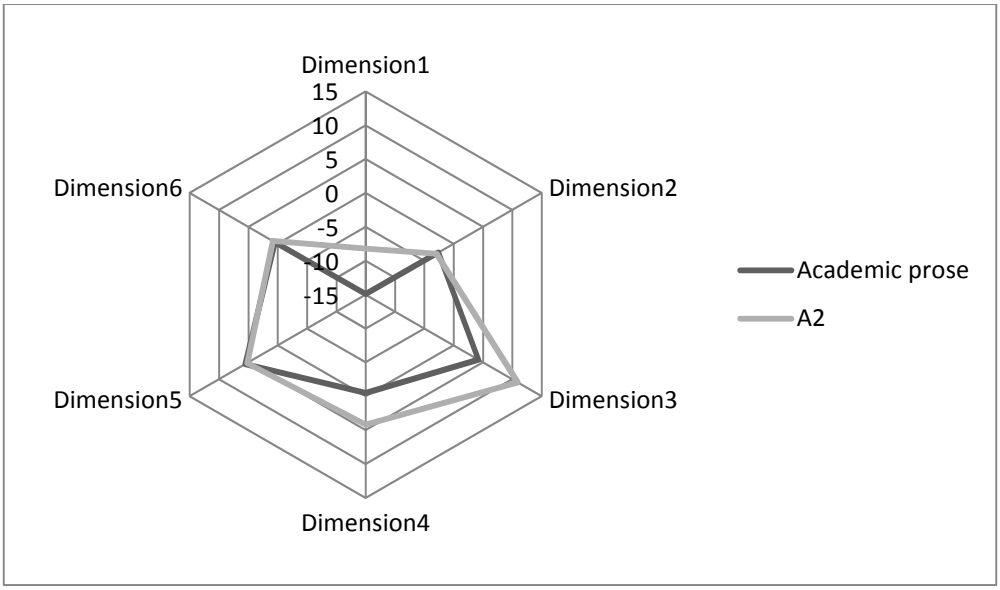
A second trial, Trial 2, was run so to use Biber's Dimensions as variables. To calculate the Dimensions, the first step was to calculate the 67 linguistic variables that Biber used. These

variables were collected by Biber from previous studies and represent linguistic features that are known to vary between genres. Examples of these are: frequencies of pronouns, frequencies of nouns, or frequencies of relative clauses (the complete list can be found in the appendix of Biber (1988)). The texts have been processed using the CLAWS part of speech tagger, which automatically tags words for parts of speech. Each of the 67 variables used by Biber (1988) was then calculated semi-automatically using WordSmith (Scott, 2011), carefully following the precise algorithms given in Biber's (1988) appendix. Once the values for each of the 67 variables were obtained for each text, the factor scores for each text for each Dimension were calculated using the guidelines provided again by Biber (1988: 93).

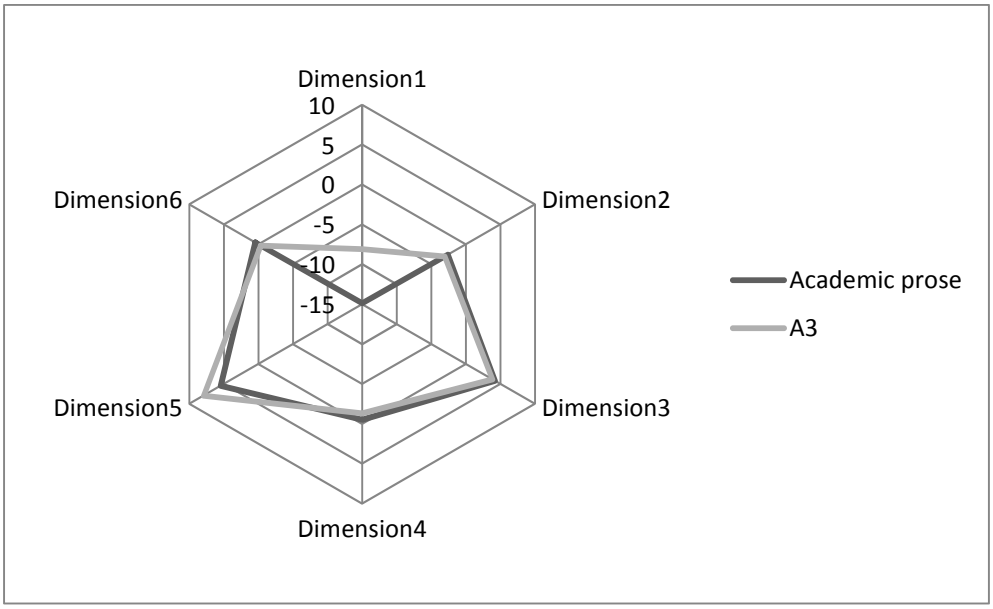
The result of the analysis was a set of six variables for each text for each author. The six variables are the factor scores for each Dimension. The radar graphs reproduced in Graph **3**, 2, 3 and 4 below show Biber's scores for the genre academic prose, and also the mean scores for each of the authors.



