Developing forensic authorship profiling

Andrea Nini

University of Manchester, UK

Abstract. *Current research into the task of determining the characteristics of an anonymous writer, authorship profiling, does not meet the demands of the forensic context, because of the lack of transparency of certain computational techniques, their requirements for large data sets, and, most importantly, since the strength of register variation does not guarantee that findings obtained in other registers will apply to forensic registers such as, for example, a threatening letter. The present article demonstrates how previously established findings related to stylistic variation in English for gender, age, and social class also apply to the kinds of texts often analysed by forensic linguists through an experiment involving 96 participants. These results constitute an example of linguistically-motivated profiling research and it is argued that the agenda to move from authorship profiling to forensic authorship profiling should be led by previously established knowledge of language variation.*

Keywords: *Authorship profiling, register variation, stylistics, threatening text, corpus linguistics.*

Resumo. *A atual investigação sobre a determinação das características de um escritor anónimo, a determinação do perfil do autor, não satisfaz as necessidades do contexto forense devido à falta de transparência de determinadas técnicas computacionais, dos seus requisitos para grandes "data sets" e, sobretudo, devido ao facto de a robustez da variação do registo não garantir que os resultados obtidos noutros registos sejam aplicáveis aos registos forenses como, por exemplo, uma carta de ameaça. Este artigo demonstra de que modo estudos prévios relacionados com a variação estilística em inglês relativamente ao género, idade e classe social também são aplicáveis aos tipos de texto muitas vezes analisados pelos linguistas forenses; para o efeito, realizou-se uma experiência que envolveu 96 participantes. Estes resultados constituem um exemplo de investigação na determinação de perfis linguisticamente motivada, defendendo-se que o plano de investigação para passar da determinação de perfis de autor para a determinação de perfis de autor forense deveria ser orientada por investigação prévia sobre variação linguística.*

Keywords: *Determinação de perfis de autor, variação de registo, estilística, texto de ameaça, linguística de corpus.*

Introduction

Authorship profiling is the task of determining the characteristics of an anonymous author, such as their demographic details, from the way they use language. Profiling questions can be of extreme importance at the investigative phases of, for example, a case involving an anonymous threatening letter or a ransom demand, when the list of suspects is too large. Despite this importance, the forensic linguistics literature on authorship profiling is very limited. Two ways of doing authorship profiling have emerged from forensic casework and research: (1) analysis of salient linguistic markers, and (2) analysis of writing style.

The first type of profiling is the application of sociolinguistic knowledge on a case by case basis to extract *ad hoc* linguistic features that are markers of a certain demographic background, as demonstrated in famous examples such as the *devil strip* case (Leonard, 2005), the Unabomber case (Shuy, 2014) or the *bad-minded* case (Grant, 2008). This type of analysis involves the linguist's experience in discovering dialectal or sociolinguistic features that can reveal clues about the background of the author.

In contrast, the second type of profiling consists in the analysis of the *stylistic variation* exhibited by the text as a whole. This analysis often involves the study of the frequency with which certain features are used, like the study of register variation (Biber, 1988) and takes as the unit of analysis the text itself. A *style*, as defined by Biber and Conrad (2009), is a variety of language associated with a particular author or social group as opposed to a situation which is constituted by linguistic features that are pervasive and frequent. It is therefore similar to a *register*, which is a variety of language associated with a particular situation, in terms of feature types that constitute it, but different in that styles are particular varieties of registers that characterise authors or social groups.

The current state of the art of authorship profiling reveals that research on the first type of analysis is virtually non-existent while the second type has become a sub-field of computer science and machine learning. It is indeed very difficult to systematise research for Type 1 profiling, as the type of markers that become important in a forensic case is often unpredictable. Analysis of Type 1 therefore relies almost entirely on the knowledge and intuition of the forensic linguist. Research on Type 2 profiling, on the other hand, has been developed by computer scientists applying machine learning techniques, for example, to automatically determine the gender or age of a writer (Argamon *et al.*, 2009). These systems usually work by taking as input an array of features, usually frequencies of words or characters, and using these arrays to train a machine to distinguish groups of texts that have been labelled already as, for instance, male or female. The outputs of these systems are the classification accuracies and, sometimes, the distinguishing features.

The fact that Type 2 authorship profiling is dominated by computer science can be quite problematic for forensic linguistics, since the needs of forensic linguists are often different from the needs of the users of computational applications. Computational authorship profiling is not necessarily interested in understanding the inner (linguistic) mechanisms of the machine, as long as the accuracy rates are outperforming previous models. This lack of linguistic understanding can however be problematic for a forensic linguist, who is ultimately called to testify about language. Similarly, most of the times these techniques require plenty of data for training and testing, which is not the standard scenario in forensic linguistics. All of these aspects of computational authorship profiling therefore make these computational techniques very good for applications where

the objective is a fast scrutiny of large data sets, for example in marketing applications, but not always useful for the typical scenario of a forensic linguist being asked by the police about the most likely profile of the author of a one page threatening letter.

The present article argues that the development of a method of *forensic* authorship profiling for anonymous written texts can only come from research in two directions: (1) the accumulation of knowledge and understanding of stylistic variation across social factors, and (2) the verification that these patterns are also found in the register of the disputed document to be profiled. The first direction addresses the need for established linguistic theory and knowledge to be applied to forensic scenarios. The second direction addresses a fact often ignored by computational research in authorship profiling, that is, the pervasive effect of register variation on language (Biber, 1995, 2012; Biber and Conrad, 2009).

This article reports on an experiment on English data aimed at identifying which stylistic patterns previously found in other studies can be used for profiling three demographic characteristics (gender, age, and social class) in a situation similar to the typical forensic linguistic scenario of an anonymous short letter.

Literature review

The pre-requisite to perform a forensic linguistic task such as profiling in a linguistically-informed way is to first carry out a complete survey of what is known about language variation and the social factors of interest. This literature review constitutes a survey of key research that could inform forensic authorship profiling for the three social factors considered: gender, age, and social class. Other social factors, such as ethnicity, could also be considered, but these three are a good starting point, given their potential investigative value as well as the existence of a large amount of linguistic research on stylistic variation associated with them. The literature review focuses only on those studies that can be used for the typical forensic linguistic scenario of the profiling of the style of an anonymous written text. The present work is not concerned with studies that looked at alternations such as *was* for *were* or *innit* as a tag question, as these features are the type of features involved in a Type 1 analysis. Instead, this review focuses on the established patterns of variation that have been found to distinguish the social groups considered in terms of, for example, the use of nouns as opposed to verbs, clausal patterns, and other lexicogrammatical features that are pervasive and therefore that are always found in any text, considering the text itself as a unit of analysis.

