Codal variation theory as a forensic tool

Andrea Nini [*]

Abstract. *This paper addresses how the Systemic Functional Linguistics (SFL) theoretical framework of* codal variation *(Hasan, 1990) can be helpful when applied to the field of forensic linguistics, especially for authorship analysis. Firstly, the framework will be introduced and discussed in the lights of traditional modern sociolinguistics. Then, it will be shown how the concept of* codal variation *can be useful for describing and understanding the idiolect, or, in SFL terms, the personalised meaning potential of an individual. An example of a successful application of this concept will be taken from the Bentley case, where the distinction between two codes proved to be of high evidential value. The discussion will then lead on to the implications that* codal variation *could have for authorship/sociolinguistic profiling, considering other examples from the literature for which an SFL interpretation could lead to an improvement. Combining the theory of* codal variation *with Biber's multidimensional framework can represent a first step towards building a method of authorship analysis that is driven by the knowledge of population base-rate of a number of linguistic variables. An example from a real case will be presented where this kind of analysis proved to be a promising first step towards a theoretically valid methodology for authorship analysis. Possible improvements and directions for more research will be illustrated and discussed.*

Keywords: Forensic linguistics, authorship analysis, authorship attribution, systemic functional linguistics, multidimensional analysis, codal variation, sociolinguistics, corpus linguistics, Bentley case.

Introduction

This paper will present the implications that *codal variation theory*, a particular branch of Systemic Functional Linguistics (SFL), can have for authorship analysis. Authorship analysis is currently dominated by two strands of methodologies: the qualitative stylistic methods and the quantitative stylometric methods. A major problem with these approaches is that they do not address the issue of providing a valid explanation of why

[*] Aston University

authorship analysis is possible. This paper proposes the theory of codal variation as a way of introducing a strong theoretical framework for authorship analysis, thus potentially solving the problem of theoretical validity (Grant and Baker, 2001). By providing a hypothesis of why authorship analysis is possible, new theories can be tested and better practice can be developed.

Semantics in SFL

Before introducing the discussion on codal variation a brief digression is needed on the model of semantics in SFL. This is necessary as many concepts introduced below depend on a particular understanding of semantics, which is marginally different from the mainstream sociolinguistic definition of the term.

SFL models the meaning of a lexicogrammar item as being the function that it serves. That being so, SFL does not consider meaning to be truth-conditional, as traditionally conceived by generativists and sociolinguists. For example, the clauses:

a) *Mary eats the apple*

b) *Is Mary eating the apple?*

c) *The apple is eaten by Mary*

in SFL do not mean the same thing. In this framework, meaning corresponds to function and function is in turn organised in a three folded division in major functional strands, or *metafunctions*: these are the *ideational metafunction*, that is, the function of language to represent things and events; the *interpersonal metafunction*, that is, the function of language to communicate interactions or to make people interact; the *textual metafunction*, that is, the function of language of distributing the information so to anchor the text to the context (Halliday and Matthiessen, 2004).

Hence, the clauses presented above are constant in their *ideational* meaning, since the same Agent (Mary) is acting in the same Process (eat) to the same Goal (apple), but do present variation at the *interpersonal* and *textual* meanings. Clause (b), as opposed to clause (a) and (c), is a yes-no question and it therefore realises the meaning of being a request for information, thus situating the speaker as the person who seeks information and the hearer as the person who is assumed by the speaker as having the information. Clause (c), as opposed to clause (a) and (b), is a passive clause which makes the Goal (apple) the starting point of the clause. In clause (c) the speaker thus expresses a *textual* meaning that consists in the assumption that the hearer knows something about an apple and that it is likely to be a new information that it is Mary who eats it (this explanation is rather simplified for reasons of space; Halliday and Matthiessen 2004: 64).

Although traditionally a division is adopted between *meaning* and *style*, where *meaning* corresponds to the *ideational metafunction* and *style* corresponds to the combination of the *interpersonal* and *textual metafunctions*, for the rest of the paper, when the term meaning is employed, it is generally adopted to intend the three metafunctions.