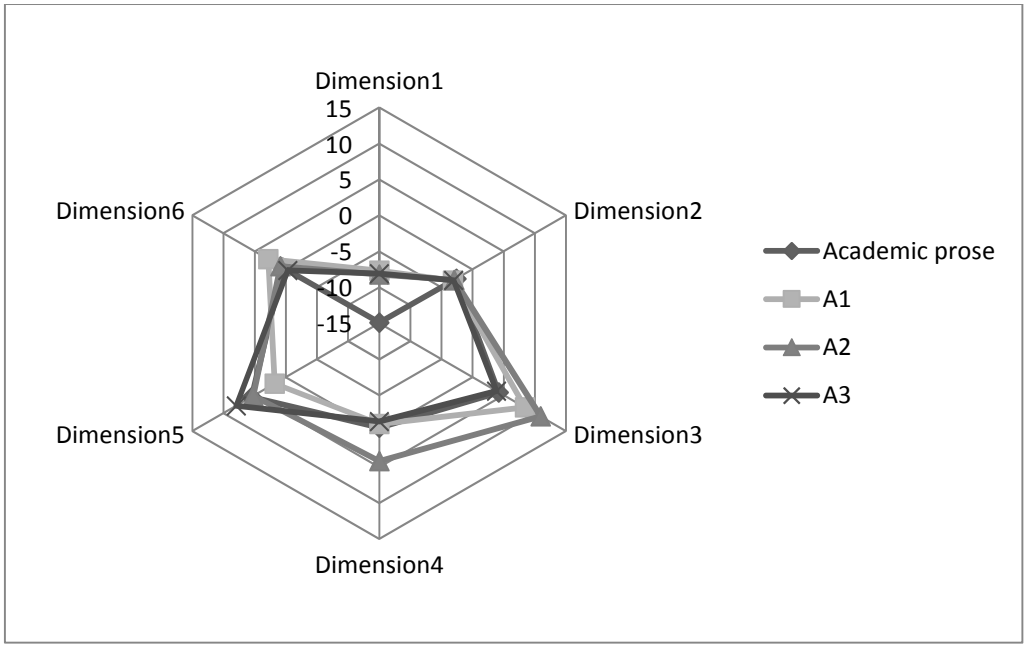
Graph 1 - Dimension scores for 'Academic prose' (as found by Biber (1988)) and for A1



Graph 2 - Dimension scores for 'Academic prose' (as found by Biber (1988)) and for A2



Graph 3 - Dimension scores for 'Academic prose' (as found by Biber (1988)) and for A3



Graph 4 – Dimension scores for 'Academic prose' (as found by Biber (1988)) and for each author