Gender

The notion and definition of the concept of gender is not trivial but despite these problems the profiling of someone's gender is a question that can be asked to forensic linguists. Although it has been often useful to draw a distinction between the socio-cultural cline of *gender* and a biological binary *sex*, there is evidence suggesting that in reality none of these constructs is binary (Bing and Bergvall, 1998). The tension in profiling work is that whereas law enforcement are interested in certain biological correlates of gender, the clues that can be found in language are more likely to reveal the socio-cultural gender of the author, which is a continuum as well as subject to variation depending on extra-linguistic context (Carothers and Reis, 2013). These issues have not been addressed extensively in stylistic research that involved gender and the research

reviewed below thus significantly simplifies the nature of this dependent variable, reducing it to a division between biological *men* and biological *women*. Despite this issue, this research is the only starting point for work on gender profiling at this stage.

The most important pattern identified by previous studies of stylistic variation and gender is in the continuum between *nominal vs. clausal style*, the former being more typical of men and the latter more typical of women. The nominal end of the continuum is more often characterised by use of features such as nouns, adjectives, prepositions, and, generally, complex noun phrases, whereas the clausal end is characterised by the use of features such as verbs, adverbs, and simple noun phrases constituted by pronouns only. This pattern has been extensively found in a large number of studies at different times and in several registers and the literature thus suggests that this is a pervasive effect, even though the reported effect sizes have been relatively small. This pattern has been found in structured sociolinguistic interviews (Poole, 1979, N = 96), casual conversations (Rayson *et al.*, 1997), personal letters (Biber *et al.*, 1998, N = 80), and large corpora of formal/informal and fiction/non-fiction written texts (Koppel *et al.*, 2002; Argamon *et al.*, 2003; Schler *et al.*, 2006; Newman *et al.*, 2008). A gender effect on the frequency of nouns and pronouns has also been observed diachronically by Säily *et al.* (2001, N = 660) in a corpus of letters dating from 1415 to 1681.

After analysing speech data from 80 participants and finding a similar effect, Heylighen and Dewaele (2002) have proposed that this pattern could be due to the level of formality, where formality indicates the level of mathematical preciseness of a text as opposed to its dependence on the extra-linguistic context. They introduce an index to measure formality defined as follows

$$F = \frac{(R_{nouns} + R_{adj} + R_{prep} + R_{art}) - (R_{pro} + R_{verbs} + R_{adv} + R_{interj}) + 100}{2}$$

where the first bracket contains the relative frequencies of the nominal/formal elements (nouns, adjectives, prepositions, and articles) and the second bracket contains the relative frequencies of the clausal/contextual elements (pronouns, verbs, adverbs, and interjections).

Despite this attempt, the literature reveals that there is far more advancement in the description as opposed to the explanation for this pattern, especially since a clear definition of *gender* is still lacking. It has been proposed in the past that the sociolinguistic effects of gender could have both biological and social explanations (Chambers, 1992) and certain elements of the patterns described above have indeed been given a psychological explanation by social psychologists who have found that increased pronoun use correlates with gender in the direction described and with a tendency for neuroticism, which is also more common among women (Pennebaker *et al.*, 2003; Rude *et al.*, 2004). In a very small pilot study of only two subjects Pennebaker *et al.* (2004) found that an increase in testosterone levels increases the level of nominal style employed. On the other hand, other plausible explanations for this gender effect can be found in the tendency for these two genders to engage with different registers (Herring and Paolillo, 2006) and in the network of relationships that they therefore establish (Bamman *et al.*, 2014), and

thus, ultimately in the different *communities of practice* that the different genders on average engage with (Eckert and McConnell-Ginet, 1992).

Age

Although the concept of ageing would intuitively seem relatively unproblematic, from a linguistic point of view it is indeed much more multi-faceted. Statistically it is convenient to reduce age to a number as has been done in several studies but this measure of biological age might not be the best predictor of linguistic variation, as *social age*, as opposed to biological age, is more likely to affect language (Eckert, 1998). Social age is marked by a series of socially-recognised landmark events in life, such as certain birthdays, marriage, entering the job market, etc., which require different linguistic varieties and which offer different registers that can and sometimes must be learned. If this is correct, then profiling age has the same tension seen above for gender: whereas law enforcement is mostly interested in biological age, linguistic variation can only reveal clues as to the social age of a person, which is only a proxy for biological age.

The most well established pattern of stylistic variation that correlates with age is the negative relationship between syntactic complexity and ageing, which has been discovered in psycholinguistics. Analysing experimental data and diary entries, Kemper (1987) discovered that as people age there is a tendency to abandon complex clausal syntax, measured by average number of clauses per sentences, and in particular left-branching complexity. The explanation that they proposed for this pattern is that it is an effect of working memory, which decreases with age and especially in situations of dementia or Alzheimer's disease. This effect of age on syntax was found in several other studies of different and large subject groups and different registers, such as oral interviews and written essays (Kemper *et al.*, 1989, N = 108), descriptive essays (Bromley, 1991, N = 240; Rabaglia and Salthouse, 2011, N = 900), speech samples (Kemper and Sumner, 2001, N = 200), and in the famous Nun Study, in which autobiographic texts of a group of 150 nuns were analysed from 1930 until 1996 (Kemper *et al.*, 2001).

Interestingly, some of these studies, such as Kemper and Sumner (2001) and Rabaglia and Salthouse (2011), have also noticed an increase in vocabulary variety with old age, measured via type-token ratio or average word length. This effect of ageing on lexical richness is in line with the recurrent finding that ageing plays a role in the nominal vs. clausal style pattern already observed for gender. For example, Pennebaker and Stone's (2003) study of emotional disclosure essays and interviews on more than 3,000 participants found generally an increase in frequencies of determiners and prepositions and a decrease in frequency of pronouns with ageing. Likewise, Schler's *et al.* (2006) analysis of blogs written by almost 40,000 writers also found that as people age they tend to adopt a more informational/nominal style.

In sum, although ageing seems to be correlated with loss in syntactic complexity, another form of complexity based on nominal structures and lexical richness seems to replace it. This is consistent with another explanation for this effect given by Kemper *et al.* (1989), who suggested that the decline in usage of complex syntactic forms might be due to older people becoming more familiar with better ways of conveying meaning that do not involve unnecessarily complicated structures, relying more on refined vocabulary. This explanation is consistent with the effect found regarding nominal style, as this style is more characteristic of literate registers, such as academic and scientific registers (Biber, 1988), which take time and experience to be acquired.

Social class

Although years of research in variationist sociolinguistics have found that social class is one of the most important predictors of language use, authorship profiling so far has not devoted much research to this factor. The problem with social class is that it is a controversial and difficult factor to quantify, a controversy made worse by the virtual disengagement between linguistics and sociological literature (Ash, 2002). Very rarely do linguistic studies adopt the same definition of social class and yet this construct very often shows effects of large magnitude. Most of the indexes used are based on occupation, but they tend to include other aspects, such as level of education, income, household, and parents' backgrounds. Despite the problems and controversies and general lack of research in computational authorship profiling, these factors are typically useful information for investigators in a case involving an anonymous text and they are therefore a necessary inclusion in the practice of forensic authorship profiling.