Codal variation

It is possible to explain codal variation by comparing how linguistic variation is modelled in traditional linguistics and in SFL. Two kinds of variation are widely recognised in both SFL and traditional modern linguistics:

1. *Situational variation* (or *registerial variation* in SFL): the variation of meanings that is found in texts produced in different contexts.

2. *Social variation* (or *dialectal variation* in SFL): the variation in the way of realising meanings, which originates by the fact that different social groups have alternative ways of expressing the same meanings.

For the sake of further explanation two examples can be used to illustrate this point: (1) *registerial variation* is the variation in the frequency of past tenses between a story and an academic paper, a variation given by the different contexts in which the writer operates and thus independent on the person who is writing; (2) *dialectal variation* is the variation in the ways of making the same meanings within different social groups, such as, for example: for *ideational meaning*: 'pail' vs. 'bucket'; or for *interpersonal meaning*: different ways of realising a tag question: 'isn't it?' vs. 'innit?'.

The assumption underlying the postulation of these two kinds of variation is that there are two causes for the variation (context and social group) and two areas in which the variation appears (meanings and realisations of the meanings). Moreover, another assumption of the models is that there can be only two combinations: context creates variation in meanings *whereas* social group creates variation in the realisations of the meanings. That is, if the context varies then there is production of different meanings but not of different realisations of meanings and, consequently, if the social group varies then there is production of different realisations of meanings but not of different meanings.

SFL, however, adds another level of variation by considering another possible combination between these parameters: social group *to* meanings. This is *codal variation*:

3. *Codal variation*: the variation of meaning in relation to the social group, when the texts considered are produced in a comparable context.

The concept of codal variation was originally developed by Hasan (1990) under the term *semantic variation.* The term was later on reframed by Matthiessen (2007) as *codal variation* to avoid ambiguity with other interpretations of the term 'semantic variation'.

The term *codal* originates from the main source of theoretical influence of this concept, Bernstein's concepts of *restricted* and *elaborated codes.* Bernstein (1962) proposed that social classes produce different meanings in similar context because of the way social classes interpret the context. In his original proposal, he claimed that this leads to individuals coming from a working class background producing different grammatical structures from individuals coming from a middle class background. Bernstein's work was extremely controversial and generated a debate with other approaches, in particular with Labovian variationist sociolinguistics (Martin, 1992: 573). The debate was generated by Bernstein's suggestion that people from different social classes produce different meanings. Such a claim was not warmly welcomed in a mainstream linguistics that at the time professed that semantics was universal and based on the truth-conditional meaning. The issues at stake in the debate can be exemplified by discussing Figure 1 below.

Figure 1 represents a series of *realisations.* The table has to be read as each layer being realised by the one below it. Therefore: context is realised by semantics, which in turn is realised by lexicogrammar, which in turn is realised by a form of expression. Although this table represents a model of language typically found almost only in SFL, certain assumptions underlying the table are indeed shared by traditional modern sociolinguistics as well. Indeed, the theoretical stance underlying this table can be also found in the concept of *sociolinguistic variable.*

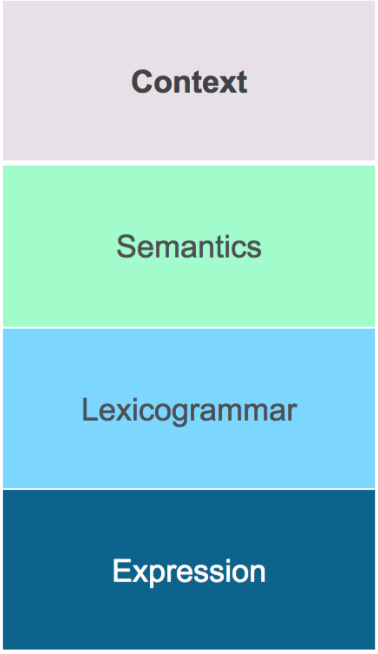


Figure 1. Levels of language according to SFL.