As the graphs suggest, the only considerable difference found was related to Dimension 1, the dimension that measures the involvedness/informational character of a text. More specifically, the difference exists between the authors' scores and Biber's (1988) average score for Dimension 1 for academic prose, with the authors being more 'involved' than expected. The authors' values clustering around -8 are distinctive for the genre, as a much lower score was expected (around -15). This difference can be observed qualitatively in Figure 4. The examples reproduced in this figure show only four of the variables accounted for by Dimension 1: nouns and attributive adjectives for the Informational pole; pronoun *it* and demonstrative pronouns for the Involved pole. These four variables were chosen as they were the most frequent ones in the samples. Figure 4 shows qualitatively that the sample from A2 is less nominal and information packaged and at the same time more contextually dependent on the knowledge of the reader, as it uses more pronominal forms. Although it could be objected that the sample in Figure 4 is taken from an engineering academic text, which would thus seem to be more informational than a typical academic text from the humanities such as A2's text, Biber (1988) found that the difference between the two sub-types of academic prose genres is rather small. Biber (1988: 185) found that the difference of the averages of natural science academic prose and humanities academic prose amounts to roughly five, with natural science academic prose being therefore just slightly more informational than the humanities one.

It is possible to conclude, therefore, that the authors in this experiment are consistent and distinctive in their style of writing in the academic prose genre. Conclusions about the

between-author difference is not possible as more data would be needed and in these case, as with much forensic linguistic casework, the conclusion to be drawn is that attribution is not possible.

Example from Biber's corpus of academic texts	A2
It follows that the **performance** of **down-draught systems** can be improved by the **influence** of **cross draughts** only if the ***thermal*** **currents** are blown into **exhaust air streams** at ***higher*** **velocities** than the **cross draughts**, so that the ***resultant*** **direction** of all ***dust-bearing*** **air streams** is towards the **grid** […] The **exhaust air volume** required by the ***6-ft. x 4-ft.*** **grid** with the ***8-in. deep hot*** and ***cold*** **moulds** and the ***16-in. deep cold*** **moulds** tested in the **absence** of ***appreciable*** **cross draughts** exceeded the **volumes** required by the ***4-ft. 6-in. x 3-ft 6-in.*** **grid** by between 25 and 40 per cent.	Leicester is a ***demonstrative*** **example** of *this*; whilst it can generally be described as having many **elements** of a ***typically*** **East Midlands accent** (Hughes, Trudgill and Watt, 2005), it has a ***high ethnic*** **population** with almost 30% of the **population** being of ***Asian*** **origin**. *This* has risen considerably within the last 10 **years** and ranks **Leicester** as having the ***highest Indian*** **population** of any ***local*** **authority area** in **England** and **Wales** (http://www.leicester.gov.uk). *This* can be explained by the **fact** that **Leicester** received over 20,000 displaced **East *African*** **Asians** when they were expelled from **Uganda** and **Kenya** in the 1970's, more than anywhere else in the **UK** (Panesar, 2005).

Figure 4 - Comparison of main feature of D1 between an example of Biber's corpus of academic texts and a sample of A2's text; Informational features: Nouns are in **bold**, attributive adjectives are in ***italicised bold***; Involved features: pronouns are underlined, demonstrative pronouns are *italicised and underlined*.

Although attribution could not be reliably assigned because of paucity of data, it is still possible to observe from this example how the method can be used to gather pieces of evidence to show distinctiveness of authorial style compared to a genre. Once the full multidimensional analysis is completed and the analyst knows the value for each of the 67 variables, it is possible to show distinctiveness by using the base-rate knowledge for the genres that Biber studied.

In addition to this application to comparative authorship analysis it is possible to note how promising this method could also be for authorship profiling. By looking at the typical scores that one expects from the genre, the analyst can objectively conclude that all these texts are more involved than expected, that is, they have an anomalously high score on Dimension 1. The statistical difference ought to be checked qualitatively in order to avoid to draw the