Previous literature has found that overall higher classes are more competent in the use of complex syntax due to their more frequent exposure to this kind of input. This pattern is very well established, with studies that found effects on various pools of subjects across decades. Syntactic complexity, measured through average sentence length or number of dependent clauses per sentence, has been found to be correlated with class by Loban (1967, N = 211) in both oral and written texts, Poole (1976, N = 80) in life-forecast essays, Johnston (1977, N = 36) in experimental elicited narratives, Poole (1979, N = 96) in structured interviews, Labov and Auger (1993, N = 10) in sociolinguistic interviews, and it was also found in Kemper *et al.*'s (1989) and Mitzner and Kemper's (2003) studies on syntax and ageing to be a good predictor of level of education.

A measure often employed to study complexity and its relationship to social class and level of education is the complexity of *t-units*, where a t-unit is defined as an independent clause with all its dependent clauses. Average t-unit length or the number of clauses per t-unit are therefore better measures of complexity than average sentence length, as sentences are instead orthographic units. Several studies have found that the management of t-units and especially the way they are punctuated is characteristic of certain levels of education and class, with both the complexity and the ratio of t-units per sentences being a good proxy to the degree of competence with standard punctuation (Loban, 1967; Hunt, 1971, 1983).

Similarly to gender and age, social class seems to participate in the nominal vs. clausal pattern, with the higher social class being more familiar and thus more frequent users of the nominal style. Heylighen and Dewaele (1999) found that their measure of formality increased with the social status of their participants. Several studies found evidence for more frequent usage of nominal parts of speech in the discourse of higher social classes, such as uncommon adjectives in essays (Poole, 1976), subject noun phrases in elicited narration (Johnston, 1977, N = 36), nouns and adjectives in elicited narration (Hawkins, 1977, N = 263), or adjectives in sociolinguistic interviews (Macaulay, 2002, N = 45). This opposition between nominal vs. clausal style mirrors very well the distinction between restricted and elaborated codes made by Bernstein (1962), which he had already associated with social class.

Finally, several studies have found a relationship between lexical richness and social class or level of education. For example, very early studies such as Bernstein (1962, N = 106) or studies on readability measures (Kitson, 1921; Dubay, 2004) found that average

word length correlates with social status, a finding confirmed by Bromley (1991, N = 240) in descriptive essays, or by Berman (2008, N = 80) for narrative speech samples. Byrd (1993, N = 200), on the other hand, found that measures such as type-token ratio and the mean rarity score of a word were higher in various essays written by higher social classes, a finding confirmed by Mollet *et al.* (2010, N = 55) in student essays, in which they used a measure called Advanced Guiraud 1000, calculated using the following formula:

$$G = \frac{V - v}{\sqrt{N}}$$

where V indicates the total word types in a text, v indicates the *common* word types of the text, that is, the most common 1000 word types of a comparison corpus such as the British National Corpus, and N is the total number of word tokens in the text.

In sum, despite its controversial status, all the evidence points to a substantial effect of social class on language and this fact alone suggests that this social factor cannot and should not be ignored when profiling.

Methodology

In order to verify to what extent these patterns are found in *malicious forensic texts* (Nini, 2017), the ideal methodology would be to compile a corpus of such texts stratified by these three social factors. However, gathering such a corpus is an impossible enterprise as malicious forensic texts are rare on their own and even rarer are texts of this kind for which the demographics of the authors are reliably known. This study therefore adopts an experimental methodology which, despite the obvious drawback of not being based on naturally-occurring data, offers the key advantage of allowing greater control of the conditions. A common problem with corpus data for sociolinguistic studies, for example, is that it is not always possible to control very accurately the conditions under which data is produced and since register is a strong source of variation, this has the potential of skewing the results if it is not carefully isolated. With an experiment, on the other hand, the researcher can control the aspects of the situation that they wish and measure their effect on the factors.

Data

Ninety-six participants, all required to be native speakers of any variety of English, were recruited from different social backgrounds, such as university students, training police officers, members of a writing group for retired people, and homeless newspaper sellers. Most of the participants were from the UK and especially from England, with the exception of three participants from North America and one from Jamaica.

54% of the subjects declared their gender to be male and 46% to be female. For age, 37.5% of the participants were between 19 and 29 years old, 38.5% were between 30 and 50, and 24% were between 51 and 78. Finally, 55% of the participants did not have a university degree, while of the remaining 45%, 16% had an undergraduate degree and 29% had a postgraduate degree.

An index of social status was calculated using mainly the occupation of the subjects averaged over the occupation of their parents. A score from 1 (lower status) to 6 (higher

status) was assigned to each participant[1] using the classification of occupations adopted for the British National Corpus (McEnery, 2006: 27) in the following way:

- A - higher managerial, administrative or professional – Score 6
- B - intermediate managerial, administrative or professional – Score 5
- C1 - supervisory or clerical, and junior managerial, administrative or professional – Score 4
- C2 - skilled manual workers – Score 3
- D - semi- and unskilled manual workers – Score 2
- E - state pensioners or widows (no other earner), casual or lowest grade workers – Score 1

For students, only the average of their parents' score was considered.

A cross-tabulation of the factors revealed that the sample is very well balanced, with only a significant association between gender and age (binarized in two categories, *Older* and *Younger* at the median age of 38) (X^2 = 8.2, df = 1, p = 0.004), as there were more younger women than younger men overall. This skew could affect some of the results and it will be further discussed below. In addition, this analysis also revealed that the social class index is a good proxy to the education of the participants as the association between having a degree or not and belonging to the *Higher* or *Lower* class (based on the median index of 3.7) was significant (X^2 = 17.9, df = 1, p = 0.00002). The distribution of the participants in the corpus according to these categories can be seen in Table 1.

Higher		
	Male	*Female*
Older	13	9
Younger	10	21

Lower		
	Male	*Female*
Older	20	5
Younger	8	7

Table 1. Distribution of number of participants in the corpus across the three categories used in the study: Gender, Age, and Class.

The subjects were asked to fill in a questionnaire with details about themselves and to carry out a writing task in a computer lab in a university room and they were compensated with an expense and participation fee of £10. The subjects were asked to write three tasks that elicit three registers (see Appendix): (1) Task 1: a letter of complaint to a holiday agent asking for compensation; (2) Task 2: a letter to the Prime Minister of the United Kingdom to complain about the economic crisis and threatening not to vote for them again; (3) Task 3: a letter to a fictitious abusive employer threatening to damage their car if their behaviour does not change. The participants completed the three writings tasks in the same session and were not given any time constraints to finish the experiment. The simulated situation of these three texts was structured in particular to capture variation in the recipient: Task 1 is addressed to a company, Task 2 is addressed to a person of higher status and power that the participants do not personally

know, and Task 3 is addressed to a person of higher status that they personally know. In addition, the three tasks can all simulate potentially threatening letters to a company, a political figure, or an employer. The experimental tasks are similar for several situational parameters (Biber, 1994), such as being written with the possibility of editing, having no audience, not being specialised, etc. but they differ greatly in topic and, most importantly, in the level of knowledge between addressor and addressee. Since audience design has already been shown to be a very important predictor of linguistic variation (Bell, 1984), this difference is important and it is predicted to have a strong influence on the style of the participants.