It is possible to reformulate the definition of a sociolinguistic variable as a socially distributed linguistic variable that measures variation at one level of language by keeping constant the level that is realised by it (or, diagrammatically speaking, the level above it). For example, a variation at expression level, such as the often studied *g clipping*, can be studied only in cases where the lexis (or lexicogrammar, in SFL terms) is constant (e.g. *singing* vs *singin'* or, grammatically speaking, *present continuous realised phonetically by* [ŋ] vs *present continuous realised phonetically by* [n]). For the same reason, lexicogrammatical variation, in the form of either lexical alternations or syntactic/morphological alternations, can be studied only when the semantics is kept constant, that is, when you have 'different ways of saying the same thing' (e.g. *car* vs *automobile*; copula deletion: *he is working* vs *he working*).

It follows that it is indeed possible to model semantic variation *only* if something can be held as constant above the semantics. In SFL, this something is the context, which is modelled as *a level of language* that is realised by the semantics. Traditional Labovian sociolinguistics lacked a model of context. However, as Hasan (2009) pointed out, a model of context is necessary to create an integrated sociolinguistic theory. The level of context on top of semantics is not just an addition to the face structure of the model. The theoretical significance of modelling an additional layer is indeed that semantics is not constrained to be the truth-conditional universal that generativists presuppose but that it is another level of language as a whole. That being so, as all the other levels of language, this level can vary sociolinguistically and it is arbitrary in the same way as lexicogrammar. According to Hasan (2009), this understanding of meaning allows the development of a new kind of sociolinguistics that takes into account how meanings together with form vary across social groups.

Hasan's (1990) work represents an empirical demonstration of this concept. Her experiment showed that the variation at the level of semantics is correlated with social groups in the same way as the variation of Labovian sociolinguistic variables at the other levels are correlated with social groups. In her study, Hasan recorded samples of Australian working class and middle class mothers in the process of talking and playing at home with their children. She then analysed these samples using SFL and used Principal Component Analysis to group her variables. The components obtained can be thought of as semantic styles in the context of mother-child home conversation in Australia. After assigning component scores to the dyads mother-child, an ANOVA revealed that the differences between the two social classes were significant. Hasan interpreted the results as pointing to the fact that the speech of mothers talking to their children is influenced by the family's social positioning in terms of social class.

In other words, Hasan's experiment is evidence that different social classes operated in the same context in different ways and therefore produced different meanings when dealing with the context of regulating children's behaviour. These different ways of dealing with the context point to a difference in how the context is interpreted and understood by the two social groups.

Similar studies were then reproduced in the SFL community by other researchers (Martin (1992: 578) cites many; Hasan (1996); Rochester and Martin (1979)).

Codal variation in the forensic context

The usefulness of this theory for forensic purposes has already been shown, although codal variation has never been defined as such. The most significant example of an application of this theory is the Derek Bentley case (Coulthard and Johnson, 2007).

The analysis of the statements involved in the case showed that police officers and lay people differed in the position of the word *then* in the context of a police statement. Coulthard's analysis showed that, in the context of a police statement, the phrase *I then* is used significantly more by police officers than by lay people, who in turn prefer *then I*. Analysing this semantically using SFL indicates that these two constructions are indeed different regarding their textual meanings.

In a clause, the structure that is used to express textual meaning is the Theme-Rheme structure. As Halliday and Matthiessen (2004: 64) propose, the Theme of a clause is '[the element of the clause] which locates and orients the clause within its context'. In the example of the Bentley case, the opposition between *I then* vs. *then I* corresponds to a shift between which element starts the clause and therefore in how the writer wants to orient the reader. For example, *I then followed* the man shifts the start of the clause on *I*, the Subject (and therefore Agent, if the clause is transitive), thus relegating the temporal orientation in the segment of the clause that is presupposed to be already known by the hearer. On the other hand, *then I followed the man* takes the adjunct *then* inside the Theme, thus adding more emphasis on the temporal orientation of the clause. Although the difference in meaning between these two clauses is rather subtle, the opposition between these two variants is undoubtedly one of meaning.