wrong inferences from the data. An anomalous score can be given by several facts and it is up to the analyst to assess the difference and express an expert opinion on what causes it. In this case, since we know that the texts analysed are sampled from the academic prose genre, a qualitative exploration of the data confirms that the most likely explanation for the highly Involved score of these texts is that the authors are not competent writers in that genre. Furthermore, this involvedness character of student writing has actually been found before by Grabe & Biber, who conclude that '[student essays] use the surface forms of academic writing (e.g. passives), but they are relatively non-informational and involved' (Grabe & Biber 1987, quoted in Biber 1988: 204). This conclusion corroborated by previous literature is a claim that involves the students' code: we can conclude that the author's experience with the genre is not that of a master of academic writing; indeed as second year undergraduate students it turns out they are all academic apprentices. As well as these case-specific conclusions, it is possible that similar reasoning around codal variation measured in this way offers the possibility of deducing a writer's gender or age. This is already suggested by Biber's work (Biber *et al.*, 1998) and theorised by Martin within the SFL framework (Martin, 1992: 578).

The problem of genre comparability

Biber's (1988) framework and SFL share a similar model of the concept of genre. This is interpreted by both as being a configuration of situational parameters, such as channel, extent of shared space, relationship between interactants and so on, or, in SFL terms, combinations and interactions across Field, Tenor and Mode (Biber, 1988: 39). Theoretically speaking, this interpretation of genre can help to solve a problem often found in forensic authorship analysis: the genre incompatibility between the texts that have to be analysed.

If a method based on codal variation theory as presented in this paper is to be applied, this can be done only when the texts taken into account are produced in comparable contexts, or, in other words, in comparable genres. However, when considering application to forensic casework one pressing issue for the analyst is cross genre comparison as the forensic linguist may have to work with, say, a threatening letter with a business letter for comparison. The question arises as to what differences in style between the two texts are accounted for by genre rather than codal differences. The way to resolve this issue can come again from Biber's work, as compatibility of contextual configuration can be measured using multidimensional analysis. If, for example, there is a significant difference in the scores of business letters *vs.* threatening letters, then the measurement of codal variation cannot be done. However, to do this, one has

to expand Biber's analysis to cover genres that have not been covered by his study, and this requires further empirical work.

Where it is more certain that two texts are indeed from incompatible genres (e.g. a diary and a threatening letter) the method suggests a possible way forward. In SFL, context is broken up into three variables: Field ('what is happening [...] the nature of the social action that is taking place'), Tenor ('who is taking part, the nature of the participants, their statuses and roles') and Mode ('what part language is playing') (Martin & Rose, 2002). Each of these variables affects *only* certain linguistic variables. For the sake of simplification, we can say that Field affects just the content words of a text, Tenor the number of declaratives/interrogatives/imperatives and the expression of attitudes, Mode affects cohesion and the Theme-Rheme patterns. Therefore, if two genres for which Field and Tenor vary but for which Mode is comparable are considered, then the measurement of the code of the author can be produced by only looking at those variables for which how much variation we expect for the genre is known, that is, cohesion and Theme-Rheme patterns. This hypothesis in SFL is known as the "Context-Metafunction Hook-up Hypothesis", which proposes that the contextual characterisation of a genre can always be broken down into the simple sum of Field, Tenor and Mode (Hasan, 1995). Confirmation of this hypothesis is being currently pursued within SFL. Although some preliminary studies showed a more conservative position (Thompson, 1999), recent empirical work indicates that there is indeed a correlation between contextual parameters and linguistic variables, at least for Mode (Clarke, 2013). It is therefore possible to expand this work by conceiving empirical tests pertinent to forensic cross genre attribution problems.