Although the experiment consisted in eliciting texts that have been designed to capture scenarios as close as possible to real forensic cases, it is reasonable to argue that these are still elicited texts and therefore they may still be different from real authentic malicious forensic texts of this kind. To address this problem, Nini (2015) compared the experimental texts against a corpus of authentic malicious forensic texts described in Nini (2017) and found that the register of these experimental texts is almost indistinguishable from the register of real malicious forensic texts. The analysis was done by testing for statistically significant difference on 135 linguistic features that vary across registers, including the features of interest for profiling identified in this article. Only 13 out of 135 linguistic features were significantly different across the data sets but a qualitative scrutiny revealed that out of these 13 features only two were due to an experimental effect: contractions and proper nouns were used much more frequently in authentic texts than in fabricated texts for reasons attributable to differences between real and experimental conditions. However, since neither of these features seems to have a role to play in profiling, it can be concluded that the experimental texts are a good approximation to the register of real malicious forensic texts.

Features

The literature review has shown that there are consistent patterns of stylistic variation that correlate with the three social factors considered. Therefore, it is possible to make certain predictions about the relationship between language and the social factors that will be observed in the simulated malicious texts:

1. The nominal vs. clausal style would pattern in the following way: male/older/higher social class participants should exhibit a more nominal style than female/younger/lower social class participants. This stylistic cline can be measured using Heylighen and Dewaele's (2002) *F* score, which includes all the features explored in a number of other studies;
2. Higher class/older participants should use a richer vocabulary than lower class/younger participants. Vocabulary richness can be measured using several indices, such as average word length, type-token ratio, etc. For this study the Advance Guiraud 1000 score presented above was chosen as it is a more direct measure for estimating *extrinsic* vocabulary richness, or the rarity of the vocabulary used (Mollet *et al.*, 2010);
3. Higher class/younger participants should use a more complex clausal syntax than lower class/older participants. Sentence complexity can also be measured in several ways, for example simply using average sentence length. However, as noticed by Hunt (1971), a sentence is an orthographic unit and this is therefore

not ideal. For this reason, this study focuses on the number of clauses per *t-unit*, where a t-unit is an independent clause with all its dependent clauses. For this analysis, T-units were identified and segmented manually but the number of clauses was determined using a computer script that counted all the main verbs in the texts.

If these features, as predicted by previous studies, are unequally distributed across the social factors in these experimental texts that set out to simulate malicious forensic texts, then this is evidence that these principles of profiling can be used in real-life forensic cases involving similar registers. The differences were tested using non-parametric significance tests as the normality of the distributions is not assumed, using the Kruskal-Wallis tests for dependent variables with more than two categories and the Wilcoxon Rank Sum test for dependent variables with only two categories.

Results

As predicted, the most important finding of this study is the pervasive effect of register, as all the features considered exhibit substantial register variation. This was expected as it has been already demonstrated that register variation is the most important predictor of linguistic variation of the kind analysed in this study and in the studies reviewed. For this reason, all the results below are plotted using a mixture of two types of graph: boxplots showing differences across tasks with overlaid dotplots and point and range plots within these boxplots to show the differences for the social factors. This way of visualising the patterns also reflects the idea that these styles associated with the social factors are indeed ways of realising a particular register.

The second most important finding is the considerable importance of the nominal vs. clausal style pattern, which affects all social factors as predicted. Figure 1 shows how *F* has a very strong register effect ($p < 0.0001$) and how this effect is reflected in the social factors. Indeed, all the social factors have a significant effect in the predicted direction for *F* but only for Task 1 (Class, $p = 0.02$; Gender, $p = 0.04$; Age, $p = 0.02$), the more formal letter of complaint, while in Task 2 this difference is less strong and only significant for Class ($p = 0.02$) and in Task 3 all categories have cut median values around the median for the register and non-significant effects. For Class, it seems evident that the difference is mostly due to the Lower category, which includes all the participants with a score between 1 and 3. For age and gender, in Task 1 older participants and male participants both scored above the median for the register, as predicted.

Because all the factors interact, it is interesting to explore the pattern emerging from these factors when they are combined. Figure 2 plots the distribution of the *F* measure for cross-categories such as Class-Gender, Class-Age, and Age-Gender. For Task 1 and 2, the predictions are all correct: the top categories that include most of the texts that are far away from the median are Higher-Male, Higher-Older, and Older-Male while the categories that score far away from the median are the opposite, Lower-Female, Lower-Younger, and Younger-Female, with the categories in between scoring in the middle and very closely to the median for the group overall. All of these differences are relatively strong and mostly significant for Task 1 (Class-Gender, $p = 0.01$; Class-Age, $p = 0.0007$; Age-Gender, $p = 0.059$), less strong for Task 2 (and all non-significant, except for Class-Age, $p = 0.02$) and they are neutralised in Task 3, with none of the effects significant. It is important to note here that these plots show how the skew noted in the Methodology