Since this difference in meaning seems to be correlated with social group, in the form of 'policemen' vs. 'lay people', it is possible to conclude that this variation found in statements is similar to what Hasan (1990) found in her studies on socialisation processes in Australia. What Coulthard and Johnson (2007) refers to as *police register* when describing the fact that police officers use *I then* could probably be regarded as an instance of codal variation and it is therefore possible to substitute the term *police register* with *police code*. Replacing the term is not just an ideological position but also a meaningful theoretical tool. If a feature is recognised as an instance of codal variation, the explanations and predictions theorised within SFL within the theory of codal variation can be extended to the particular instance under analysis. Since codal variation exists because of the different interpretation of the context that social groups develop as part of their sub-culture and the experience they have with a particular genre, in this case it is possible to hypothesise that the difference between *I then* and *then I* is originated from a different interpretation of the context of a police statement given by the social group *police officers*. Lay people who do not experience statements in the same way as policemen seem to interpret the genre as a form of narrative, thus providing elements of narrative like the focus on temporal sequences such as the thematisation of *then*. On the other hand, the experience of policemen and their community of practice trains them to focus on more important things, such as the Subject (most of the times also Agent) of the clause.

The usefulness of this theoretical construct is immediately apparent for authorship profiling and attribution and it is indeed already applied in forensic contexts, although never recognised as such. Chaski (2001) and her syntactic markers are an example of codal variation used for attribution, as the differences between preferences of determiners or use of different verb phrases in the context of emails is indeed an instance of codal variation. Studies such as Koppel *et al.* (2002) or Argamon *et al.* (2009) are again finding out differences in coding orientations between social groups such as age or gender. By knowing why these differences are found and by contextualising this practice in a broader theory, practice can be informed and improved in order to produce more accurate analyses.

Towards a method: Biber's multidimensional analysis

Whether the theory produces valid hypotheses or not is only partially tested. However, the results so far are promising enough to justify the proposal of a general method of authorship analysis. The method can then be tested to validate the theory or reformulate it.

In general, the method consists in the analysis of the known sets of texts for semantic features and then in the identification of those features that do not vary because of registerial reasons but because of codal reasons. In other words, the method consists in finding those linguistic variables that present more intra-author variation than intra-genre variation. If the theory is correct, those variables will represent how the social group(s) to which the writer belongs understand(s) or interact(s) with that particular genre.

Since this hypothetical method requires a set of linguistic variables, the ones that are most obviously suitable would seem to be the classic SFL ones, such as frequencies of types of transitivity, frequencies of types of mood, frequencies of types of theme and so forth. This, however desirable and optimal in theory, has turned out to be unpractical in recent pilot studies (Nini and Grant, 2013). There are two major reasons for this impracticality: (a) at the present time, there are no reliable parsers for SFL features; this implies that the analysis has to be carried out manually, which in turn implies that at times the analysis can be too subjective to be used for forensic purposes; (b) there is no knowledge of how SFL variables systematically vary in different contexts to allow the analyst to understand to what extent the variation observed for a particular variable in a particular genre is distinctive or normal for that genre. This prerequisite for the variables is necessary as otherwise an assessment of the intra-genre variation is difficult to obtain. Mainly for these two reasons, the option considered in this paper is to introduce another framework that is compatible with codal variation theory as well as satisfying the two above-mentioned criteria. This framework is Biber's multidimensional analysis (Biber, 1988).