Conclusions

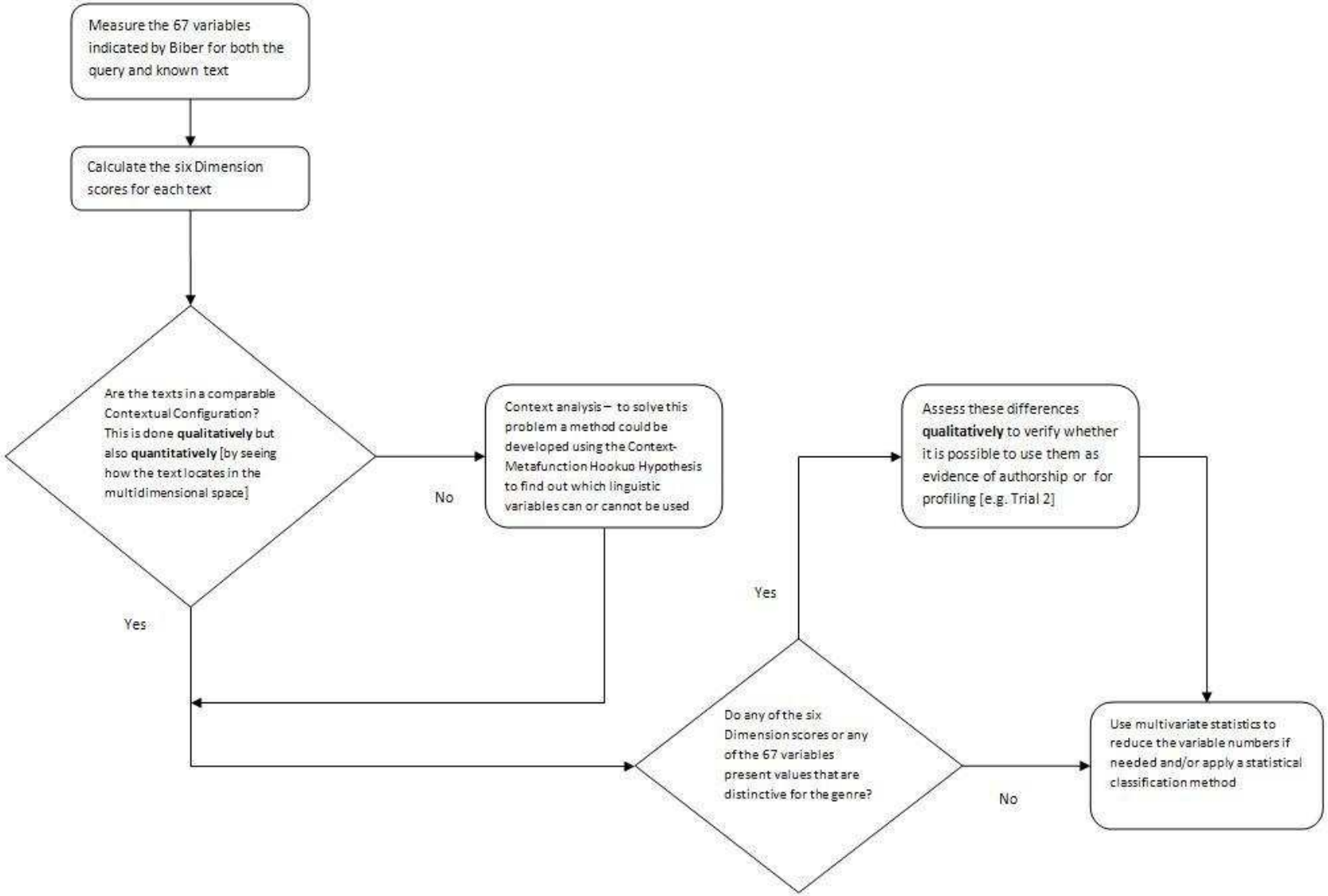
This paper started with the argument between stylistic and the cognitive approaches to forensic authorship analysis and their associated strengths and weaknesses. A theoretical discussion showed that these two methods are not in opposition but that can be seen as complementary if we observe them through the lenses of SFL codal variation. The summary of the theoretical points is given in Table 2 below:

Cognitive approach	Stylistic approach	SFL – codal variation theory
Theory of idiolect is fundamental.	Theory of idiolect is not necessary.	Individual linguistic variation is theorised.
The individual is unique because of cognitive differences.	The individual is unique because of stylistic choices.	The individual is unique because of stylistic choices given by their code, which in turn is influenced by cognitive differences.
Analysis of syntactic structures.	Analysis of style markers.	Since in SFL the syntactic structures can be imagined as the results of choices in the systems, and since the systems are network of choices (therefore, style markers), those two positions of analysis coincide.
Variability caused by differences in cognitive structures.	Variability caused by socio-historical differences.	Codal variation theory points to a view of the mind as personalised brain shaped by the context: it therefore provides a theoretical stance that reconciles the two kinds of variability presupposed by the two approaches