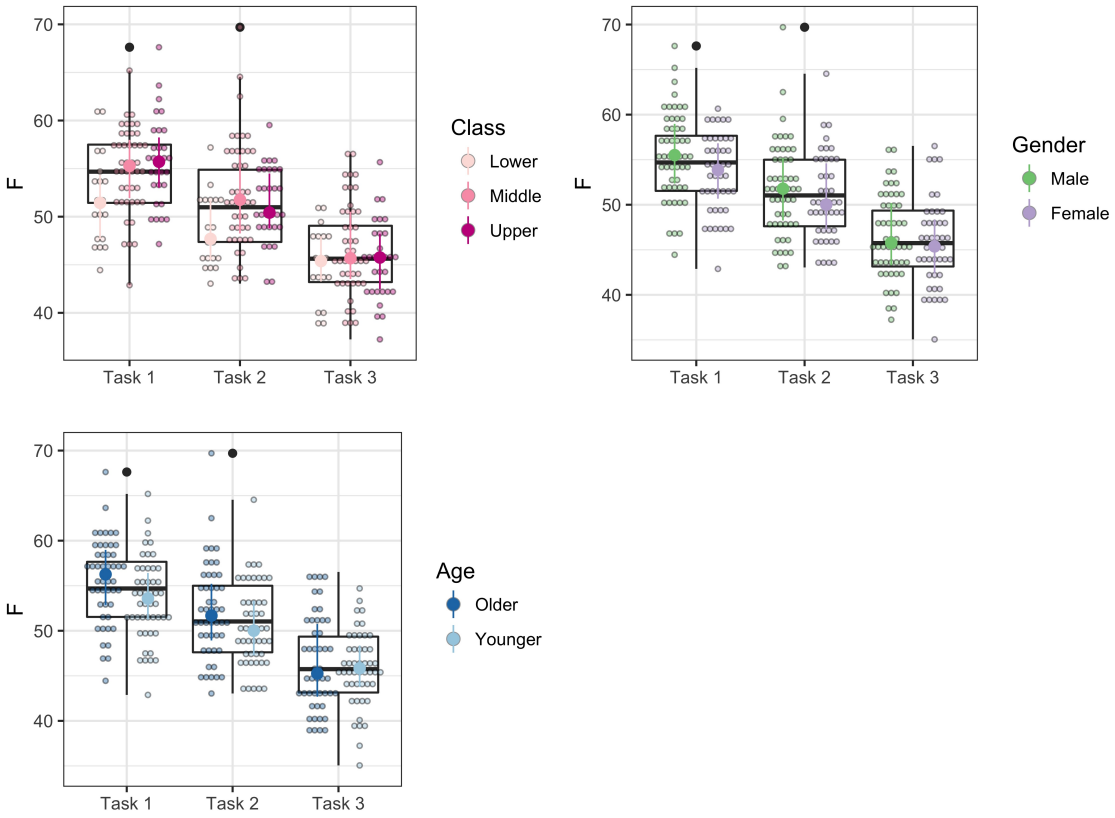

Figure 1. Boxplots showing the distribution of the *F* measure across Tasks. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot).

section does not have a strong confounding effect in these results for *F*, as despite the relatively higher number of younger females in the sample, the predicted pattern is still observed.

Vocabulary richness measured through the Advanced Guiraud 1000 score also shows the predicted pattern, with a strong register effect ($p < 0.0001$) (Figure 3). Similarity to F, the predicted direction is observed here for all the tasks, except perhaps Task 3, with the category Higher-Older scoring the highest, the middle categories situated along the median for the registers, and the lowest category being Lower-Younger. However, in this case the effects are significant only for Task 2 ($p = 0.02$).

Finally, again the analysis of syntactic complexity using the measure of clauses per t-units confirms previous findings (Figure 4). For this measure of syntactic complexity, however, the register differences are far less accentuated, although still very significant ($p = 0.001$). The difference seems to involve mostly Task 2, which has a higher syntactic complexity overall than the other two registers.

In this case, the literature would predict that the highest scores for syntactic complexity would be obtained by the youngest members of higher social classes and the analysis reveals that this is the case, with a cline that follows the predictions. However, this effect is significant again only for Task 2 ($p = 0.03$), the register characterised by the highest median syntactic complexity overall.

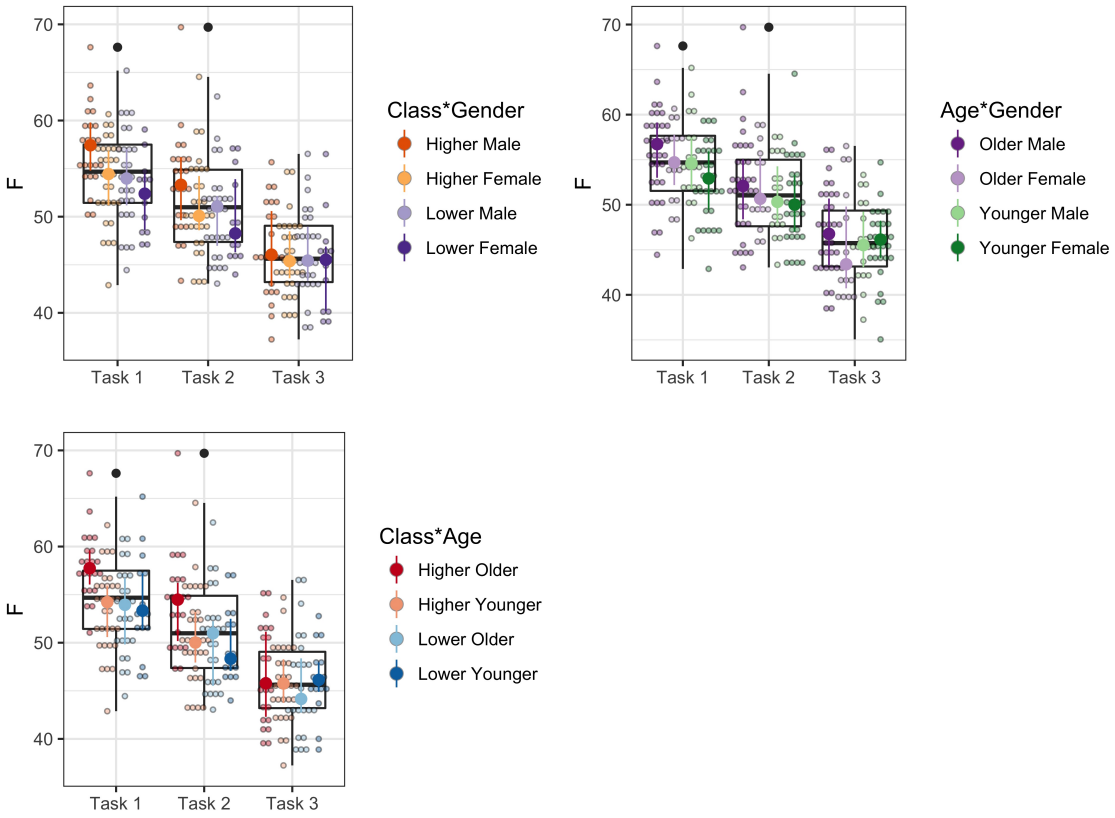

Figure 2. Boxplots showing the distribution of the *F* measure across Tasks for the social factors combined. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot).

Discussion

These results firstly indicate that the nature of the linguistic features considered requires that an analysis of the register of the text in question is conducted before any profiling, as the register effect of these features is generally much stronger than any social factor effect. However, provided that this is done, the results reported in this paper suggest that the relationship between stylistic variation and social factors previously identified are generalisable to registers similar to malicious forensic texts. These findings also suggest that even though the effects seem stable and unlikely to reverse direction, they do not necessarily appear in all registers. Therefore, although it would be very unlikely to find, for example, younger women to have a higher *F* score than older men in any register, it is possible that the predicted effect is neutralised by register effects. In other words, these findings suggest that register gives the space for stylistic variation of this kind to occur, as can be seen in the analysis of syntactic variation, which presents a social effect only for Task 2 where the amount of clausal complexity is overall higher.