In a large scale experiment, Biber (1988) applied a multivariate statistical test called factor analysis to study how linguistic variables (such as frequency of past tenses, frequency of nouns, frequency of mental verbs) that are known to vary from register to register co-vary altogether to create functional orientations. Biber (1988) examined a general corpus representative of the most important genres of the English language and measured automatically 68 linguistic variables. Once the frequencies for these variables were calculated, they were arranged by the factor analysis along factors. The process consists in trying to explain the co-variation between these variables so that the variables that contribute to the same function can be grouped together. In this way it is possible to reduce a set of 68 variables to a more manageable smaller set of factors that can be interpreted for the linguistic function that they realise. In Biber (1988), the analysis generated six factors, which therefore created a six-dimensional space where the genres of the English language can be located. Mapping a genre on this space means calculating the score for each of the factors for each of the texts in a genre, finding the average for the genre and then comparing the figure obtained with the averages calculated for other genres.

Theoretically, the compatibility between this analytical framework and codal variation theory can be noticed in many points of overlap between the two. First of all, in both of these works, it seems evident that the authors start from a model that is based on an analysis of underlying functional orientations. Secondly, it is possible to notice that both Biber (1988) and Hasan (1990) employ a multivariate statistical analysis to find the semantic styles or functional orientations of some language varieties. Finally and most importantly, codal variation theory and the analysis presented in Biber (1988) are compatible with Finegan & Biber's (2001) *register axiom*. Finegan and Biber (2001), in a fashion very similar to Hasan's codal variation, postulate that different social groups possess different degrees of competence of different registers. This competence is formed by exposure to the registers and it therefore varies from social group to social group because different social groups are exposed to different registers.

These theoretical and practical advantages make Biber's multidimensional analysis a good candidate for transforming codal variation theory into a tool for authorship analysis. The method here proposed can be explained by looking at an example taken from a real case.

An example from a real case

The case examined in the present paper is an inclusion/exclusion attribution case where the analyst is provided with an email spreading malicious information (500 words) and four known emails (500 words in total) authored by the main suspect.

The first assumption that has to be met for the method to work is that the known set and the questioned set be compatible in terms of register. This can be assessed qualitatively, by looking at the recipient(s) and the medium, for example. It can also be tested quantitatively, by measuring Biber (1988) variables, calculating the factor scores, plotting the texts in the multidimensional space and finally verifying to what extent the texts analysed fall within the same genre. Both pieces of evidence can be collected to conclude whether the contextual comparability assumption is met.

If this assumption is met, it is reasonable to assume that the variation observed in the two sets is mainly given by the coding orientations of the authors, that is, on the way the authors of the questioned set and the known set interpreted the context in which they operated. In this particular inclusion/exclusion case, the question that is asked is: 'is the questioned set compatible with the known set?', which is a subset of the question: 'is there linguistic evidence for common authorship?'. Showing that the coding orientation of the author of the questioned set is compatible with the coding orientation of the known set is a piece of evidence that contributes to answering this question.

Using a notion introduced by Grant (2010), the coding orientation of one author could be thought of as those variables that are *consistently* and *distinctively* used in the author's texts when compared to the genre analysed. To find these variables, one can simply replicate Biber's (1988) study for the known set and questioned set and determine the values for each of the 68 variables, as well as the factor scores for each text. Furthermore, what is needed to validate these counts is a theoretical explanation of why these variables vary in accordance with codal variation theory.

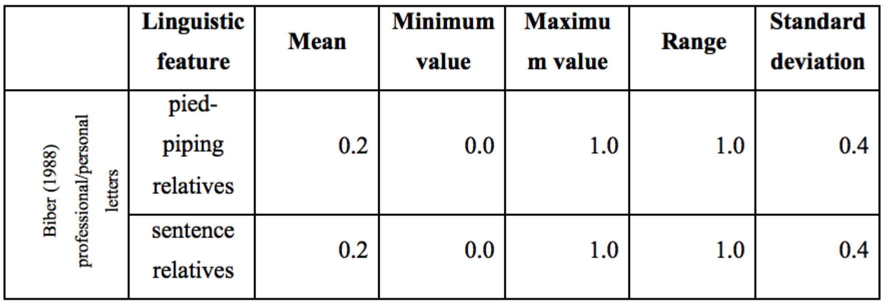
In the case presented as example, after examining all the variables, including the factor scores, two variables were found that equally occurred in the known set and in the questioned set and that, in addition, occurred significantly more often than expected for the genre, that is, that showed *consistency* and *distinctiveness.* These variables were *sentence relatives*, defined as the normalised frequency of: <COMMA + which>, and *pied-piping relatives*, defined as the normalised frequency of: <PREPOSITION + (who|whom|whose|which)>. Without a base-rate knowledge of these variables, comparing the raw or normalised frequencies of these two variables between the questioned and known set and obtaining similar figures is not enough on its own to claim compatibility. In other words, if it is impossible to know what the normal frequency for these variables is, it is equally impossible to gather a piece of evidence to claim common authorship. Looking at Biber's (1988) analysis, the two variables considered present the following frequencies in the genre that is closest to emails, 'professional/personal letters':