Table 2 - Summary of the characteristics of the three theoretical stances

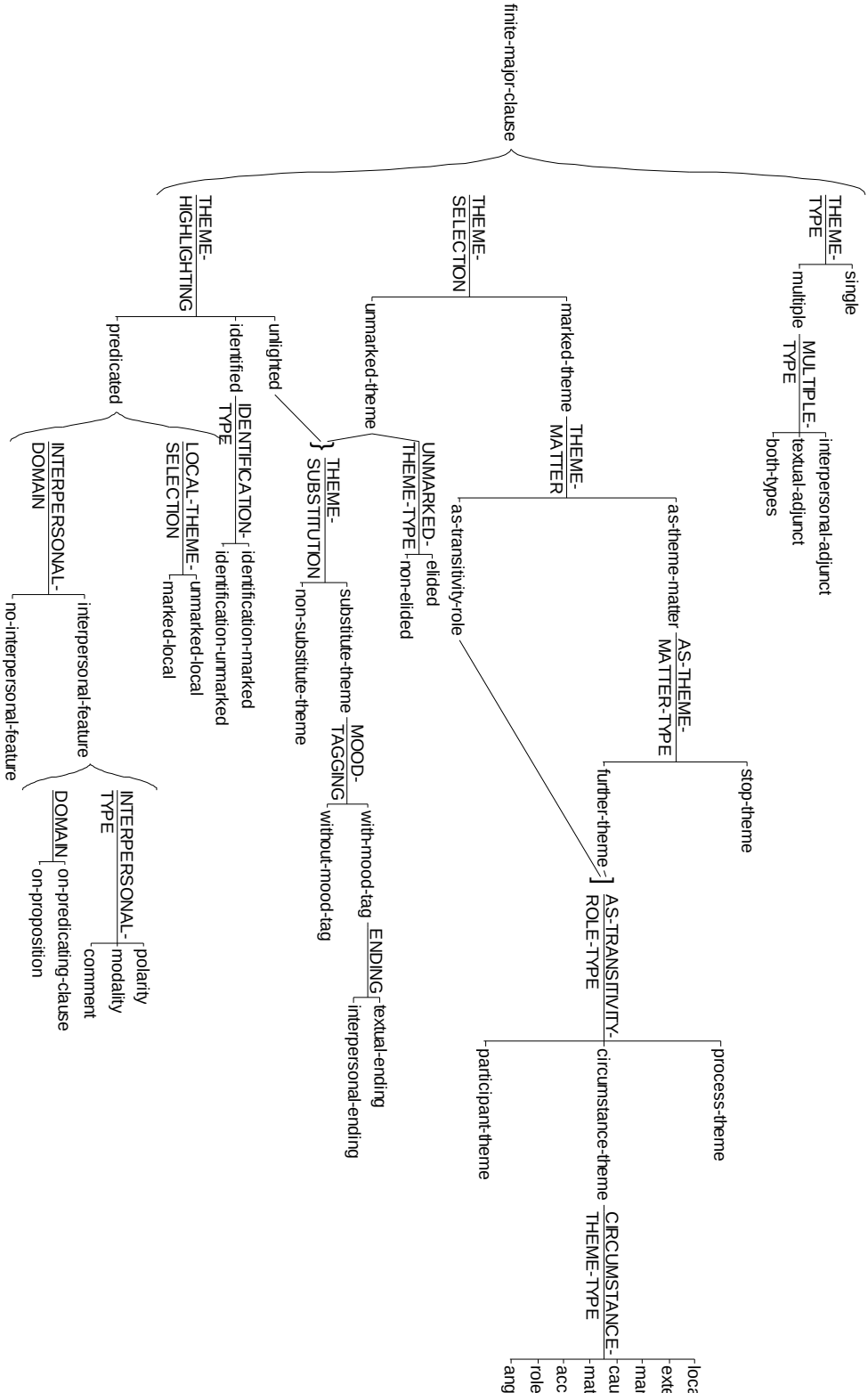
After the theoretical discussion, however, a pilot study showed that SFL in itself is not enough to support a new method of authorship attribution. Biber's multidimensional analysis was then proposed so to solve the issues raised by Trial 1. With Trial 2 it has been shown that the application of Biber's multidimensional analysis combined with the SFL theory of codal variation can be a way to provide a theory to the stylistic approach. If this methodology of doing forensic stylistics proves valid with further experimentation and application, it could be potentially showed that a codal variation theoretically motivated forensic stylistics indeed coincides with cognitivists' syntactic approach, although interpreted through the lenses of stylistics and sociolinguistics. Finally, this kind of analysis provides forensic stylistics with methods that resolve potential biases in replication and scientific method.