Because of this strong register effect, it is fair to conclude that it is unlikely that any of these effects are exclusively the results of biological or psychological factors such as working memory. For example, if syntactic complexity decreased only because of a decrease in working memory capacity, then the same effects observed for Task 1 should be observed in Task 3. Explanations should instead be sought in particular in the reasons

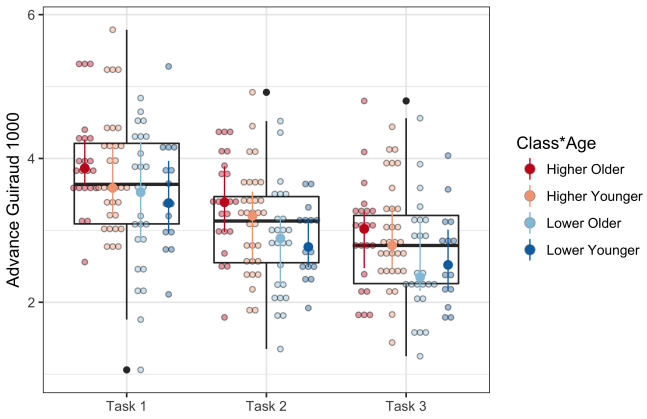

Figure 3. Boxplots showing the distribution of the Advance Guiraud 1000 score across Tasks for Social class and Age. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot).

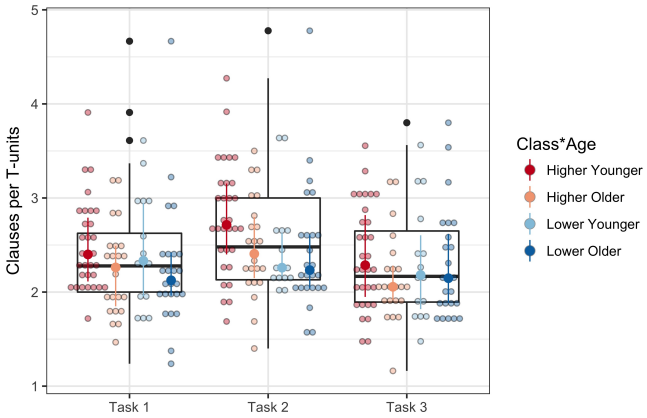

Figure 4. Boxplots showing the distribution of the number of clauses per T-units across Tasks for Social class and Age. For each boxplot, a dotplot per social factor value is plotted containing information about the median (larger dot) and quartile (the range crossing the median dot)

why these social categories employed these styles and on the nature of the relationship between styles and registers.

To explain the relationship between stylistic variation and social factors, let us consider the case of *F*, the most important linguistic feature of this study representing the opposition between nominal vs. clausal style. This stylistic contrast has been found across several studies in English and other languages and has been named in different ways. Heylighen and Dewaele (1999) expressed this contrast in terms of reliance on context and formality, while Biber's (1988) multidimensional study named this stylistic contrast functionally as the opposition between informational and involved discourse. More recently, Biber (2014) has renamed this opposition as the contrast between clausal vs. phrasal discourse. The findings of this study completely support previous findings: the *F* measure increases as the personal knowledge between interactants decreases because

a higher degree of distance between addressor and addresse(s) requires less reliance on context and thus a more pervasive adoption of nominal features of elaboration.

Crucially, these register differences for *F* and for the other features might also be responsible for the social differences observed in this and other studies, as explained by the *register axiom*. Finegan and Biber (2001: 265) define the register axiom as follows:

> If a linguistic feature is distributed across social groups and communicative situations or registers, then the social group with greater access to the situations and registers in which the features occur more frequently will exhibit more frequent use of those features in their social dialects.

In Systemic Functional Linguistics, this is expressed in the theory of *codal variation* or *semantic variation*, for which social groups differ in terms of the meanings that they make as well as the linguistic features they use and that this difference is due to the different degree of access that social groups have to certain registers (Hasan and Cloran, 1990; Hasan, 1996, 2009). This theory can help in explaining, at least partially, the effects observed, and in particular the results regarding the clausal vs. nominal style, as the nominal style is more frequently encountered in written formal writings and only members of the higher classes who work in occupations in which they often encounter this nominal style can therefore develop competence with it. This theory can also help explaining the neutralisation effect of register: Task 3, which does not require a nominal style, does not lead to social differences because even the higher social classes, who are capable of using the nominal style, still choose to use the clausal style as it is the most appropriate for the context. Suggestions along similar lines have been proposed by Bernstein (1962), who proposed that there are two codes of expression, the *restricted* and the *elaborated* code, and that social inequality arises as only certain occupations have access to both codes. If the nominal style can be compared to the elaborated code, then these results are compatible with his theory and provide a linguistically justified explanation for the social effect that can inform the profiling task.

However, although richer vocabulary and familiarity with the nominal style for certain participants in certain occupations are both explainable with their greater familiarity with certain registers, other effects cannot be easily explained using only the register axiom. For example, the difference in the *F* score found for males vs. females, although modest, cannot be explained by the unequal gender access to certain occupations, since occupation and social class were controlled in this experiment. It seems that even with equal access to certain registers, men tend to score on average higher on *F* than women and therefore there must be other factors at play. Similarly, the psycholinguistic literature has demonstrated that older speakers tend to use less complex syntax partially because of decrease in working memory and this effect cannot be completely discounted. These considerations lead to the more general conclusion that authorship profiling might not be an exclusive *sociolinguistic* phenomenon and it would therefore be partially erroneous or misleading to refer to the task of authorship profiling as simply *sociolinguistic profiling*, as certain linguistic patterns might have explanations outside of the field of sociolinguistics, for example in psycholinguistics.

The last consideration is about the importance of taking a measure of social class or occupation into account, as previous linguistics literature and this study show how this social factor has the largest effect on language. Virtually no computational authorship profiling research has been devoted to profiling social class or occupation, and this is

problematic as these results show how much an impact this social factor has, even if the goal is the profiling of other demographics.

Conclusions: how to develop forensic authorship profiling?

In sum, it seems very difficult at the present stage for an automatic system to be able to untangle all of the factors that this study has outlined, from the importance of register to the interaction of the social factors, especially if the text to be profiled is very short, as is common in forensic linguistics. Carrying out profiling for forensic purposes means, in essence, estimating the most likely demographics of the author of one of the dots in any of the four figures above. As an example, let us assume a questioned text has been analysed and its *F* score is 45. Looking at Figures 1 and 2 it is evident that whereas a score of 45 is completely the norm for a text like Task 3, it is definitely outside the norm for a text like Task 1, for which this score is very unlikely and only found in the lower classes. The understanding of the register is therefore a precursory step for profiling. However, even an analysis of the register does not substantially help in the majority of cases. The clouds of points in those graphs makes it evident that there is a great degree of overlap between the categories and, consequently, not very much discriminatory potential. Profiling of the general demographics is therefore a very difficult task, which might be possible only in certain extreme circumstances, such as when the questioned texts behave in ways that are substantially outside the norm. For example, the results of this study show how an *F* score of 60 for Task 1 is very unlikely for the average Lower-Female but typical for the average Higher-Male.