Assuming that the emails examined for the case and the personal/professional letters used by Biber (1988) are compatible in their registers, a comparison of those features with the observed frequencies for those two variables for the known and questioned sets is reported below:

The difference between a typical personalprofessional letter and the two known and questioned sets is strikingly significant. Whereas in a typical letter it is normal to find about 1 instance of both variables every 1,000 words, in the known and questioned sets an average of 0.6 per 100 words is observed.

By confronting the observed frequencies with the frequencies expected for the genre it is possible to assess how much *distinctive* the known and questioned sets are when

Figure 2. Frequency for pied-piping relatives and sentence relatives. Frequencies per 1000 words.



	Linguistic feature	Mean	Minimum value	Maximu m value	Range	Standard deviation
Biber (1988) professional/personal letters	pied-piping relatives	0.2	0.0	1.0	1.0	0.4
	sentence relatives	0.2	0.0	1.0	1.0	0.4

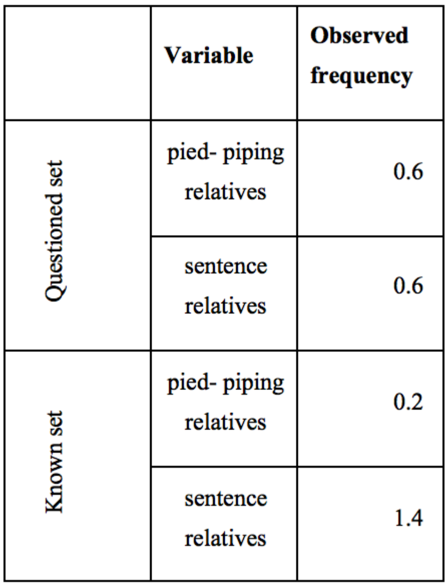
compared to the norm. The shared *distinctiveness*, given the fact that register has been controlled for, is a piece of evidence to corroborate the hypothesis of common authorship of the two sets.

Qualitatively speaking, an assessment of these two variables in the texts shows that the authors of the known and questioned sets use far more *pied-piping relatives* and *sentence relatives* because they employ a more convoluted and complex syntax full of sub-specification and interpersonal comments on previous sentences. This analysis allows for a quantitative estimation of the variation observed, as well as the possibility of qualitatively explaining the stylistic difference.

An objection to this analysis is that *pied-piping relatives* and *sentence relatives* are two variables that rarely occur in texts. They do not follow a linear distribution that increases with the size of the corpus and therefore plenty of data is needed to establish the typical distributions of these variables (Biber, 1993). Nonetheless, the theoretical qualitative explanation of the difference observed in the two samples corroborates the quantitative finding and compensates for the problem of non-linearity of these two variables.

The sample gathered by Biber (1988) is biased towards middle class educated writers and reflects the coding orientation of that social group. The convoluted syntactic style given by a high frequency of relatives seems to be stigmatised in high education, as it increases sentence length but focusing on clausal complexity rather than nominal complexity (Hunt, 1983). On the other hand, previous studies have shown that individuals with lower education background tend to use more clauses per sentence and more sentences per t-units (Hunt, 1971, 1983). In other words, the figures given above by Biber are skewed towards a particular social group and do not represent a normal distribution for the English language as a whole. However, having this knowledge of how these particular variables are distributed in terms of genre and social groups, that is, in terms of registerial and codal variation, can be useful in forensic cases. In this example, for instance, if the research on relatives and level of education is confirmed, there is reason to believe that both the author of the known set and the author of the questioned set belong to a low education level social group and this, in turn, can be used as a piece of evidence for the sets being produced by the same author.