To conclude, we provide here a description of the SFL-MD theoretically based method of authorship attribution that we propose for further studies. The procedure is described in Figure 5 below using a flowchart.


Measure the 67 variables indicated by Biber for both the query and known text
Calculate the six Dimension scores for each text
Are the texts in a comparable Contextual Configuration? This is done **qualitatively** but also **quantitatively** [by seeing how the text locates in the multidimensional space]
No
Context analysis – to solve this problem a method could be developed using the Context-Metafunction Hookup Hypothesis to find out which linguistic variables can or cannot be used
Yes
Assess these differences **qualitatively** to verify whether it is possible to use them as evidence of authorship or for profiling [e.g. Trial 2]
Yes
Do any of the six Dimension scores or any of the 67 variables present values that are distinctive for the genre?
No
Use multivariate statistics to reduce the variable numbers if needed and/or apply a statistical classification method

Further experiments are carried out at the moment to improve this methodology and to find out the qualitative significance of the variables used by Biber for the purpose of authorship profiling and attribution.

Appendix 1 - Sample of system network: the system of THEME



References

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press;

Biber, D., 1995. *Dimensions of Register Variation: a Cross-Linguistic Comparison*, Cambridge; New York: Cambridge University Press;

Biber, D., 2003. 'Variation among university spoken and written registers: A new multi-dimensional analysis'. *Language and Computers*, Vol. 46(1), pp.47–70;

Biber, D., Conrad, S., & Reppen, R. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press;

Chaski, C. 2001. 'Empirical evaluations of language-based author identification techniques'. *International Journal of Speech, Language and the Law*, Vol. 8, pp 1-64;

Clarke, B. P. 2013. 'The differential patterned occurrence of ellipsis in texts varied for contextual mode. Some support for the 'mode of discourse – textual metafunction' hook-up'. In G. O'Grady, L. Fontaine & T. Bartlett (eds) *Choice in Language: Applications in Text Analysis*. Equinox;

Coulthard, M. & Johnson, A. 2007. *An Introduction to Forensic Linguistics*. London: Routledge;

Dror, I. E., Charlton, D., Peron, A. 2006. 'Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications'. *Forensic Science International*, Vol. 156, pp 74-78;

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., Rosenthal, R. 2011. 'Cognitive issues in fingerprint analysis: Inter- and intra-expert consistency and the effect of a "target" comparison'. *Forensic Science International*, Vol. 208, pp 10-17;

Eggins, S. 2004. *An Introduction to Systemic Functional Linguistics*. New York/London: Continuum;

Finegan, E. & Biber, D. 1994. 'Register and social dialect variation: An integrated approach'. In D. Biber & E. Finegan (eds) *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press, pp 315-347;

Finegan, E. & Biber, D., 2001. 'Register variation and social dialect variation: The register axiom'. In P. Eckert & J. R. Rickford (eds) *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press, pp. 235-267;

Grant, T. 2007. 'Quantifying evidence in forensic authorship analysis'. *International Journal of Speech Language and the Law*, Vol. 14(1), pp.1-25;

Grant, T. 2010. ‘Txt 4n6: Idiolect free authorship analysis’ in M. Coulthard (ed.), *The Routledge Handbook of Forensic Linguistics*. London: Routledge, pp. 508–523.