The crucial step for carrying out profiling right now thus seems to be the identification of deviation from a norm. For example, although it now seems established that higher social classes/men/older individuals use a more nominal style than lower social classes/women/younger individuals, what *more* and *less* mean depends on the register of the questioned text, which should therefore be analysed before carrying out profiling. My proposal for an algorithm for the forensic authorship profiling of writing style based on these considerations is therefore as follows:

1. Study the extra-linguistic situation of the questioned text, for example using Biber's (1994) Situational Parameters;
2. Collect and analyse a corpus with comparable situational parameters to establish the norm for the linguistic features that will be analysed, the set of which should be based on previously established literature on stylistic variation. If possible, the corpus should contain texts written by a stratified sample of the population to verify that the previously established stylistics patterns are present and whether they follow the predicted direction and to what extent;
3. Check the position of the disputed text in the register space given by the comparison corpus, similarly to the graphs presented above, so that the position of the text in relation to the distribution for the register can be assessed;
4. Bearing in mind previous literature, of which this article is an initial survey, compare the linguistic behaviour of the disputed text against the norm;
5. Very importantly, the meaning of the numbers should not be ignored, especially for short texts. Knowledge from previous literature is useful because it provides an explanation for the linguistic patterns that we observe but only if the linguistic patterns can be explained by the same principles can these be used to infer the characteristics of the anonymous author.

The most challenging component of this algorithm is probably step (2), as it might be difficult or impossible to collect a stratified sample of certain registers. However, this is what core research in *forensic* authorship profiling should do: focus on expanding on the present work so that a forensic linguist does not have to collect an *ad hoc* corpus for every case and can therefore use previous studies for direct comparison. For example, the study reported here could be used as a baseline for forensic work on a questioned text with situational parameters similar to one of the three Tasks, even though replications of this study are also, of course, highly encouraged.

For the future, two items are particularly urgent in the agenda: (1) to increase understanding of the social factors that are profiled, and (2) to develop new computational techniques that are aware of these issues and that include linguistic theory.

The first point concerns the issues raised in the literature reviews above regarding the definition of the three social factors, gender, age, and social class. It is unquestionable that these categories cannot be simply defined in the way that has been used in previous studies and, consequently, in this present study. However, there is a problematic tension between the requirements of law enforcement and what an analysis of language can reveal. In all likelihood linguistic analyses can only profile social factors that are proxies to the type of social information that law enforcement needs and future research into *forensic* authorship profiling should address this tension. For example, more studies should focus on untangling the elements of gender that correlate with stylistic variation, so that it is clear, for instance, what the *F* score is actually measuring. Equally, studies are needed to verify whether biological age is indeed a proxy to social age in terms of stylistic variation. This knowledge can inform the type of inference that a forensic linguist can make when faced with a profiling problem.

The second point concerns the direction of research and the collaboration between computer scientists and linguists. There is no doubt that more sophistication in the analysis can help with the issues outlined in this article and this level of sophistication can certainly only come from the fields of computer science and computational statistics. However, the research in these fields should be guided both by the needs and, more importantly, by the previous knowledge already available in the fields of enquiry in which these statistical and computational techniques are applied, that is, linguistics. This collaboration can ensure that sophistication of method is paired with a high degree of interpretability and that it is also contextualised within the field of linguistics. It is likely that a method based on machine learning, such as that of Argamon *et al.* (2009), if applied to the present data sets would still return good accuracy rates and, if trained with appropriate awareness of register issues, even achieve better performance. However, it is still debatable to what extent these results would be useful in a forensic context without a proper linguistic interpretation.

The understanding of the underlying linguistic patterns responsible for the predictions is a pre-requisite for *forensic* authorship profiling because, ultimately, the evidence analysed is linguistic and not statistical. Therefore, although computational methods can and should be employed to aid the analysis, this must not be done at the expense of the underlying linguistic explanations, which should remain the primary focus within forensic linguistics.

In conclusion, because of what is at stake in a forensic setting, authorship profiling can be developed into *forensic* authorship profiling only when linguistics and computer science work side by side keeping the focus not on techniques but on linguistic explanations, theories, and knowledge, with particular attention to the forensic context.

Notes

[1]With the exception of three participants for whom it was not possible to obtain occupation information about their parents.

References

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3), 321–346.

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), 119–123.

Ash, S. (2002). Social class. In J. K. Chambers, P. Trudgill and N. Schilling-Estes, Eds., *The Handbook of Language Variation and Change*. Malden, MA; Oxford: Blackwell Publishers, 402–423.

Bamman, D., Eisenstein, J. and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2), 135–160.

Bell, A. (1984). Language style as audience design. *Language in Society*, 13, 145–204.

Berman, R. (2008). The psycholinguistics of developing text construction. *Journal of child language*, 35(4), 735–71.

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. (1994). Register and social dialect variation: An integrated approach. In D. Biber and E. Finegan, Eds., *Sociolinguistic Perspectives on Register*. Oxford: Oxford University Press, 315–347.

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

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1).

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1), 7–34.

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge; New York: Cambridge University Press.

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

Bing, J. M. and Bergvall, V. L. (1998). The question of questions: beyond binary thinking. In J. Coates and P. Pichler, Eds., *Language and Gender: a Reader*. Chichester, West Sussex, U.K.; Malden, MA: Wiley-Blackwell, 495–511.

Bromley, D. B. (1991). Aspects of written language production over adult life. *Psychology and Aging*, 6(2), 296–308.

Byrd, M. (1993). Adult age differences in the ability to write prose passages. *Educational Gerontology: An International Quarterly*, 19, 375–396.

Carothers, B. J. and Reis, H. T. (2013). Men and women are from Earth: Examining the latent structure of gender. *Journal of Personality and Social Psychology*, 104(2), 385–407.

Chambers, J. K. (1992). Linguistic correlates of gender and sex. *English World-Wide*, 13(2), 173–218.

Dubay, W. H. (2004). *The Principles of Readability*. Costa Mesa: Impact Information.

Eckert, P. (1998). Age as a sociolinguistic variable. In F. Coulmas, Ed., *The Handbook of Sociolinguistics*. Oxford, UK; Cambridge, Mass: Blackwell Publishers, 151–167.

Eckert, P. and McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology*, 21, 461–490.

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*. Cambridge: Cambridge University Press, 235–267.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In J. Gibbons and M. T. Turell, Eds., *Dimensions of Forensic Linguistics*. Amsterdam: John Benjamins Publishing Company, 215–231.

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*. London: Cassell, 191–242.

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*. London: Equinox, 5–40.