Figure 3. Observed frequencies for pied-piping relatives and sentence relatives for the Q and K sets. Frequencies per 100 words.



	Variable	Observed frequency
Questioned set	pied- piping relatives	0.6
	sentence relatives	0.6
Known set	pied- piping relatives	0.2
	sentence relatives	1.4

Conclusions

In conclusion, it is proposed that codal variation theory can be a valid tool for forensic authorship analysis, provided that there is enough research on the variables for the genres that are typically involved in a forensic scenario. A method based on this theory is theoretically grounded in SFL but analytically based on Biber's (1988) multidimensional framework.

The example considered might seem just another application of corpus methods to a forensic case and this is indeed true. However, the significant difference with other case reports of this kind is that the corpus analysis carried out is entirely theory based and theory driven. The lack of theory in forensic authorship analysis might become a danger that could threaten the field. Without the knowledge that explains why a particular system or tool works, it is neither possible to be sure that the results given by this tool can be replicated in other cases apart from the experimented ones, nor is it possible to improve them. It is likely that the field can move forward only thanks to theories that can explain why a particular tool is useful and, most importantly, that generate hypothesis that can be validated or rejected in future research.

References

Argamon, S., Koppel, M., Pennebaker, J. W. and Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), 119.

Bernstein, B. (1962). Linguistic codes, hesitation phenomena and intelligence. *Language and Speech*, 5(4), 221–240.

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

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, 8(1), 1–65.

Coulthard, M. and Johnson, A. (2007). *An Introduction to Forensic Linguistics*. London: Routledge.

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

Grant, T. (2010). Txt 4n6: Idiolect free authorship analysis. In M. Coulthard and A. Johnson, Eds., *Routledge Handbook of Forensic Linguistics*, 508–523. London: Routledge.

Grant, T. and Baker, K. (2001). Identifying reliable, valid markers of authorship: a response to Chaski. *Forensic Linguistics*, 8(1), 66–79.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (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*, 73–118. London: Equinox.

Hasan, R. (1996). Ways of saying: ways of meaning. In C. Cloran, D. Butt and G. Williams, Eds., *Ways of Saying, Ways of Meaning: Selected Papers of Ruqaiya Hasan*, 191–242. London: Cassell.

Hasan, R. (2009). Wanted: a theory for integrated sociolinguistics. In J. Webster, Ed., *The Collected Works of Ruqaiya Hasan Vol. 2: Semantic Variation: Meaning in Society and in Sociolinguistics*, 5–40. London: Equinox.

Hunt, K. (1971). Teaching syntactic maturity. In *Applications of Linguistics: Selected Papers of the Second International Congress of Applied Linguistics*, 287–301. Cambridge: Cambridge University Press.

Hunt, K. (1983). Sentence combining and the teaching of writing. In M. Martlew, Ed., *The Psychology of Written Language: Developmental and Educational Perspectives*, 99–125. New York: John Wiley.

Koppel, M., Argamon, S. and Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), 401–412.

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

Matthiessen, C. M. I. M. (2007). The "architecture" of language according to systemic functional theory: developments since the 1970s. In R. Hasan, C. Matthiessen and J. Webster, Eds., *Continuing Discourse on Language*, 505–561. London: Equinox.

Nini, A. and Grant, T. (2013). Bridging the gap between stylistic and cognitive approaches to authorship analysis using systemic functional linguistics and multidimensional analysis. *International Journal of Speech, Language and the Law*, 20(2), 173–202.

Rochester, S. and Martin, J. R. (1979). *Crazy Talk: a Study of the Discourse of Schizophrenic Speakers*. Norwood, N.J.: Ablex.