Grant, T. & Baker, K. 2001. ‘Identifying reliable, valid markers of authorship: a response to Chaski’. *Forensic Linguistics*, Vol. 8(1), pp.66–79;

Grieve, J. 2007. ‘Quantitative authorship attribution: An evaluation of techniques’. *Literary and Languistic Computing*, Vol. 22(3), pp 251-270;

Grieve, J., Biber, D. & Friginal, E., 2011. ‘Variation Among Blogs: A Multi-dimensional Analysis’. *Genres on the Web*, Vol. 42, pp.303-322;

Halliday, M. A. K. & Hasan, R. 1989. *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford: Oxford University Press;

Halliday, M. A. K. & Matthiessen, C. 2004. *An Introduction to Functional Grammar*. London: Arnold;

Hasan, R. 1990. ‘A sociolinguistic interpretation of everyday talk between mothers and children’. In J. Webster (ed.) *The Collected Works of Ruqaiya Hasan Vol. 2: Semantic Variation: Meaning in Society and in Sociolinguistics*. London: Equinox Publishing, pp 73-118;

Hasan, R. 1995. 'The conception of context in text'. In P. Fries & M. Gregory (eds) *Discourse in Society: Functional Perspectives*. Norwood, N.J.: Ablex, pp 183-283;

Hasan, R. 2009. ‘On semantic variation’. In J. Webster (ed.) *The Collected Works of Ruqaiya Hasan Vol. 2: Semantic Variation: Meaning in Society and in Sociolinguistics*. London: Equinox Publishing, pp 41-72;

Howald, B. 2009. ‘Authorship attribution under the rules of evidence: Empirical approaches in a layperson’s legal system’. *International Journal of Speech, Language and the Law*, Vol. 15, pp 219-247;

Labov, W. 1966. 'The linguistic variable as a structural unit'. *Washington Linguistic Review* 3, pp 4-22;

Lamb, S. 2013, forthcoming. ‘Systemic networks, relational networks, and choice’. In L. Fontaine, T. Bartlett & G. O’Grady (eds) *Systemic Functional Linguistics: Exploring Choice*, Cambridge: Cambridge University Press;

Martin, J. R. 1992. *English Text: System and Structure*. Philadelphia: John Benjamins;

Martin, J. R. & Rose, D. 2002. *Working with Discourse: Meaning beyond the Clause*. London: Continuum;

Matthiessen, C. 1995. *Lexicogrammatical Cartography: English Systems*. Tokyo: International Language Sciences;

Matthiessen, C. 2007. ‘The “architecture” of language according to systemic functional theory: developments since the 1970s’. In R. Hasan, C. Matthiessen & J. Webster (eds) *Continuing Discourse on Language*, Vol. 2. London: Equinox, pp 505-61;

McMenamin, G. R. 2001. ‘Style markers in authorship studies’. *International Journal of Speech, Language and the Law*, Vol. 8, pp 93-97;

McMenamin, G. R. 2002. *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton, Fl/London: CRC;

O'Donnell, M. 2010. *UAM CorpusTool, Version 2.6*;

O’Donnell, M., Zappavigna, M. & Whitelaw, C. 2008. 'A survey of process type classification over difficult cases", in C. Jones & E. Ventola (eds) *From Language to Multimodality: New Developments in the Study of Ideational Meaning*. London: Continuum;

Scott, M. 2011. *WordSmith Tools, Version 5.0*;

Stamatatos, E. 2009. ‘A survey of modern authorship attribution methods’. *Journal of the American Society for Information Science and Technology*, Vol. 60(3), pp 538-556;

Thompson, G. 1999. 'Acting the part: lexico-grammatical choices and contextual factors'. In M. Ghadessy (ed.) *Text and Context in Functional Linguistics*. Amsterdam: Benjamins, pp 103-126;

Vygotsky, L. S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Edited by M. Cole, V. John-Steiner, S. Scribner & E. Souberman. Cambridge, MA: Harvard University Press.