Hasan, R. and Cloran, C. (1990). A sociolinguistic interpretation of everyday talk between mothers and children. In M. A. K. Halliday, J. Gibbons and H. Nicholas, Eds., *Learning, Keeping, and Using Language. Volume 1*. Amsterdam; Philadelphia: John Benjamins Publishing, 67–100.

Hawkins, P. R. (1977). *Social Class, the Nominal Group and Verbal Strategies*. London: Routledge and Kegan Paul.

Herring, S. C. and Paolillo, J. C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4), 439–459.

Heylighen, F. and Dewaele, J. (1999). *Variation in the contextuality of language: an empirical measure, Center "Leo Apostel"*. Brussels: Free University of Brussels.

Heylighen, F. and Dewaele, J. (2002). Variation in the contextuality of language: an empirical measure. *Foundations of Science*, 1–27.

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*. New York: John Wiley, 99–125.

Johnston, R. (1977). Social class and grammatical development: A comparison of the speech of five year olds from middle and working class backgrounds. *Language and Speech. SAGE Publications*, 20(4), 317.

Kemper, S. (1987). Life-span changes in syntactic complexity. *Journal of Gerontology*, 42(3), 323–328.

Kemper, S., Greiner, L. H., Marquis, J. G., Prenovost, K. and Mitzner, T. L. (2001). Language decline across the life span: Findings from the Nun Study. *Psychology and Aging*, 16(2), 227–39.

Kemper, S., Kynette, D., Rash, S., O'Brien, K. and Sprott, R. (1989). Life-span changes to adults' language: Effects of memory and genre. *Applied Psycholinguistics*, 10(01), 49–66.

Kemper, S. and Sumner, A. (2001). The structure of verbal abilities in young and older adults. *Psychology and Aging*, 16(2), 312–322.

Kitson, H. D. (1921). *The Mind of the Buyer*. New York: MacMillan.

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

Labov, W. and Auger, J. (1993). The effect of normal aging on discourse: A sociolinguistic approach. In H. H. Brownell and Y. Joanette, Eds., *Narrative Discourse in Neurologically Impaired and Normal Aging Adults*. San Diego, California: Singular Pub Group, 115–135.

Leonard, R. (2005). Forensic Linguistics: Applying the Scientific Principles of Language Analysis to Issues of the law. *The International Journal of the Humanities*, 3.

Loban, W. (1967). *Language Ability - Grades Ten, Eleven, and Twelve. Final Report*. Rapport interne, Berkeley.

Macaulay, R. (2002). Extremely interesting, very interesting, or only quite interesting? Adverbs and social class. *Journal of Sociolinguistics*, 6(3), 398–417.

McEnery, T. (2006). *Swearing in English*. London: Routledge.

Mitzner, T. and Kemper, S. (2003). Oral and written language in late adulthood: Findings from the Nun Study. *Experimental Aging Research*, 29, 457–474.

Mollet, E., Wray, A., Fitzpatrick, T., Wray, N. R. and Wright, M. J. (2010). Choosing the best tools for comparative analyses of texts. *International Journal of Corpus Linguistics*, 15(4), 429–473.

Newman, L. M., Groom, C. J., Handelman, L. D. and Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211–236.

Nini, A. (2015). *Authorship Profiling in a Forensic Context*. Phd thesis, Aston University, UK.

Nini, A. (2017). Register variation in malicious forensic texts. *International Journal of Speech Language and the Law*, 24(1), 99–126.

Pennebaker, J. W., Groom, C. J., Loew, D. and Dabbs, J. M. (2004). Testosterone as a social inhibitor: Two case studies of the effect of testosterone treatment on language. *Journal of Abnormal Psychology*, 113(1), 172–175.

Pennebaker, J. W., Mehl, M. R. and Niederhoffer, K. G. (2003). Psychological aspects of natural language use: our words, our selves. *Annual Review of Psychology*, 54, 547–77.

Pennebaker, J. W. and Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, 85(2), 291–301.

Poole, M. (1976). *Social Class and Language Utilization at the Tertiary Level*. St. Lucia, Q: University of Queensland Press.

Poole, M. E. (1979). Social-class, sex and linguistic coding. *Language and Speech*, 22, 49–67.

Rabaglia, C. and Salthouse, T. (2011). Natural and constrained language production as a function of age and cognitive abilities. *Language and Cognitive Processes*, April 2013, 37–41.

Rayson, P., Leech, G. and Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1), 133–152.

Rude, S., Gortner, E.-M. and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, 18(8), 1121–1133.

Säily, T., Siirtola, H. and Nevalainen, T. (2011). Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing*, 26(2), 167–188.

Schler, J., Koppel, M., Argamon, S. and Pennebaker, J. (2006). Effects of age and gender on blogging. In *2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, 199–205, Stanford, CA.

Shuy, R. (2014). *The Language of Murder Cases: Intentionality, Predisposition, and Voluntariness.* Oxford: Oxford University Press.

Appendix – The Experiment Tasks

Thank you for agreeing to participate in this experiment. The study is concerned with cases of interaction that are unfavourable or undesirable for the addressee.

The experiment consists of three tasks. For each, you will need to put yourself imaginatively in the situation that is described and then write a short text (at least 300 words) according to the guidelines provided.

The information you provide will be treated confidentially and will not be used for purposes other than the statistical measurement required for the present study.

SITUATION (1): Last year you bought a travel package from the FirstHoliday travel agency. Unfortunately, the holiday was totally unsatisfactory and you feel that it was not worth the price you paid. Indeed, you feel that the company should give you a refund.

TASK (1): Write a letter to the agency. You must not only express your feelings of dissatisfaction, but also describe how and why the situation made you very upset and angry. Warn them about possible legal action and ask for a partial refund of £500.

SITUATION (2): The economic crisis is making your life significantly more difficult. You feel frustrated that the coalition government is not addressing the issue as seriously as it deserves and you are worried that you might lose your job in the next few months if the planned cuts are not rescinded. You therefore think it is time to send a letter to them to make sure they understand that voters like you are unhappy and desperate.

TASK (2): Write an anonymous letter, signed as "A disappointed voter", to the Prime Minister showing your disappointment in how the government is managing the economic crisis. Express how the recession has hit you and that you are very angry that nothing has been done to prevent the situation. Make it very clear that you won't vote for them again if they don't change policies.

SITUATION (3): You are an employee of a company where you have been working for a long time. You have a newly appointed boss who is extremely abusive to you and to your colleagues and apparently does not value your work. To scare your boss, you are planning to make him think that if he does not change his unreasonable behaviour, someone will damage his car.

TASK (3): Write an anonymous letter, signed as "An angry employee", where you express your thoughts and feelings about his abusive behaviour. As well as expressing your views, scare your boss by using one of the following options for each category:

(a) car parts to be damaged: bodywork mirrors – tyres – lights
(b) object used to damage: baseball bat – jack – nail – spray paint
(c) time: early morning – lunch break – night